ACADEMIA IN INFORMATION TECHNOLOGY PROFESSION (AITP)

(An Interest group of Nigeria Computer Society)

Website: *www.aitp.org.ng*

Email: *academia.infotech@gmail.com*

PROCEEDINGS

of the

2020 INTERNATIONAL CONFERENCE

on

28-30TH JULY, 2020

INFORMATION TECHNOLOGY IN EDUCATION AND DEVELOPMENT (ITED)

THEME:

FOSTERING IT ECOSYSTEM FOR EFFECTIVE REALIZATION OF THE 4TH INDUSTRIAL REVOLUTION

VENUE: ZOOM PLATFORM

EDITED BY:

Prof. Afolayan. A. Obiniyi

Dr. (Mrs.) Uyinomen O. Ekong

ACADEMIA IN INFORMATION TECHNOLOGY PROFESSION (AITP)
(An interest group of Nigeria Computer Society)
Website: www.aitp.org.ng
Email: academia.infotech@gmail.com

PROCEEDINGS

of the

2020 INTERNATIONAL CONFERENCE

on

INFORMATION TECHNOLOGY IN EDUCATION AND DEVELOPMENT (ITED)

THEME:
FOSTERING IT ECOSYSTEM FOR EFFECTIVE REALIZATION OF THE 4TH INDUSTRIAL REVOLUTION

DATE:
28th - 30th July, 2020

VENUE: ZOOM PLATFORM

EDITED BY:
Prof. Afolayan. A. Obiniyi
Dr. (Mrs.) Uyinomen O. Ekong

PUBLISHED BY ACADEMIA IN INFORMATION TECHNOLOGY PROFESSION (AITP): AN INTEREST GROUP OF NIGERIA COMPUTER SOCIETY (NCS)

The President,
Academia in Information Technology Profession (AITP),
Nigeria Computer Society,
10, Jobi Fele Way, Behind MKO Garden,
Alausa, Ikeja, Lagos.
www.ncs.org.ng

ISBN 978012679-1

TABLE OF CONTENTS

LIST OF PAPERS ACCORDING TO SECTIONS

SECTION A: INTELLIGENT COMPUTING

S/N	PAPER TITLE	AUTHOR(S)	PAPER CODE	PAGES
1.	Predictive Analysis Of Human Activities Using Supervised Learning	Cecelia Ajowho Adenusi, Olufunke Rebecca Vincent, Abiodun Folurera Ajayi	ITED20073	225-233
2.	Emerging Trends In Artificial Intelligence And Machine Learning: Historical And State-Of-The-Art Perspectives	Edward E. Ogheneovo	ITED20078	33-41
3.	Appraisal of Selected Principal Component Analysis-Based Methods for Face Recognition System's Accuracy.	Madandola Tajudeen Niyi	ITED200724	83-88
4.	Comparative Analysis of Machine Learning Classifiers for Detecting Malware in Portable Executable	Faden David Nanven, Morufu Olalere, Adebayo Olawale Surajudeen, Joseph Ojeniyi	ITED200748	146-152
5.	An Intelligent System For Student Placement Using Fuzzy Logic	Femi Temitope Johnson, Olufunke Rebecca Vincent	ITED200726	100-108
6.	A Predictive Model for Student Performance in Examination using Supervised Machine Learning Algorithm	Bukola Taibat Adebiyi, Olufunke Rebecca Vincent	ITED200740	135-140
7.	Intelligent Road Traffic Control System (A Case Study Of Sango Intersection, Ibadan)	Abdulhameed, I.A., Azeez, S.A., Mamudu, A.O. Oluwaseun, S.A.	ITED200757	195-197
8.	A Framework For Intelligent Malware Detection In Mobile Devices	Ekong U.O., Oyong S.B., Obot, O.U. and Ekong V.E.	ITED200710	46-50
9.	Staff Performance Evaluation using Artificial Neural Network	A.U. Rufai and I. A. Adeboye	ITED200713	58-62
10.	Performance Evaluation of ANOVA and Recursive Feature Elimination (RFE) Algorithms with Support Vector Machine Classifier using Microarray Dataset	Sulaiman Olaniyi Abdulsalam	ITED200721	76-82

SECTION B EDUCATION TECHNOLOGY

SECTION C CLOUD COMPUTING

SECTION D NETWORK SYSTEM SECURITY

SECTION E BIOTECHNOLOGY/ OPERATIONAL TECHNOLOGY

SECTION F NETWORK SYSTEM SECURITY

	Remainder Theorem			
3.	Probing Attack Detection Using JRIP Classifier.	Olomi Isaiah Aladesote, Ajayi Ebenezer Akinyemi	ITED200756	191-194
4.	A Survey on Slow DDoS Attack Detection Techniques	Oluwatobi Shadrach Akanji	ITED200754	182-190
5.	Approximate Probability of Satisfactory State of varying Number of Peers in P2P Live Streaming Network using Poisson Distribution	Dima R.M., Aina S.K., Bashir A.J., and Yunus A.A.	ITED200732	109-116

LIST OF REVIEWERS

1. Dr. (Mrs) O. R. Vincent

   Department of Computer Sciences
   College of Physical Sciences,
   Federal University of Agriculture, Abeokuta
   Ogun State.

2. Prof. Oludele Awodele

   Department of Computer Sciences,
   School of Computing and Engineering Science,
   Babcock University,
   Ilishan,

3. Dr. B.Y. Baha

   Department of Information Technology,
   School of Management Technology,
   Modibbo Adama University of Technology
   Yola, Adamawa State.

4. Dr. Adewole Rufai

   Department of Computer Sciences
   University of Lagos
   Lagos, Nigeria.

5. Prof. A.A. Obiniyi

   Department of Computer Science
   Ahmadu Bello University
   Zaria.

6. Prof. Rasheed G. Jimoh

   Department of Computer Science
   University of Ilorin
   Kwara State.

7. Dr. Victor E. Ekong

   Department of Computer Science
   Faculty of Science,
   University of Uyo, Uyo
   Akwa Ibom State.

8. Dr. Uyinomen O.Ekong

   Department of Computer Science
   Faculty of Science,
   University of Uyo, Uyo
   Akwa Ibom State.

9. Mr. Ebenezer Ajayi Akinyemi

Computer Science Department,
Kebbi State Polytechnic,
P.M.B. 1158,
Dakingari, Kebbi State

10. Dr. Oluwakemi C. Abikoye

Department of Computer Science,
University of Ilorin,
Kwara State.

11. Dr. Abdulrauf U. Tosho

Department of Physical Science,
Al-Hikmah University,
Kwara State

12. Dr. Kayode Sakariyah Adewole

Department of Computer Science,
University of Ilorin,
Kwara State.

13. Dr. Agaji Iorshase,

Department of Mathematics/Statistics/Computer Science,
Federal University of Agriculture,
Makurdi, Benue State.

14. Dr. Kamaldeen Ayodele Raji

Kwara State Polytechnic (Kwarapoly)
Ilorin, Kwara State.

FORWARD

On behalf of the 1st National Executive Council of Academia in Information Technology Profession (AITP): an umbrella body of all academia in Colleges of Education, Polytechnics, Monotechnics, Universities and all academia in ministries and parastatals, I welcome you all to the 1st Virtual International Conference. This year's International Conference on Information Technology in Education and Development (ITED 2020) takes place via zoom platform from 28th to 30th July, 2020.

Indeed, the conference supposed to have been held in University of Abuja on March 17th -19th 2020 with all arrangement made for the conference to be held, however, two weeks to the conference, the Academic Staff Union of Universities (ASUU) strike started, hence, the conference was postponed. Preparations were made to find a convenient date for the conference to be held, then, the announcement of Corona virus pandemic (COVID-19) was made with all the attendant protocols or rules such as social distancing, no gathering of people to avoid personal contacts and so on, became pronounced. Glory be to God that we were able to hold the conference despite all odds today.

COVID-19 pandemic is changing the way we live, work and interact, thus bringing a new normal to the world. We are now adapting to a new way of life in which we have to work and do everything from home, in other words, from a distance. Hitherto, the theme of this year's International Conference "Fostering IT Ecosystem for effective Realization of the Fourth Industrial Revolution" is well chosen. It is pertinent to key into this theme since we as a nation could not partake in the three earlier industrial revolutions. The time has come to transform and initiate discussions on how to make teaching and research impactful and more visible for scholars with attendant advantages of innovative development that can engender IT transfer and aid global growth and development. It is imperative to tinker on the fact that IT is fast becoming a new oil production for Nigeria. This, we must leverage on, and proffer solutions to the present economic problems emanating from COVID-19. It is time we think of automating our education system, our government and governance in Nigeria and virtually all our ways of life. This will go a long way in improving our Gross Domestic Products (GDP) and alleviate the poverty in the country.

There are diverse stakeholders from the government, academia, industry and a host of other youth groups in this conference in which we should leverage on, to engender us, to be in the forefront of the 4th Industrial Revolution.

On this note, we want to thank the Speaker of the House of Representatives; Honourable, Femi Gbajebiamila, the Honourable Minister of Communication and Digital Economy; Dr. Isa Ali Pantami, the Director General of National Information Technology Development Agency; Mr. Kashifu Inuwa Abdullahi, the President, Digital Bridge Institute; Professor Mohammad Ajiya, the Vice Chancellor of Federal University of Agriculture; Professor Musa Isiyaku Ahmed, all our awardees and all the special guests too numerous to mention for making it to this conference.

I cannot but recap the effort of the Vice Chancellor of University of Abuja; Professor Abdul-Rasheed Na'Allah for all his contribution to holding the conference in his University which was, however, aborted by ASUU strike. Sir, we are grateful. The local organizing committee headed by Professor Gbolagade's effort to holding the conference in the University of Abuja is well appreciated.

The indefatigable effort of Dr (Mrs) Uyinomen Ekong (the Proceeding Editor) who spent sleepless night in making sure that the proceeding was ready despite all odds is well appreciated.

We are indebted to thank Professor R. G. Jimoh who chaired the Virtual Conference Event and Planning Committee to making it a success. We are grateful to all the reviewers of the proceeding who spent time to make sure the proceedings were thoroughly reviewed.

The organizers of this conference thank our partners-Nigeria Computer Society (the umbrella of all computer groups in Nigeria), the Computer Professional Registration Council of Nigeria (CPN), National Information Technology Agency of Nigeria (NITDA) and individuals for the great roles they played to make this conference a success.

Finally, I want to thank the 1st National Executive Council members of AITP for working together not only as Council member but as a family. In fact, your corporation will forever live an indelible mark in my memory.
Thank you all and remain blessed.

Professor A. A. Obiniyi, *MAIT, MITSSP, MCPN, FNCS*
President, Academia in Information Technology Profession (AITP)

Wireless Fingerprint Attendance Recording System

Temilola Adijat Okunade Felix Olutokunbo Idepefo Idris Abiodun Aremu
Department of Computer Science
Lagos State Polytechnic
Lagos, Nigeria
taokunade@yahoo.com felixidepefo@gmail.com, aremu.i@mylaspotech.edu.ng

ABSTRACT— Attendance monitoring is a common phenomenon in higher institutions of learning. Reasons for monitoring attendance may include the need to improve performance of students (by regulating the acceptable percentages of student absenteeism that qualify them to sit for an examination), ensuring sponsoring bodies do not waste funds on students that do not attend classes and assessing the commitment of lecturers. Methods that have been applied in recording attendance in lecture halls include manually signing attendance registers, face recognition technologies, use of RFID tags and queuing up to scan fingerprints at entrance of lecture halls. These methods have the challenges of students signing for each other, overlapping of faces with crowded classes, wearable tags can be held or worn by another student and time wastage respectively. This research presents a wireless attendance recording system with minutiae point matching for authentication. Fingerprint impression of students were captured and minutiae points of the fingerprints extracted and stored in the database during enrolment. During the attendance recording, the fingerprints of students were matched with those enrolled in database. In case of a match, the name, matriculation number and other details of the students are sent wirelessly to the institution's electronic database. Results of trial run demonstrates the effectiveness and efficiency of the system with average authentication time of 12.34s. The system was able to overcome the challenge of overcrowding at lecture hall entrance, prevent impersonation and falsification of attendance. The reports generated are also valuable in assessing lecturers and monitoring students.

Keywords— Attendance recording, Attendance monitoring, Fingerprint matching, RFID

I. INTRODUCTION

A higher institution of learning is regarded as an 'adult' environment. It is expected that students that gain admission into these institutions must have attained some level of maturity. Such students should not require a monitoring body to enforce attendance at classes. They are regarded to be old enough to decide whether they want to attend lectures or not. The freedom of choice is also highly recognised. Research has shown that regular attendance at lectures has a positive and appreciable impact on students' performance in examinations [12]. [9] in his research, made a distinction between seminar attendance and lecture attendance. He was able to show that seminar attendance has high correlation with academic performance. Seminar attendance is also a strong determinant of students' academic performance. The research of [18] also demonstrated the positive impact of attendance in the performance of students. The need for sponsoring bodies to regularly ascertain that inadequate funds are not being wasted is also a major factor in the decision to monitor attendance of students at lectures. A good record of attendance was shown to support increased retention [26, 27] A good record of attendance has been identified as an essential part of the learning process but it is the follow up that students that require support are identified. Students with specific issues are identified and assistance is rendered where necessary (for example, counselling) to correct the attitude. There are also situations that may warrant disciplinary action, appeals or situations that may lead to withdrawal of scholarships, etc. The accurate record of absence can be used to justify such actions. A good record of attendance encourages a better student experience because the staff will be more aware of any welfare issues. It can help improve student concentration on their programme of study. Monitoring attendance at lectures helps students to build a disciplined attitude that will be useful for their work after university [27].

Varieties of approaches have been applied in keeping records of attendance at lectures. These vary from the crude manual method of students signing an attendance register to wearable cards and face recognition systems. The tendency for students to write names for their friends has made the manual approach unacceptable in most cases. Other disadvantages include the loss of the lecturer's time while attendance is being taken; the extra burden on lecturers decreases his ability to perform his primary job; wasteful distraction of students' attention and inaccurate reporting for large sessions, [16].

Monitoring attendance requires a secure and effective identification system. A good identity management system requires the addition of an efficient biometric identification system. Biometrics technologies confirm identity using any of fingerprint, face, iris, retinal pattern, palm print, voice, handwriting, and so on. These techniques are more reliable than conventional methods like password or ID cards because they use physical data. The data used by these methods are unique to individuals and remains the same for the person's lifetime.

Section 4 takes a look at related research in attendance management and the technologies that have been applied. Section 5 presents the methodology applied in this research. Section 6 presents the experiment and results. Section 7 is a conclusion of the research.

II. STATEMENT OF PROBLEM

The importance of records of class attendance in higher institutions cannot be over emphasized. Quite a number of institutions still rely on manual method of recording attendance. Students sign on a sheet of paper that is circulated by the lecturer. This is flawed by problems such as loss or destruction of paper, signing on behalf of an absent student or even alteration or addition to the list (at a later date). There have also been attempts to use fingerprints as a means of taking attendance at classroom entrances. This resulted in long queues of students at the entrance of classrooms. A wireless attendance recording system will eradicate these problems.

III. OBJECTIVE

The main objective of this paper is to design and implement a wireless attendance recording system for classroom use.

IV. IV. LITERATURE

The related literature is made up of a review of techniques used in attendance recording and a review of fingerprint matching techniques.

A. *Techniques Applied in Attendance Recording*

Reference [6] proposed a cheap method of recording student attendance using face detection technique called Image Based Attendance System (IBAtS). In the solution proposed, attendance was taken on class photographs. Students' faces were automatically located, and students then registered their attendance by identifying their faces on the records. Lecturers and students were able to interface with the system through Mobile applications. The system performed better than older methods it was more cost efficient. The limitation of this approach is that the system requires all attendees to already be seated when images are taken by the lecturer. This can cannot accommodate late attendance. [5] proposed a face recognition attendance system based on deep learning. It uses convolutional neural networks (CNNs) cascade for face detection and also for generating face embedding. An overall accuracy of 95.02% was realised on a small dataset of face images of employees in real-time. However, the data sets used for the training of the CNNs is too small. [4] presented an integrated system for monitoring attendance using face recognition system. A classroom camera captures pictures of students during lecture, recognises the students and keeps records of attendance. The challenge with this system is that it is not effective in crowded classrooms as face recognition cannot be efficient in such an environment due to overlap in images.

Reference [3] proposed a smart mobile application that automatically records attendance of students upon detecting their university ID cards. Each lecturer would use his/ her own mobile phone to access the proposed system to check the attendance of students. The students' ID cards are detected as soon as they come within range of the lecturer's reader. A challenge of this system is that students' ID not detected by the RFID reader are marked absent. Also, RFID technology is not suitable for crowded classrooms. The identity card can also be given to another student. [19] designed and implemented an attendance recording system based on the RFID technology. It is a very efficient system with simplicity. The challenge with this approach is that RFID technology is more expensive, partly because it requires system integration and an integrator. In [24] and [13] RFID-based attendance management systems were proposed. The systems were plagued with the challenges of using wearable devices. This implies that the card can easily be passed to another student, thereby defeating the purpose of monitoring attendance. RFID-based systems are also more expensive. Also, the constraint of distance comes to play here as students will have to pass close enough to a sensor to be able to record their attendance. This may get very rowdy in a large class.

Reference [22] proposed an automated approach that was implemented using fingerprint recognition. The developed system has a timekeeping feature that will register every student in a database. It has very high accuracy, is quite economical and requires small storage space for biometric template. [10] developed a smart attendance system that uses a fingerprint sensor in and Arduino UNO. The fingerprints of all staff and students are enrolled into a database. The device can only be activated by the fingerprint of a lecturer or other authorized personnel. [7] developed a prototype attendance system based on a microcontroller. The system consists of a fingerprint sensor and face recognition module. It uses GSM modem to send details of attendance as SMS to the parents. [2] presented a fingerprint-based attendance management system to combat impersonation and the menace of ghost workers. [21] also developed an Academic attendance monitoring system that takes the fingerprint of students before gaining entry to classrooms. After authentication, the student's matriculation number and lecture attended are sent to the database for record purposes. The challenge with their development is the crowd that resulted at the entrance area of a lecture hall due to students queuing up to mark attendance. [17] proposed an improved attendance management system using fingerprint technology in a higher institution. It consists of two procedures; enrolment and identification. Results show that the system is secured, reliable, and capable of averting impersonation but there was no assessment of the system's performance. [15] proposed Smart Mobile Attendance system for employees using QR scanner. The system uses a Mobile App which scans the QR code which acts as user ID while user verification uses fingerprint or voice recognition. The benefit of this system is that it enables employees to update their attendance using their smartphone instead of standing in queue. [23] proposed a Fingerprint Based Approach for Examination Clearance in Higher Institutions. The system recognizes an individual by comparing his/her fingerprint with every record in the database.

In summary, manual method are time wasting and also the records of attendance (sheets of paper) may me lost. Face recognition systems are inefficient in crowded rooms due to overlapping of faces. RFID systems as efficient as they are, have the challenge that wearable devices can be given to another student to hold. Thereby recording attendance for other students. Fingerprint systems result in long queues hence, this research proposes a wireless system that will allow students to take their seats while the device is passed round.

B. *Fingerprint Matching Techniques*

Some research works applied a combination of types of techniques. Among these is the research of [11]. They combined phase-based (a correlation-based technique) with feature-based technique in fingerprint matching. Experimental results demonstrated that the combination works better than a typical minutiae-based technique. [20] made a case for combining image-based fingerprint matching technique with minutiae-based technique to improve verification performance. They also drew attention to the advantage that image-based techniques produce fixed length feature vectors, which make it easy to index. Experiment results show that the proposed system performs better than the other image-based fingerprint matching techniques. Another experiment shows that the proposed image-based technique cannot be compromised using only the knowledge of the minutiae position and orientation. A successful attack will require in the least, the original orientation image. This makes the system less prone to attacks. [1] employed a hybrid of shape and orientation descriptor for fingerprint matching.

Reference [8] in a survey of fingerprint matching techniques classified existing fingerprint matching techniques into correlation-based, minutiae-based and non-minutiae feature based. They concluded that within the past two decades, the minutiae-based techniques were 'the most common'. The research of [25] is also an affirmation of this believe. A comparative study of fingerprint matching techniques was carried out by [14]. They applied the three types of techniques to electronic voting. Experiment results confirms the claim that minutiae-based technique offers the best performance in terms of total election time (while not compromising accuracy) and also requires the least amount of memory.

This research proposes a mobile, wireless fingerprinting solution. The device can be passed from student to student while lecture goes on. This will minimise the disruption that may occur during lectures.

V. METHODOLOGY

The study focused on the design and implementation of a wireless fingerprint-based attendance recording system. The system was designed to help the institution overcome the previously mentioned challenges being experienced.

The system was designed using the client/server architecture (Fig. 1). The central database resides on the server. It stores records on students, lecturers, courses, course allocation, timetable and records of attendance. The clients (lecturers' systems) connect to the server via a wireless network. The fingerprint scanner connects to the lecturers' system via Bluetooth to ease mobility within the lecture room. Data was stored in a MS Access database while the user interface and fingerprint scanner interaction were programmed in MATLAB.

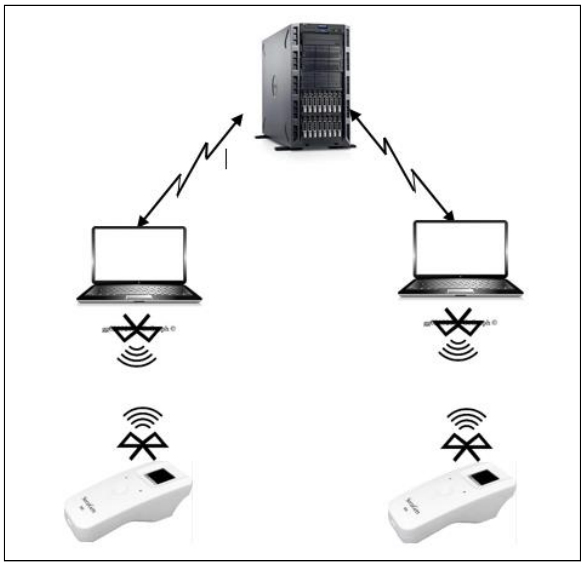


Fig. 1. Wireless fingerprint attendance

Upon enrolment, minutiae points of a fingerprint were extracted and stored in form of a matrix. Each matrix contains the x and y coordinates and the orientation of each point. These features were stored in the database.

Authentication of fingerprints occurs at the point of recording attendance. The positions and orientations of minutiae points extracted from a student's fingerprint are compared to those already in the database using Euclidean distance. If a match is found, the attendance is recorded. Otherwise, the lecturer is alerted to record attendance for the student manually.

A. The Functional Model of the System

The functional model of the system is shown in Fig. 2. The actors are Student, Lecturer and Administrator.

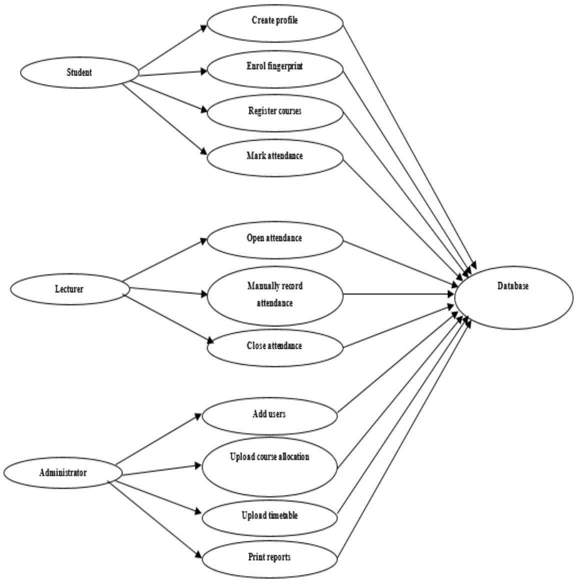


Fig. 2. Functional Diagram of Wireless Fingerprint-based Attendance System

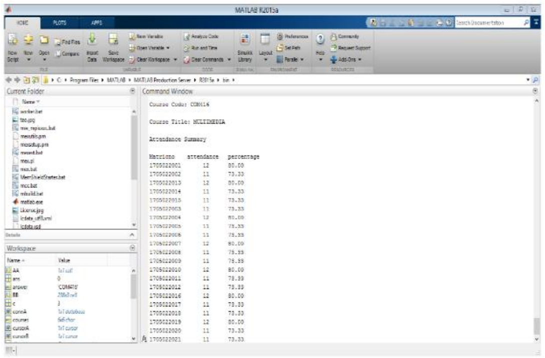
VI. RESULTS AND DISCUSSION

The system was evaluated with a class of 60 students. The average time taken to authenticate a student ranged from 9.55s to 14.53s, with an average time of 12.34s. The interval between authentications of students was difficult to measure because it depended on each student's level of concentration. After the initial fascination with the wireless scanner wore off it took less time to record attendance for a class. For class size of approximately 60 students, the attendance recording took approximately 45 minutes on average.

The attendance summary for COM416 (Introduction to Multimedia) is shown in Fig. 3. The report shows the course code and title. It also shows the matriculation numbers of students that attended lectures, number of times they attended and percentage attendance for the semester.

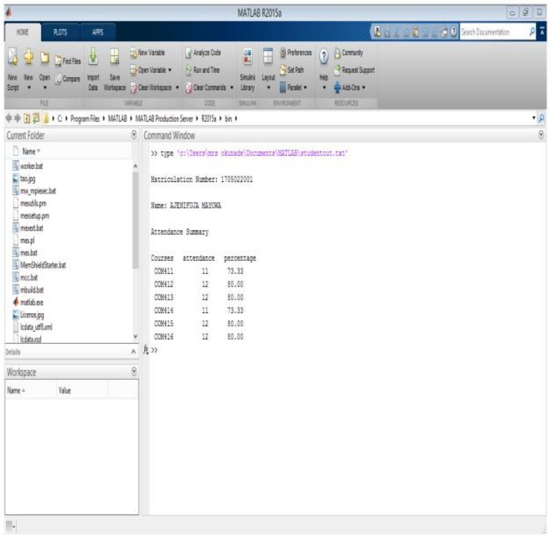
A sample report for students attendance summary is shown in Fig. 4. The report shows the list of courses the student registered for, number and percentage of attendance for the semester.

Just like students' attendance can be monitored, a lecturers' attendance can also be monitored. A sample report that shows list of courses taken by the lecturer and the number of time lectures were held is shown in Fig. 5.



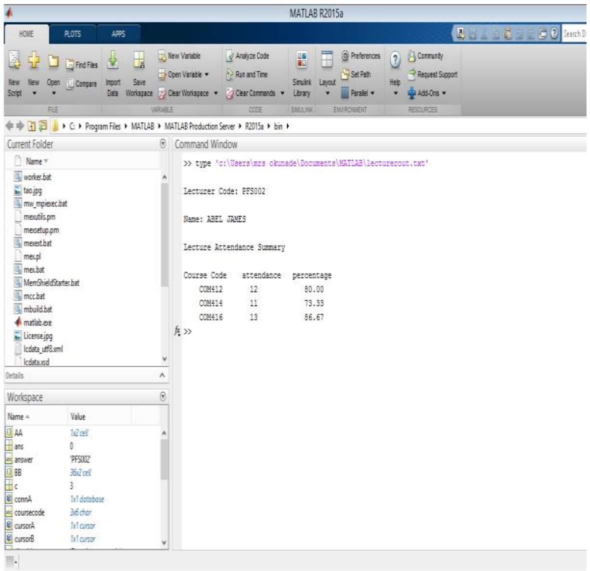
```
Course Code: COM416

Course Title: MULTIMEDIA

Attendance Summary

Matricno    attendance   percentage
1705022001      12       80.00
1705022002      11       73.33
1705022013      12       80.00
1705022014      11       73.33
1705022015      11       73.33
1705022003      11       73.33
1705022004      12       80.00
1705022005      11       73.33
1705022006      11       73.33
1705022007      12       80.00
1705022008      11       73.33
1705022009      11       73.33
1705022010      12       80.00
1705022011      11       73.33
1705022012      11       73.33
1705022016      12       80.00
1705022017      11       73.33
1705022018      11       73.33
1705022019      12       80.00
1705022020      11       73.33
1705022021      11       73.33
```

Fig. 3. Attendance summary for COM416



```
>> type 'c:\Users\mrs okunade\Documents\MATLAB\studentout.txt'

Matriculation Number: 1705022001

Name: AJENIFUJA MAYOWA

Attendance Summary

Courses   attendance   percentage
 COM411        11       73.33
 COM412        12       80.00
 COM413        12       80.00
 COM414        11       73.33
 COM415        12       80.00
 COM416        12       80.00
>>
```

Fig. 4. Attendance summary for a student



```
>> type 'c:\Users\mrs okunade\Documents\MATLAB\lecturerout.txt'

Lecturer Code: PFS002

Name: ABEL JAMES

Lecture Attendance Summary

Course Code   attendance   percentage
    COM412      12            80.00
    COM414      11            73.33
    COM416      13            86.67
>>
```

Fig. 5. Lecturer's attendance summary

VII. CONCLUSSION

This research presents and a wireless attendance recording system for an academic environment. Minutiae points of fingerprints were extracted and stored during enrolment. The authentication process involved minutiae matching. The system showed high level of efficiency. It was able to overcome the challenge of overcrowding at lecture hall entrance and falsification of attendance. The reports generated are also valuable in assessing lecturers and monitoring students.

REFERENCES

[1] Abraham, J, Kwan, P. and J. Gao, “Fingerprint matching using a hybrid shape and orientation descriptor”, State of the Art in Biometrics, 2009, pp. 25 - 56, 2011

[2] Akinduyite C.O, Adetunmbi A.O, Olabode O.O and E. O. Ibidunmoye, “Fingerprint-based attendance management system”. Journal of Computer Sciences and Applications, vol 1, no. 5, 2013, pp. 100-105.

[3] S. Alghamdi, “Monitoring student attendance using a smart system at Taif University”. *International Journal of Computer Science & Information Technology (IJCSIT) vol, 11*. Available at SSRN: https://ssrn.com/abstract=3439186, 2019.

[4] Alia, M. A.; Tamimi, A. A. and O. N. A. AL-Allaf, “Integrated system for monitoring and recognizing students during class session”, International Journal of Multimedia and its Applications (IJMA) vol.5, no.6, 2013.

[5] Arsenovic, M., Sladojevic, S., Anderla, A., and D. Stefanovic, “FaceTime—Deep learning based face recognition attendance system”. in *2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY)* (pp. 000053-000058). IEEE. Available on https://ieeexplore.ieee.org/abstract/document/8080587 , September 2017.

[6] Budi, S., Karnalim, O., Handoyo, E. D., Santoso, S., Toba, H., Nguyen, H., and V. Malhotra, “IBAtS-Image based attendance system: a low cost solution to record student attendance in a classroom”, In *2018 IEEE International Symposium on Multimedia (ISM)* (pp. 259-266). IEEE. Available at https://ieeexplore.ieee.org/abstract/document/8603299 /metrics#metrics, 2018.

[7] Chandramohan, J., Nagarajan, R., Dineshkumar, T., Kannan, G., and R. Prakash, “Attendance monitoring system of students based on biometric and GPS tracking system”, *International Journal of Advanced engineering, Management and Science*, *3*(3). Available on https://dx.doi.org/10.24001/ijaems.3.3.16, 2017.

[8] Dyre, S. and C. P. Sumathi, “A survey on various approaches to fingerprint matching for personal verification and identification”. International Journal of Computer Science & Engineering Survey (IJCSES) vol.7, no.4, 2016.

[9] G. Gbadamosi, “Should we bother improving students’ attendance at seminars?”. Innovations in Education and Teaching International, 52(2), pp. 196-206. [Published Online in 2013].

[10] Ghosh, S., Mohammed, S. K., Mogal, N., Nayak, P. K., and B. Champaty, “Smart attendance system”. In *2018 International Conference on Smart City and Emerging Technology (ICSCET)* (pp. 1-5). IEEE. Available on https://ieeexplore.ieee.org, January 2018.

[11] Ito, K., Morita, A., Aoki, T., Nakajima, H., Kobayashi, K., and T. Higuchi, “A fingerprint recognition algorithm combining phase-based image matching and feature-based matching”, in International Conference on Biometrics (pp. 316-325). Springer, Berlin, Heidelberg, January 2006.

[12] D. Ishita, “Class attendance and academic performance: a subgroup analysis”. International Review of Economics Education https://doi.org/10.1016/j.iree.2018.03.003, 2018.

[13] Kassem, Z. C. A., Hamad, M. and S. E. Dahdaah, “An RFID attendance and monitoring system for university applications”. Proceedings of the 17th IEEE International Conference on Electronics, Circuits, and Systems (ICECS), 2010, pp. 851– 854, 2010.

[14] Kumar, D. A. and T. U. S. Begum, “A comparative study on fingerprint matching algorithms for EVM”. Journal of Computer Sciences and Applications, 1(4), 55-60, 2013.

[15] Kumar, B. D., and S. Kareemulla, “Smart mobile attendance system for employees using QR Scanner”. *Asian Journal of Applied Science and Technology (AJAST)*, *1*(5), 35-39, 2017.

[16] Mohamed, B. K., and C. V. Raghu, “Fingerprint attendance system for classroom needs”, in *2012 Annual IEEE India Conference (INDICON)* (pp. 433-438). IEEE, 2012.

[17] Monday, H. N., Dike, I. D., Li, J. P., Agomuo, D., Nneji, G. U., and A. Ogungbile, “Enhanced attendance management system: a biometrics system of identification based on fingerprint”. in *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)* (pp. 500-505). IEEE. Available on https://ieeexplore.ieee.org, 2018.

[18] J. Muir, “Student attendance: is it important, and what do students think?”, CEBE Transactions, vol. 6, issue 2, September 2009, pp 50-69, (20)DOI: 10.11120/tran.2009.06020050.

[19] Nagwani, N., Sharma, N. K., Khan, M. R., and D. K. Gautam, “An attendances system unit using the Radio Frequency Identification concept”. *Journal of Computer and Mathematical Sciences*, *10*(4), pp. 946-953. Available on http://www.compmath-journal.org/dnload, 2019.

[20] Nanni, L., and A. Lumini, “Descriptors for image-based fingerprint matchers”. Expert Systems with Applications, 36(10), pp. 12414-12422. 2009.

[21] Nawaz, T.; Pervaiz, S.; Korrani, A. and Azhar-Ud-Din, “Development of academic attendence monitoring system using fingerprint identification”, International Journal of Computer Science and Network Security, vol.9 no.5, 2009.

[22] Oyebola, B. O., Olabisi, K. O., and O. S. Adewale, “Fingerprint for personal identification: a developed system for students attendance information management”, *American Journal of Embedded Systems and Applications*, *6*(1), 1-10. Available on http://www.sciencepublishinggroup.com/j/ajesa, 2018.

[23] Saheed, Y. K., Hambali, M. A., Adeniji, I. A., and A. F. Kadri, “Fingerprint based approach for examination clearance in higher institutions”. *FUOYE Journal of Engineering and Technology*, *2*(1). Available on http://engineering.fuoye.edu.ng/journal, 2017.

[24] Shoewu O., Olaniyi O.M. and A. Lawson, “Embedded computer-based lecture attendance management system. African Journal of Computing and ICT, vol 4, no. 3. Pp. 27- 36, 2011.

[25] Ravi. J., Raja, K. B. and, K. R. Venugopal. “Fingerprint recognition using minutia score matching”, International Journal of Engineering Science and Technology. vol.1(2), 2009, pp. 35-42.

[26] University of West London, “Student attendance monitoring and engagement policy and procedure”. Retrieved December 31, 2017 from https://www.uwl.ac.uk/sites/default/files/attendance_monitoring_and_engagement_policy_april_17.pdf.

[27] University of Leeds, “Attendance monitoring policy, guidance and examples of good practice for schools”. Retrieved December 31, 2018 from, https://ses.leeds.ac.uk › download › attendance_monitoring_policy, (n.d.).

Cloud Model for Academia Contact Database and SMS Search Engine using USSD Code

U. R. Alo [1]   C. I. Akobundu [2]   M. S. Julius [3]
[1] Dept. of Computer Science, Alex Ekwueme Federal University, Ndufu-Alike, Ikwo Ebonyi State, Nigeria.
Alo.ur@funai.edu.ng
[2, 3] Dept. of Computer Science, Evangel University, Akaeze, Ebonyi State, Nigeria.
chinyerekingsleyfavor@evangeluniversity.edu.ng. julius.michael@evangeluniversity.edu.ng

ABSTRACT

Research basically has been for knowledge creation and incessant flow of innovative ideas. Individual researchers come up with ideas but collaboration with researchers of like best-and-bright minds would help to amplify the impact of such ideas. Academic collaboration has been noted to foster innovative ideas to meet with the dynamic global world. The Academia globe is so wide such that researchers of like minds most times cannot easily reach themselves despite the glowing technology. This paper presents cloud model for academia contact database and SMS search engine using USSD code, where the contacts of all subscribed researchers are banked in the database of the external application. Interested researcher can either dial the short code in a pre-defined system format or send a message through the Short Message Service (SMS) to an USSD code using the system pre-defined format over the Global System for Mobile Communication (GSM) network. The Mobile Switching Centre (MSC) sets up the Mobile Application Part that delivers the USSD message via the Short Message Peer-to-Peer Protocol (SMPP). The SMPP is the protocol that creates an adaptable interface for message transferring between a message centre and external application. The concept will provide secured unified platform that will greatly support and advance academia research collaboration and communication when fully implemented

Keywords: Cloud model, Research collaboration, Search engine, USSD code

I INTRODUCTION

The quintessence of the world of research is nowledge creation and the incessant flow of new ideas while the collaboration of best and brightest minds (the researchers) definitely would amplify the research impact and further enhance its reputation. The need to ameliorate and sustain academia collaboration is paramount to realize more robust throughput and maneuver complex tasks with ease.

Oxford Learners' Dictionary defined collaboration as the act of working with another person or group of people to create or produce something[1]. Academia Collaboration with industry is critical mostly for the academic because it helps to create scientific knowledge and easily obtain data from the industries[2]. Further worthy of note is that collaboration with universities is crucial for organizations in joint scientific-base research projects in order to develop solutions for production-sourced problems[2]. High standard projects can be more realistic with better ability to scrutinize, debate and share experience(s) with great positive criticisms which are constructively challenging and acceptable opinions and ideas fostered by collaboration [3]. This is essential for academic and scientific accomplishment. There has been different guises of collaboration in the learning environment such as students-to-students (peer learning), Lecturer-to-students, Lecturer-to-Lecturer (team teaching), collaborative research and collaborative curriculum [4].

In the rapidly changing global society, there is need for academic partners to exchange ideas and methods. Mohammed Yasir affirmed that there is abundant evidence that research partnerships and academic collaboration have become the norm in the modern academia [5]. The advantages of collaboration in the area of research can never be over emphasized as ninety-one percent (91%) of researchers have agreed that collaboration increases research impact[6]. Academic collaboration has given rise to new applications for research methodologies, inspiration of new interdisciplinary fields, meeting with the need of the rapidly changing global society and boost efficiency. Effective collaboration leads to easy circumvention of shared obstacles in the research world[7]. Great collaborations are rooted in a shared vision of how to advance the individual ideas. To further advance the vision with an effective throughput, there is need to further strategic objectives of each partner, as the individual interest alone cannot animate truly transformative ideologies. When the true picture of the factors that stimulate researchers to collaborate are brought to limelight, the essence of collaboration will not be evitable, because these factors cannot be meet unless there is a sustained commitment of capable, realistic and trustworthy partners who through partnership come up with innovative solutions [8][9][14].

In order to achieve a productive collaboration, partnership in research has to go beyond the circular of same caucus, known friends, known researcher or clicks in the Universities. Researchers with common interest and objective(s) can affiliate with each other as individual-individual or as groups through an easily accessible means. Some models have been proposed by researchers foster collaboration but the researcher tried to tackle the hurdles of collaboration be deploying a cloud based model in conjunction with API.

The adoption of cloud computing models have cut off so many barriers in the world today, IBM stated it categorically that cloud models have really proven resourceful and have granted accesses to powerful services. It has emerged as the most prominent open-source computer technology [10].

The simplest means of communication between Mobile phones and application programs in the network has been the Unstructured Supplementary Service Data (USSD), a global system for mobile (GSM) communication technology.

The objective of this paper is to build a cloud model for academia contact database and SMS search engine using USSD code to facilitate academia research idea collaborations.

The significance of this paper is evidently undeniable because the dynamic nature of the ever changing world has made it an issue of concern for researchers to work on meeting with the dynamic demands evolving with time. Researchers' collaboration is a better idea for the challenges of the day, and bringing researchers of like-best-minds closer no matter the distance each is operating from will help unit researchers and bring out the best in them.

II RELATED WORKS

Research has shown that academic collaboration has often been in the forefront of Scientific progress, whether amongst prominent established researchers or upcoming one[11]. In reality, there are lots of similar research ideas which are done independently from different researchers that has not been able to locate themselves for greater achievement leading to the need to develop an effective mechanism to identify and peer potential researcher [12]. Qiang Ma et al built a model called h-Reinvestment model which shows how researchers split their effort in realizing a given task. They employed the tools from the field of Game Theory, they studied researchers' collaborative behavior over time under this model, with the premise that each researcher wants to maximize his or her academic success. They observed analytically that there is a strong incentive to collaborate rather than work in isolation, and that studying collaborative behavior through a game-theoretic lens is a promising approach to help better understand the nature and dynamics of academic collaboration, and one of their results showed that two researchers perform asymptotically better by collaborating than by publishing only independent work, reflecting that collaboration is preferable to isolation [13].

Recent academic research as depicted that work involving scientific research tends to be more prolific through collaboration and cooperation among researchers and research group. There are difficulties in discovering new collaborators who are smart enough to conduct joint-research work[14]. Another group of researcher Dinh T. H. et al proposed a hybrid collaboration recommendation method that accounts for research similarities and the previous research cooperation network. Their model measure the extent of individual commitment by combining the collaboration time and the number of co-authors who already collaborated with at least one scientist. Research similarity is based on authors' previous publications and academic events they attended. They used a weighted directed graph to discover new collaborators by using direct and indirect connections between scientists. While a consensus-based system was built to integrate bibliography data from different sources[15].

Researcher Muhammed Y. A. called out a real-life study to examine how research partnership and academic collaboration between different Universities can improve the research impacts. He carried out the research between University of Canberra (UC)-ACT and Khalifa University (KU)-UAE. The result of the research generally encourages academic researcher to explore new areas of focused and result oriented research which will facilitate, enhance and improve its global ranking[16].

Barry B. et al carried out a detailed evaluation on the research collaboration to develop a framework for organizing the research collaboration literature. They studied different data and methods to provide a strong evidence that collaboration tends to enhance the productivity of scientific knowledge. They identified three main attribute categories that are consistently analysed in the literature including collaborator attributes such as personal (race, national, origin), human capital (degree, field of training, tacit knowledge, network ties) and career (career stage, administrative, role); attributes about the collaboration in general such as process (openness, management style and structure) and composition (statuses, demographic mix, roles); and specific organizational or institutional attributes like (resource providers, regulations, competitions, organizational actors). They feel that the “evidence is clear that collaboration provides benefits [16]

III METHODOLOGY ARCHITECTURE

The methodology adopted for the research is qualitative, specifically the Object Oriented Analysis and Design (OOAD) was used for the analyses. Object Oriented Analysis (OOAD) is a Model-driven technique that integrates data and process concerns into constructs called objects. OOAD models are pictures that illustrate the system’s objects from various perspectives such as structure and behavior. OOAD) is a popular technical approach to analyzing, designing an application by applying the object-oriented paradigm and visual modeling throughout the development life cycle to promote communication and product quality.

A. Proposed System Analysis

There are six basic components that illustrate the proposed system. These help to describe the functions and operation of the system including the relationship of the components to essential processes. Cloud model for academia contact database and SMS search engine using USSD comprises of the followings components:

1. Academia components - the components that accommodate researchers where a Researcher 1 can initiates a collaboration request while Researcher 2 who responds to the collaboration request.
2. Mobile phone component – the interface for sending the short code.
3. Mobile Switching Centre (MSC) component - it setups the MAP transaction to the HLR or USSD gateway.
4. Mobile Application Part (MAP) component - is used by the elements, and is located at the application layer that help deliver USSD messages.
5. Home Location Register (HLR) component - this is the main database that houses the researchers’ information.
6. Short Message Peer-to-peer (SMPP) component - this is a protocol component. Protocol is the open industry standard protocol that offers an adaptable data communication interface for transferring short messages between a message centre and the developed application (external application).

B. Use Case Diagram

The Use Case diagram are commonly used for high level requirement analysis of a system to illustrate functionalities of the system and identify actors that perform the functions. It helps in visualizing the functions of the system including the relationship that exist among the identified Actors to essential process. In the cloud model for academia contact database and SMS search engine using USSD, three (3) actors are identified and their relationships as depicted in the use case diagram in figure 1. The actors include Academia – researchers that are members of registered institutions who can either initiate research collaboration request within or outside their institution or respond to the collaboration request; Institution's Admin – a member of registered institutions who creates account, assigns privileges, manages and maintains the account of all the academia in their respective institutions and System Admin - creates account and assigns roles and privileges to registered Institutions' Admin', attends to requests; generates reports; do backup; manages and maintains the system activities and processes.

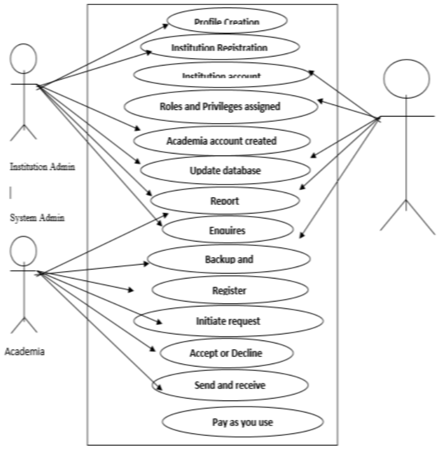


Figure 1: Use case diagram of Cloud model for academia contact database and SMS search engine using USSD code

The architectural design of Cloud model for academia contact database and SMS search engine using the Unstructured Supplementary Service Data (USSD), a Global System Mobile (GSM) communication technology is depicted in figure 2.

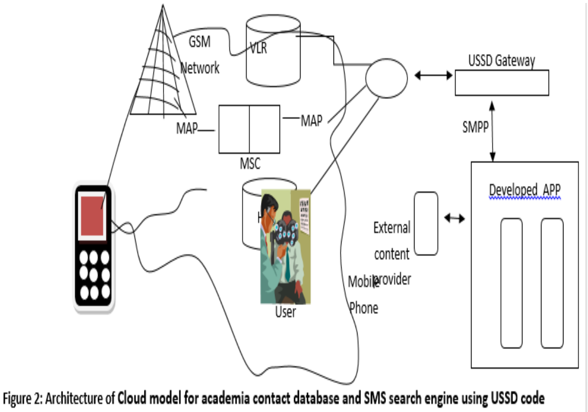


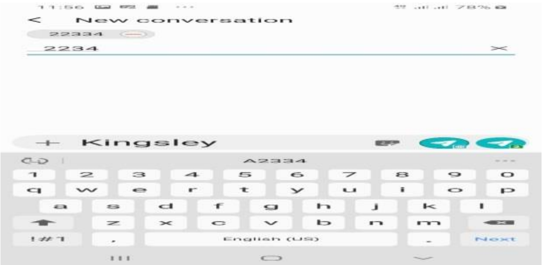
Figure 2: Architecture of Cloud model for academia contact database and SMS search engine using USSD code

IV SYSTEM IMPLEMENTATION

The prototype of Cloud model for academia contact database and SMS search engine using the unstructured Supplementary Service Data (USSD), where the contact of a researcher can be searched is presented in this section.

A. System Input

Cloud model for academia contact database and SMS search engine provides unrestricted access to researcher with common research interest for easy collaboration. The researcher develops interest in the area of research with another researcher and seeks for collaboration. As a registered user of Cloud model for academia contact database and SMS search engine, dials the short code *2334*ResearchersName#, or type the first name of the researcher and sends to the short code 2334. The system explores the full names of all the registered researchers with such names entered, the user picks one option from the listed names, and send back to the same short code. The system sends message to the researcher in search to confirm the release of his/her contact to the researcher seeking for research collaboration. On confirmation, the system sends the contact to the seeker, see figure 3.


New conversation

22334

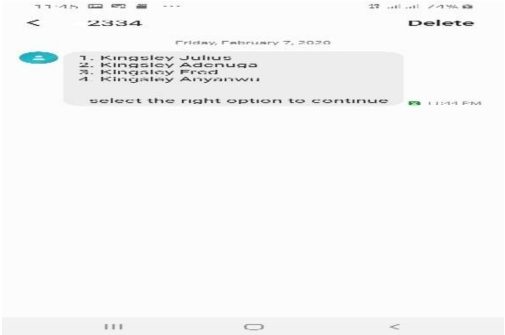
2234

Kingsley

Figure 3: Screenshot of Researcher1 requesting

B. System Output

The system output is the message received from the USSD code that is 2334, which displays the lists of related name in search or the contact of the researcher, see figures 4 to 7.


2334

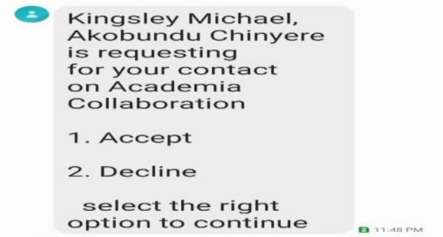
Friday, February 7, 2020

1. Kingsley Julius
2. Kingsley Adenuga
3. Kingsley Fred
4. Kingsley Anyanwu

select the right option to continue

11:44 PM

Figure 4: List of the names with Kingsley in the Academia database.


Kingsley Michael, Akobundu Chinyere is requesting for your contact on Academia Collaboration

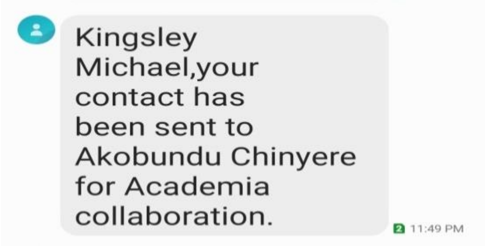
1. Accept

2. Decline

select the right option to continue

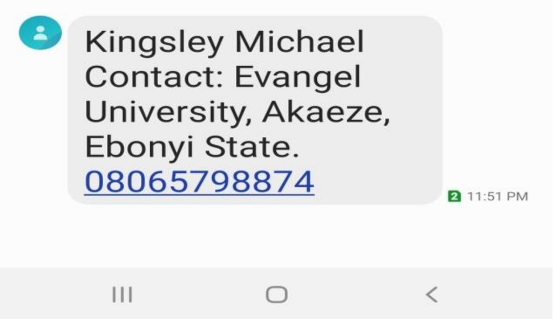
11:48 PM

Figure 5: Notification to researcher2 for acceptance or decline


Kingsley Michael,your contact has been sent to Akobundu Chinyere for Academia collaboration.

11:49 PM

Figure 6: Contact detail release notification


Kingsley Michael Contact: Evangel University, Akaeze, Ebonyi State. 08065798874

11:51 PM

Figure 7: Contact details of Researcher2

V. CONCLUSION AND FUTURE WORKS

The essence of collaboration in the world of academia can never be over emphasized. Studies have been carried out on academia collaboration but no researcher has geared the interest on the unrestricted access of researcher with common interest as how to reach fellow researchers easily for research collaboration. This research built a prototype model of storing the contacts of all subscribed academia in a cloud database through an application system and the contact is accessible using predefined USSD codes technology. A random sample was taken to portray the feasibility of the research and the result turned out positive and will go a long way to unit researchers for the best. The implementation of this research will lead to effective collaboration thereby meeting with the need of the rapidly changing global society and boost efficiency.

Worthy of note is that this paper is part of a large research; the researchers are currently working on the full implementation of the concept which promises to greatly support and advance academia research collaborations and communications.

VI. REFERENCES

[1]. A. S. Hornby, *Oxford Advanced Learner's Dictionary of Current English*. Oxford University Press, Eighth edition, Oxford New York, 2015.

[2]. Kaymaz K. and Eryigit, *Determining Factors Hindering University-Industry Collaboration: An Analysis from the Perspective of Academicians in the context of Entrepreneurial Science Paradigm*. International Journal of Social Inquiry. Vol. 4, issue 1. 2011.

[3]. https://www.qs.com accessed 21/01/2020

[4]. A. Barfield, *Collaboration*. ELT Journal, Vol. 70, Issue 2, 2016, pp. 222 – 224.

[5]. Muhammad Y. A. *Research Partnership and Academic Collaboration between University of Camberra-ACT and the Khalifa University- UAE*. Student-power Jetzt Arbeiten hochladen. www.grin.com/document/427714. Retrieved 21/01/2020

[6]. Jobs.ac.uk. *Survey.* Retrieved 21/01/2020

[7]. *Digital Science*, https//www.eaie.org. Retrieved 21/01/2020

[8]. H. Iglic, P. Doreian, L. Kronegger and A. Ferligoj, *With Whom Do Researchers Collaboration and Why*? Scientometerics, Vol 112, Issue 1. 2017, Pp 153-174.

[9]. M. Hill, A. Hudson, S. Mckendry, N. Raven, S. Saunders, J. Storan and T. Ward, *Partnership to Participation, Collaboration do Widen Participation: To, through and Beyond Higher Education*. FACE, University of East London, London.UK, 2015.

[10]. IBM Academic Initiative. www.ibm.com

[11]. N. Johri, D. Ramage, D. A. Mcfarland and D. Jurafsky, *A Study of Academic Collaboration in Computational Linguistics with Latent Mixtures of Authors.* LaTeCH '11: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Science and Humanities Portland OR, USA. June 2011, pp 124 – 132.

[12]. Z. Liu, X. Xie and L. Chen, *Context-aware Academic Collaboration Recommendation.* In KDD '18: The 28th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, United Kingdom ACM New York, NY, USA. August, 2018.

[13]. M. Qiang, S. Muthukrishan, T. Brian and C. Granham, *Modeling Collaboration in Academia: A Game Theory Approach*. WWW '14 Companion: Proceedings of the 23rd International Conference on World Wide Web Conference Committee (IW3C2). April, 2014. Pp 1177 – 1182.

[14]. J. Li, F. Xa, W. Wang, Z. Chen, N. Y. Asabere and H. Jiang, *ACRec: A Co-authorship based Random Walk Model for Academic Collaboration Recommendation.* WWW '14 Companion: Proceedings of the 23rd International Conference on World Wide Web. April, 2014. Pp 1209 - 1214

[15]. T. H. Ding, T. N. Ngoc, C. T. Van, and H. Dosan, *Research Collaboration Model in Academic Social Networks*. Journal of Enterprise Information Systems, Vol. 13, Issue 8, July, 2019. Pp 1023 – 1045.

[16]. L. E. Weber and J. J. Duderstadf, *Partnering on a Global Scale. The Globalization of Higher Education*. Economic Ltd. London. 2008

.

Fingerprint Based Student Attendance Management System

Amanze Bethran C[1] Nwoke Bethel C[2] Udegbe Valentine, I[3] Durunna Lilian I[4]
[1,2]Department of Computer science, Imo State University, Owerri, Nigeria.
[3,4]Department of Computer Science, Imo State Polytechnic, Umuagwo, Nigeria.
amanzebethran@yahoo.com bethelnwoke@gmail.com udegbevalentine@gmail.com durunnalilian@gmail.com.

ABSTRACT - A finger print biometrics is a security system design with interactive forms. Users have to register their data before they can be authenticated to mark the attendance for a specific course. The lecturer activates the attendance on entering the class. After successfully marking the attendance, the lecturer deactivates the attendance and view attendance history. In this work, Object Oriented Analysis Design and Methodology (OOADM) was very efficient. It can be concluded from the above discussion that a reliable, secure, fast and an efficient system has been developed replacing a manual and unreliable system. Results have shown that this system can be implemented in academic institutes for better results regarding the management of attendance. This system will save time, reduce the amount of work the administration has to do and will replace the stationery material with electronic apparatus. Hence a system with expected results has been developed but there is still some room for improvement.

Keywords- Student, Lecturer, OOADM, Fingerprint scanning machine.

I INTRODUCTION

The approach of using paper sheets and the old file system to confirm students has been in use for years. There are so many constrictions with this old method, one of such problem is the hassles of roll calling, the difficulty for the management to compute the percentage of student attendance in classes and frequently modify their information.

Biometrics is seen as a solution for detecting user's identity and security challenges emanating in this modern day. Biometric identification is any automatically measurable, robust and distinctive physical characteristic or personal trait that can be used to identify an individual or verify the claimed identify of an individual. Biometric science utilizes the measurements of a person's behavioral characteristics (keyboard strokes, mouse movement) or biological characteristics (fingerprint, iris, nose, eyes, jaw, voice pattern, etc.). It is the features captures that is being transformed digitally into a template. The recognition software can then be used to discover an individual as the person they claim to be. Fingerprint recognition is the most common biometric method adopted in identification of a person (Yuanyuan 2012).

Biometric is a field of technology that uses automated methods for identifying and verifying a person based on physiological and behavioral traits. Because some parts of the human body is used in biometrics, the issue of getting lost is not possible and for password to be easily guess can be easily avoided. Also, utilizing biometrics in most cases can be said to be more efficient when speed is considered and convenient than employing password and ID cards method (Chikkerur et al., 2005).

Attendance management of students in institution can be rigorous using the conventional method of paper sheets and old file system method. Every academic institution poses some standards concerning how attendance is to be confirmed for student in classes, laboratory sessions and examination halls. That is why keeping the accurate record of attendance is very important. The approach of using paper sheets and the old file system to confirmed students has been in use for years. There are so many bottlenecks with the conventional method, one of such problem is the difficulty for the management to compute the percentage of student attendance in classes and frequently modify their information. Also in institution, tracking and monitoring student time of attendance could be tedious task, time consuming and as well prone to errors. As an alternative to traditional manual clocking process by students in classes or during examination, biometrics characteristics can be used for authenticating students. This research will focus on developing Fingerprint based Biometric Student Attendance Monitoring System. The fingerprint Biometrics is adopted in this research work for the fact it is one of the most successful applications of biometric technology. In the manual signing processes, where lecturer give a sheet of paper to student to write their names and signature as a form of confirming their presence for a particular class session, falsification in student attendance mostly occur a situation where by a student can sign on behalf of his or her colleague as being present in the class when not true can be so difficult to prevent from happening especially for large classes where row count can takes longer time International Journal of Computer Science and Network Security (2009).

The manual method being used has shown to be inadequate in handling the activity of students' attendance with regards to the importance attached to it as result of this problem encountered.

(1) Impersonation of student attendance occurs frequently and cannot be tracked

(2) Very slow to mark attendance and heavy workload to lecturers in compiling the semester attendance for student.

(3) Breached of security as individual can easily steal attendance sheet and write their name all through.

The aim of the study is to design and develop a student attendance system using fingerprint.

The objectives are:

(1)To carry out the analysis of manual processes involved in class attendance and examination attendance.

(2) To improve existing algorithms to make the fingerprint recognition accurately.

(3) To evaluate the performance of the system.

(4) To develop the system using PHP, OOADIM.

(5) develop the software for student attendance system.

STRUCTURE OF FINGERPRINT

A fingerprint is the pattern of ridges and valleys on the surface of a fingertip. The endpoints and crossing points of ridges are called minutiae. It is a widely accepted assumption that the minutiae pattern of each finger is unique and does not change during one's life.

There are three basic types of Minutiae features:

i) **Ridge endings:** are the points where the ridge curve terminates.

ii) **Bifurcations:** are where a ridge splits from a single path to two paths at a Y-junction.

iii) **Dot:** This is a short ridge also called as Dot.

The diagram below illustrates an example of a ridge ending and a bifurcation. In this example, the black pixels correspond to the ridges, and the white pixels correspond to the valleys.

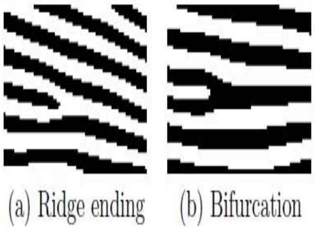


Figure 1: Example of a ridge ending and a bifurcation (Mishra&Trivedi, 2011)

When human fingerprint experts determine if two fingerprints are from the same finger, the matching degree between two minutiae pattern is one of the most important factors. Thanks to the similarity to the way of human fingerprint experts and compactness of templates, the minutiae-based matching method is the most widely studied matching method.

The three fundamental patterns of fingerprint ridges are presented below.

(i) **Arch:** In arch, the ridges will enter from one side of the finger then rise in the center forming an arc, and then exit the other side of the finger.

(ii) **Loop:** The ridges enter from one side of a finger, form a curve, and then exit on that same side.

(iii) **Whorl:** Ridges form circularly around a central point on the finger. The arch pattern the loop pattern the whorl pattern

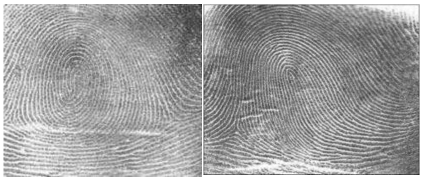


Figure 2: Fingerprint Patterns (Gabor, 1946)

II REVIEW OF RELATED WORKS

There are some existing related works on the application of different methods in managing attendance of students.

Chikkeruret *al.*, (2005) proposed fingerprint enhancement using Short Term Fourier Transforms (STFT), which is based on not stationary signals. In this paper, researchers extended the properties of STFT to two-dimensional (2D) fingerprint images. They proposed a new algorithm for image enhancement process based on contextual filtering in Fourier domain. The new algorithm simultaneously yields local ridge orientation and local ridge frequency level 1 feature. The intrinsic features of the fingerprint image can be computed using single unified approach rather than multiple algorithms. Compare to other image processing algorithm like local/windowed processing, more formal approach for analyzing the non-stationary fingerprint image. Hsieh *et al.*, (2003) proposed an effective and efficient algorithm for fingerprint image enhancement, which not only improves the quality of the clarity of the image but also improves the continuity of the ridge structure based on global texture and local orientation. The global texture is exposed using multi resolution analysis and local orientation through wavelet transforms. In wavelet-based fingerprint analysis first input image is converted into normalized image. Normalized image is decomposed using wavelet decomposition. Wavelet decomposition image is processed again using global texture filtering. Next Local directional compensation is done and finally, wavelet reconstruction process is achieved. Normalization, Wavelet decomposition, Global texture filtering, Local directional compensation, wavelet reconstruction are the flowchart components of proposed enhancement algorithm. Their Experiment results show that enhanced image using wavelet-based enhancement algorithm out performed in terms of efficiency and execution time in improving minutiae detection. Paul and Lourde (2006) proposed the new method for image enhancement using the applications of wavelet transforms. Before the inventions of these techniques, popular other techniques were Gabor filtering and Fourier filtering. The new method outperformed compared to this method in terms of efficiency and execution time. Ye *et al.,* (2007) additionally utilized a 2D discrete wavelet transform to digitally compress fingerprint and to reconstruct the original image, whenever necessary using some reconstructing attributes. Few quantitative measurements are used to evaluate the quality of wavelet transform, which helps in image enhancement process. In this paper researcher also used a different measure to evaluate the performance of wavelet transform and obtained higher efficiency. Farina *et al.,* (2008) worked on a binary image, the input is either already taken as a binary image or converted into binary from the greyscale image and also the image is skeletonized. Due to differences in a number of minutiae occur in real, there is a necessity of post pre-processing, in order to maintain the consistency of image and to reduce the computational cost. They also proposed a new method for ridge cleaning based on ridge positions. In order to validate endpoints and bifurcation, they used two novel approaches and related algorithm. The presented minutiae extraction algorithm performs well in dirty areas and on the backgrounds. Maio & Maltoni, (2007) focused on the extraction of ridge ending and bifurcation called as minutiae directly from the gray-scale image rather than converting it into the binary image and then extracting minutiae. The new techniques were based on ridge line following algorithms and algorithm follows or goes parallel along with ridge line until ridge ending or bifurcation occurs. They compared their algorithm with those known approaches, which converts the original image into binary image and new method showed superiority in terms of efficiency and robustness. Jain*et al.,* (2005) presented a fast fingerprint enhancement algorithm based on level features like fingerprint ridge pattern and orientation and substantially improve the quality of ridge and furrow structures on the estimated local ridge orientation and frequency. This is one of most cited journal paper in image enhancement process. They have evaluated the performance of the image enhancement algorithm using goodness index evaluation criteria of minutiae and by comparing the accuracy of online fingerprint system for verification purpose. They used Gabor filter to tune local ridge orientation and ridge frequency.

Teddy & Martin (2002) demonstrated spatial analysis techniques for latent fingerprint image enhancement. The latent fingerprint is not good in quality which includes some degrade quality like blurred, incomplete or partial and also their spatial definition is not clear. In order improve the quality and thereby by achieving classification or comparison, they used some nonlinear filters and frequency domain filters along with high-pass Butterworth filter with the aid of adaptive fast Fourier transform for enhancement of the degraded image. Fingerprint captured using ink or live scan usually requires only spatial filtering like brightness, contrast, and color map adjustment to examine the level 2 features. Yun *et al.,* (2006) proposed an adaptive filter according to different conditions of the input image which are oil, dry, and neutral instead of the uniform image. To identify oil/dry/neutral image five features are used which are mean, variance, block directional difference, ridge and valley thickness ratio, and orientation change. In adaptive filtering first, several features of the image are extracted and then it is fed into clustering module and then adaptive filtering is applied on clustering to produce a good quality image. For clustering wards clustering method is used. After clustering, once the image is processed depending on the image characteristics, for oily images valleys are improved by expanding thin and detached one, for dry images ridges are enhanced by extracting their center lines and removing white pixels. Chengpu *et al.,* (2008) proposed an effective robust algorithm for fingerprint enhancement and firstly, used contrast stretching approach to improve the clarity between foreground and background of the fingerprint image. Secondly, to improve the orientation estimation utilized the structure of the tensor property. Finally, in order to take the advantages both Gabor filter and diffusion filter, they are combined and adopted low pass filter at the direction that is parallel to the ridge and used band pass filter at the direction perpendicular to the ridge. Wang *et al.,* (2008) introduced log Gabor filter in order to overcome drawbacks of the traditional Gabor filter and to promote and improve fingerprint enhancement performance. The result showed good performance and efficiency compare to traditional Gabor filter. Yuanyuan (2012) proposed new image enhancement algorithm based on elliptical Gabor filter. The ridge information on the fingerprint is used for determining the range of filtering dynamically. Estimating the degree of curvature and the frequency of fingerprint ridge in local areas are used for accomplishing elliptical Gabor filter. To correct errors in the input image and to obtain more precise enhancement, elliptical Gabor filter is used. The experimental results show that the precision of minutiae extraction is significantly improved and which results in good and higher accuracy rate of the subsequent operations are also improved. Babatunde (2012) modified some of the existing sub-models mathematical algorithms for fingerprint image enhancement and obtained new version.

He *et al.,* (2003) developed fingerprint image enhancement algorithm based on orientation fields with three aspects as ridge information for minutiae matching process in a simple and effective way, use of variable sized boundary boxes, and use of simpler alignment method. The first aspect overcomes the problem of reference point per selection with low computational cost. The second aspect makes the algorithm more robust to nonlinear deformation between fingerprints. The third approach reduces the complexity of alignment. Umu*et al.,* (2004) proposed a Biometric template selection and update: a case study in fingerprints.

III METHODOLOGY

A. ANALYSIS OF THE EXISTING SYSTEM

Attendance is an important aspect in institutions, regular attendance will not only ensure full exposure to the scope of majors and opportunities available at institution, and it is also one of the criteria used in determining your final grade. The system currently used in the organization is manual attendance. Tracking and monitoring student time of attendance using the manual attendance in colleges and universities could be tedious, time consuming and more prone to errors. The manual attendance system that is use in classroom (signature system) is not too secure because some students can copy other student's signature.

For manual attendance signing process, the most common problem is the lecturer need to take student daily attendance and manually filled the record in attendance sheet or book for every lecture. If the attendance sheet is missing or misplace, it could lead to big problem because the lecturer need the attendance record to make analysis and generate an attendance report. Another problem is the lecturer will need more time to analyze and generate the attendance report because the lecturer needs to search and refer the old attendance record first. Besides that, an error could happen when the teacher make the calculations to generate the attendance report by themselves. Even though the attendance record is hassle to keep by the lecturer, management report is required in urgent basis. Analyzed attendance record is required by the school management for future actions is normally being delay because of the lack of precise. Moreover, delay analyzes would leads to prolong the time to inform the parents about the truancy students.

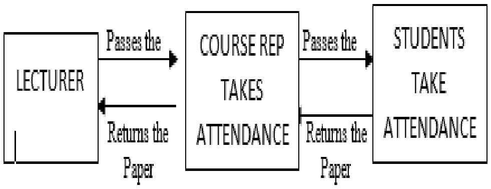


Figure 3: Manual Method of Marking Attendance

B. ANALYSIS OF THE SYSTEM

The system is more reliable, secure, fast and efficient. When this system is implemented, it will eradicate impersonation of student. On entering the class, the lecturer activates the attendance while the students login in their details and mark their attendance.

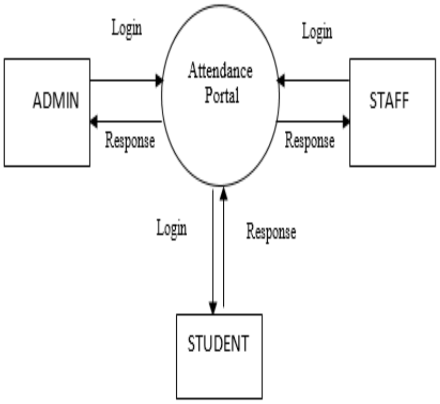


Figure 4: 0–Level Data Flow Diagram

1. USE CASE DIAGRAMS OF THE SYSTEM

A use case diagram is a representation of a user interaction with the system that shows the relationship between the user and the different use cases in which the user is involved. Figure 5 below show the use case diagram of the proposed system.

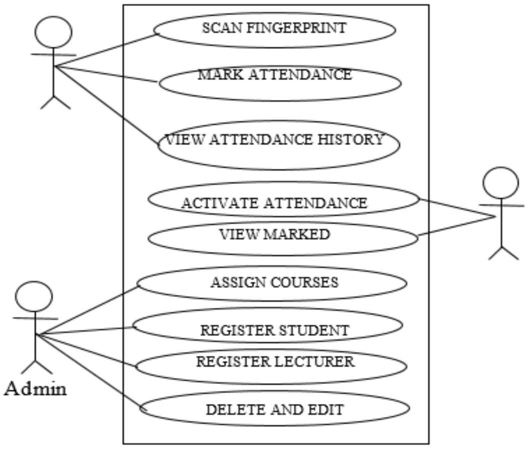


Figure 5: Use Case Diagram

C. ENTERPRISE ARCHITECTURE (EA) OF STUDENTS' ATTENDANCE USING FINGERPRINT

The proposed system is a web based application to be hosted on a web server thaty communuicates to a database server. The user on a web interface makes a web request which is received by the web server. The web server processes the request and interacts with the database server using SQL embedded in PHP scripts. The response is a web page data sent on the web interface for the user. The diagram below shows the enterprise architecture of students' attendance using fingerprint.

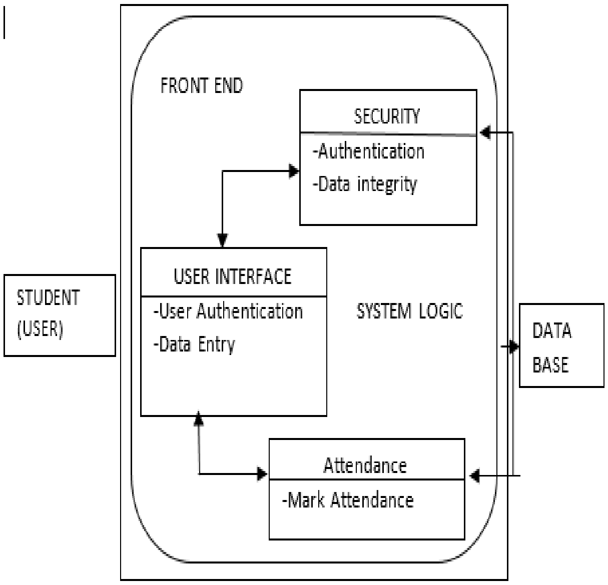


Figure 6: Architecture of the Student Attendance System

D. HIGH LEVEL MODEL OF THE PROPOSED SYSTEM

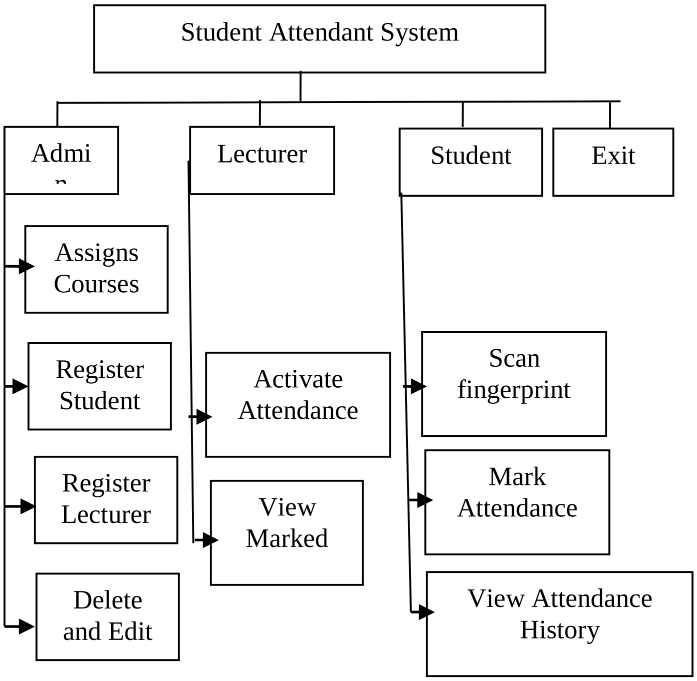


Figure 7: High Level Model

E. CLASS DIAGRAM

This is a static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (methods) and the relationships among objects.

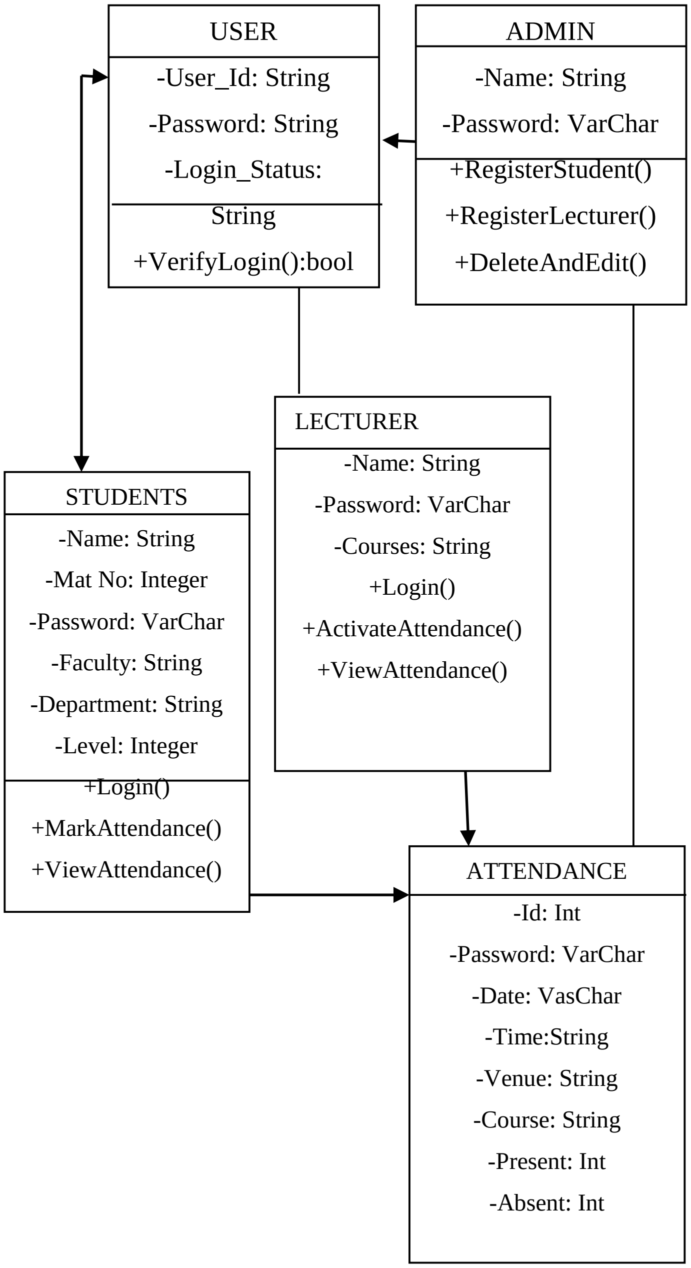
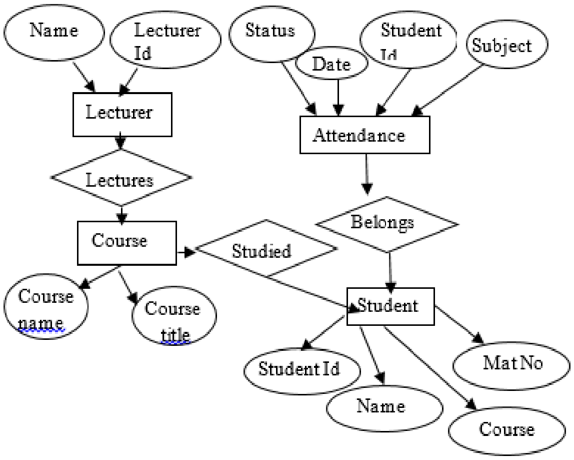


Figure 8: Class Diagram

F. ENTITY RELATIONSHIP DIAGRAM (ERM).

This is a graphical representation of the data requirements for a database. It takes all of the parts of a database and puts them in a box and line form.

This is a physical realization of the logical design. Tables, forms and reports were captured and relationships defined among these tables and security constraints set. During the physical design process, the researchers were able to translate the expected schemas into the actual database structures. At this stage, the designer concentrated on the use of the relation data model where data and relationships are represented as tables each of which has a number of columns with unique names like entities became tables in SQL; attributes became columns in SQL database.



IV. CONCLUSION

Having presented a biometric identity-based fingerprint scheme. One have utilized, extended and implemented ideas in the areas of error corrected string construction from biometric data, key generation, and pairing based fingerprint schemes to form the components of the system. The research presented the application of such a scheme to repudiation situations. Discussion on advantage of using the biometric data in the public key and described the utility of using biometric evidence in disputes that may arise. This work has been an insight into the hidden problems; the manual attendance system tends within daily activities. The problems are fair and need computerized authentication system to replace the manual student attendance system.

V. RECOMMENDATION

Through analysis of the data and research conducted for this study, the school district maintains or develop strict guidelines for student attendance and monitor factors that could hinder a student from attending school on a regular basis. The use of encryption for files in the database transit is an area of protection that should be visited. One strongly believe in protection. Window based authentication system is an important management tool which reduces the lecturers/teachers work load of colleges and university. Therefore, is highly recommended that all schools should adopt it. The system was designed to ease the lecturer work and also allow lecturer and students to use the system without taking special training for it. Should any modification or upgrading arises it should be done with the idea of making it a user friendly so as to make it easily accessible to users, efficient and readily available to specified user.

VI. REFERENCES

Chikkerur, S., Cartwright, A. N., &Govindaraju, V. (2005). Fingerprint enhancement using STFT analysis. Pattern recognition, 40(1), 198-211.

Hsieh, C. T., Lai, E., & Wang, Y. C. (2003). An effective algorithm for fingerprint image enhancement based on wavelet transform. Pattern Recognition, 36(2), 303-312.

Paul, A., &Lourde, R. (2006). A Study on Image Enhancement Techniques for Fingerprint Identification. 2006 IEEE International Conference on Video and Signal Based Surveillance. https://doi.org/10.1109/AVSS.2006.14

Ye, Z., Mohamadian, H. and Ye, Y. (2007) Information Measures for Biometric identification via 2D Discrete Wavelet Transform, Proceedings of the 3rd Annual IEEE Conference on Automation Science and Engineering Scottsdale, AZ, USA, Pp. 22- 25

Farina, A., Kovacs-Vajna, Z. M., & Leone, A. (2008). Fingerprint minutiae extraction from skeletonized binary images. Pattern recognition, 32(5), 877-889.

Maio, D., &Maltoni, D. (2007). Direct gray-scale minutiae detection in fingerprints. IEEE transactions on pattern analysis and machine intelligence, 19(1), 27-40.

Jain, A. K., Hong, L., Pankanti, S., &Bolle, R. (1997). An identity-authentication system using fingerprints. Proceedings of the IEEE, 85(9), 1365-1388.

Gabor, D. (2002) Theory of communication, Journal of IEE, 92, 429-457.

Hong, L., Jain, A.K., Pankanti, S. and Bolle, R. (2004) Fingerprint enhancement IEEE, 5, 202-207.

Yang, J., Liu, L., Jiang, T., & Fan, Y. (2003). A modified Gabor filter design method for fingerprint image enhancement. Pattern Recognition Letters, 24(12), 1805-1817.

Greenberg, S., Aladjem, M., Kogan, D., & Dimitrov, I. (2000). Fingerprint image enhancement using filtering techniques. In Pattern Recognition, 2000. Proceedings. 15th International Conference on (Vol. 3, pp. 322325). IEEE.

Wu, C., Shi, Z., &Govindaraju, V. (2004). Fingerprint image enhancement method using directional median filter. In Proceeding of SPIE (Vol. 5404, p. 67).

Teddy, K. and Martin, L. (2002) Fingerprint Enhancement by spectral analysis techniques, Proceedings of 31st Applied Imagery Pattern Recognition workshop, Pp 133-139.

Yun, E. K., & Cho, S. B. (2006). Adaptive fingerprint image enhancement with fingerprint image quality analysis. Image and Vision Computing, 24(1), 101-110.

Chengpu, Y., Mei, X. and Jin, Q. (2008). An effective and robust fingerprint enhancement method, IEEE International Symposium on computational Intelligence and Design, 7, 110113.

Wang, W., Li, J., Huang, F., & Feng, H. (2008). Design and implementation of LogGabor filter in fingerprint image enhancement. Pattern Recognition Letters, 29(3), 301–308. https://doi.org/10.1016/j.patrec.2007.10.004.

Yuanyuan, Z. (2012). Fingerprint image enhancement based on elliptical shape Gabor filter. 2012 6th IEEE International Conference Intelligent Systems, 344–348. https://doi.org/10.1109/IS.2012.6335240.

Babatunde, I. G. (2012). Fingerprint Image Enhancement: Segmentation to Thinning.(IJACSA) International Journal of Advanced Computer Science and Applications, 3(1), 15–24.

He, Y., Tian, J., Luo, X., & Zhang, T. (2003). Image enhancement and minutiae matching in fingerprint verification. Pattern Recognition Letters, 24(9–10), 1349–1360. https://doi.org/10.1016/S0167-8655 (02)00376-8.

Emerging Trends in Artificial Intelligence and Machine Learning: Historical and State-of-the-Art Perspectives

Edward E. Ogheneovo
Department of Computer Science
University of Port Harcourt
Port Harcourt, Nigeria.
edward.ogheneovo@uniport.edu.ng

ABSTRACT

The growing importance and relevance of artificial intelligence (AI) in every field of human endeavor ranging from business, healthcare, education, agriculture, bioinformatics, cybersecurity, military, etc., is on the increase. Since the start of the 21st century, many businesses have realized that artificial intelligence and machine learning will increase the potential at which people will do their businesses. This is why businesses and organizations such as Google, Facebook, Amazon, etc., are investing heavily in these areas especially with the advent of Big data regime, so that they can stay ahead of their competitors. The application and use of Artificial intelligence have become so important that one cannot but discuss it. AI has become more popular these days due largely to increased data volumes, advanced algorithms, and improvements in computing powers and storage. AI makes it possible for machines to learn from experience, adjust to new inputs and perform human-like tasks such as chess-playing, robotics used in war zones and manufacturing processes, self-driving cars, all of which rely on deep learning and natural language processing where computers are trained to accomplish specific tasks by processing large amounts of data and recognizing patterns in the data. This paper discusses the emerging trends in artificial intelligence (AI) and Machine learning by tracing the history their histories and state-of-the-art of the two concepts. It also discusses other concepts such as deep learning, machine learning, facial recognition, privacy policy, AI-enabled and embedded chips, cloud especially cloud AI, brainware and improved data analysis.

Keywords: artificial intelligence, machine learning, deep learning, cloud computing, brainware.

I. INTRODUCTION

The growing importance and relevance of artificial intelligence (AI) in every field of human endeavor ranging from business, healthcare, education, agriculture, bioinformatics, cybersecurity, military, etc., is on the increase [1] [2]. The application and use of Artificial intelligence have become so important that one cannot but discuss it. AI has become more popular these days due largely to increased data volumes, advanced algorithms, and improvements in computing powers and storage. AI makes it possible for machines to learn from experience, adjust to new inputs and perform human-like tasks such as chess-playing, robotics used in war zones and manufacturing processes, self-driving (autonomous) cars, all of which rely on deep learning and natural language processing where computers are trained to accomplish specific tasks by processing large amounts of data and recognizing patterns in the data [3].

Artificial Intelligence (AI) is defined as any technique that enables computers to mimic human intelligence using logic, if-then rules, decision trees, and machine learning (including deep learning) [4]. Machine Learning (ML) is a subset of AI that includes abstract statistical techniques that enable machines to improve at tasks with experience. The category includes deep learning [5]. Deep Learning is the subset of machine learning composed of algorithms that permit software to train itself to perform tasks like speech and image recognition, exposing multi-layered neural networks to vast amount of data [6]. The relationship between these concepts is shown in figure 1.

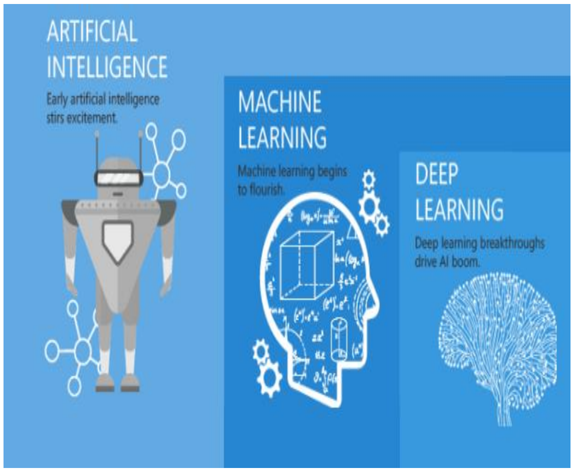

ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.
MACHINE LEARNING
Machine learning begins to flourish.
DEEP LEARNING
Deep learning breakthroughs drive AI boom.

Fig. 1: Relationship between Artificial Intelligence, Machine Learning, and Deep Learning, **Source:** Kumar, S. (2018)

The term machine learning was coined by Arthur Samuel in 1959 while working at International Business Management (IBM) [7][8]. Prior to this time, in the Pre – 1940's, there were lots of mathematical theorems in statistics underpinning modern machine learning. Among these are the works of Thomas Bayes (Bayes Theorem) in 1812, which defines the probability of an event based on prior knowledge, the least squares of conditions that might be related to it, which was developed by Adrien – Marie Legendre for data fitting in 1905, and Adrey Markov who developed the Markov Chains in 1913 for analysis techniques. These techniques are all fundamental to modern day artificial intelligence and machine learning. In the late 1940s the concept of stored program computers that holds instructions (programs) in the same memory used for data were developed. These computers include EDVAC, and Mark 1 both developed in 1949, EDVAC in 1951 [9].

In 1950, Alan Turing published Computing Machinery and Intelligence with the question "Can machine think?" This paper was one of the first attempt to describe how 'artificial' intelligence could be developed based on the power of computers. Turing's paper proposed the "imitation game", a test game to determine whether a computer is intelligent by asking a person to distinguish between a human and a computer when communicating with both of them through typed messages [10]. In 1951, Marvin Minsky and Dean Edmonds **[12]** built the first artificial neural network – a computer-based simulation of the organic brains work. In 1952, Arthur Samuel at IBM laboratory started working on some of the first machine learning programs by creating programs checker player. However, it was in 1959 that Arthur Samuel finally coined the name 'Machine Learning.' In 1957, Frank Resenblatt invented the perceptron while working at the Cornell Aeronautical Laboratory. The invention of the perception generated great excitement and was widely used in the media at the time. In 1967, the nearest neighbor algorithm was created, allowing computers to begin using basic pattern recognition. The nearest neighbor algorithm was used to map a route for traveling salesmen, starting at a random city but ensuring that all cities are visited during a short tour [13][14].

In 1970, Seppor Linnainmaa publishes the general method for automatic differentiation (AD) of discrete connected networks of nested differentiable functions which is similar to the concept of present day back – propagation. In 1979, students at Stanford University developed a Cart which was able to navigate and avoid obstacles in a room, and in 1980, the concept of Neocognition was proposed by Kenihiko Fukushima. Neocognition is a type of artificial neural network (ANN). The concept of neocognition later inspired researchers in the area of convolutional neural networks (CNNs) [15]. Gerald Dejon in 1981 introduced the concept of Explanation Based Learning (EBL), in which a computer was able to analyze training data and creates a general rule it can follow by discarding unimportant data. In 1985, Terry Sejnowski invented NETalk, a program that learns to pronounce words the same way a baby does. In 1989, Q – learning was used to greatly improve the practicality and feasibility of reinforcement learning. In 1989, Ancelis Inc. releases Evolver, the first software package to commercialize the use of genetic algorithms on personal computers. This marked the beginning of commercialization of machine learning on personal computers [16][17].

In 1990s works on AI and machine learning shifted from a knowledge – driven approach to a data – driven approach. Scientists begin to create programs for computer to analyze large amounts of data and draw conclusion – or "learn" – from the results. In 1992, Gerald Tesauro developed TD – Gammon, a computer backgammon program that uses an artificial neural network (ANN) trained using temporal – difference learning. In 1995, the Support Vector machines (SVMs) was developed. SVMs are supervised machine learning algorithms for analyzing data used for classification and regression analysis. In 1997, IBM developed the IBM Deep Blue a chess – playing computer was able to beat the world champion in a chess competition. In 1998, at AT & T Laboratories, a team of researchers developed MNIST database, a dataset comprising a mix of handwritten digits, a digit recognition which was later used for good accuracy in detecting handwritten postcodes from the US Postal Service. The software used back – propagation, a neutral network model for optimizing learning by feeding “training data” with correct output into the neural network [18].

II. ARTIFICIAL INTELLIGENCE (AI) AND MACHINE LEARNING (ML) IN THE 21ST CENTURY

Since the start of the 21st century, many businesses have realized that artificial intelligence and machine learning will increase the potential at which people will do their businesses. This is why businesses and organizations such as Google, Facebook, Amazon, etc., are investing heavily in these areas especially with the advent of Big data regime, so that they can stay ahead of their competitors. In 2006, Netfix released a software that use machine learning to beat Netfix’s own recommendation software’s accuracy in predicting a user’s rating for a film. Also, in 2006, Geoffrey Hiton coined the term Deep Learning to explain new algorithms that let computers “see” and distinguish objects and text in images and videos. In 2009, ImageNet, a large visual database envisioned by Fei – Fei Li from Stanford University was created. Fei – Fei Li realized that the best machine learning algorithms would not work well for real world data [19][20].

Today, ImageNet has been regarded by many AI and many machine learning researches as the catalyst for the AI boom in the 21st century. In 2012, the Google Brain team, led by Andrew Ng and Jeff Dean, developed a neural network that learns to recognize cats by watching unlabeled images. Thus this deep neural network focused mainly on pattern detection in images and video. The Google Brain was able to use Google’s resources, which made it incomparable to much smaller neural networks. It was later used to detect objects in YouTube videos. Also, in 2012, the use of Graphical Processing Unit (GPU) and Convoluted Neural Networks (CNNs) in machine learning was used by AlexNet to win ImageNet competition by a large margin. They also created ReLU, an activation function that greatly improves efficiency of CNNs. GPUs are Graphical Processing Units that have about 200 times more processors per chip than CPUs. They are very important in the world of machine learning.[21]

In 2014, Deep Neural Network was created by Facebook, it is a tool used for face recognition called DeepFace. According to them, DeepFace can recognize people with the same precision as a human can. In that same year, DeepMind, a company bought by Google, is capable of playing basic video games to the same levels as humans [22]. In 2015, Amazon Machine Learning Platform, a part of Amazon Web Services, showed how most big companies want to get involved in machine learning. According to them, machine learning drives many of their internal systems, from regularly used services such as search recommendations and Alexa, to more experimental ones like Prime Air and Amazon Go. U–net was created in 2015. It is a CNN architecture that specialized in biomedical image segmentation. It used an equal amount of up sampling and down sampling layers as well as skipping of connections. In 2016, it was able to beat a professional at the game Go, which is considered to be one of the world’s most difficult board game. Computer Go program to bear an unhandicapped professional player used combination of machine learning and tree search technique. It was later improved as Alpha Zero and then in 2017 to chess and more two – player game with Alpha Zero [23].

Also, in 2016, natural language processing gives life to a digital personal shopper. The North face became the first retailer to use IBM Watson’s natural language processing in a mobile application. The Expert Personal Shopper helps consumers find what they are searching for through conversation just as human sales representative would. In 2018, Generative Adversarial Networks (GAN), a type of unsupervised deep learning system implemented as two competing neural networks was produced. In GAN, one network, the generator, creates take data that looks exactly like the real data set, while the other, the discriminator ingests real and synthetic data. Thus each network improves, thus enabling the pair to learn the entire distribution of the given data sets. In 2018, deep learning theory was used to explain the principle of deep neural networks and how it works by trying to mimic the human brain and their ability to “learn” from images, audio, and text-data. Deep learning also enables greater development and use by providing insight into optimal network design and architecture choices, while providing increased transparency for safety – critical or regulatory applications.

The concept of capsule networks was also proposed in 2018. Capsule network is a new type of deep neural network. It is used to process visual information in a similar way to human brain, thereby, maintaining hierarchical relationships. This a huge leap from convolutional neural networks which fails to take into account important spatial hierarchies between simple and complex objects, resulting in mis-classification and a high error rate. Another major development in AI and machine learning is the concept of deep reinforcement learning, (DRL) a technique which combines reinforcement learning with deep neural networks to learn by interacting with the environment. Deep reinforcement is a type of neural network that learns by interacting with the environment through observation, actions, and reward. Deep reinforcement learning has been used to learn gaming strategies such as Atari, Go, and AlphaGo program that beat a human champion. DRL is the most general purpose among all learning techniques and has been used in most business applications. It requires less data the other learning techniques. Also, it can be trained via simulation, thus eliminating the need for labeled data completely.

Another major technological breakthrough of machine learning in 2018 is the concept of Lean and Augmented Learning (LADL). LADL is a combination of techniques that enable a model to learn from less data or synthetic data. The biggest challenge in machine learning, deep learning in particular, is the availability of large volumes of labeled data to train the system. Two major techniques are used to address this challenge: 1) synthesizing new data, and 2) transferring a model trained for one task or domain to another. Techniques such as transfer learning (i.e., transferring the insights learned from one task/domain to another) or one – shot learning (i.e., transfer Learning taken to the extreme with learning occurring with just one or no relevant examples) making them "learn data" learning technique. Similarly, synthesizing new data through simulations or interpolations helps obtain more data, thereby augmenting existing data to improve learning. Using these techniques, it is possible to address a wider variety of problems that especially those with less historical data [23].

Probabilistic Programming, a high – level language that makes it easy for developers to define probability models is another major technological breakthrough in 2018. Probabilistic programming is a high – lever programming language that more easily enables a developer to design probability models and then automatically "Solve" these models. Probabilistic programming languages make it possible to reuse model liberaries, support interactive modeling and formal verification, and provide the abstraction layer necessary to foster generic, efficient inference in universal model classes. Probabilistic programming languages have the ability to accommodate the uncertain and incomplete information that is very common in the business domain. Hybrid Learning Models, an approach that combines different types of deep neural \ networks with probabilistic approaches to model uncertainty is one of the highlights in machine learning in 2018.

In hybrid learning models, different deep neural networks such as generative adversarial network (GANs) or deep reinforcement learning (DRL) have shown great promise in terms of their performance and widespread application with different types of data. However, deep learning models do not model uncertainty, the way Bayesian, or probabilistic approaches do. Hybrid learning models combines the two approaches to leverage the strengths of each. Some examples of hybrid models are Bayesian deep learning, Bayesian GANs, and Bayesian conditional GANs. Hybrid learning models make it possible to expand the variety of business problems to include deep learning with uncertainty. Thus can help business owners achieve better performance and explain ability of models, which in turn could encourage more widespread adoption. Another major breakthrough in machine learning in 2018 is the concept of automated machine learning, a technique for automating the standard workflow of machine learning. Developing machine learning models is time consuming and expert driven workflow, which includes data preparation, feature selection, model or technique selection, training, and tuning. Automated Machine Learning (Auto ML) aims to automate the workflow using a number of different statistical and deep learning techniques. Auto ML is part of what is seen as a democratization of AI tools, enabling business users to develop machine learning models without a deep programming background which will help speed up the time required for data scientists to create models [24].

Digital Twin, a virtual model used for facilitating detailed analysis and monitoring of physical or psychological systems was another milestone achievement in the development of machine learning in 2018. The concept of the digital twin originated in the industrial world where it has been widely used to analyze and monitor things like windmill farms or industrial systems. With agent – based modeling (i.e. computational models for simulating the actions and interactions of autonomous agents) and system dynamics (a computer – aided approach to policy analysis and design), digital twins are being applied to nonphysical objects and processes including predicting customer behavior. Digital twins can help spar the development and broader adoption of the internet of things (IoTs), providing a way to predictable diagnosis and maintain IoT systems [25].

Explainable Artificial Intelligence (Explainable AI), a machine learning technique that produce more explainable models while maintaining high performance was another major breakthrough in the development of machine learning in 2018. Today, there are scores of machine learning algorithms in use that sense, think, and act in a variety of applications. Yet many of these algorithms are considered as "black boxes" offering little or no insight into how they reached their outcome. Explainable AI is a movement to develop machine learning techniques that produce more explanations to models while maintaining prediction accuracy. AI that is explainable, provable, and transparent will be critical to establishing trust in the technology and will encourage wider adoption of machine learning techniques. Enterprises will adopt explainable AI as a requirement or best practice before embarking on widespread deployment of AI, while governments may make explainable AI a regulatory requirement in the future.

A. New Trends in Artificial Intelligence and Machine Learning

In recent times, artificial intelligence and machine learning have deeply involved in a number of groundbreaking researches and discoveries in many areas. These include facial recognition, security and privacy policies, AI-enabled chips, cloud maturity, deep learning, etc.

Facial Recognition

Facial recognition is a technology that enables the recognition of human faces and detects facial features and expressions. This technology helps to identify or verify a person from a digital image or a video frame from video. It does this by creating a template of someone's image as shown in figure 2. Facial recognition helps capture the picture of a person by creating an image or template

Fig. 2: Facial Recognition, **Source: [26]**

B. Improved Cyber Security and Privacy Policy

Artificial intelligence and machine learning have greatly improved cybersecurity and privacy policies of businesses and organizations in recent time through the automation of complex processes that detect cyber-attacks and cyber-crimes and other security breaches. This is made possible as a result of improved incident monitoring of cyber criminals. With improved incident monitoring, speed of detection and the resulting monitoring are very crucial to mitigating damages that may be caused by these cyber criminals. Figure 3 shows a typical example of privacy policy of an organization.

Fig. 3: Privacy Policy, **Source:** [27]

Thus artificial intelligence and machine learning provides automated responses to various forms of cyber-attacks without human intervention. Artificial intelligence and machine learning will also provide countermeasures to reduce the incidence of counterfeiting for both online and physical commerce to spot incidence of face and adulterated products.

III. CONVERGENCE OF ARTIFICIAL INTELLIGENCE (AI) AND INTERNET OF THINGS (IOTS)

There convergence of AI and IoTs in recent time. Presently, AI and IoTs are very important research areas and virtually every technology is embedded with AI and IoTs software. AI and IoTs have met at the edge of computing layer and have converged. AI and Machine learning are already used for root cause analysis for automatic detection of device problems by advanced learning machine models and neural networks are being used for speech synthesis. They are also used for analyzing video frames, time-series data and unstructured data from IoTs devices. The collaboration is expected to leverage the use of distributed systems in the near future [28].

AI-Enabled Chips

Artificial Intelligence (AI) depends on specialized processors. Applications such as speech recognition and natural language processing (NLP) requires high processor speeds. Ordinary central processing units (CPUs) cannot run these high-speed AI applications. They require AI-enabled chips that are very robust and fast to be able to perform optimally and efficiently. These specialized chips are now being produced by International Business Management (IBM) and Intel by optimizing them to deal with specific applications that require high processor speed. Figure 4 shows an example of AI-Enabled Chips [29].

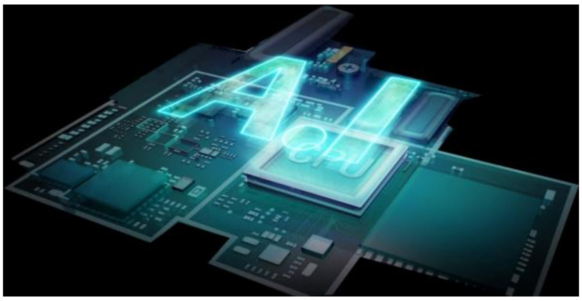


Fig. 4: AI-Enabled Chips

Improved Data Analysis

We live in the era of big data. The amount of data collected on daily basis in businesses worldwide is massive. These data need to be proper analyses and interpretation to guide manager and business executives make proper decisions as it concerns their businesses. Emerging AI and machine learning trends have made it possible and easier for large organizations to collect, store and analyze these data. During the process of data analysis, massive computation and patterns determined and recognized on the data and recommendations for the most relevant data. From the patterns discovered, such data provide information for reuse and business prediction in future analysis and management purposes [29].

Cloud Maturity

Amazon EC2 C5/C5d and M5/M5d instances are built on the Nitro system. This is a collection of AWS-built hardware and software components that enable high performance, high availability, high security, and bare metal capabilities to eliminate virtualization overhead. Based on these enhancements, Nitro system EBS-optimized bandwidth has increased significantly from 9Gbps for C5/C5d and 10Gbps for M5M5d respectively to 14Gbps. Other metrics have also increased considerably. This performance increase enables a user to speed up the parts that are workflows dependent on EBS-optimized instance performance thereby saving costs and users are now able to handle unplanned tasks in EBS-optimized instance demand without major impacts on user's application performance. Figure 5 shows a typical Google cloud AI. It is built on AI tensor process technology [30].

Fig. 5: Google Cloud AI
Source: Cloud TPU, https://cloud.google.com/tpu/

Machine learning has produced business and research breakthroughs ranging from network security to medical diagnoses. Today, Tensor Processing Unit (TPU) are built to make sure that everyone is able to achieve similar breakthroughs. Cloud TPU is the custom-designed machine learning ASIC that powers Google products like Translate, Photos, Search, Assistant, and Gmail. Here's how you can put the TPU and machine learning to work to increase speed and scalability.

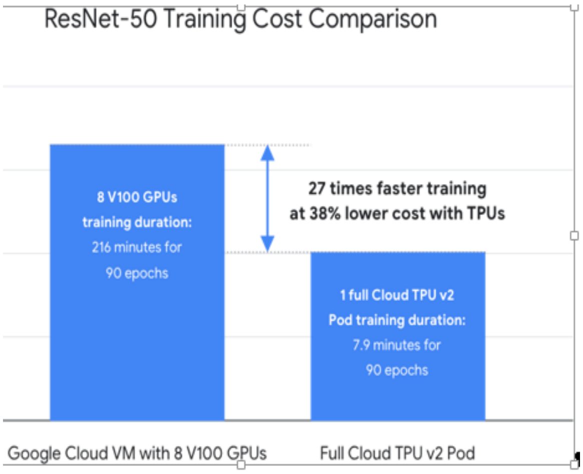


Fig. 6: Google Cloud Virtual Machine
Source: Cloud TPU

Figure 6 shows the Google Cloud Virtual Machine with 8 V100 GPUs versus 1 full Cloud TPU v2 Pod. With the Google Cloud Virtual Machine with 8 V100 GPUs, it can be seen that the training performance is 27 times faster and a much lower training costs of 38% reduction. The implication is that it will become cheaper and affordable foe companies and small organizations to adopt cloud computing more and its deployment and adoption has made it cheaper thus making AI to become customers friendly and machine learning user friendly.

IV. MACHINE LEARNING AND DEEP LEARNING

Deep Learning is the subset of machine learning composed of algorithms that permit software to train itself to perform tasks like speech and image recognition, exposing multi-layered neural networks to vast amount of data. Figure 7 shows the diagram of how deep learning works. With the introduction of the concepts such as artificial intelligence, machine learning and deep learning, there virtual agents in company's websites which now responds to customer's need. These agents provide human-like customized help to clients. These virtual agents rely on AI systems to provide answers to customer's frequently asked questions using the concept of machine learning and deep learning. The answers to these questions are used to predict future conversations and market trends.

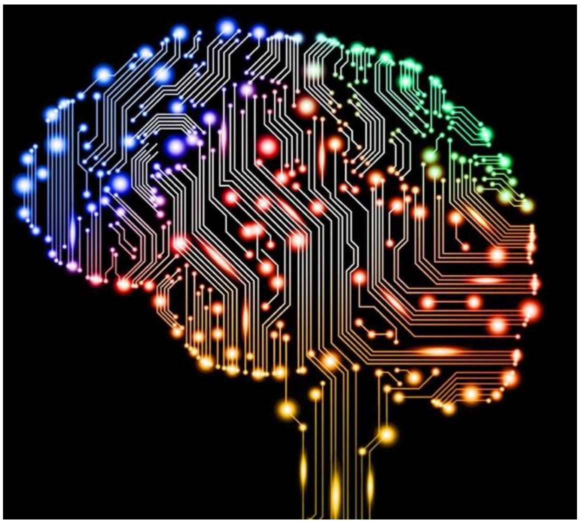


Fig. 7: Deep learning

Figure 7 shows an example of a deep learning. Deep learning help customer to be able to understand and predict business patterns more reliably. With machine learning and deep learning, customers companies now assign repetitive customer service tasks to virtual agents which has helped to reduce the costs of doing businesses drastically. Also, machine learning and deep learning are expected to increase the rate innovation and enable virtual agents to perform better in terms of query handling, customer service, better performance, costs reduction, etc.

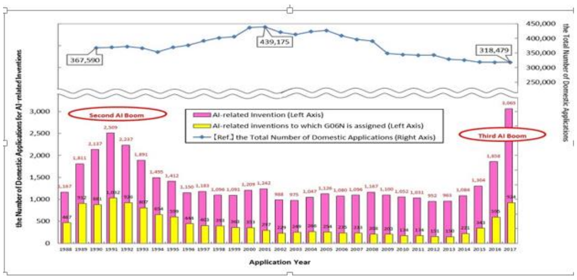


Fig. 8: AI-related inventions (1988-2017)
Source: [31]

Figure 8 shows the graph of AI-related inventions between 1988 and 2017. There was the second AI boom in 1991 and again in 2017 during which time there was drastic increase in the number of domestic applications of AI related inventions.

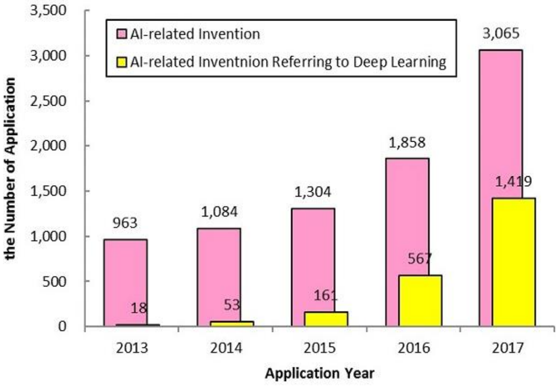


Fig. 9: AI-related inventions (2013-2017)
Source: [31]

Figure 9 shows the actual number of these applications and their years. It can be seen that there was a gradual rise of domestic applications of AI-enabled or AI-related inventions from 963 in 2013 to 1, 858 in 2016 and a drastic increase in 2019. Also, in deep learning, there was a drastic increase from 2014 to 2017 and the trend is expected to continue due to the importance of artificial intelligence, machine learning and deep learning in businesses and organizations due to global interest in these fields.

V. CONCLUSION

In recent years, there has been a growing trend in AI of attention from policy makers, universities, researchers, corporations, media, and the public. Driven by advances in big data and computing power, breakthroughs in AI and machine learning research and technology seem to happen about on daily basis. The growing importance and relevance of artificial intelligence (AI) in every field of human endeavor ranging from business, healthcare, education, agriculture, bioinformatics, cybersecurity, military, etc., is on the increase. Since the start of the 21st century, many businesses have realized that artificial intelligence and machine learning will increase the potential at which people will do their businesses. This is why businesses and organizations such as Google, Facebook, Amazon, etc., are investing heavily in these areas especially with the advent of big data regime, so that they can stay ahead of their competitors. The application and use of Artificial intelligence have become so important that one cannot but discuss it. AI has become more popular these days due largely to increased data volumes, advanced algorithms, and improvements in computing powers and storage. AI makes it possible for machines to learn from experience, adjust to new inputs and perform human-like tasks such as chess-playing, robotics used in war zones and manufacturing processes, self-driving cars, all of which rely on deep learning and natural language processing where computers are trained to accomplish specific tasks by processing large amounts of data and recognizing patterns in the data. This paper discusses the emerging trends in artificial intelligence (AI) and Machine learning by tracing the history their histories and state-of-the-art of the two concepts. It also discusses other concepts such as deep learning, machine learning, facial recognition, privacy policy, AI-enabled and embedded chips, cloud especially cloud AI, brainware, etc.

REFERENCES

1. M. S. V, Janakrina (2018). "5 artificial intelligence trends to watch out for in 2019." **https://www.forbes.com/sites/janakirammsv/2018/12/09/5-artificial-intelligence-trends-to-watch-out-for-in-2019/#65f468d05618.** Retrieved 20th September, 2019.
2. Techopedia: artificial intelligence (AI). https://www.techopedia.com/definition/190/artificial-intelligence-ai. Retrieved 12/09/2019.
3. M. A. Aceves-Fernandez, (2018). "Artificial intelligence - emerging Trends and applications". DOI: 10.5772/intechopen.71805.
4. Patent Examination Office, "Recent trends in AI-related inventions – report (2017)". patent examination department (electronic technology). Japan Patent Office. https://www.jpo.go.jp/e/system/patent/gaiyo/ai/ai_shutsugan_chosa.html. Retrieved 20/10/2019.
5. A. Gaggiolo (2018). "Virtual personal assistants: an emerging trend in artificial intelligence." *Cyberpsychology, Behavior, and Social Networking*, vol. 21, no. 12, pp. 803–804.
6. C. Erik and Bebo, W. (2014). "Jumping NLP curves: a review of natural language processing research." IEEE Computational Intelligence Magazine, vol. 9, no. 2, pp. 48-57.
7. *C. Cesar, Luca, C., Carrillo, H., Yasir, L., Davide, S., Jose, N., Ian, R., Leonard, J. J. (2016). "Past, present, and future of simultaneous localization and mapping: toward the robust-perception age". IEEE Transactions on Robotics. vol. 32, no. 6, pp. 1309-1332.*
8. B. Scassellati (2002). "Theory of mind for a humanoid robot". Autonomous Robots, vol. 12, no. 1, pp. 13-24.
9. C. Yongcan, Y. Wenwu, R. Wei and C. Guanrong (2013). "An overview of recent progress in the study of Distributed multi-agent coordination". IEEE Transactions on Industrial Informatics, vol. 9, no. 1, pp. 427-438.
10. P. Soujanya, C. Erik, B. Rajiv and H. Amir (2017). "A review of affective computing: from unimodal analysis to multimodal fusion". Information Fusion, vol. 37, pp. 98-125.
11. C. Pennachin, C.; Goertzel, B. (2007). "Contemporary approaches to artificial general intelligence. Artificial General Intelligence. Cognitive Technologies. Cognitive Technologies". Berlin, Heidelberg: Springer.
12. J. Roberts (2016). "Thinking machines: the search for artificial intelligence". Distillations, vol. 2, no. 2, pp. 14-23.
13. *M. Volodymyr, K, Koray, K., David, S., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, vol. 518, no. 7540, pp. 529-533.*
14. B. Goertzel, Lian, R., Arel, I., de Garis, H. and Chen, S. (2010). "A world survey of artificial brain projects, Part II: Biologically Inspired Cognitive Architectures. Neurocomputing". vol. 74, issue (1-3), pp. 30-49.
15. M. Hutson (2018). "Artificial intelligence faces reproducibility crisis". Science, pp. 725-726.
16. A. Lieto, Lebiere, C. and Oltramari, A. (2018). "The knowledge level in cognitive architectures: current iimitations and possible developments". Cognitive Systems Research, vol. 48, pp. 39-55.
17. A. Lieto, Bhatt, M., Oltramari, A. and Vernon, D. (2018). "The role of cognitive architectures in general artificial intelligence". Cognitive Systems Research, vol. 48, pp. 1-3.
18. A. Miller (1986). "Expert systems: The structure, history, and future of successful AI application." Artificial Intelligence: A Modern Approach. Simon & Schuster pp. 22 – 23.
20. B. K. Wong, J. A. Monaco (1995). "Expert system application in business: A review and analysis of the literature (1977 – 1993)." Information and Management, vol. 29, no. 3, pp. 141 – 152. DOI: 10.1016/0378 – 7206 (95) 00023 – P.
21. M. S. V. Janakiram (2018). "5 artificial intelligence trends to watch out for in 2019." https://www.forbes.com/sites/janakirammsv/2018/12/09/5-artificial-intelligence-trends-to-watch-out-for-in-2019/#65f468d05618. 11/10/2019.
22. S. Nadella (2018). "Vision keynote: intelligent cloud and intelligent edge." https://www.microsoft.com/en-us/research/video/vision-keynote-intelligent-cloud-and-intelligent-edge/.Retrieved 13/10/2019.
23. B. Nushi and E. Kumar (2019). "Creating better AI partners: A case for backward compatibility.".https://www.microsoft.com/en-us/research/blog/creating-better-ai-partners-a-case-for-backward-compatibility/. Retrieved 11/10/2019.

24. S. Young (2019). "10 Trends of artificial intelligence (AI) in 2019." https://becominghuman.ai/10-trends-of-artificial-intelligence-ai-in-2019-65d8a373b6e6. Retrieved 12/10/2019.
25. A. Sheidon (2019). "Top 5 trends of artificial intelligence (AI) in 2019." https://hackernoon.com/top-5-trends-of-artificial-intelligence-ai-2019-693f7a5a0f7b. Retrieved 12/10/2019.
26. D. Weldon (2018). "8 top artificial intelligence and analytics trend for 2019." https://www.information-management.com/list/8-top-ai-and-analytics-trends-for-2019.
27.S. Kumar (208)." Importance of artificial intelligence." https://mindmajix.com/why-ai-is-more-important'than-you-think/.
28. D. Burger and T. Way (2018). "Hyperscale hardware: ML at scale on top of Azure + FPGA: Build 2018," Microsoft Developer, May 9, 2018. https://www.microsoft.com/en-us/research/video/hyperscale-hardware-ml-scale-top-azure-fpga/.
29. J. Fowers, K. Ovtcharov, M. Papamicahel, T. Massengill, M. Liu, D. Lo, S. Alkalay, M. Haselman, L. Adama, M. Ghandi, S. Heil, P. Patel, A. Sapek, G. Weisz, L. Woods, S. Lanka, S. Reinhardt, A. Caulfield, E. Chang and D. Burger (2018). "A configurable cloud-scale DNN processor for real-time AI". In Proceedings of the 45th Int'l Symposium on Computer Architecture (ISCA 2018), June 2018.
30. E. Chung, J. Fowers, K. Ovtcharov, A. Caulfield, T. Massengill, M. Liu, M. Ghandi, D. Lo, S. Reinhardt, S. Alkalay, H. Angepart, D. Chiou, A. Fourin, D. Burger, L. Woods, G. Weisz, M. Haselman and D. Zhang (2018). "Serving DNNs in real-time at datacenter scale with project brainwave". IEEE Micro, vol. 38, pp. 8-20.
31. D. Burgar and T. Way (2018). "Hyperscale hardwered: ML at scale on top of Azure + FPGA: Build 2018 Microsoft Developer", May 9, 2018.

A Review on Intrusion Detection System Using Chinese Remainder Theorem

Bukola Fatimah Balogun
Computer Science Department
Kwara State University
Malete, Nigeria.
arinolafula87@gmail.com

Kazeem Alagbe Gbolagade
Computer Science Department
Kwara State University
Malete, Nigeria.
gkazy1@yahoo.com

***ABSTRACT*— Information security has become an indispensable part of an Organization and critical national infrastructure. In today contemporary society, information is prone to attack as a result of the ever evolving and migration of activities to the internet. As network-based computer systems have important roles in modern society, they have become the targets of intruders. Therefore, we need to build the best possible rules and methods to protect our systems. Information security practitioners have recommended Intrusion Detection System (IDS) as a solution to network based computer system security. However, the suggested IDS system suffered a lot from higher detection accuracy. How to improve the detection accuracy of IDS has been serious concern for information security professionals and researchers. This paper reviews the techniques for IDS to provide better understanding among the existing techniques that may help interested researchers to work future in this direction. The empirical results of our reviewed suggested that CRT would assist in improving the detection accuracy of IDS.**

Keywords*— Network based computer, Information security, Intrusion detection system, CRT.

VIII. INTRODUCTION

An intrusion detection system (IDS) is composed of hardware and software elements that work together to find unexpected events that may indicate an attack will happen, is happening, or has happened [1]. Security is becoming a critical part of organizational information systems and security of a computer system or network is compromised when an intrusion takes place. In the field of computer networks security, the detection of threats or attacks is nowadays a critical problem to solve [2]. With the enormous growth of computer networks usage and internet accessibility, more organizations are becoming susceptible to a wide variety of attacks and threats.

As a result, Intrusion Detection System (IDS) proposed by Denning have become an essential element of security infrastructure which is useful to detect, identify these threats and track the intruders. Since then, many research works have been focused on, how to effectively and accurately construct detection models.

As network-based computer systems have important roles in modern society, they have become the targets of intruders. Therefore, we need to build the best possible rules to protect our systems. The security of a computer system is vulnerable when an intrusion takes place. An intrusion can be defined as any action done that harms the integrity, confidentiality or availability of the system. There are some intrusion prevention techniques which can be used to prevent computer systems as a first line of defense. A firewall is also one of it. But only intrusion prevention is not enough. As systems become more complex, there are always exploitable weaknesses in the systems due to design and programming errors, or various penetration techniques. Therefore Intrusion detection is required as another measure to protect our computer systems from such type of vulnerabilities.

Most people use the Internet to communicate these days. Thus, secure communication channel or secure network is highly expected, especially when it involves communicating confidential or sensitive data. Over the last few years, much research has been done on network security area to ensure the data stored and/or transmitted are secure and trustworthy. One of the most common techniques that have been used to ensure secure network is by implementing the intrusion detection system (IDS). IDS is a system or tool that helps the administrator to protect the network from malicious activity. In general, IDS can be classified into two categories: signature based and anomaly detection systems.

In signature-based systems, attack patterns or behaviors of intruder are modeled and the system will alert once a match is detected [3]. The limitation of such system is that it is able to detect known attacks only and requires the attack signatures to be updated frequently. On the other hand, anomaly detection systems first have to create a baseline profile of the normal behavior of the system or network. Afterwards, if any activity deviates from the normal profile, it will be treated as an intrusion. Anomaly detection systems are able to detect previously unknown attacks and thus more efficient than the signature-based. However, in anomaly-based IDS, there is too many attributes need to be considered when identifying specific attacks. The system has to deal with huge amounts of network traffic and with highly imbalanced data distribution, thus making recognizing normal versus abnormal behavior a challenge. With huge traffics, more data has to be considered when forming patterns, otherwise there is increased risk for a high false positive rate. However, despite these challenges, machine learning technique has been proven able to yield high accuracy results in detecting anomalies and thus has been used for the past decades to detect attacks [4]. IDS is one of the components of the defense system used to identify abnormal activities happening in the computer system. Furthermore, IDS is basically a security software that has the capability to detect and automatically warn administrators in the event of someone or something attempting to compromise the information system by using malicious activities or by violating security policy. False positive and false negative are two types of alarm in IDS [5]. The false-positive alarm refers to normal behaviors or instances that are erroneously indicated as an attack. On the other hand, false negative indicates the opposite of false positive where true attacks are detected as normal behaviors or instances. Combinations of these two alarms are measured as the false alarm rate and used to estimate the effectiveness of a particular IDS model during evaluation of the performance.

This paper is organized as follows; Section 2 highlights the related work, intrusion detection approach and is taxonomy is explained in section. The detection approaches and its types are presented in section 3. Section 4 concludes the paper.

IX. RELATED WORKS

In [6] proposed an approach based on a Culture Particle Swarm Optimization algorithm (CPSO). They used the algorithm to optimize the parameters of SVM classifier. They used the colony aptitude of particle swarm and the ability to conserve the evolving knowledge of the culture algorithm to construct the population space based on the particle swarm and the knowledge space.

The authors in [7] presented an IDS based on SVM combined with Particle Swarm Optimization. They used two different PSO algorithms: Standard PSO to seek optimal SVM parameters and Binary PSO to extract the best feature subset. Their model improved detection rate with a high accuracy.

[8] used Support Vector Machine and classification tree Data mining technique for intrusion detection in network. They compared C4.5 and Support Vector Machine by experimental result and found that C4.5 algorithm has better performance in term of detection rate and false alarm rate than SVM, but for U2R attack SVM performs better.

In [9] proposed a SVM-based IDS model, which used hierarchical clustering algorithm and SVM classifier. It was able to minimize the training time and improve the performance of SVM classifier. They applied a simple feature selection procedure to eliminate irrelevant features from the training set. The results showed that the SVM model could classify attacks more accurately.

In [10] used Ant Colony Algorithm to optimize SVM parameters. They tested their model on the KDD CUP99 Dataset and reported that the anomaly detection rate can reach a high accuracy.

[11] proposed Intrusion Detection System using data mining technique SVM (Support Vector Machine) and in their proposed system SVM is used for classification and verification regarding the effectiveness of the proposed system is done by conducting some experiments using NSL-KDD Cup’99 datasets which is improved version of KDD Cup’99 data set. The SVM is one of the most noticeable classification algorithms in the data mining area, but its disadvantage is its large training time. For this drawback they proposed system that carried out some experiments using NSL-KDD Cup’99 data set. The experimental results show that it can reduce extensive time required to build SVM model by performing proper data set pre-processing. Also when they do proper selection of SVM kernel function such as Gaussian Radial Basis Function because of this it produces a result that shows attack detection rate of SVM is increased and False Positive Rate (FPR) is decrease.

In [12] proposed a genetic algorithm to search the genetic principal components that offer a subset of features with optimal sensitivity and the highest discriminatory power. They used SVM for classification. The results show that the proposed model improves SVM performance in Intrusion Detection.

[13] proposed two techniques, C4.5 Decision tree algorithm and C4.5 Decision tree with Pruning, using feature selection approach to IDS. Due to the conventional intrusion detection technology indicates more limitation like low detection rate, high false alarm rate and so on. Performance of the classifier is an essential concern in terms of its effectiveness and also number of feature to be examined by the IDS

should be improved. So, they used the two proposed techniques, of which during their cause of the experiment they projected the following result. In C4.5 Decision tree with pruning they considered only discrete value attributes for classification. They used both the KDDCup'99 and the NSL_KDD dataset to train and test the classifier. The Experimental Result shows that C4.5 decision tree with pruning approach is giving better results with all most 98% of accuracy. As can be seen from the above reviews, none of the work presented showed that CRT algorithm can be adopted to improve the detection rate of NAIDS. Thus, this review recommends CRT as a novel technique for improving the detection accuracy of IDS.

X. INTRUSION DETECTION

Intrusion Detection Systems (IDS) Intrusion activities of computer systems are increasing due to the commercialization of the internet & local networks. An IDS, collects the information and analyze it for uncommon or unexpected events. ID is the process of monitoring and analyzing the events which occurred in a computer system in order to detect signs of security problems[12]. Figure 1 shows a taxonomy of Intrusion Detection Systems. IDS first introduced by James P. Anderson in 1980.

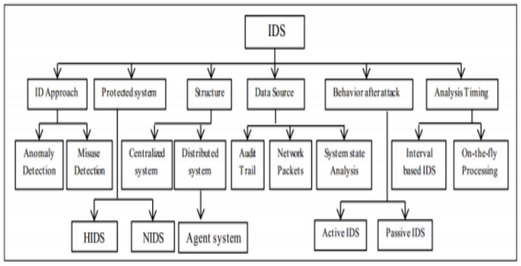


Figure 1. Taxonomy of Intrusion Detection System

A. *Detection Approaches*

Traditionally, intrusion detection techniques are classified into two broad categories:

1) Misuse or signature based IDS: Misuse detection searches for the traces or patterns of well-known attacks which are stored as signatures. These signatures are provided by a human expert based on their extensive knowledge of intrusion techniques. In this process if a pattern matched is found, this signals an event for which an alarm raised [12].

2) Anomaly based IDS: Anomaly detection uses a model of normal user or system behavior and flags significant deviations from this model as potentially malicious. This model of normal user or system behavior is commonly known as the user or system profile [12]. Three broad categories of anomaly detection techniques exist: supervised, semi-supervised and unsupervised anomaly detection techniques [12].

3) Hybrid IDS or compound detection system: It's the combination of both misuse and anomaly detection techniques. It makes decisions using a “hybrid model” that is based on both the normal behavior of the system and the intrusive behavior of the intruders.

B. *Types of Intrusion Detection Systems*

IDSs are categorized according to the kind of input information they analyze. So, this is classified into host-based and network-based intrusion detection system [14]**.**

1) Host-based IDS (HIDS): Host based intrusion detection system mainly deals with single computer and perform intrusion detection based on the system call, kernel, firewall and system logs.

2) Network-based IDS (NIDS): Network intrusion detection system works on a large scale. It monitors network traffic and examines the traffic, and based upon the observation it categorizes the traffic into normal or suspicious. Traffic monitoring is done at firewall, hub and switch etc.

As shown in the Table 1, each of the approaches and types of IDS has its own advantages and disadvantages.

Type	**Advantages**	**Disadvantages**
HIDS	1. It can judge whether or not the host is intruded more accurately. 2. It can detect attacks under encrypted network environment. 3. It does not need additional hardware.	1. Higher cost. 2. It may affect system efficiency of monitoring hosts.
NIDS	1. Low cost. 2. It can detect attacks that cannot be done by host based IDS, such as: Dos, DDos.	1. The flux is large, and some packets may be lost. 2. In a large-scale network, it requires more rapid CPU and more memory space, to analyze bulk data 3. It cannot deal with encrypted packets
Anomaly	1.Ability to detect novel attacks or unknown attacks. 2. Low false negative.	1. Slow Timely Notifications. 2. High false alarm rate. 3. Low detection rate (for known attacks).

Misuse	1.High detection rate and low false alarm rate (for known attacks). 2. Fast Timely Notifications.	1. Detection capacity is low for unknown detection methods. 2. Attack database should be renewed on a regular basis.

XI. DRAWBACKS OF IDS

IDS have become a standard component in security infrastructures as they allow network administrators to detect policy violations. These policy violations range from external attackers trying to gain unauthorized access to insiders abusing their access. Current IDS has a number of significant drawbacks:

i)False positives: A false positive occurs when normal attack is mistakenly classified as malicious and treated accordingly.

ii) False negatives: This is the case where an IDS does not generate an alert when an intrusion is actually taking place.

iii) Data overload: Another aspect which does not relate directly to misuse detection but is extremely important is how much data an analyst can efficiently analyze. That amount of data he needs to look at seems to be growing rapidly.

CONCLUSION

In this paper various techniques are described for the NAIDS that had been proposed in the past few years. This review will be helpful to researchers for gaining a basic insight of various approaches for the IDS. Although, much work had been done using independent algorithms, hybrid approaches are being vastly used as they provide better results and overcome the drawback of one approach over the other. Every day new unknown attacks are witnessed and thus there is a need of those approaches that can detect the unknown behaviour with higher accuracy. In this paper, CRT algorithm is recommended to improve the detection accuracy in future work.

REFERENCES

[1] A. Singh and A. Goyal, "Intrusion Detection System Based on Hybrid Optimization and using Neural Network: A Review," *Ijrece*, vol. 6, no. 3, pp. 1138–1143, 2018, doi: 10.13140/RG.2.2.23285.83681.

[2] P. Amudha, S. Karthik, and S. Sivakumari, "Classification Techniques for Intrusion Detection An Overview," *Int. J. Comput. Appl.*, vol. 76, no. 16, pp. 33–40, 2013, doi: 10.5120/13334-0928.

[3] J. Kim and H. S. Kim, "Intrusion Detection Based on Spatiotemporal Characterization of Cyberattacks," 2020, doi: 10.3390/electronics9030460.

[4] A. Tamilarasan, S. Mukkamala, A. H. Sung, and K. Yendrapalli, "Feature ranking and selection for intrusion detection using artificial neural networks and statistical methods," *IEEE Int. Conf. Neural Networks - Conf. Proc.*, pp. 4754–4761, 2006, doi: 10.1109/ijcnn.2006.247131.

[5] S. Anwar *et al.*, "From intrusion detection to an intrusion response system: Fundamentals, requirements, and future directions," *Algorithms*, vol. 10, no. 2, 2017, doi: 10.3390/a10020039.

[6] E. M. Chakir, M. Moughit, and Y. I. Khamlichi, "An effective intrusion detection model based on svm with feature selection and parameters optimization," *J. Theor. Appl. Inf. Technol.*, vol. 96, no. 12, pp. 3873–3885, 2018.

[7] J. Wang, X. Hong, and R. Ren, "A real-time intrusion detection system based on PSO-SVM," *Int. Work.*, no. Iwisa, pp. 319–321, 2009.

[8] P. Anitha and B. Kaarthick, "Oppositional based Laplacian grey wolf optimization algorithm with SVM for data mining in intrusion detection system," *J. Ambient Intell. Humaniz. Comput.*, 2019, doi: 10.1007/s12652-019-01606-6.

[9] S. J. Horng *et al.*, "A novel intrusion detection system based on hierarchical clustering and support vector machines," *Expert Syst. Appl.*, vol. 38, no. 1, pp. 306–313, 2011, doi: 10.1016/j.eswa.2010.06.066.

[10] J. Pu, Y. Li, L. Xiao, and X. Dong, "A detection method of network intrusion based on SVM and ant colony algorithm," *Proc. 2012 Natl. Conf. Inf. Technol. Comput. Sci. CITCS 2012*, no. Citcs, pp. 153–156, 2012, doi: 10.2991/citcs.2012.237.

[11] G. Hochman, A. Glöckner, and E. Yechiam, "Physiological measures in identifying decision strategies," *Found. Tracing Intuit. Challenges Methods*, vol. 3, no. 3, pp. 139–159, 2009, doi: 10.4324/9780203861936.

[12] I. Ahmad, A. Abdullah, A. Alghamdi, K. Alnfajan, and M. Hussain, "Intrusion detection using feature subset selection based on MLP," *Sci. Res. Essays*, vol. 6, no. 34, pp. 6804–6810, 2011, doi: 10.5897/SRE11.142.

[13] N. G. Relan and D. R. Patil, "Implementation of network intrusion detection system using variant of decision tree algorithm," *2015 Int. Conf. Nascent Technol. Eng. Field, ICNTE 2015 - Proc.*, pp. 3–7, 2015, doi: 10.1109/ICNTE.2015.7029925.

[14] S. Singh, "DOI : 10.5281/zenodo.1218603," vol. 7, no. 4, pp. 339–343, 2018.

A Framework for Intelligent Malware Detection in Mobile Devices

Ekong U.O., Oyong S.B., Obot O.U., Ekong V.E.
Department of Computer Science, University of Uyo, Uyo, Akwa Ibom State, Nigeria.
uyiekong@yahoo.com, samueloyong@gmail.com, abatakure@yahoo.com, victoreekong@uniuyo.edu.ng

ABSTRACT - Malware developers are becoming very slippery in the proliferation of malware and evasion techniques; with consequential damage to systems, data and information theft. To rise to the challenge of malware intrusion, a framework that will analyse, detect and mitigate malware intrusions without human intervention is proposed. Data used in the implementation of the framework is gotten from Kaggle, a public repository. Support vector machine (SVM), k-nearest neighbour (KNN) and Adaboost algorithms are trained to analyse, detect and mitigate intrusion. Preliminary results from the implementation of the framework shows an accuracy of 99.23%. A result much better compared to that of the component classifiers of SVM (97%) and KNN (96%).

Keywords: Intelligent system, Malware Intrusion, SVM, KNN, Adaboost, IDS, IRS.E-commerce

I. INTRODUCTION

An intelligent system is one that interacts with its environment to generate plans, executes those plans, and communicates with other agents [1]. Mobile devices such as cellular phones, smart phones, Tablets, PDAs, sensors are among such systems [2]. Mobile devices provide the following services: Wi-Fi, global system for mobile communication (GSM), Bluetooth, e-commerce, online shopping and sales, e-banking [3], which has made them ubiquitous, portable and convenient household properties. E-commerce is an online business transaction (buying and selling of products and services) carried out using mobile devices and the internet; it is however threatened by malware intrusions [4].

Android operating platform a market lead of 84% share, endeared them to not only individual users, but company users and malware developers. Malware are programs written to exploit the network and standalone systems for monetary gains. A market survey by Privacy Right Clearing House [5] shows that more than 245 million customer records have been stolen since 2005. In [6], it is reported that more than 285 million systems have been compromised into zombies and used to perpetrate the nefarious acts. For instance, in 2010 the 'Zombie' virus infected more than one million smart phones in China, and caused a loss amounting to $300,000 per day [7].

The attacks are perpetrated by agents such as viruses, worms, Trojans, rootkits, key loggers and categorized into denial of service (DOS), probe or surveillance, user to root (U2R) attack, and remote to local (U2L) attacks [8, 9, 10]. Traditional systems such as firewalls, anti-virus programs, and cryptography have not succeeded in providing the much needed protection from these attacks [11, 12, 13]. The objective of this paper is to design, a framework that combines Support Vector Machine (SVM), K-nearest Neighbor (KNN) and Adaboost to not only analyze and detect, but also to mitigate the intrusions. Data that will be used to implement the framework will be gotten from kaggle repository. Min-Max normalization technique will be used to normalize the data before the removal of noisy, redundant data and insertion of missing data by principal component analysis (PCA).

The combined efforts of machine learning techniques could be used to mitigate malware, especially in today's big data collection, analysis and storage that supports decision making in governments, companies, and individuals, with analytic being the driving force. The modern machine learning techniques of interest include intrusion detection systems (IDS) and intrusion response system (IRS), which detects and mitigates attacks in real time.

Intrusion detection systems (IDS) are used to monitor and identify attacks and intrusions in standalone systems, networks and mobile devices. The detected intrusions are reported to the system Administrator or IRS. On the other hand, intrusion response systems (IRS) are designed to monitor system performance, identify and mitigate potential incidents of attack. Unfortunately, existing IRSs adopt static approaches in selecting optimum response and lacks the dynamic approach to solving the problem in real-time [14]. The use of single classifiers has also been identified as a loose way of providing intrusion detection, since unknown malware are not detected in all cases; hybridization is therefore needed to overcome the problem of single models.

Most machine learning (ML) algorithms are designed to use specific data set or tasks; but combining multiple machine learning algorithms can greatly improve the overall result by either helping to tune one another, or generalize or adapt to unknown tasks.

Hybridization is known to improve the functionality of the individual classifiers, however the combinations with intrusion response systems (IRS) to mitigate obfuscation and emulation detection has not been explored Banks and telecommunication operators have enormous potentials for m-commerce, but awareness and patronage is low and the cost of high quality phones is high; and the lack of basic infrastructure and security issues pose a major threat to m-commerce's wide scale implementation. Traditional classification algorithms are designed to work with static features and a set sample size. These algorithms tend to perform very well with structured data, but when data becomes unstructured, the same classifiers become harder to apply. The remaining part of the work are organized as follows; Section II discusses reviewed related works, Section III discusses the methodology utilized, Section IV discusses the preliminary result obtained and Section V draws conclusions and further work.

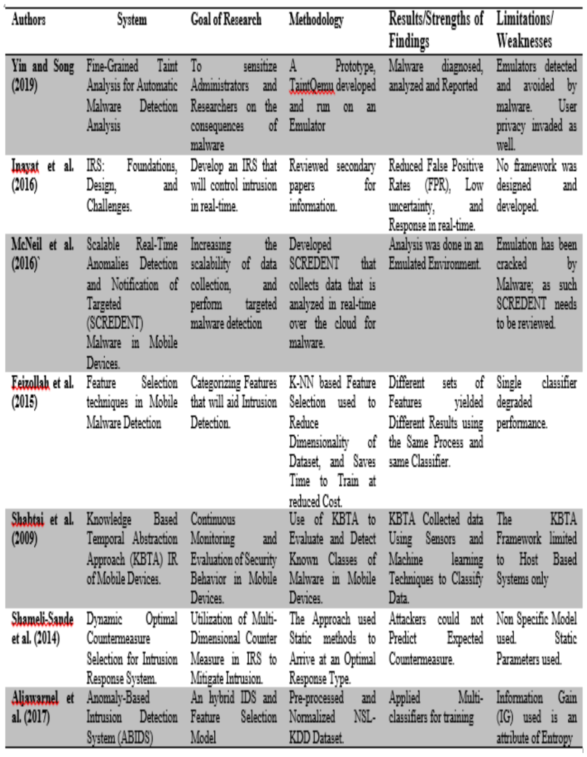
II. RELATED WORKS

Security is the process that ensures that the system's vulnerable points are patched, and the data there-in protected from unauthorized users [15]. Security has the following attributes: confidentiality, integrity, authentication, non-repudiation, availability and access control. Confidentiality is keeping another person's or entity's information private; integrity ensures that data is not modified; authentication verifies the identities of the communicating parties in a transaction. Non-repudiation ensures that communicating parties do not disown their involvement or commitment to a transaction; availability implies that information is available to the authorised parties when required; and access control ensures that only rightful communicating parties have access to the data, through user name and password, digital signature stamped on the file, and finger printing.

Security of communication systems, data, and transmission channels remains a disturbing issue in recent trends of technological developments. This is because malware developers and hackers employ various means of intrusion using sophisticated tools, most of which are freely available on the net. More so, cyber threats to critical infrastructure continue unabated, posing a serious national security challenge. Undetected cyber-attacks force users, companies and governments to incur serious financial losses, information loss; and even the reputations of affected industries are destroyed [16, 17].

To protect and prevent the destruction of security characteristics, anti-malware developers and research community are fighting back using various techniques. One way to fight the menace of malware and its developers is the use of machine learning techniques. A review of some of these approaches are given in Table 1.

Table 1: Reviewed related works



Authors	System	Goal of Research	Methodology	Results/Strengths of Findings	Limitations/ Weaknesses
Yin and Song (2019)	Fine-Grained Taint Analysis for Automatic Malware Detection Analysis	To sensitize Administrators and Researchers on the consequences of malware	A Prototype, TaintQemu developed and run on an Emulator	Malware diagnosed, analyzed and Reported	Emulators detected and avoided by malware. User privacy invaded as well.
Inayat et al. (2016)	IRS: Foundations, Design, and Challenges.	Develop an IRS that will control intrusion in real-time.	Reviewed secondary papers for information.	Reduced False Positive Rates (FPR), Low uncertainty, and Response in real-time.	No framework was designed and developed.
McNeil et al. (2016)`	Scalable Real-Time Anomalies Detection and Notification of Targeted (SCREDENT) Malware in Mobile Devices.	Increasing the scalability of data collection, and perform targeted malware detection	Developed SCREDENT that collects data that is analyzed in real-time over the cloud for malware.	Analysis was done in an Emulated Environment.	Emulation has been cracked by Malware; as such SCREDENT needs to be reviewed.
Feizollah et al. (2015)	Feature Selection techniques in Mobile Malware Detection	Categorizing Features that will aid Intrusion Detection.	K-NN based Feature Selection used to Reduce Dimensionality of Dataset, and Saves Time to Train at reduced Cost.	Different sets of Features yielded Different Results using the Same Process and same Classifier.	Single classifier degraded performance.
Shabtai et al. (2009)	Knowledge Based Temporal Abstraction Approach (KBTA) IR of Mobile Devices.	Continuous Monitoring and Evaluation of Security Behavior in Mobile Devices.	Use of KBTA to Evaluate and Detect Known Classes of Malware in Mobile Devices.	KBTA Collected data Using Sensors and Machine learning Techniques to Classify Data.	The KBTA Framework limited to Host Based Systems only
Shameli-Sande et al. (2014)	Dynamic Optimal Countermeasure Selection for Intrusion Response System.	Utilization of Multi-Dimensional Counter Measure in IRS to Mitigate Intrusion.	The Approach used Static methods to Arrive at an Optimal Response Type.	Attackers could not Predict Expected Countermeasure.	Non Specific Model used. Static Parameters used.
Aljawarnel et al. (2017)	Anomaly-Based Intrusion Detection System (ABIDS)	An hybrid IDS and Feature Selection Model	Pre-processed and Normalized NSL-KDD Dataset.	Applied Multi-classifiers for training	Information Gain (IG) used is an attribute of Entropy

III. METHODOLOGY

In this study, a mobile device is used to communicate with a server, over the internet, for necessary analysis, detection and mitigation of malware intrusion, as illustrated in Figure 1.

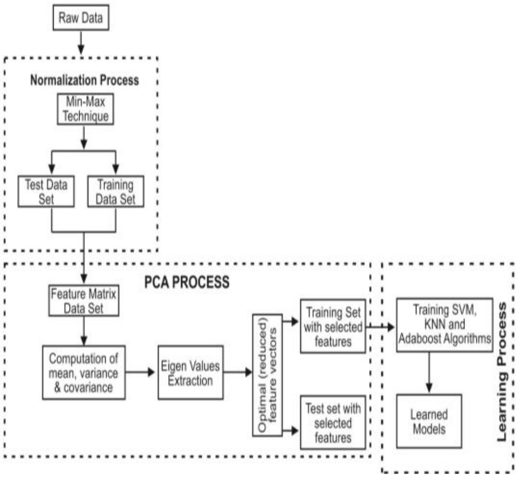

Raw Data
Normalization Process
Min-Max Technique
Test Data Set
Training Data Set
PCA PROCESS
Feature Matrix Data Set
Computation of mean, variance & covariance
Eigen Values Extraction
Optimal (reduced) feature vectors
Training Set with selected features
Test set with selected features
Training SVM, KNN and Adaboost Algorithms
Learned Models
Learning Process

Figure 1: Malware detection and mitigation framework

The data used to test the system for generality and performance efficiency is gotten from Kaggle, a public data repository. This data is National Scientific Laboratory-knowledge Discovery Data-mining (NSL-KDD) dataset. NSL-KDD dataset may not be a perfect dataset, but it is an improvement on the KDD'CUP 99 dataset, for it does not include redundant, noisy and missing records [24, 25]. It consists of five data types: normal data, and four attack types: denial of service (DOS) attack, probe or surveillance attack, user to root (U2R) attack, and remote to local (R2L) attack. The dataset contains 262,178 attack records and 812,814 normal records in the training dataset; and 29,378 attack records and 49,911 normal records in the test dataset. In all, 1,074,992 records are contained in the training dataset; and 77,289 records are used in the test dataset (GitHub Inc., 2020), with forty one (41) attributes. The attributes are contained in attribute relations file format (ARFF). ARFF is an ASCII text file that describes the list of instances that shares a set of attributes. It is divided into two: header and data. The header describes the attributes of the dataset, while the data section describes the observations of the dataset. The labeling of the application records is either zero (0) or one (1) for malicious (anomaly) or benign (normal) data respectively, which is achieved through normalization and dimensionality reduction (feature extraction and selection). The pre-processed data is then used to train SVM, KNN and Adaboost for effective analysis, detection and reporting. The dataset is loaded into the smart phone memory, normalized using min-max normalization technique, to reduce the numeric differences and make them uniform within a certain range, say [-1, 1]. Symbolic features such as protocol types, service and flag (non-numeric data) will be converted to numeric data by assigning numeric weights to them, without loss of efficiency [24].

Normalization becomes necessary because principal component analysis (PCA) an unsupervised learning algorithm, K-NN and all algorithms that use Euclidean distance measure are usually biased towards the data class with more features, to the detriment of the class with fewer features. Normalization avoids over fitting. The normalized data is then split into training (80%) and testing data sets (20%) and sent to PCA processing unit, as shown in Figure 1. Data loaded into the memory is raw and contains redundant features, noisy features and even missing data. To remove these unwanted features and reduce the dimension of the dataset, PCA is used. The extracted data of the training dataset are sent to the learning process for training. SVM, KNN and Adaboost are used to build models that select malicious features from the benign features.

The trained models are then combined to provide a more efficient analysis in the detection and mitigation of intrusions.

SVM-K-NN –Adaboost Hybrid

The fusion of SVM and K-NN produces a better classification accuracy than each single classifier, using NSL-KDD dataset. However, the success of hybridized methods for intrusion classification is dependent on the choice of a method to combine the diversified outputs of the individual classifiers into a single output [25, 26]. Adaboost is used to filter the results of the two component classifiers, as shown in Figure 2. Adaboost is an iterative classifier that runs other programs multiple times in order to reduce the error. For the first iteration, all algorithms have the same weights. As the iterations continue, the boosting process adds weights to the unclassified features from the component classifiers [27].

First, the individual models submit their detection results to the Adaboost module for a final classification result. The final result is analyzed for performances (accuracy, specificity and sensitivity) as shown in Figure 2.

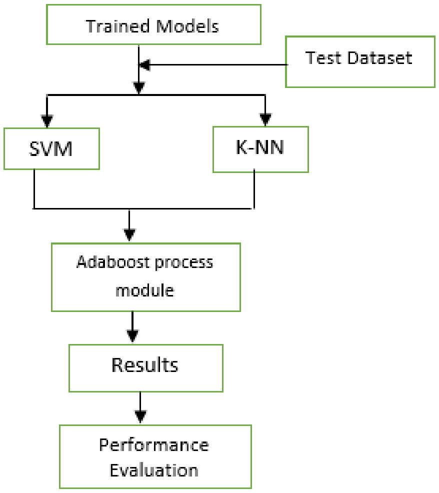


Figure 2: Performance Analysis and Detection of Malware

The result is then segregated and benign data allowed to terminate; while the malicious data is sent to the Intrusion Response System (IRS) for mitigation processes, as shown in Figure 3. From the policy rule base, the IRS mitigation process is fed with appropriate mitigation option in real-time, as shown in Figure 3. The process is coded using Python library modules such as Scikit Learn and Numpy.

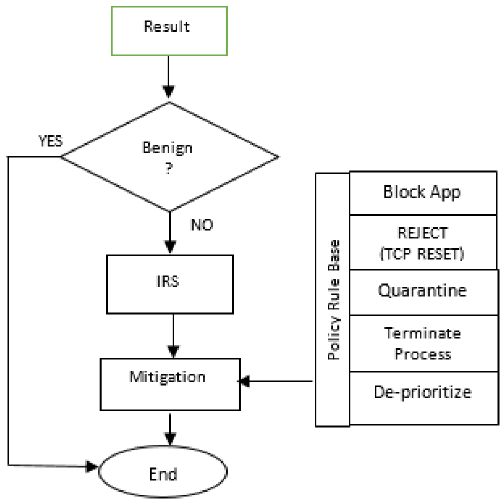


Figure 3: Mitigation Process of IRS

IV. RESULTS/DISCUSSION

Initial results obtained were analyzed for their performances and the results shown in Table 2.

Table 2: Performance Analysis of KNN, SVM and Adaboost

Model	Precision (%)	Recall (%)	Accuracy (%)		F1 Score (%)
			Train	Test	
K-NN	92	99	92.24	91.60	96
SVM	90	100	94.37	91.82	97
Adaboost					99.23

It is observed that Adaboost took a long time to run, as at the time of this report, it was yet to complete its analysis, but only provided the F1-score accuracy of 99.23%, which is an improvement compared to the results of KNN (96%) and SVM (97%). From Table 2 it is clear that SVM did better than KNN in all measurements, except precision, where KNN had 92% and SVM 90%. F1 specifies the entire system's (model's) performance. Since Adaboost filters the results of KNN and SVM, then it stands that its result (99.23%) is the overall performance of the framework.

V. CONCLUSION

The proposed framework integrated SVM, KNN and Adaboost to achieve the desired purpose of not only detecting intrusions, but mitigating their effects in real-time. It achieved an accuracy of 99.23%, a much better result compared to that of SVM and KNN. The results of the system are not conclusive yet, especially on the aspect of intrusion response system. Further work will attempts to incorporate extreme learning machine (ELM), Random forest and other unsupervised algorithm to achieve higher accuracy. It will take advantage of system log files, which is dynamic in nature for the detection and mitigation of malware intrusions.

REFERENCES

[1] Langley, P. Machine Learning for Intelligent System. AAAI-97 Proceedings, 1997.

[2] Dietrich, D., Heller, B., and Yang, B. *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*, John Wiley and Sons Inc., Indianapolis, Indiana, 2015.

[3] Arshad, S; Khan, A., Shah, M.A., and Ahmed, M., Android Malware Detection and Protection: A survey. *International Journal of Advanced Computer Science and Applications. (IJACSA),* 7(2), 2016.

[4] Ayo, C. K., Ekong, U.O., Adebiyi, A. A., Mobile Commerce. *Journal of Internet Banking and Commerce. ISSN: 1204-5357,* 2007.

[5] Privacy Rights Clearing House, *Chronology of data breach*,2010. http://www.privacyrights.org/

[6] Baker, W.H., Hutton, A., Hyulender, C. D., Novak, C., Sartin, B., Tippett, P., Valentine, J.A., 2009. *The 2009 Breach investigations Report. Verizon Business.* *http://www.verizxonbusiness.com/resources/security/reports/2009_databreach_rp.pdf*

[7] Yesilyurt, M and Yalman, Y., Security threats on Mobile Devices and Their Effects; Estimation for the Future. *International Journal of Security and its Applications.* 10(2):13–26, 2016. http://dx.doi.org/10.14257/ljsla.2016.10.2.02

[8] Koshal J. and Bag, M., Cascading of C4.5 Decision Tree and Support Vector Machines for Rule Based Intrusion Detection System. *International Journal of Computer Network and Information Security,* 8: 8-2, 2012. *DOI: 10.5815/ijenis.2012.08.02,*[9] Wang, G., Hao, J; Ma, J. and Huang, L., A new approach to Intrusion Detection using Artificial Neural Networks and Fuzzy clustering. *ELSEVIER: Expert Systems with Applications*. 37:6225 – 6232, 2010.

[10] Bhuyan, M. H., Bhattacharyya, D. K., and Kalita, J. K., Network Anomaly Detection: Methods, Systems and Tools. *IEEE Communications Surveys and Tutorials,* 16(1), 2014.

[11] Tesfahum, A. and Bhaskari, D. L., Effective Hybrid Intrusion Detection System: A layered Approach. *International Journal of Computer Network and Information Security.* 3(3):35 – 41, 2015.

[12] Rathi, D. and Jindal, R., DroidMark--A Tool for Android Malware Detection using Taint Analysis and Bayesian network. *International Journal of Recent and Innovation Trends in Computing and Communication,* 6(5), 2018.

[13] Petsas, T., Vogatzis, G., Athenasopoulos, E., Polychronakis, M., and Loannidas, S., Rage Against the virtual machine: Hindering Dynamic Analysis of

Android Malware, 2014. http://dx.doi.org/10.1145/2592791.2592796

[14] Inayat, Z; Gani, A., Anuar, N. B; Khan, M. K., and Anwar, S., Intrusion Response System: Foundations, design, and Challenges. ELSEVIER *Journal of Network and Computer Applications*, 62:53 – 74, 2016.

[15] Bishop, M., What is Computer Security? *IEEE Security Privacy,* 1(1):67 – 69, 2003.

[16] Sullivan, D.T., Survey of Malware Threats and Recommendations to Improve Cyber Security for Industrial Control Systems version 1.0. US Army Research Laboratory (ARL), 2015.

[17] Rajyaguru, M.H., CRYPTOGRAPHY – Combination of Cryptogrphy and Steganography with Rapid Changing key. *International Journal of Engineering Technology and Advanced Engineering.* 2(10), 2012.

[18] Yin, H. and Song, D., Whole-System Fine-grained Taint Analysis for Automatic Malware Detection and Analysis, 2019. https://www.researchgate.net/publication/2287857-54

[19] McNeil, P., Shetty, S., Guntu, D., and Barve, G., SCREDENT: Scalable, Real-time Anomalies Detection and Notification of Targeted malware in Mobile Devices. ELSEVIER: *Second International Workshop on Mobile Cloud computing systems, Management and Security (MCSMS)*, 2016.

[20] Feizollah, A., Anuar, N. B., Salleh, R., Wahab, A., A Review of Feature Selection in Mobile Malware Detection. *ELSEVIER: Digital Investigation*. 13:22–37, 2015. http://dx.doi.org/10.1016/j-diin.2015.02.001.

[21] Shabtai, A., Kanonou, U., Elovici, Y., Detection, Alert and Response to Malicious behavior in Mobile Devices: Knowledge based Approach, *In Proceedings of RAID 2009, LNCS 5758, Springer, Saint-Malo Britany, France,* 35 – 358, 2009.

[22] Ikram, S.T. and Cherukuri, A.K., Improving Accuracy of Intrusion Detection Model Using PCA and Optimized SVM. *CIT Journal of Computing and Information Technology,* 24(2):133 – 148, 2016.

[23] Aljawarneh, S., Aldwairi, M., and Yassrin, M. B., Anomaly-based Intrusion Detection System through Feature Selection analysis and Building Hybrid Efficient Model, 2017.

[24] McHugh, J., "Testing Intrusion Detection Systems". A Critique of the 1998 and 1999 DARPA IDS Evaluations as Performed by Lincoln Laboratory" *ACM Translations on Information and System Security,* 3(4):262–294, 2000.

[25] Dada, E.G., A Hybridized SVM-KNN-pdAPSO Approach to Intrusion Detection System, University of Maiduguri, Faculty of Engineering Seminar Series, Nigeria, 8, 2017.

[26] Altaher, A., Fishing Website Classification Using Hybrid SSVM and KNN Approach. (IJACSA) *International Journal of Advanced Computer Science and Applications*, 8(6), 2017, ELSEVIER. *Journal of Computer Science. https://dx.doi.org/10.1016/j.jocs.2017.03.006*

[27] Freund, Y. and Schapire, R.E., Experiments with new Boosting Algorithm. *Proceedings of the Thirteenth International Conference on Machine Learning*, 1996, 148–156.

Cloud Based Anti-Terrorism System for Social Media Applications

Amanze Bethran .C[1*] Nwoke Bethel .C[2]., Ononiwu Chamberlyn .C[3].
[1, 2] Department of Computer Science, Faculty of Science, Imo State University, Owerri, Nigeria.
[3]Department of Computer Science, Imo State Polytechnic, Umuagwo, Nigeria.
amanzebethran@yahoo.com.

ABSTRACT
The aim of the work was to design a centralized cloud based anti-terrorism system for dictating suspected terrorist on social media platforms. The recruitment of terrorists through social media which has helped in the spread of terrorism motivated this work. Terrorism, using the social media, has become one of the most concerning issues across the world. Terrorist organizations are using social media platforms for recruiting, training and communicating with their followers, supporters, donors, as it is cheaper, easier, faster and effective method of communication. Object-oriented analysis and design methodology (OOADM) was adopted for the analysis and development of the anti-terrorism.

Keywords- Facebook, OOADM, I.T, sample Crime

I INTRODUCTION

There is interplay between home-grown terrorist groups and international terrorist organizations which is playing the central role in accelerating the situation. The members of the terrorist organizations are spreading their ideological thoughts, propaganda and their activities to the world using social media platforms. This research will focus on how terrorist groups are using social media platforms, especially Facebook and Twitter to threaten the peace and security of countries and how to reduce the situation to its barest minimum. Communication between terrorist organizations, vulnerable persons and potential Jihadis is greatly benefited through the availability of cyber environments such as forums, social media platforms, web pages, blogs, post boards and email (Denning, 2001). The advantage of digital technology to terrorist groups is clear, social media crosses borders and allows groups to connect associates and reach out to persons vulnerable to radicalization (Torok, 2013). Although the term terrorism is not subject to a universally agreed definition, UN (2018) defined terrorism as a method of coercion that utilizes or threatens to utilize violence in order to spread fear and thereby attain political or ideological goals. Notwithstanding the absence of a globally agreed, legal definition of terrorism, an effective and prevention-focused international response to terrorism is highly desirable, particularly one guided by a normative legal framework and embedded in the core principles of the rule of law, due process and respect for human rights. Many international and regional legal security agencies already exist which are dedicated to countering and deterring terrorism, primarily through the investigation and prosecution of those suspected of committing related crimes by means of state criminal justice processes. While such international and regional security agencies provide for effective prevention mechanisms, including interventions, targeting specific types of criminal acts like hostage-taking, the hijacking of planes or ships, terrorist bombings and the funding of terrorism. It was observed that tracking of terrorists by law agencies are not properly carried out. Some of the problems identified in the process include:

1. Loss of data/information due to manual processing.
2. Delay in tracking suspected terrorists.
3. Inability to prevent suspected terrorists from carrying out attacks.
4. Poor security of crime/terrorist records.
5. Small and limited database for crime/terrorist information storage.

The aim of this paper is to design a cloud based anti-terrorism system for international security agencies which will help them to detect suspected terrorists on social media and recommend a preventive measures before they carry out their attacks. The Objectives are:

1. To identify suspected terrorists before they carry out their attacks.
2. To provide a centralized security database for international security agencies.
3. To provide encryption/decryption technique as a security measure to secure stored data/information from unauthorized personnel.

2.0 Organized Crime and Terrorism

The definition of organized criminal group in the Organized Crime Convention only includes groups that through their activities seek to obtain, directly or indirectly, a financial or other material benefits. This would not, in principle, include groups such as some terrorist or insurgent groups, provided that their goals were purely non-material. However, the Convention may still apply to crimes committed by those groups in the event that they commit crimes covered by the convention in order to raise financial or other material benefits. Organized crime is actually one type of several categories of organized criminal behavior. White-collar crime, for instance, is related to and overlaps with organized crime and the definition contained in the organized crime convention allows capturing many cases of white collar crime. Nonetheless, the two crimes have significant differences in that white-collar crime occurs as a deviation from legitimate business activity, whereas organized crime occurs as a continuing criminal enterprise that exists to profit primarily from illicit activity. White-collar crime can be carried out by an individual whereas organized crime requires more people and planning in order to carry out offences on a more systematic basis. White-collar crime can also be carried out by organized criminal groups. (Kego et al, 2011). Governmental bodies and politicians can also be considered offenders of organized crime if the elements of the general definition are met. It is also important to notice that not only individuals but also legal persons, such as corporations, can carry out crimes during the course of business. Serious crime is frequently committed through or under the cover of legal entities. Terrorism is another form of "organized" criminal behavior, but it is distinct from organized crime. In general terms, terrorism involves crimes committed with the objective of intimidating a population or compelling a government or international organization with a view to achieving political or social objectives. Examples would include hostage-taking in order to secure freedom for those seen as imprisoned unjustly or acts of violence done in retribution for perceived past injustices. An act of terrorism has a political objective. Organized crime, on the other hand, always seeks to obtain a financial or other material benefit, whereas power and control can be secondary motives. Organized crime can involve violence and coercion, but the objective in organized crime remains profit. Furthermore, while they generally pursue different objectives, activities of terrorists and organized criminal groups can overlap (Bassiouni, 1990). A clear example is when terrorist groups use organized crime activity to fund their political objectives. Terrorist organizations can thus adapt the conventional tactics of organized criminal groups, such as generating profits from drug trafficking, or other types of illicit trade. Another important element of distinction between these two crimes is that by definition, organized crime cannot be committed by a single person while a terrorist act can be.

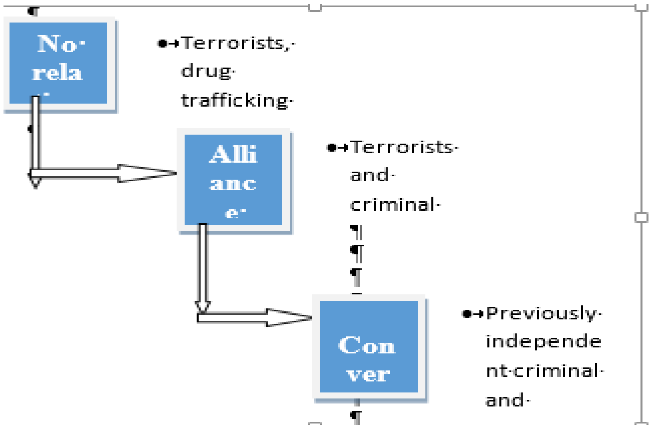


Figure 1: Merging of Interests between Criminal and Terrorist groups (Kunreuther et al., 2013)

Effects of Terrorism on Developed and Developing Countries

Terrorism has terrible consequences on the countries where it exists and most times extends to the neighboring countries. A good example is the Boko Haram terrorism in Nigeria which has extended to Cameroon and Chad. A rational terrorist organization can, in principle, reach its goal quicker if it is able to augment the consequences of its campaign. These consequences can assume many forms including casualties, destroyed buildings, a heightened anxiety level, and numerous economic costs. Clearly, the United States of America attacks on September 11, 2001 had significant costs that have been estimated to be in the range of $80 to $90 billion when subsequent economic losses in lost wages, workman's compensation, and reduced commerce are included (Kunreuther et al., 2003). Terrorist incidents divert foreign direct investment (FDI) funds to security, which is not a good plan for a developing country. If a developing country loses enough FDI, which is an important source of savings, then it may also experience reduced economic growth. A sufficiently intense terrorist campaign may greatly reduce capital inflows (Enders and Sandler, 2008). Terrorism, like civil conflicts, may cause spillover costs among neighboring countries as a terrorist campaign in a neighbor country discourages capital inflows, or a regional multiplier causes lost economic activity in the terrorism-ridden country to resonate throughout the region. In some instances, terrorism may impact specific industries as the United States September 11, 2001 attacks did on airlines and tourism (Drakos, 2004). Another cost is the expensive security measures that must be instituted following large attacks. An example is the massive homeland security outlays since the September 11, 2001 attacks on America (Enders and Sandler, 2008). Terrorism also raises the costs of doing business in terms of higher insurance premiums, expensive security precautions, and larger salaries to at-risk employees.

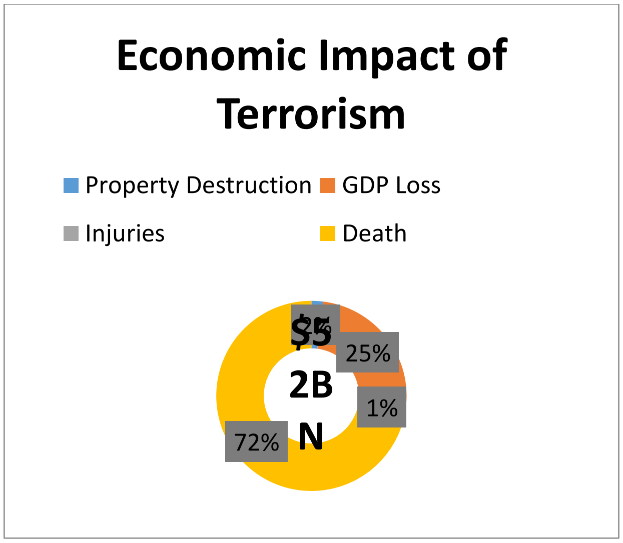


Figure 2: Global Economic Impact of Terrorism **(Enders and Sander, 2008)**

Domestic and Transnational Terrorism

Terrorism comes in two essential types: domestic and transnational. Domestic terrorism is homegrown with consequences for just the host country, its institutions, citizens, property, and policies. In a domestic terrorist incident, the victim and perpetrators are from the host country (Enders and Sandler, 2008). The Oklahoma City bombing on April 19, 1995 was a domestic terrorist event as was the 276 female students kidnapped in Chibok, Borno state, Nigeria by Boko Haram terrorists. Many ethno-nationalist conflicts in Africa are associated with mostly domestic terrorism, unless the rebels desire to target citizens from other countries to publicize their cause to the world. Domestic events tend to outnumber transnational terrorist events by eight to one (Enders and Sandler, 2008). In contrast, transnational terrorism involves more than one country. This international aspect can stem from the victims, targets, institutions, supporters, terrorists, or implications. For example, September 11, 2001 attacks on America is a transnational terrorist event because the victims were from many different countries, the mission was financed and planned from abroad, the terrorists were foreigners, and the implications of the events like financial and security were global. A hijacking that originates in one country but terminates in another country is another instance of transnational terrorism. Transnational terrorist attacks often entail trans-boundary externalities: actions or authorities in one country impose uncompensated consequences on person or property in another country. Thus, spillover costs can result so that the economic impact of a terrorist event may transcend the host country. The collapse of the World Trade Center towers on 9/11 killed many British nationals and had ramifications for British financial institutions. The distinction between domestic and transnational terrorism is of utmost importance when determining the right data for calculating the economic consequences of terrorism.

Figure 3: World Trade Center in New York which was attacked by Terrorists on September 11, 2001.

Cloud Delivery Services

Software as a Service (SaaS)

The software-as-a-service (SaaS) service-model involves the cloud provider installing and maintaining software in the cloud and users running the software from cloud over the Internet (or Intranet). The users' client machines require no installation of any application-specific software since cloud applications run in the cloud. SaaS is scalable, and system administrators may load the applications on several servers. In the past, each customer would purchase and load their own copy of the application to each of their own servers, but with the SaaS the customer can access the application without installing the software locally. SaaS typically involves a monthly or annual fee. Software as a service provides the equivalent of installed applications in the traditional (non-cloud computing) delivery of applications.

Development as a Service (DaaS)

According to Wang, (2010), Development as a service is web based, community shared development tools. This is the equivalent to locally installed development tools in the traditional (non-cloud computing) delivery of development tools.

Platform as a Service (PaaS)

Platform as a service is cloud computing service which provides the users with application platforms and databases as a service (Mell and Grance, 2011). This is equivalent to middleware in the traditional (non-cloud computing) delivery of application platforms and databases.

Infrastructure as a service (IaaS)

Infrastructure as a service involves converting the physical hardware resources like servers, networks, storage, and system management to virtual resources on the cloud. This is the equivalent to infrastructure and hardware in the traditional (non-cloud computing) method running in the cloud. In other words, businesses pay a fee (monthly or annually) to run virtual servers, networks, storage from the cloud. This will mitigate the need for a data center, heating, cooling, and maintaining hardware the local level (Wang, 2010).

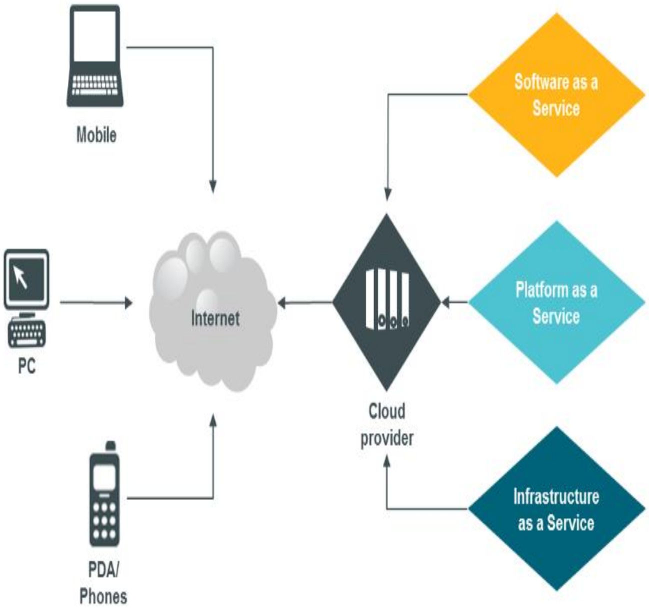


Figure 4: Cloud Services Delivered over the Internet (Wang, 2010)

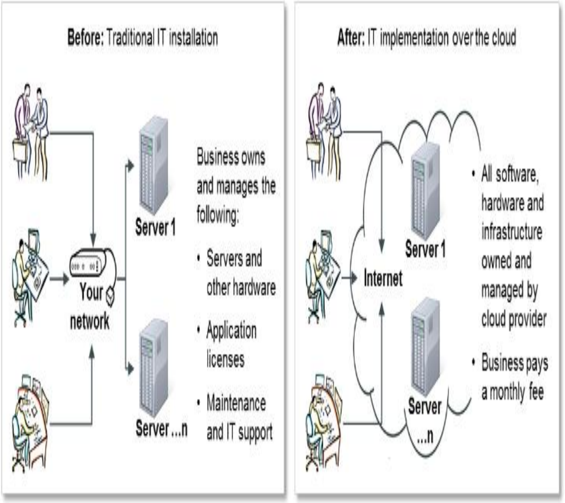

Before: Traditional IT installation
Your network
Server 1
Server ...n
Business owns and manages the following:
• Servers and other hardware
• Application licenses
• Maintenance and IT support
After: IT implementation over the cloud
Internet
Server 1
Server ...n
• All software, hardware and infrastructure owned and managed by cloud provider
• Business pays a monthly fee

Figure 5: Traditional IT Installation vs Cloud Implementation (Wang, 2010).

3.0 Analysis of the Present System

The present system focuses on terrorism tracking on social media by cyber security agents or technicians employed by the social media firms. It is a manual system which is controlled by a technician at the back end. After users of the application register, they make posts or comment on friends posts. The technician manually checks posts and comments for terror related words or words that encourage terrorism or waits for another user to report to them before they check it out. Once confirmed to be terror related, the technician removes the comment or post from the social media platform and sometimes suspends/deletes the related account from its database.

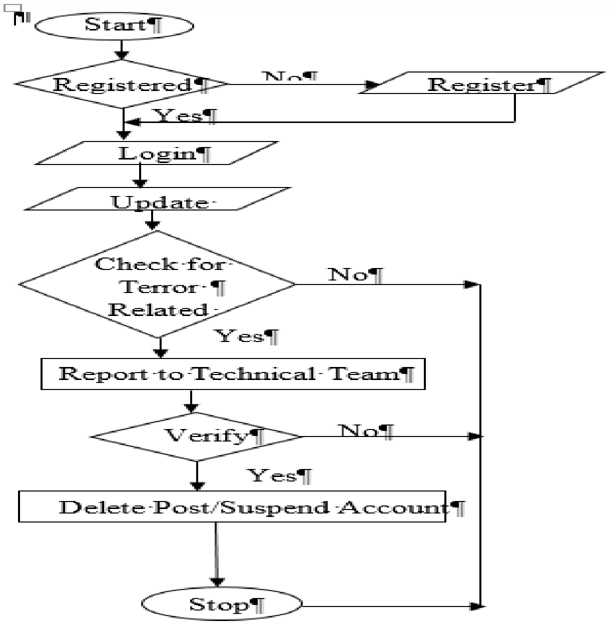

Start¶
Registered¶
No¶
Register¶
Yes¶
Login¶
Update·
Check·for·Terror·¶Related·
No¶
Yes¶
Report·to·Technical·Team¶
Verify¶
No¶
Yes¶
Delete·Post/Suspend·Account¶
Stop¶

Figure 6: Analysis of the Existing System

Analysis of the New System

The new system is an enhanced version of the existing system. The system is aimed at detecting terror related posts, broadcasts or group creation by checking for keywords already defined to be related to terrorism. In delay of tracking terrorists, the system will prompt the admin about the post/comment of the user by capturing the sentence immediately and submitting to database. This feature will help security agents to take action without delay. The system has a central cloud based database which is accessible to the social media enterprise and certified security agencies. The suspected terrorists' information is stored on the database for easy retrieval. This eliminates the disadvantage of poor storage system for terror related information. The new system storage will be hosted on a cloud database and this feature will improve the security of the system as the cloud service providers will provide the necessary techniques to ensure the data are secured. The security agents who are the administrators of the system will make it easy for the tracking and monitoring of the user posts and comments and possibly carry out arrests when the need arises. This feature makes it uneasy for suspected terrorists to evade arrest. One difficulty the social media companies face is, if a suspected person is blocked from one platform, he might simply move to a different one. In response to this, this proposed system has a database of "hashes". A hash is a unique digital fingerprint that can be used to track digital activity. When pro-terrorist content is removed by one social media firm, its hash is shared with the other participating companies to enable them to block the content on their own platforms and the details of the person are submitted to the central database accessible to security agencies.

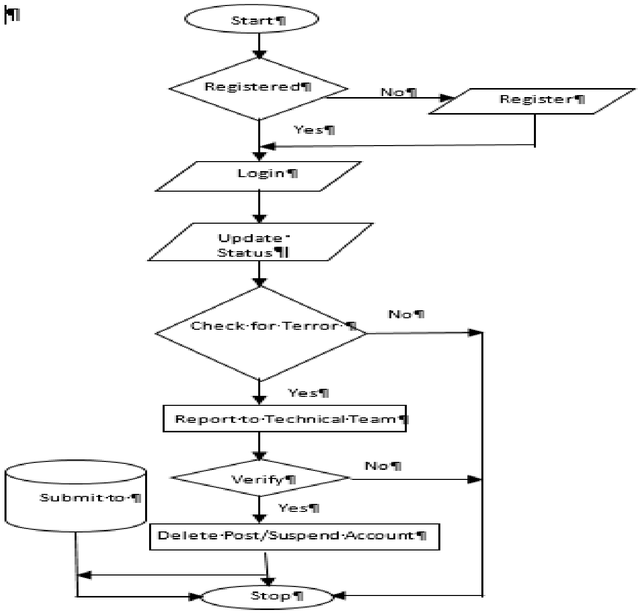


Figure 7: Analysis of the New System

Class Diagram of the New System

A class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among objects.

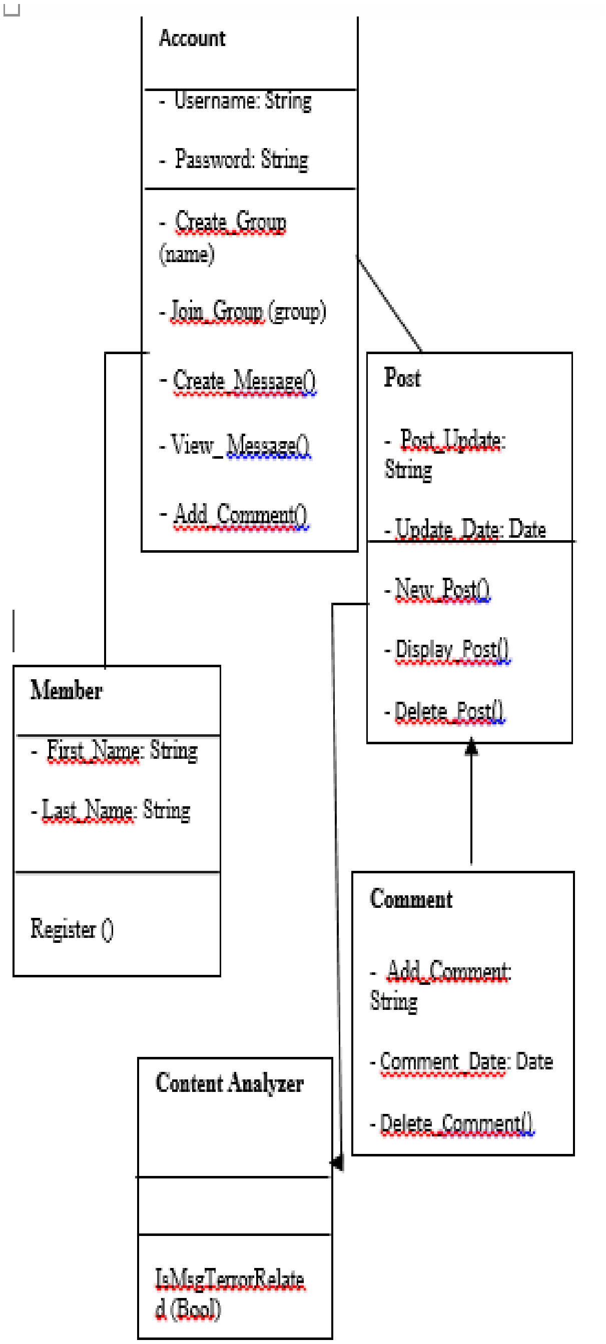


Figure 8 Class Diagram

Activity Diagram

An activity diagram is a behavioral diagram i.e. it depicts the behavior of a system. An activity diagram portrays the control flow from a start point to a finish point showing the various decision paths that exist while the activity is being executed.

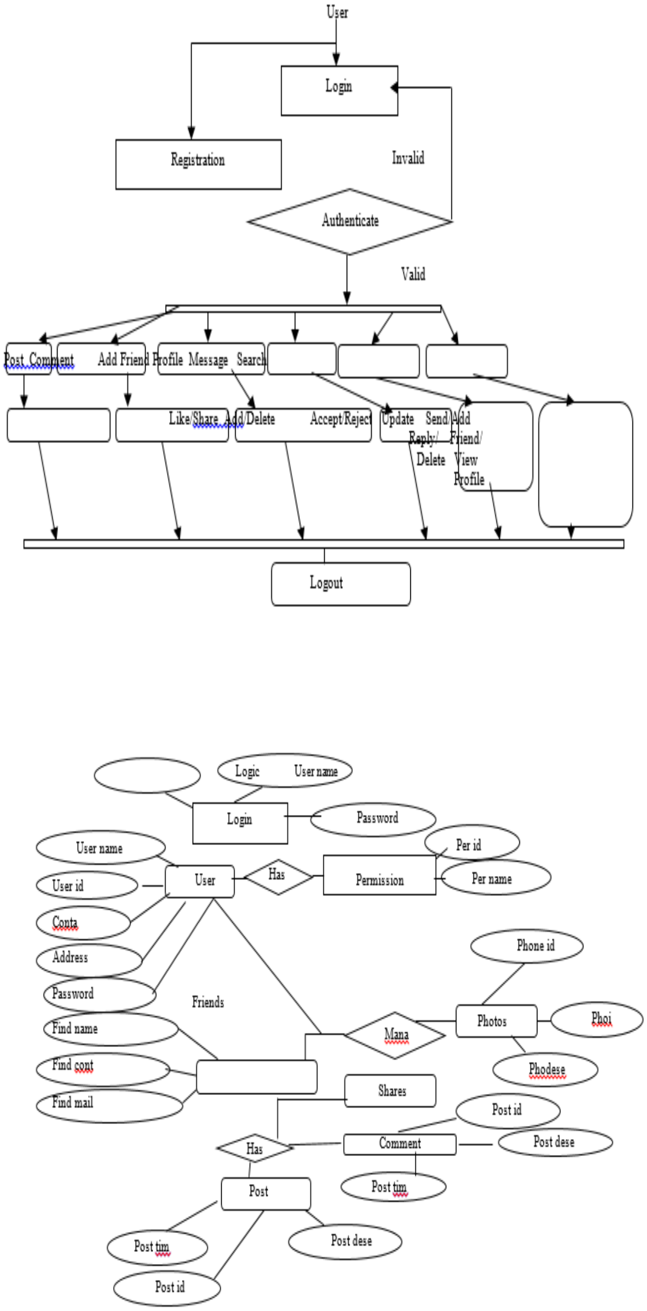


Figure 9: Activity Diagram

Entity Relationship Diagram

The entity-relationship diagram shows all the visual instrument of database tables and the relations between all components of the system. It is used to define the **r**elationships between structured data groups of the system functionalities.

Use Case Diagram

A use case diagram is a graphic depiction of the interactions among the elements of a system. A use case is a methodology used in system analysis to identify, clarify, and organize system requirements.

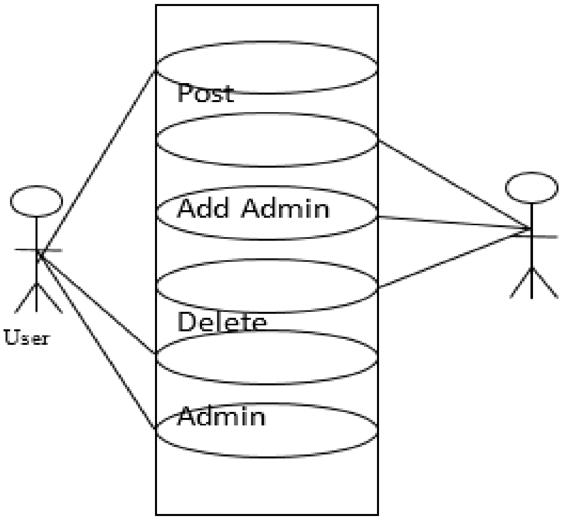
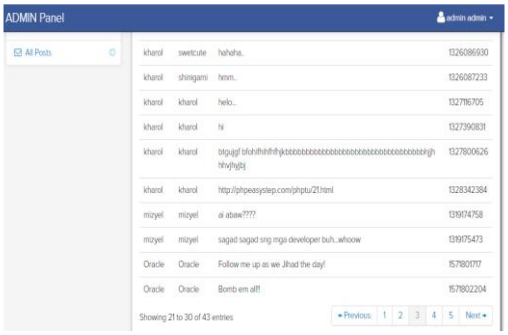


Figure 11: Use Case Diagram

Test Results The admin of the system logs into the admin account and clicks on the "**Detected Suspected Terrorist Posts"** tab and the sentences will be retrieved from the database.



kharol	swetcute	hahaha..	1326086930
kharol	shinigami	hmm..	1326087233
kharol	kharol	helo..	1327116705
kharol	kharol	hi	1327390831
kharol	kharol	btgujgf bfohifhihfhfhjkbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbhjh hhvjhyjbj	1327800626
kharol	kharol	http://phpeasystep.com/phptu/21.html	1328342384
mizyel	mizyel	ai abaw????	1319174758
mizyel	mizyel	sagad sagad sng mga developer buh...whoow	1319175473
Oracle	Oracle	Follow me up as we Jihad the day!	1571801717
Oracle	Oracle	Bomb em all!!	1571802204

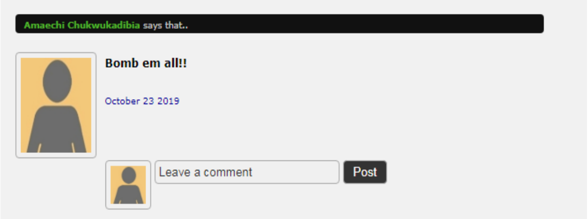
Showing 21 to 30 of 43 entries

Result 1:

Fig. 12 showing terror related input by the user

Figure. 12: Terror Related Input

The system allows the user to make post and captures the sentence if it contains any terror related word.



Amaechi Chukwukadibia says that..

Bomb em all!!

October 23 2019

Result 2:

Fig. 13 showing admin retrieval of terror related words captured by the system.

Figure. 13: Terror Related Word Captured by the System.

The suspected post or comments captured by the system is saved to the database and can be retrieved by the admin.

CONCLUSION

Terrorism is a crime against humanity and therefore requires attention from everyone especially security agencies. The anti-terrorism campaign needs to be extended to the social media in other to reduce terrorists' recruitment and reduce their communication channel. Adopting available technologies like Cloud computing and combination of security techniques can be implemented in the anti-terrorism campaign. This paper proposes a solution on social media applications like Facebook and Twitter that do not implement end to end encryption.

REFERENCES

1. Adamou L., and Sharifi S. (2018). Taliban threaten 70% of Afghanistan, BBC finds. Retrieved August 31, 2019, from https://www.bbc.com/news/world-asia-42863116.
2. Amnesty. (2017). Lake Chad region: Boko Haram's renewed campaign sparks sharp rise in civilian deaths. Retrieved September 5, 2019, from https://www.amnesty.org/en/latest/news/2017/09/lake-chad-region-boko-haramsrenewed-campaign-sparks-sharp-rise-in-civilian-deaths/
3. Bassiouni, M. Cherif. (1990). Effective National and International Action against Organized Crime and Terrorist Criminal Activities. Emory International Law Review, vol. 4, 9-42.
4. Dolvara Gunatilaka (2016). "A survey of Privacy and Security Issues in Social Networks". Retrieved September 30, 2019, from http://www.cse.wustl.edu/~jain/cse571-11/ftp/social/index.html
5. Drakos, Konstantinos (2004), "Terrorism-Induced Structural Shifts in Financial Risk: Airline Stocks in the Aftermath of the September 11th Terror Attacks," European Journal of Political Economy, 20(2).
6. Enders, Walter and Todd Sandler (2008). The Political Economy of Terrorism Cambridge: Cambridge University Press.
7. Felter C. (2018). Nigeria's battle With Boko Haram. Retrieved August 8, 2019, from

https://www.cfr.org/backgrounder/nigerias-battle-boko-haram.

8. Global Terrorism Index (2018). Available online at: www.gti.org
9. Judith, H.; Marcia, K. and Fern, H. (2012). Cloud Services for Dummies. John Wiley & Sons, Inc., Hoboken, New Jersey.
10. Kegö, Walter, Erik Leijonmarck, and Alexandru Molcean, eds. (2011). Organized Crime and the Financial Crisis: Recent Trends in the Baltic Sea Region. Stockholm: Institute for Security and Development Policy.
11. Kunreuther, Howard, Erwann Michel-Kerjan, and Beverly Porter (2013), “Assessing, Managing and Financing Extreme Events: Dealing with Terrorism,” Working Paper 10179, National Bureau of Economic Research, Cambridge, MA.
12. Luke Bertram (2016). “Terrorism, the Internet and the Social Media Advantage”.
    International Journal of De-radicalization. Published August 2016.
13. Mell, P.; Grance, T. (2011). "The NIST Definition of Cloud Computing. NIST Special Publication 800-145 (September 2011). National Institute of Standards and Technology, U.S. Department of Commerce". Online: http://csrc.nist.gov/publications/nistpubs/800-145/sp800-145.pdf. Accessed: June, 2019.
14. Schmid, Alex (2006). “Magnitudes and Focus of Terrorist Victimization: Large-Scale Victimisation as a Potential Source of Terrorist Activities. IOS Press.
15. Sherbak,; T., Sweere, N. and Belapurkar, V. (2012). Virtualized Enterprise Storage for Flexible, Scalable Private Clouds. Dell Power Solutions, 2012. Online:http://i.dell.com/sites/content/business/solutions/power/en/Documents/Ps1q12-20120209-Sherbak.pdf. Accessed: June, 2019.
16. Wang, R. (2013). Developing an explanatory model for the process of online radicalization and terrorism. Security Infomatics, a Springer Open Journal, volume 2, No. 6. Security Research Institute, Edith Cowan University. Perth, Australia. Online: http://www.security-informatics.com/content/2/1/6 Accessed: June, 2019.
17. United Nations Office on Drugs and Crime (2018). The Doha Declaration: Promoting a culture of Lawfulness. Online: https://www.unodc.org/e4j/en/tertiary/organized-crime.html. Accessed: June, 2019.

Staff Performance Evaluation Using Artificial Neural Network

Adewole Rufai and Ibrahim Adeboye
Department of Computer Sciences
University of Lagos
Lagos, Nigeria
arufai@unilag.edu.ng comdayo@gmail.com

Abstract*—This paper proposed the use of Artificial Neural Network (ANN) for Staff Performance Evaluation. To achieve this aim, the IBM HR Employee Attrition dataset was adapted. This paper proposes the use of Artificial Neural Network in assessing the performance of the employees for aiding the process of promotion. A neural network was trained and evaluated. The data collected were entered into the model after identifying the appropriate variables. Result of this study shows that certain factors have the most important effect in predicting promotion for a given staff. The training loss is significantly low, 0.018 while the training and test accuracies are 99% and 97% respectively. It means that organizations and human resources (HR) departments should focus on these factors when evaluating the performance of an employee. This study helps to find out how to improve the efficiency of staff and aids to eliminate inefficiencies with a relatively easy to employ framework.

***Keywords*—Sigmoid function, Artificial Neural Network (ANN), ReLU, NN Architecture, Staff performance, prediction.**

I Introduction

As people congregate to work, issues surrounding human resources become important. Management of human resources is responsible for combining human resource preparation into overall organizational strategy that involves career development, training, and rewards. Others include performance reviews, promotions and transfers. An employee's performance record is generally defined by their biological data, academic and professional credentials, specialization or generalization area, history and job experience. Assessment of every organization's workforce is a subtask in managing human resources. In other ways, to achieve the mission of the organization and support its strategies, goals and strategies for human resource need to be developed.

The goals of human resources are to: (a) Develop staff recruitment and selection schemes to hire the best possible staff in accordance with the needs of the company. (b) Optimize each employee's potential to achieve the goal of the organization and guarantee individual career development and personal integrity. (c) Retain workers whose performance assists the organization in achieving its objectives and relieving those whose performance is unconvincing; (d) Monitor organizational conformity with federal and state regulations related to the purpose of human resource management.

Man, materials and money are the resources of an organization. Man is the determining factor in the use of other resources. Man is characterized by two major components, namely, administrative staff and technical staff. Personnel evaluation in most establishments done manually carried out by the Human Resource Management (HRM) department. The assessment process usually begins with an evaluation by each head of divisions who also makes recommendations under his supervision for promoting the workers. This recommendation is then sent to the Department of Human Resource, which has the responsibility to schedule minutes for each possible candidate. In turn, it will be passed on to the promotion board. Ultimately, the panel implements legislation on matters of selection and promotion. Nonetheless, information technology (IT) is believed to be necessary to strategic human resource (HR) department decisions such as recruitment, assessment, rewards, training and development. The conventional system of Staff evaluation by Human Resource (HR) department in most organizations are manual and in reality, real-system charade, with a method that is quite slow and error-prone.

A lot of research have been carried out in the evaluation of human resource performance in the Human Resource Management domain. Most of these studies are subjective in nature. The overarching goal of human resource performance evaluation is based on the need to have an objective appraisal that would lead to the motivation of staff in organizations. The aim of this work, is to provide a much more scientific way of evaluating staff performances in an organization.

For a long time, Artificial Intelligence (AI) researchers have been searching for ways to establish and experiment intelligent computer systems that can perform a wide variety of tasks which conventional computing systems find difficult or impossible to do. Quantitative and qualitative information are required as factors for the decision of computer system to function mimic human intelligence.

II RELATED WORKS

Pawlyta *et al.* (2020) worked on Deep Recurrent Neural Networks for Human Activity Recognition during Skiing. The work focused on identifying skiing activity. A database containing information from three inertial body sensors - accelerometer, gyroscope and barometer - was created. Then two deep models based on Long Short-Term Memory were created and compared.

Mohammad and Jianguo (2019) carried out the Evaluation of Factors Affecting Employees' Performance Using Artificial Neural Network (ANN) Algorithm. The study indicates that the payroll indicators, environmental conditions and reporting culture are the strengths of the system under review, and are now at an appropriate level. The results show a negative impact on the indicators of awareness, system planning and preparation for critical situations, amenities, training, and job security.

Kharola *et al*. (2018) proposed a soft computing model of Artificial Neural Network (ANN) based on a novel Performance evaluation technique. Comparing the proposed approach to adaptive neuro fuzzy inference system (ANFIS) based appraisal approach has got an inherent advantage of learning through supplied data samples. The results showed that the proposed approach is effectively applicable in carrying out the evaluation of multiple employees considering several performance factors.

Dalia *et al.* (2018) investigated the performance of the neural network for Human Activity Recognition (HAR). They proposed a multi-layer (Neural Network) NN for (Human Activity Recognition (HAR) and concluded that the desired and actual outputs values differences can be calculated at the completion of each iteration which is called the error.

Murad and Pyun (2017) proposed Deep Neural Networks to Recognize Human Activity. The proposed models deliver better output than other machine learning approaches including deep belief Networks (DBNs) and convolutional neural networks (CNNs). Health care organization extend their reach without fear of cost or insufficient facilities.

A comparison of the logic-based classifications Non-Parametric and Fuzzy in consideration of human everyday activities was carried out by Namdari *et al.* (2017). K-nearest neighbour (KNN) classifier as a non-parametric classifier encompasses a very high accuracy rate of about 99.758% was used. Recognition of activities using multiple small body-worn accelerometers as well as the implementation of an engineered unobtrusive system will carry a very significant result to us.

Okoye-Ubaka *et al.* (2013) agreed with the fact that Neural Network model's success is greatly trained dataset dependent. Also, the degree of performance assessment is defined by appropriate data available.

III METHODS

This study built a Neural Network by using TensorFlow to predict the promotion of Staff with specific work environment characteristics. The NN is trained by providing examples for the network and changing the weight parameters appropriately with some learning principles until the output of the NN corresponds acceptably to the desired output.

The Artificial Intelligence (AI) tool, Neural Network shall be employed in this study to train the proposed system to recognize patterns and suggest whether staff should be promoted or not. It reviews provided staff details and records, and make predictions to the Human Resources Management based on trends provided by the learning process.

A. Data Collection

Datasets used for this analysis was extracted from the Human Resource (HR) employee attrition data set that comes from the IBM Watson website. In IBM(2020) optional imprecise linguistic values are given to each of the decision parameters, including, Significantly Exceeds Requirements, Exceeds Requirements, Meets Requirements, Partially Meets Requirements, Requirements Not Met. Thus these data points were split into two: *train data* (80%), and *test data* (20%). Appropriate weight matrix were attached to the linguistic meaning.

The NN is formed by multiple neurons and diverse representations of neural networks which have distinct learning rules. The back-propagation neural network (BPNN) model chooses sigmoid activation function which is given in equation 1:

$$f(x) = {}^{1}\!/_{(1+e^{-x})} \qquad (1)$$

The learning rule implements the reverse propagation algorithm for the BP error.

Input Layer

Hidden Layers Output Layer

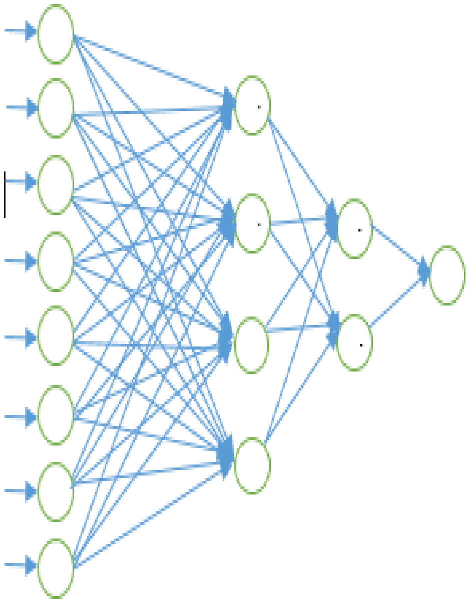


Fig. 1: Neural Network Architectural Design

The following are identified as input variables for the input layer:

X_1=Age
X_2=BusinessTravel
X_3=Department
X_4=Attendance
X_5=Education
X_6 = Gender
X_7 = JobLevel
X_8 =JobRole
X_9 = JobSatisfaction
X_{10} = MaritalStatus
X_{11} = OverTime
X_{12} = PerformanceRating
X_{13} = TotalWorkingYears
X_{14} = WorkLifeBalance
X_{15} = YearsAtCompany
X_{16} =YearsSinceLastPromotion

The hidden layer has double layers with eight and four nodes or units: h1-h8, h1-h4

The output layer has a single value, O_i, which can be a zero (0) or one (1). One stands for Yes which indicate turnover and zero stands for No.

i. Input Layer $x_i : i = 1 \ldots 16$

ii. Hidden Layer processing element y_i : $i = 1,2,3,4$

iii. Output Layer processing element z_k : $k = 0,1$

B. *The Learning Process*

The learning procedure is as follows:

i. We initialize the connection weights of random numbers between 0 and 1 for all neurons.

ii. We use the input vectors x_i and d_k as the output from the input layer of the processing element.

iii. Next is selecting the first training input and output pair (x, d) from the training pair vectors (x_i, d_k)

iv. Then we compute the value of hidden layer h_j, defined by equation below:

$$h_j = f\left[\sum_{i=1}^{16}\sum_{j=1}^{4}[w_{ij}x_i + O_j]\right] \qquad (2)$$

Where *f* is the activation (sigmoid) function, w_{ij} are the edge weights (input to hidden layer), x_i is the input layer and O_j output of the hidden layer.

v. Compute the value of the output layer O_k defined by the equation below:

$$O_k = f\left[\sum_{j=1}^{4}[w_{kj}h_j] + O_k\right] \qquad (3)$$

Where *f* is the activation (sigmoid) function, w_{kj} are the edge weights (hidden to output layer), h_j is the hidden layer and O_k output layer.

vi. Compute the result and obtain the difference

$$difference = d_k - O_k \qquad (4)$$

Where d_k is the desired output while O_k is the calculated output.

vii. Obtain output error defined by:

$$e_k = (d_k - O_k)(O_k)(1 - O_k) \qquad (5)$$

Where e_k is the output error.

viii. Computer the error at the hidden layer using the equation below:

$$e_k = h_j\left[1 - h\left[\sum_{j=1}^{4} w_{kj}e_k\right]\right] \qquad (6)$$

If the error is greater than a certain predefined value, propagate the error back into the network by means of adjustment of weights which link the output to hidden layers and weights connecting to the output from hidden layer.

ix. Adjust the weights between the kth output neuron and the jth hidden neuron defined by

$$w_{kj}(n+1) = w_{kj}(n) + B(e_k h_j) \tag{7}$$

Where n is the step in the training cycle, B is the learning rate.

x. We now adjust the weights between *jth* hidden neuron and *ith* input neuron.

$$w_{ij}(n+1) = w_{ij}(n) + B(e_i h x_i) \tag{8}$$

Where n is the step in the training cycle, B is the learning rate.

xi. Finally repeat steps ix through as many cycles as required until the sum of squared error (SSE) is within a prescribed tolerance using the formula:

$$\text{SSE} = \sum_{k=1}^{n}(e_k)^2 \qquad (9\}$$

C. *Degree of Staff Performance Variables*

Different factors are considered in evaluating the performance of staff. The most important ones for predicting promotion of staff include BusinessTravel, Attendance, JobSatisfaction, OverTime, LastPromotion and PerformanceRating. These information and other attributes such as Age, MaritalStatus, Gender, Department, JobRole, and TotalWorkingYears are obtained from employee with the aid of performance appraisal form.

PerformaceRating has subfactors which include Skills, Initiative, GeneralConduct, Discipline, TeamWork, all combined under a scale of 1-5. However, in order to pass the information into a Neural Network for computing, it must be represented in numbers for easy manipulation. A scale of 1-4, 1-5 are chosen to represent the degree of each variable. A staff that substantially exceed job requirements will get a rating of 5, 4 for Exceed Requirements, 3 for Meet Requirements, 2 for Partially Meet Requirements while 1 for staff that fails to Meet Requirements. A rating of 1-4 for JobSatisfaction and WorkLifeBalance.

XII. Results and Discussion

It is the implementation of ANN model which include the programming tools used, coding of the application, creation of the sample with test prediction. The coding with done in Python and its libraries. Anaconda prompt was used for the frontend where the user can run the testClassifier as well as the SinglePrediction model. Testing was done to validate the accuracy of the trained network with a random selection of data from the 20% test dataset.

After the training, the Neural Network is tested with real-life data to ascertain whether the produced output matches the expected output. In testing, the result obtained was compared with the target result. The target result is as obtained during data collection. It was observed that of the random selection out of 20% test data, the network produced a similar result as those fed into the network for training. After the simulation was done, the result below was recorded. The neural network is a good model for decision making, therefore, the result in this research is suitable for the development of staff performance evaluation system which could help organizations predict promotion for staff after analyzing the criteria for efficiency. It can also help decide either to keep a certain staff or not based on the performance evaluation scale.

A. *Training Results*

Training loss = 0.0184
Training accuracy = 0.99 (99%)
Test accuracy = 0.97 (97%)

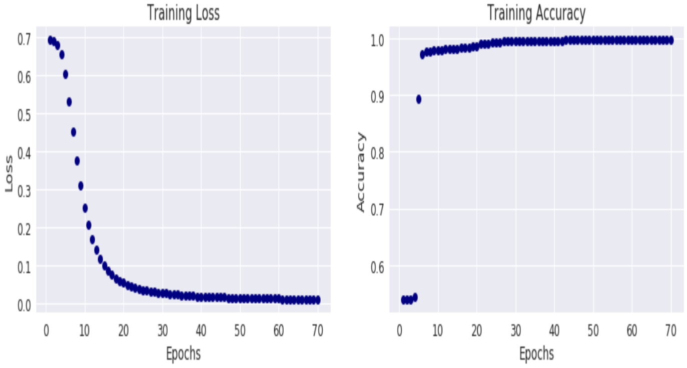


Fig. 2: Training loss and training accuracy

Training parameters
Neural Network Architecture: 3 layers. The first two hidden layers have 8 and 4 nodes respectively. The output layer has one node. Each node in the hidden layer makes use of the rectified linear unit activation function (**ReLu**). The output node makes use of the **sigmoid** function.
Hyperparameters: Training optimizer used is the **ADAM** optimizer with **binary cross-entropy** loss.

Number of training epochs: 70
Batch size: 32

XIII. Conclusion

This work presented a novel artificial intelligence (AI) approach to staff performance evaluation using a dataset obtained from the employee. The prediction was made possible by the application of Neural Network, useful for learning and decision making. Artificial Neural Network System Design and Employee Performance evaluation importance were discussed in the work.

The data collected were entered into the model after identifying the appropriate variables. Result of this study shows that PerformanceRating, OverTime, LastPromotion, Attendance, BusinessTravel have the most important effect in predicting promotion for a given staff. It means that organization HR and other stakeholders should focus on these factors when evaluating the performance of an employee. It is hoped that this research will help in simplifying the staff evaluation process and help reduce inefficiency and attrition rate of an employee in every organization.

This study addresses the prediction of measurement of employee performance using the model Neural Network. This included experiments using the back-propagation algorithm on the use of supervised Neural Network models with different network parameters and training end parameters (accuracy and epochs).

The Independent variables used are components of the dataset obtained from IBM HR Employee Attrition. The results presented above reveals the importance of these variables in the effective prediction on performance assessment. The Neural Network model demonstrated the ability to provide an adequate model for predicting promotion for staff without stress and bias or compromise. Neural Network model's efficiency is largely dependent on the data used for training and the availability of suitable dataset dictates the degree to which performance evaluation models are to be created.

References

[28] Daila K., (2018) Deep neural network for human activity recognition. International Journal of Computer Application. 180(21): 44-50.

[29] Kharola A., Mamgain R., & Jain A. (2018) Artificial Neural Network based novel performance evaluation techniques. PM World Journal 8(3): 1-12.

[30] Mohammad R. & Jianguo Z., (2019).”Evaluation of Factors Affecting Employees’ Performance Using Artifical Neural Networks Algorithm: The Case Study of Fajr Jam." International Business Research, Canadian Center of Science and Education, vol. 12(10): 86-97.

[31] Murad, A & Pyun, J. (2017) Deep Neural Network for Human Activity Recognition. Sensors (Basel Switzerland) 17(11): 2556.

[32] Namdari H., Tahami E., & Far, F. H. (2017). A comparision between the non-parametric and fuzzy logic-based classifications in recognition of human daily activities 29(01): 1750003.

[33] Okoye-Ubaka M. N., Adewole A. P., Folorunso O. & Ezike, J. O. J. (2013). Neural Network Model for Performance Evaluation of Academic Staff of Tertiary Institutions. International Journal of Applied Information Systems 5(1): 1-9.

[34] Pawlyta M., Hermansa M., Szczęsna A., Janiak M., Wojciechowski K. (2020) Deep Recurrent Neural Networks for Human Activity Recognition During Skiing. In: Gruca A., Czachórski T., Deorowicz S., Harężlak K., Piotrowska A. (eds) Man-Machine Interactions 6. ICMMI 2019. Advances in Intelligent Systems and Computing, Springer, Cham 1061(3): 231-246.

[35] IBM (2020). *HR Analytics Employee Attrition & Performance*[Data set].www.kaggle.com

Implementation of RRNS Based Architecture in DNA Computing for Single Bit Error Detection and Correction

Olatunbosun Lukumon Olawale.
ICT. Department of Computer science
Federal University of Agriculture
Abeokuta. Nigeria.
Tel: +2348029290875
Latunbosunol@funaab.edu.ng

Gbolagade Kareem Alagbe
ICT.Department of Computer science
Kwara state University
Malete. Nigeria
Tel: +2348136273074
Kazeem.gbolagade@Kwasu.edu.ng

Abstract - In this paper a comprehensive study of Residue Number System (RNS) and Redundant Residue number system (RRNS) were investigated. Emphasis was made on the propagation of (RRNS) architecture for the implementation in detection and correction of error in gene sequencing computing, application via communication line. A new RNS algorithm is premised on, with specially selected Redundant Modulus Set Projection (SSRMSP), using the Chinese Remainder Theorem CRT and moduli projection (MP) to accomplish the goal. Simple adders and Carry-Save Adders (CSAs) for the hardware architecture were implemented and compares other works with our proposed scheme, in which the number of iterations in the error correction scheme has been drastically decreases the propagation delay by 48.7% and also in turn reduces the hardware design complexity by 8.9.%

Keywords: DNA Sequence, RRNS, Selected Moduli Projections (SMP), RNS, Mutation, HC, CRT. Data transmission.

I. INTRODUCTION

Sequence computing in DNA requires great efficiency, integrity, speed and reliability which are paramount characteristics in Residue number system (RNS) as an unweighted number system has been found effective in this research area. This paper present a novel algorithm that is used with respect to specially selected moduli set projections (SSMPS) and the Chinese remainder theorem CRT for the detection and correction of single bit error data transmission in Bioinformatics computing which may be generated due to various factors like mutation, insertion and deletion in the LCS base pair, noise, heat, fault during the transmission process and other disturbances from the close system. RNS which ensures parallelism and unweighted in nature is recently been implemented to ensure error free high performance and speed via digital transmission channel, because of carry free nature. RNS with its potential features in reducing power supply voltage and hardware cost etc. **[1]** this make RNS more efficient in many applications such as DSP, communication engineering, Bio medical sciences. Cryptography, etc [2][4] involving the inclusion of specially selected Redundant moduli set (SSRMS) to aid in performing the task."Detection and correction of single bit error via digital transmission channel"
Mutation is any activity be it Chemical, Biological and Physical states that can alter the order of the sequence in the genome cell of a living organism that is more or less permanent and that can be transmitted to the cells descendants by DNA replication resulting in cell abnormal function (Biological state). Mutation also result from accidents during the normal chemical transactions of DNA like replication, exposure to high-energy electromagnetic radiation., heat, ultraviolet light (X-ray) (Chemical state) and Physical interference such as noise in the communication circuits, and along the transmission lines circuits or insertion and deletion of single base pairs in LCS. Given the complexity of DNA and the vast number of cell division and other physical activities that take place within the lifetime of multi cellular organisms, copying errors are likely to occur, if uncorrected, such errors will change the sequences alignment order of the DNA bases and alter the genetic sequences.

II. LITERATURE REVIEW

Part of DNA computing process which requires digital data transmission via communication channel can run safe premises but also prone to error. [1], [3], [11], Error result in changes into the content of data transferred. Researchers have discovered wide range logic to detect and correct the error. One way is by using hamming code an error detective technique that can detect some errors but, it is only capable of single error correction and where there is a few randomized mistake [4] which inserts (n + 1) check bit into 2n data bits with the use of logic operation XOR

(Exclusive or) in the process of detecting and correction of error, while input and output of data from the HC in the form of binary number [1], [3], [4], ,[8]. This method hamming code was invented by Richard W. Hamming in 1940s [3] but can detect errors in situation where errors are occurring randomly [4].

Another method which is adopted in this work is the use and exploitation of redundant residue number system (RRNS). Modulus projection and Chinese remainder Theorem CRT which was proposed by Gold rich el at and Yang yang et al. implementation based on read only memories (ROMs) and latches. However the cost of implementing ROMs and latches affect the speed and size of the architecture [8],[9]. Pontarelli et al [10] proposed Q novel technique for the detection and correction of single bit error in RRNS using CRT based on the FIR technique where look up table (LUT) is implemented for the purpose of single bit error detection and correction.

RNS is defined in terms of a set of relatively prime moduli. If P denotes the moduli set, then $\{P = m_1, m_2 .. mk\}, GCD\ (mi\ ,mj\ ) = 1$, for $i \neq j$. Any integer X in the range [0, M-1] such that X € (0, M-1) where

$$M = \overset{k}{\underset{I=j}{\Pi}} = m_I \qquad [1]$$

can be uniquely and unambiguously represented by the residue sequence: X↔ (x1, x2 xk) where

$$x_i = X \bmod m_i \, , \qquad [2]$$

$i = 1, 2 \ldots k.$ is the residue modulus mi of X. The range [0, M) is called dynamic range or the legitimate range of X. [6] [7].

Given a residue sequence (x1, x2, xk), the corresponding integer X in [0, M) can be uniquely recovered from the k residues using the Chinese Remainder Theorem (CRT). According to the CRT, for any given k-tuple (x1, x2, xk), where $0 \leq xi < mi$; i = 1, 2,.k, then the value of X can be found from the residue using;

$$X = \left| \sum_{i=1} M_i \left| M_i^{-1} r_i \right|_{m_i} \right|^{k}_{M} \qquad [3]$$

Where from eq 1 $M = \overset{k}{\underset{i=1}{\Pi}} = m_i$

$$Mi = \frac{M}{mi} \qquad .[4]$$

Mi^{-1} is the multiplicative inverse of the corresponding to m_i [8]

III. REDUNDANT RESIDUE NUMBER SYSTEM (RRNS)

Redundant Residue Number System (RRNS} is defined and emanates as a sub set from RNS with additional redundant moduli. Each redundant modulus is generally greater than any of the moduli of the chosen moduli set. Assuming the standard RNS consists of the moduli set of $\{m_1, m_2, m_k\}$, the corresponding RRNS consists of a moduli set of $\{m_1, m_2, m_k, m_{k+2r}\}$ $(r \geq 1)$ [3] [9] [10].The RRNS has error detection and correction capability. By using 2r $(r \geq 1)$ redundant moduli, r errors can be detected and corrected [3].

Forward conversion in any extended moduli-set consisting of more than four moduli can be accomplished equally easily by partitioning X appropriately for each modulus in the set. When the number of moduli in the set increases, the circuit-complexity will increase linearly. The complexity can be reduced if conversion is done sequentially that is, if one residue is determined at a time but doing so will result in a linear increase in conversion time, with delays introduced in the converters adding to the overheads in the overall system

A. ERROR DETECTION AND CORRECTION BASED RRNS.

These are technique used for a reliable transmission of digital data (DNA sequence) over unreliable communication channel. Noise may be subjected to communication line thereby introducing errors during transmission from source to a receiver. Error detection technique allow detecting such error while error correction enables reconstruction of the original data, this is accomplished by adding some redundancy to a message/data where receiver can use to check for the consistency and discover data determined to be corrupted for correction where an algorithm to the received data bit are compared with received output and check bits if the value do not match, an error has occurred (Ben-Gal et al, 2003)

B. PROPOSED ALGORITHM:

The algorithm has been proposed using the moduli set: $\{2^n - 1, 2^n, 2^n+1, 2^{2n} - 5, 2^{2n} + 3, \}$ for n even Where $\{2^n - 1, 2^n, 2^n+1 \}$ is the information moduli set and $\{2^{2n} - 5, 2^{2n} + 3\}$ is the redundant moduli set.

1. *INPUTE an integer number X*
2. *Compute the integer number x using the CRT*
3. *Perform interactions using factorial approach*

$$C_r^n = \frac{n^!}{(n-r)^! \mathrm{r}^!}$$

4. *Discarding a residue at time*
5. *If integer X : $x >= 1$_when x = 3 x for x = 1,2,3.4…n the illegitimate range but not found within the legitimate range*
6. *Declare the error in the residence digit*
7. *ELSE 3*
8. *END.*

The computing technique convert the moduli projection into integers by decoding the algorithm used which is premised on the moduli projection and the Chinese remainder theorem thus:

$$X_i = X(mod\,)\frac{M_{m_r}}{m_i} \qquad [5]$$

The moduli set define as $\mu = (2^n - 1, 2^n, 2^n + 1,\ 2^{2n} - 5, 2^{2n} + 3$
Where $m_1 = 2^n - 1, m_2 = 2^n, m_3 = 2^n + 1,\ m_4 = 2^{2n} - 5, m_5 = 2^{2n} + 3.$

The serial m_i projections are:

$$M_1^1 = (m_2 m_3 m_4 m_5) \ \rightarrow\ (2^n)(2^n + 1)(\ 2^{2n} - 5)\left(2^{2n} + 3 \quad\right) \qquad [6]$$

$$M_2^1 = (m_1 m_3 m_4 m_5) \ \rightarrow (2^n - 1)(2^n + 1)\ (2^{2n} - 5)(2^{2n} + 3) \qquad [7]$$

$$M_3^1 = (m_1 m_2 m_4 m_5) \ \rightarrow \quad (2^n - 1)(2^n)\ (2^{2n} - 5)(2^{2n} + 3) \qquad [8]$$

$$M_4^1 = (m_1 m_2 m_3 m_5) \rightarrow (2^n - 1)(2^n)(2^n + 1)(2^{2n} + 3) \qquad [9]$$

$$M_5^1 = (m_1 m_2 m_3 m_4) \ \rightarrow \quad (2^n - 1)(2^n)(2^n + 1)(2^{2n} - 5) \qquad [10]$$

The projections for the respective moduli are given as;

$$X_1 = \left|X\right|_{m_1^1} = \left|X\right|_{(2^n)(2^n+1)\left(2^{\,2n}-5\right)\,(2^{2n}+3)} \ [11]$$

$$X_2 = \ \left|\,X\right|_{\,m_2^1} = \left|X\right|_{(2^n-1)(2^n+1)\left(2^{\,2n}-5\right)\,(2^{2n}+3)} \cdot \qquad [12]$$

$$X_3 = \ \left|X\right|_{m_3^1} = \left|X\right|_{(2^n-1)(2^n)\left(2^{\,2n}-5\right)\,(2^{2n}+3)} \qquad [13]$$

$$X_4 = \left|X\right|_{m_4^1} = \left|X\right|_{(2^n-1)(2^n)(2^n+1)(2^{2n}+3)} .[14]$$

$$X_5 = \left|X\right|_{m_5^1} = \left|X\right|_{(2^n-1)(2^n)(2^n+1)(2^{2n}-5)} [15]$$

Both the Hardware resources implementation and the error detection emanated from the projected CRT i.e. where the corrupted integer is detected, as in the equation (17) below:

$$X_i^n = \Big| \textstyle\sum_{i=1}^{n} \quad \left|M_1^{-1}.x_i\right|_{mi} * Mi \Big|\ M_n \qquad [16]$$

Where i = 1, 5 an integer 1, 2, 3, 4, 5 represented in the moduli set M = $\{2^n - 1, 2^n,\ 2^n + 1, 2^{2n} + 5\ \ 2^{2n} - 3.\ \}$

This gives $X_{1\,2\,3\,4\,5} = \ \Big| \sum_{i=1}^{n=3} \left|xiM_i^{-1}\ \right|_{mi} * M_i\Big|_{\ M}$

[17]

IV HARDWARE IMPLEMENTATION:

(i). the moduli set of the architecture is implemented in two segments:

(i). for the first three non redundant part
$m = (2^n - 1, 2^n, 2^n + 1)$……($x_{1.}$,$x_2$,$x_{3)}$

(ii). for the last two Redundant part
$m = (2^{2n} + 5)\ (2^{2n} - 3).)$
with respect to the corresponding …(x_4,$x_{5)}$

The binary representation of the residues for the combine moduli set:

$$x_1 = x_{1,}\, n - 1 \ldots \ldots .\, x_{1\,1} x_{1,0} \qquad [18]$$

$$x_2 = x_{2,}\, n - 1 \ldots\ x_{2,1}\ x_{2,0} \qquad [19]$$

$$x_3 = x_{3,n,n-1} \ldots\ x_{3,1} x_{3,0} \qquad [20]$$

A. REDUNDANT MODULI SET

X_4 = X4.2n-1,X4.2n-2 ….X4.0 [21]

$x_{5=}\, x_{5,2n,2n-1,2n-2,\ldots x5,0}$ [22]

Applying the CRT for the all moduli sets

$$X_{12345} = \Big|\ _{x_1} \left|M_1^{-1}\right| m_1 * M_1 + x_2 \left|M_2^{-1}\right|_{m_2} * M_2 +\ x_3 \left|M_3^{-1}\right|_{m_3} * M_3 + x_4 \left|M_4^{-1}\right|_{m_4} * M_4 + x_5 \left|M_5^{-1}\right|_{m_5} * M_5 \Big|\ _{M} \qquad [23]$$

Bit-level representation for the non-redundant residue NRR: is given as:

$$x_1 = (x_{1,n-1}\, x_{1,n-2} \ldots. x_{1,1} x_{1,0})_2 \qquad [24]$$

$$x_2 = \left(x_{2,n-1} x_{2,n-2} \ldots x_{2,1} x_{2,0}\right)2 \ldots \qquad [25]$$

$$x_3 = \left(x_{3,n-1}\right)..\,(_{3,1} x_{3,0})_2 \ldots \ldots \ldots \qquad [26]$$

This non redundant value $x_1 x_2 x_3$ can be calculated from CRT in equation (3) thus;

$$X_{1,2,3} = \left| M_1 \left| M_1^{-1} \right|_{m_1} * M_1 + x_2 \left| M_2^{-1} \right| m_2 * m_2 + x_3 \left| M_3^{-1} \right| m_3 * M_3 \right|_{m} \quad [27]$$

The Dynamic range M for (NRM) is given as;

$$\mathrm{M} = m_1\, m_2\, m_3 = (2^n - 1)(2^n)(2^n + 1).. \quad [28]$$

$$M_1 = m_2\, m_3 = (2^n)(2^n + 1) \quad [29]$$
$$M_2 = m_1 m_3 = (2^n - 1)(2^n + 1) \quad [30]$$
$$M_3\ m_1 m_2 (2^{n-}1)(2^n) \quad [31]$$

The multiplicative inverses for the (NRM)

$$\left| M_1^{-1} \right| m_1 = (m_2 m_3)^{-1}_{m_1} = \left| \left((2^n)(2^n - 1)\right)^{-1} \right| 2^n - 1 = \mathbf{2^n - 2}$$
[32]

$$\left| M_2^{-1} \right|_{m_2 =} (m_1 m_3)^{-1}_{m2} = \left| [(2^n - 1)(2^n + 1)]^{-1} \right| 2^n = \mathbf{2^n - 1} \quad [33]$$

$$\left| M_3^{-1} \right|_{m_3 =} (m_1 m_2)^{-1}_{m3} = \left| [(2^n - 1)(2^n)]^{-1} \right| 2^n + 1 = \mathbf{2^n - 1} \quad [34]$$

Expand equation. (17) With respect to the NRM dynamic range of the multiplicative inverse progression yield:

$$\boldsymbol{X_{123} = \left| x_1\left((2^n - 2)(2^n + 1)\right) + x_2\left((2^n - 1)(2^n - 1)(2^n + 1)\right) + x_3\left((2^n - 1)(2^n - 1)(2^n)\right) \right| (2^n - 1)(2^n)(2^n + 1)} \quad [35]$$

The X_{123} from equation **27** above, NRM correct result is achieved where an error occurs in the NRM part; this will be detected in the RMS part (2)

V. HARDWARE ARCHITECTURE

The architecture is built on two fundamental techniques used to attain better performance: the first is the use of carry-save adders; and the second is the exploitation of more parallelism. The scheme were well_implemented to the derivatives of the 2^n moduli sets of the form (2^n - 1*;* 2^n*;* 2^n + 1*;*2^{2n}-5 and 2^{2n} +3) for even n and odd respectively. In the proposed scheme shown in schematic Figure1.below is a design based on these techniques. A carry-save adder (CSA) imposes a Delay of D_{FA} with a sum area of (12n+3) ΔFA as compare with 12n+5) DFA total delay require in proposed scheme. The CSA takes three operands and produces two outputs, a partial-sum (PS) and a partial-carry (PC), with just a sequence of full adders. The addition time is constant and independent of the number of digits with a series of additions, the carries is saved for propagation at the last addition that is fed finally into a carry-propagate adder (CPA) with a bit length of 3n+1 bits such that the carries are propagated between digits to produce a result in conventional form. The addition time requires a function of the number of digits for n digits, the operational time with an asymptotic complexity between the order O(*n*) and O(log *n*). The computation of r_1 requires a corrective subtraction of the modulus $\boldsymbol{m_3}$, and that of $\boldsymbol{r_3}$ require the subtraction $\boldsymbol{2m_1}$. All the different possible results are computed in parallel and the correct one then selected through a multiplexer.

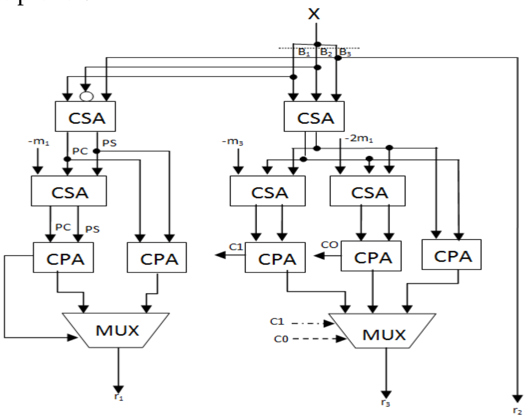


Fig. 1. The schematic Architecture of Hardware Performance ($_2{}^n$ - 1, 2^n, 2^n $_+$1)

In the computation of the residue corresponding to 2^n+1, the binary representation of X is partitioned into blocks of *n*+ 1 bit each. The procedures illustrates beneath is used to compute the residues for each of the two sets [7],[14].

Example: Consider the moduli-set {3; 4; 5 ;}: which are relatively prime, and the third modulus is of the form $\mathbf{2^n}$ + 1, with *n* = 2. If X is 6689, in order to obtain the residues with respect to 3,4 and 5, Proceeding by partitioning the binary representation of 6689, which is **10 00 00 10 00 01**, into 2-bit blocks, since *n* = 2:

$|6689|_5 = |10 + 00 + 00 + 10 + 00 + 01|2^n + 1$
$|6689|5 = |10+00 + 00 + 10 + 00 + 01|_5$
$= |3+0+ 0 + 3 + 0 + 1|5 \rightarrow \mathbf{2}$
$\mathbf{= 2}$

Similarly, the residue with respect to 3 is obtained as

$|6689|_3 = |10 + 00 + 00 + 10 + 00 + 01|2^n - 1$
$|6689|3 = |10+00 + 00 + 10 + 00 + 01|_3$
$= |3+0+ 0 + 3 + 0 + 1|3 \rightarrow \mathbf{1}$

= 1

The proposed scheme is very simple and straightforward and the time it will take to correct an error is relatively less than other detection and correction schemes and error in one digit does not corrupt any other digits. Consiqentlly, the use of redundant moduli that play no role in determining the dynamic range but facilitate both error detection and correction which is done at the reverse converse of the general RNS processor where the algorithm is incorporated in a reverse converter.

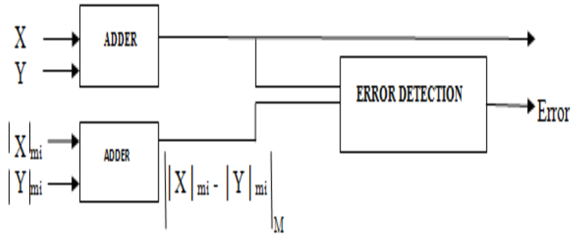


Fig. 2. Circuit for Error Detection Scheme (Source: Ben – Gal et al) **[3]**

A. PERFORMANCE ANALYSIS

Results of this paper were compared with the results in [12] where the number of iterations performed in order to correct a single error with 5- moduli set is twelve; that is a complete recombination. However, if we decide to generalize our observation such that we assume we have an RRNS with R redundant moduli, with R ≥ 3 and the moduli satisfy the magnitude-condition stated above. Then a single residue error in the ith position in the representation of X, a number in the legitimate range, can be unambiguously located. Whence, there will be only one **mi** projection of the resulting illegitimate number that lies in the legitimate range, and this number is the correct, legitimate value.

Suppose the moduli of an RRNS (R = 3) satisfy the magnitude condition given above and that the representation of a legitimate number, *X*, has been corrupted into that of an illegitimate number, x̶ Also, suppose that x̶*i* is a legitimate projection of x̶. Then all other projections, x̶ j , j ≠ i are illegitimate. Furthermore, x̶i will be the correct value for **X**. Therefore, it must be the case that the error is in the residue *xi* and that the correct value for that digit is │x̶i│mi

VI. NUMERICAL RESULT

Suppose the longest common sub sequence LCS in DNA sequence alignment from nucleotide is to be transmitted via communication line from the source to the receiver for further investigation implementation, data transmission which could be in form of data, voice or image there is the possibility of error detection and correction process performed by the transmitter by adding a few extra bits into the data to be transmitted in this case, as in High speed modems ,copying errors such as noise, interference in the communication circuits, and along the transmission lines circuits or insertion and deletion of single base pairs in LCS are likely to occur, and if undetected, uncorrected such errors will effect a change or modification in the implementation sequences alignment order of in the DNA bases.

Example: Take the RRNS with the moduli-set {3; 4; 5; 11; 19}, where 11 and 19 are redundant the legitimate range is [0*;* 60), and the total range is [0*;* 12540). Suppose the representation of the legitimate number 1673 ≡ (2; 1; 3; 1; 1) has an error in the residue-digit corresponding to the modulus 3. Then the new number is x̶ ≡ *(*2*;* 1*;* 6*;* 1*;* 1*)*. The value of x̶ is 6689, and the projections of 6689 with respect to each modulus are:

Table 4 : Numerical Result in MP and x̶-Value

x̶-Value	Modulus Projection [MP]	Numerical Result
6689_3	$\left|6689\right|_{4180}$	2509
668_4	$\left|6689\right|_{2135}$	419
$\mathbf{6689}_5$	$\left|\mathbf{6689}\right|_{2508}$	**1673**
6689_{11}	$\left|6689\right|_{1140}$	989
6689_{19}	$\left|6689\right|_{660}$	89

Observation form table 4 above now shows that there is only one legitimate projection the one with respect to the modulus 5. We therefore conclude that the erroneous digit is in the third residue and that the correct value for that digit is │1673│5, which is 3.

ITERATION PROCESSES				
C RT for X_{Mn-i}			**MP for $X_{NRM} = 6689$ constant**	
	$X_M{}^1{}_{n-i}$	**Integer-value**	**X_{123}**	**Integer-value**
1	$X_M{}^1{}_5$	2514	$\left|X\right|_{m^1 5}$	2509
2	$X_M{}^1{}_4$	419	$\left|X\right| m^1{}_4$	419
3	$Xm^1{}_3$	**1673**	$\left|X\right| m^1{}_3$	**1673**
4	$Xm^1{}_2$	1169	$\left|X\right| m^1{}_2$	989
5	$X_M{}^1 1$	89	$\left|X\right| m^1{}_1$	89

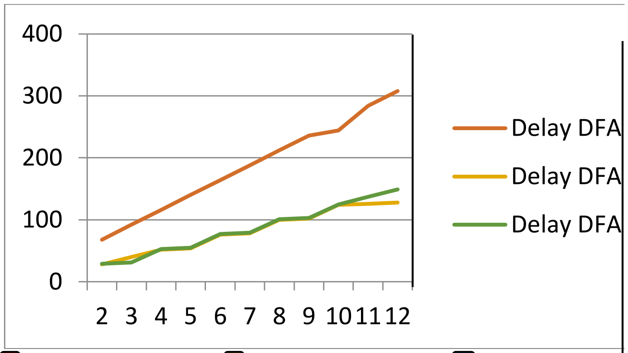
Table 1 : Iteration Processes for CRT and Moduli Projection

Schemes	**Area (Δ_{FA})**	**Delay D_{FA}**
Related work [10]	(12n + 12) Δ_{FA}	24n + 20 D_{FA}
Related work [12]	(12n + 2) Δ_{FA}	(12n + 4) D_{FA}
Proposed	(12n + 3) Δ_{FA}	(12n + 5) D_{FA}

Table 2: comparison illustration of Area and Delay

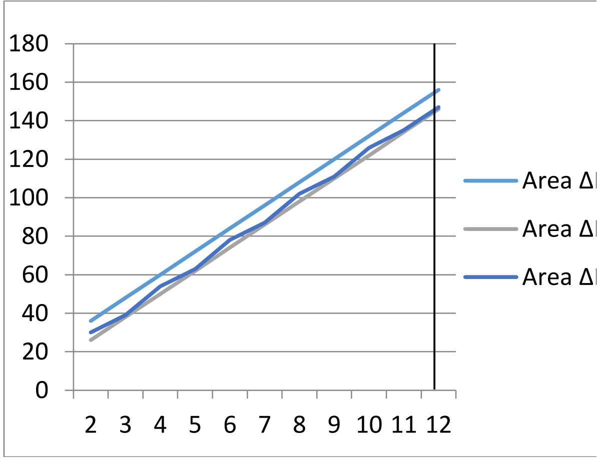
	[10]		[12]		Proposed	
s/n	Area Δ_{FA}	Delay D_{FA}	Area Δ_{FA}	Delay D_{FA}	Area Δ_{FA}	Delay D_{FA}
2	36	68	26	28	30	29
3	48	92	38	40	39	31
4	60	116	50	52	54	53
5	72	140	62	54	63	55
6	84	164	74	76	78	77
7	96	188	86	78	87	79
8	108	212	98	100	102	101
9	120	236	110	102	111	103
10	132	244	122	124	126	125
11	144	284	134	126	135	137
12	156	308	146	128	147	149

Table3: Performance comparison using different n- value



🟥 Related work [🟨0] Related wor🟦[12] Proposed

Fig.3: Delay comparisms between [10],[12] and proposed Scheme.



🟦 Related work [1🟩] : Related work [⬜2] : Proposed Scheme.

Fig.4: Area comparism between [10],[12] and proposed Schemes.

In comparison with [10] and [12] and proposed scheme, it can be seen theoretically as illustrated in the table (3) and fig. (3, 4) that that the proposed scheme perform better in terms of speed and hardware complexity with significant difference of 158 $_{DFA}$ and 9 **Δ_{FA} respectively.**

VII. CONCLUSION

This paper proposes a new scheme on modular set that detects and corrects single bit errors with 2^n-1, 2^n, 2^n+1, and 2^{2n}-5, 2^2+3 based on MP and traditional CRT technique. The MP reduces the iteration process and enhanced the speed of the architecture. The reliability of data transmission in DNA computing over communication channel is enhanced as compared to [10] and [12] such that if an error occur in the record integer message using redundant modulus, the error control scheme in this work detect and correct the single bit error.

REFERENCES

[1]. Lalitha, K.V and Sailaja,V.(2014) "High Performance Adder using Residue Number System". International Journal of VLSI and Embedded Systems (IJVES), ISSN 2249-6556, Vol. 5, No. 5, pp. 1323-133.

[2] Taylor F. A. (1985)"Single Modulus ALU for Signal Processing", IEEE Transactions on Acoustics, Speech, Signal Processing, Vol. 33, pp. 1302-1315.

[3] Deepika, A.Kumarand Gurusiddayya ,(2016)"A Study on Error codding Techniques "International Journal for Research in Applied Science and Engineering Technology (IJRASET), vol.4,no.4, pp.825-8828,.

[4] Fitriani.W. and Siahaan.A.P.U.(2016) 'Single-Bit Parity Detection and Correction using Hamming code 7-Bit Model, International Journal of Computer Application vol. 154, no.2, pp.12-16,

[5] Hamming.(1950) "error Detection and Error Correction Codes,"The Bell System Technical Journal, vol.29,no.2,

[6].Gbolagade.K.A.and.Daabo,M.I (2014)"An Overflow Detection Scheme with a Reverse Converter for the Moduli set $\{2^n - 1, 2^n, 2^n + 1\}$ ".Journal of Emerging Trends in Computing and Information Sciences, ISSN 2079-8407, Vol. 5, No. 12, pp. 931-935.

[7] Omondi, A. and Premkumar.B.(2007) "Residue Number System theory and Implementation" Imperial College, Press.

[8]Roshanzadeh,M.Ghaffari.A.andSaqaeeyan,S.(2011) Using Residue Number Systems for Improving QoS and Error Detection and Correction in Wireless Sensor Networks, Communication Software and Networks (ICCSN), IEEE 3rd International Conference on Page: 1-5 .3

[9] Gbolagade, K. A. (2011)An Efficient MRC based RNS-to-Binary Converter for the moduli set, {22n+1-1, 2n, 22n-1}, AIMS SA,

[10] Pontarelli S., Cardarilli, G.C., Re, M., & Salsano A. “A Novel Error Detection and Correction Technique for RNS Based FIR Filters”, IEEE International Symposium on Defect and Fault Tolerance of VLSI Systems, 2008, pp. 436-444. Doi:10.1109/DFT.2008.32

[11] Duc-Minh Pham, A. B. Premkumar and A. S. Madhukumar, Error Detection and Correction in Communication Channels Using Inverse Gray RSNS Code, IEE Transactions on Communications 59(4): 975-986 April 2011.8

[12] Jenn-Dong Sun and Hari Krishna, Fast Algorithm for Multiple Errors Detection and Correction in Redundant Residue Number System Journal of Circuit, Systems and Signal Processing December Volume 12, Issue Issue 4, pp 503-531, 1993.

[13] Hari Krisna, Kuo-Yu Lin, and Jenn-Dong Sun,(1992) “A coding Theory Approach to Error Control in Redundant Residue Number Systems” Part I: Theory and Single Error Correction, IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing Vol 39 issue 1 pp 8-17 .

[14] Mohan. A.(2002).“Residue Number Systems: Algorithms and Architectures”. Kluwer Academic Publishers, Dordrecht.

An Improved RNS Based Data Security by Combination of Compression, Cryptography and Steganography

Eseyin Joseph B.
ICT Directorate, University of Jos,
eseyinjb@unijos.edu.ng

Kazeem A. Gbolagade
Department of Computer Science, College
of Library and Information Technology,
Kwara State University, Malete, Ilorin,
kazeem.gbolagade@kwasu.edu.ng

ABSTRACT
Compression, Cryptography and steganography are distinct ways of sending and receiving information in a secured manner. Their functionalities differ in the way of securing data and information. One hides the existence of the message while the other scrambles the message. There are many cryptographic techniques existing and among them RSA is one of the most powerful techniques. In Steganography, various techniques in different domains like spatial domain, frequency domain are also available to hide messages. It is a herculean task to perceive a hidden message in a frequency domain.

In this paper we propose a new technique in which Compression, Cryptography and Steganography are combined along with an enhanced RNS based security module for transmitting text messages.

In this paper our new system is proposing the use of Compression with the application of Residue Number System on cryptography and steganography for better confidentiality, security and achieves a faster encryption and decryption method. This system will hide more data in an image and lessening the transmission time. In this approach, the input plain text will be converted into its Residue Numbers System equivalent and compressed using Huffman compression and the compressed text will be encrypted using RSA algorithm, after encryption the compressed cipher text is hidden in a DCT of an image. The application of a compression techniques will improve data hiding capability and cipher hiding in steganographic image and application of CRT make the system more secured and better deal with. Complexity issues and enhanced the speed of execution and reduced computational cost.

Keyword: Cryptography, Steganography, Stego- image, Compression, Threshold Value, DCT Coefficient

I INTRODUCTION

Cryptography [1] is a universally known scheme that is used to deploy messages in other to obtain a secret message. And steganography [1] is the science and art of communication which hides the presence of a message in a medium so that the third party would not be able to deduce its existence. Cryptography distort the message in a way that it cannot be understood by the non-intended recipient. But steganography furs the message as not to be seen nor noticing the presence of such message.

In this paper we are proposing a Residue Number Based system which combines compression with cryptography and steganography for improved confidentiality, security and accomplish a quicker decryption method. In doing this we can hide a greater amount of data in an image and lessen the decryption time. Currently RSA algorithm is a well-protected and secured cryptographic method and the techniques that uses frequency domain is well-thought-out to be secured for steganography. But considering the desperation of the attackers coalescing these techniques directly may give room for the intruder to detect the original message after several attempts. Thus, our proposal is to use the Huffman compression method [2] to compress the text message and the compressed message in encrypted using the RSA algorithm which will be converted to its Residue Number System equivalent using Chinese Remainder Theorem. The output of the encryption will be hidden in and image through the DCT techniques of steganography with more security to get a more secured system for data hiding. In this approach we would develop a new system with extra security features where a meaningful communication can be hidden by an RNS based combination of the three techniques of compression, cryptography and steganography. Our main goal of the paper is to propose a new system which is highly secured and even if the message is intercepted the message from stego image [3] becomes worthless for any prevailing cryptographic techniques.

II CONCEPTS AND RELATED WORK

There are various categories of security applications and one important aspect of this secured communications is cryptography. It is often observed that considering the desperation and sophistication of attackers' cryptography nowadays is not in itself sufficient. It is not easy to hide the whole encrypted data in an image using DCT technique in steganography; therefore, we use compression technique first before encrypting the data.

There are several compression methods:

- Huffman Compression
- LZW compression
- Arithmetic coding and so on.

But in this paper, we are using Huffman compression method.

A. Huffman Compression

Huffman coding [4] deals with the frequency of incidence of a data item which can be the number of pixels in images this method made use of the lesser number of bits in encoding data that appears most often. This is a lossless data compression algorithm that was technologically advanced by David Huffman in the early fifties. The algorithm was based on a binary-tree frequency-sorting that allow encoding any message sequence into petite encoded messages and reassemble it into the original message without loss of data. Huffman compression is an algorithm with a variable code word length. Huffman compression fit in into a household of algorithms with a variable codeword length. In this algorithm individual characters in a text file are swapped by bit sequences that have a distinct length. Therefore, characters that occur most often in a file are given a short sequence while other that are used infrequently get a longer bit sequence. The basic idea behind the algorithm is to build the tree bottom-up whose leaves are labeled with the weights. In the construction of Huffman algorithm, the weights denote the probabilities that are linked with the spring letters. Primarily there is a set of singleton trees, one for each weight in the list.

At each step in the algorithm the trees corresponding to the two smallest weights, $w(i)$ and $w(j)$, are merged into a new tree whose weight is $w(i)+w(j)$ and whose root has two children which are the sub trees represented by $w(i)$ and $w(j)$. The weights $w(i)$ and $w(j)$ are detached from the list and $w(i)+w(j)$ is fixed into the list. This process goes on until the weight list contains a single value. If we have more options to a smallest pair of weights, any such pair may be chosen. In the Huffman algorithm, the process usually begins with a non-increasing list of weights since it provides an efficient implementation.

Example:

Table 1

(a)A NON-INCREASING LIST OF WEIGHT

A1	0.25	0.25	0.25	0.33	0.42	0.58	1.0
A2	0.20	0.20	0.22	0.25	0.33	0.42	
A3	0.15	0.18	0.20	0.22	0.25		
A4	0.12	0.15	0.18	0.20			
A5	0.10	0.12	0.15				
A6	0.10	0.10					
A7	0.08						

(b)

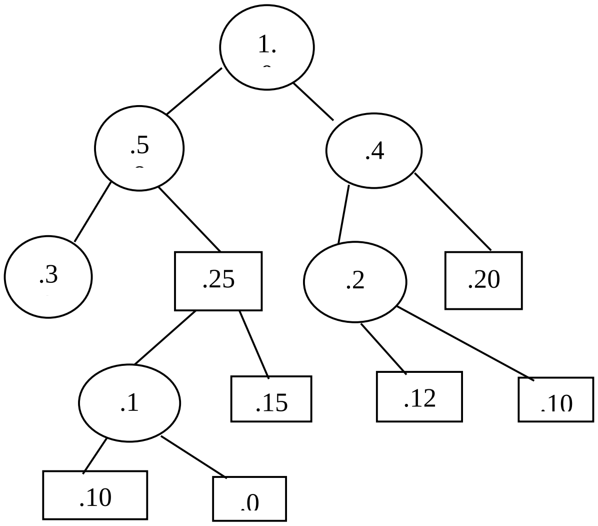


Fig. 1 Sample Huffman binary tree

In our proposal compressing the plaintext before RSA will ensure more hidden plain text volume in an image indirectly.

B. Cryptography Techniques

Since we have discussed the Huffman compression algorithm there is need to briefly talk about Cryptographic algorithms. There are some specific security requirements [5] for cryptography, such as Authentication, Privacy/confidentiality, Integrity/Non-repudiation.

There are three well known cryptographic algorithms such as:

(i)Secret Key Cryptography (SKC) which made use of a single key for both encryption and decryption

(ii)Public Key Cryptography (PKC): Uses one key for encryption and the other for decryption while

(iii)Hash Functions: Use a mathematical transformation to irreversibly "encrypt" information. But our discussion shall only concentrate on the public key cryptography.

1. Public Key Cryptography

Public key cryptography has its own public key which is well known to everyone and has a matching private key which is only known to the projected recipient. Most of the public key cryptosystems are based on one-way trapdoor function, where the encryption rule is easy to compute, but decryption rule is computationally infeasible without any additional information. Thus, the security of a public key cryptosystem is based on the intractability of hard mathematical problems such as integer factorization problem (IFP), discrete logarithm problem.[1]

C. Steganography Techniques

Steganography is the other technique for secured communication. It encompasses methods of transmitting secret messages through innocuous cover carriers in such a manner that the very existence of the embedded messages is undetectable. Information can be hidden in images [6], audio, video, text, or some other digitally represented code. Steganography systems [7] can be grouped by the type of covers [8] used (graphics, sound, text, executables) or by the techniques used to modify the covers

- Substitution system [9].
- Transform domain techniques [10]
- Spread spectrum techniques [10]
- Statistical method [11]
- Distortion techniques [12]
- Cover generation methods [13]

We have used DCT frequency domain algorithm in our proposed system.

D. DCT-frequency domain algorithm for Steganography [13]

According to the method presented in this paper, the compressed encrypted message is inserted into the DCT domain of the host image. The hidden message is a stream of “1” and “0” giving a total number of 56 bits. The transform is applied to the image as a multiple factor of 8x8 blocks. The next step of the technique after the DCT is to select the 56 larger positive coefficients, in the low-mid frequency range. The high frequency coefficients represent the image details and are vulnerable to most common image manipulation like filtering, compression [15] etc. Our scheme is applied to the whole image and since robustness is the main issue, the low and mid frequency coefficients are the most appropriate. The selected coefficients *ci* are ordered by magnitude and then modified by the corresponding bit in the message stream.

If the *i*th message bit ***s(i)*** to be embedded is “1”, a quantity *D* is added to the coefficient.

This *D* quantity represents the persistence factor.

If the message bit is “0”, the same quantity is subtracted from the coefficient.

Thus, the replaced DCT coefficients are

DCT (new) = DCT+1*D for s(i)=1;

Else

DCT (new) =DCT-1*D

for s(i)=0.

DCT can separate the Image into High, Middle and Low Frequency components. To hide information, we need to set a threshold value [14] for the DCT coefficients depending on the quality of the images.

1. Advantages of using frequency domain Steganography

Very secure, hard to detect, flexible, different techniques for manipulation of DCT coefficients values

III METHODOLOGY

A. A propose technique for combination

The design for combining the two different techniques and applying the Residue number system is purely based on the idea – converting the message using Chinese Remainder Theorem, distort the converted message, and hide the existence of the distorted message in an image. And to get back the original message – retrieve the distorted message and regain the actual message by reversal of the distortion process.

Here we design the system with four modules-

For Compression – Compression Module, for Cryptography - Crypto Module, for Residue Number System – RNS Module, for Steganography - Stego Module For extra security - Security Module

The extra security module that we are providing make this system highly secured. The process flow for the system is as follows-

B. Hiding the Text

Compressed Module:

For Compressed Module basically we compressed the Input plain text using Huffman algorithm as shown in Figure**1**.

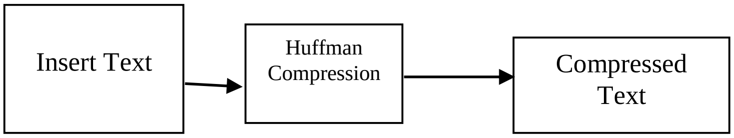


Fig. 2 Compressed Module

C. Residue Number System Module

This will enhance another measure of security to the system. The Residue Number System architectures offer an inherent protection mechanism. The key to security is the Base Conversion algorithm. It will render the hardware fault attack useless.

The complexity of RSA Cryptosystem depends heavily on the size of decryption exponent, d and the modulus, n regarding the decryption process. This exponent indicates the number of modular multiplications that is needed to perform the exponentiation. Therefore, to reduce the size of both d and n we employ the use of the Chinese Remainder Theorem for faster computation and less propagation delay.[15]

Given pairwise coprime positive integers n1,n2...nk And integers a1,a2,...ak, the system of simultaneous congruences are as follows:

$x \equiv a1(mod(n1))$

$x \equiv a2\ (mod(n2))$

.

.

.

.

$x \equiv ak\ (mod(nk))$

These congruences have a solution, and the solution is unique modulo:

N = n1 n2…nk

The following is a general method to find a solution to the system of congruences using the Chinese remainder theorem:

1. Compute N=n1×n2×…nk .
2. For each i=1,2,…k, compute
   Y1 =n1 n2…ni-1 ni+1 …nk
3. For each i=1,2,3…k , Compute
   Zi =yi(-1) mod ni
4. By using Euclid's extended algorithm
   (zi exists since n1 n2 …nk are pairwise coprime).
5. The integer $x= \sum^{k}_{i=1}$ a1, y1,z1
   will be a solution to the system of congruences, and x mod N is the unique solution modulo N. [17]

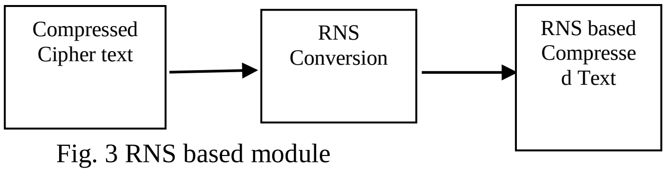


Fig. 3 RNS based module

D. Cryptographic Module:

For Crypto Module the following steps are considered for encrypting the data as shown in

- Insert compressed text for encryption.
- Apply RSA algorithm using 128 bit key (Key 1).
- Generate Cipher Text.

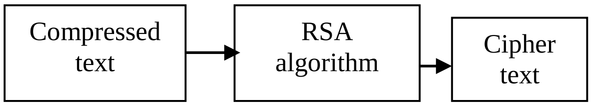


Fig. 4 Cryptographic Module

E. Steganographic Module:

For Stego Module the following steps are considered for hiding the above generated Cipher text.

- Take a Gray Scale Image.
- Find the DCT of the Image.
- Hide the Cipher by altering DCTs.
- Apply Inverse DCT.
- Find the Stego Image.

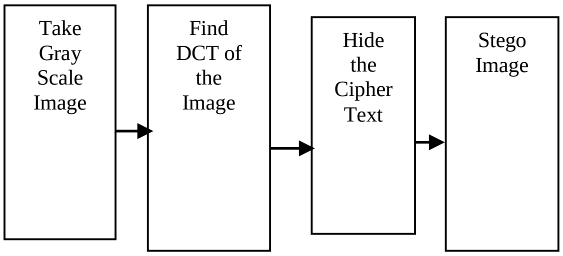


Fig. 5 Steganographic Module

Retrieving Text

Steganographic Module (Reverse Process)

For Steganographic Module the following steps are considered for retrieving the cipher text

- Take DCT of the Original Image.
- Take DCT of the Stego Image.
- Take difference of DCT coefficients.
- Retrieve the original image.

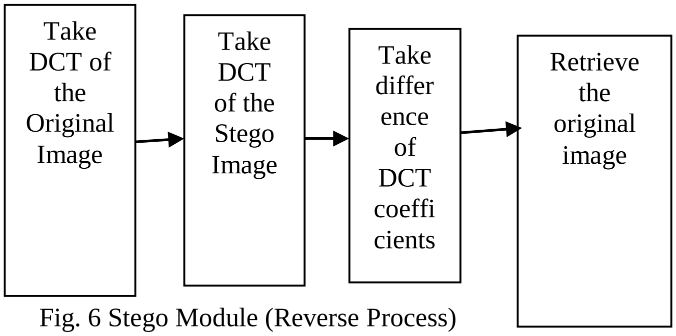

Take DCT of the Original Image
Take DCT of the Stego Image
Take difference of DCT coefficients
Retrieve the original image

Fig. 6 Stego Module (Reverse Process)

Cryptographic Module (Reverse Process):

For Cryptographic Module the following steps are considered for retrieving the original text.

- Get the above retrieved cipher text.
- Reverse RSA algorithm by using Key 1.
- Get the Compressed message.

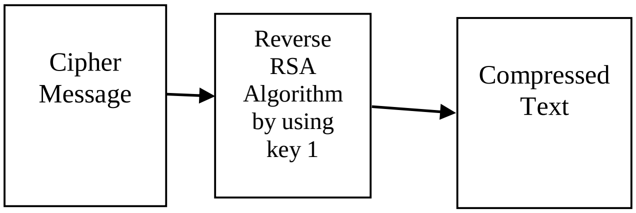


Fig. 7 Cryptographic Module (Reverse Process)

Decompressed Module

- Apply Huffman decompression technique to the text retrieved from the Cryptographic Module.
- Get the above retrieved text.
- Apply decompression using Huffman Decompression technique.
- Retrieve the Residue Number based compressed text.

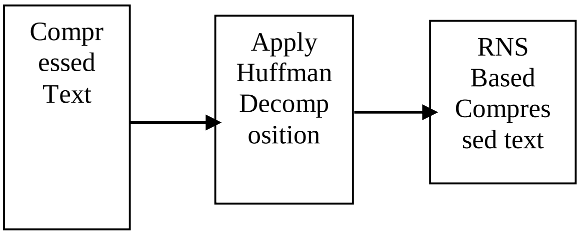


Fig. 8 Decompressed Module

Reversed Residue Number System Conversion Module

- Apply CRT to the text retrieved from the decompressed module
- Perform reversed conversion
- Get the original text

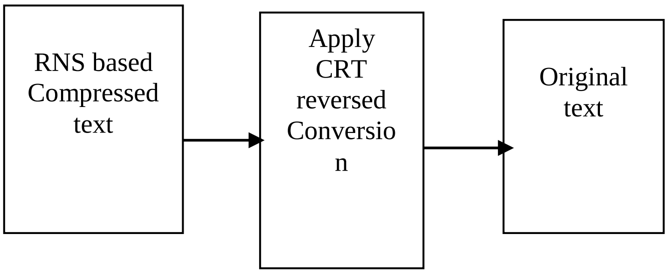


Fig. 9 RNS reversed conversion module.

4. Implementation Details

This new technique is developed with VC6.0 platform using VC++.

The four modules involved –

- Compression Module – Huffman Compression Module
- Residue Number System Module – Conversion to RNS using CRT
- Cryptographic Module - RSA Implementation Module
- Steganographic Module - DCT Techniques Implementation Module

These modules are designed and coded as reusable components and can work independently.

4.1 **Tools and Libraries used**

i. Arisimage Routines [16]

ii. Cximage599c [17] these libraries are available with free licenses.

5. The Security of the proposed system.

The proposed solution is highly secure since it is a combination of four techniques**:**

- Huffman Compression
- RNS conversion using CRT
- RSA for cryptography
- DCT manipulation for steganography

It would be very difficult if not impossible for an attacker to gain access to the message. Since hiding the data in an image is better than sending the encrypted message directly. This was done by hiding the encrypted message in a popular object that will not attract any attention. And peradventure, if the data is extracted it will be in the compressed form which will still be very hard to crack. After decompression, there is still the challenge of reverse conversion using the Chinese Remainder Theorem. Infact, only the intended recipient can have the clue of how the message can be encrypted. In this system to get the original message one should know, along with keys for Compression, Cryptography and Steganography and be conversant with reverse conversion using Residue Number System.

6. Conclusion

In this paper we have presented a new system for a Residue Number System based combination of Compression, cryptography and Steganography which could be proven as highly secured method for data communication in near future. Steganography especially combined with cryptography and compression is a powerful tool which enables people to communicate without possible eavesdroppers. The proposed method provides acceptable image quality with very little distortion in the image and highly secured means of communication in the net.

REFERENCES

[1] Domenico Daniele Bloisi , Luca Iocchi, "Image based Steganography and cryptography, Computer Vision theory and applications" volume 1, pp. 127-134 . August, 2010

[2] Mamta Sharma, S.L. Bawa "Compression Using Huffman Coding", IJCSNS International Journal of Computer Science and Network Security, VOL.10 No.5, May 2010.

[3] Kharrazi, M., Sencar, H. T., and Memon, N.. "Image Steganography: Concepts and practice". In WSPC Lecture Notes Series.

[4] David A. Huffman+, Associate, A, "Method for the Construction of Minimum Redundancy Codes" proceedings of the IRE. 1952

[5] Daemen, Joan; Rijmen, Vincent. "AES Proposal": Rijndael. Source:

http://csrc.nist.gov/encryption/aes/rijndael/Rijndael.pdf.1991.

[6] Owens, M., "A discussion of covert channels and steganography", SANS Institute. 2002.

[7] Chandramouli, R., Kharrazi, M. & Memon, N., "Image Steganography and steganalysis: Concepts and Practice", Proceedings of the 2nd International Workshop on Digital Watermarking, October 2003.

[8] Jamil, T., "Steganography, " The art of hiding information is plain sight", IEEE Potentials, 18:01, 1999.

[9] Stefan Katznbeisser, Fabien.A, P.Petitcolas editors, "Information Hiding Techniques for Steganography and Digital Watermarking", Artech House, Boston. London, 2000.

[10] Wang, H & Wang, S, "Cyber warfare: Steganography vs. Steganalysis", Communications of the ACM, 47:10, October 2004.

[12] Dunbar, B. , "Steganography techniques and their use in an Open-Systems environment", SANS Institute, January 2002.

[13] C.E., Shannon, "Communication theory of secrecy systems", Bell System Technical Journal, 28, 656-715. 1949

[14] Marvel, L.M., Boncelet Jr., C.G. & Retter, C., "Spread Spectrum Steganography", IEEE Transactions on image processing, 8:08, 1999.

[15] Currie, D.L. & Irvine, C.E., "Surmounting the effects of lossy compression on Steganography", 19th National Information Systems Security Conference, 1996.

[16] Gbolagade K.A. and Saheed Y.K."RSA Cryptosystem Encryption Based on Three Moduli set With Common Factors {2n+2, 2n+1, 2n}.Computing and Information SystemsJournal, University of the West of Scotland, USA. 2018.

[17] Curtis Clement, Satvik Golechha, Pi Han Goh "Chinese remainder theorem and proof " https://brilliant.org/wiki/chinese-r emainder-theorem/. 2004

[18] Aris Adrianto S http://www.codeproject.com/KB/library/ArisFFTDFTLibrary.aspx. 2004

[19] http://www.xdp.it

Performance Evaluation of ANOVA and RFE Algorithms for Classifying Microarray Dataset Using SVM

Sulaimon Olaniyi Abdulsalam, Jumoke F. Ajao, Micheal Olaolu Arowolo , Ronke Babatunde and Abubakar Adamu
Dept.of Computer Science
Kwara State University, Malete,
Nigeria
abdulsalamny@gmail.com jumoke.ajao@kwasu.edu.ng arowolo.olaolu@gmail.com ronke.babatunde@kwasu.edu.ng
mohammedabubakaradamu@gmail.comju

ABSTRACT— A significant application of microarray gene expression data is the classification and prediction of biological models. An essential component of data analysis is dimension reduction. This study presents a comparison study on a reduced data using Analysis of Variance (ANOVA) and Recursive Feature Elimination (RFE) feature selection dimension reduction techniques, and evaluates the relative performance evaluation of classification procedures of Support Vector Machine (SVM) classification technique. In this study, an accuracy and computational performance metrics of the processes were carried out on a microarray colon cancer dataset for classification. SVM-RFE achieved 93% compared to ANOVA with 87% accuracy in the classification output result.

Keywords—SVM-RFE, ANOVA, Microarray, SVM, Cancer.

I Introduction

In biological learning, Next-generation sequencing (NGS) has been expansively utilized. General NGS information is the Ribonucleic Acid sequencing (RNA-seq); it is utilized to test the anomalies of mRNA expression in ailments. In difference with microarray advancements, microarray talks about significant data that presents explicit inventiveness of narrative protein isoforms with various compound scopes of uncovered qualities.

Microarray has become an expansively utilized genome-wide expression profile for figuring substance cells, because of their capacity of determining potential heterogeneities in cell populaces (Aaron, Davis, John, 2016).

Since the advancement of RNA tasks as a notable intermediary among genome and proteome, finding and estimating gene expression have been the unmistakable conduct in biological science (Ana, Pedro, Sonia, David, Alejandra, Michał, Daniel, Laura, Xuegong and Ali, 2016).

There is no foremost prospective or good channel for the assorted variety of claims and analysis state in which microarray can be utilized. Researches and adoption of systematic methodologies on living being and their objectives have advanced (Levin,Yassour, Adiconis, Nusbaum, Thompson, Friedman, 2010).

A flourishing microarray study must have a major prerequisite of creating information with the possibilities of responding to biological inquiries of concern. This is practiced by characterizing an investigational aim, series intensity and replicating reasonable biological plans under examination and by advancement of sequencing research, ensuring that information achievement does not end up being tainted with redundant views. One critical part of the microarray information is the expulsion of the scourge of high-dimension, for example, noises, commotions, repetition, redundancy, immaterial as well as irrelevant data, among others (Pierson, and Yua, 2015). Because of high-measurement of biological information challenges, dimension reduction techniques are vital. Microarray information has turned out to be a potential high-throughput procedure to simultaneously profile transcriptomes of substantial information (Dongfang and Jin, 2018). Microarray has key advantages, for example, the capacity to spot narrative transcripts, precision, and dynamic range (Junhyong, 2012). Thousands of quality genes are simultaneously communicated and expressed in microarray, expression levels of genes are usually difficult, finding an effective low-dimensional representation of microarray information is important. A few dimension reduction methods utilized for gene expression data analysis and information investigation to expel noises related to explicit information exist (Bacher, and Kendziorski, 2016).

This study proposes a computational dimensionality reduction technique using ANOVA and RFE, to deal with the issue of curse of high dimensionality in gene expression space and analyzes SVM kernel classification methods. This study exhibits the robustness of this technique regarding to noises and sampling on RNA-Seq Anopheles Gambiae dataset..

II Related Works

In 2015, Pierson and Yau worked on a dimensionality reduction model for zero inflated single cell gene expression analysis, they built a dimensionality reduced technique, zero inflated factor analysis (ZIFA), which expressly models the dropout attributes, and demonstrate that it improves modelling precision on biological and simulated datasets. They modified the PPCA and FA framework to represent dropout and deliver a safe technique for the dimensionality reduction of single-cell gene expression data that gives robustness against such vulnerabilities.

Without dropouts, the method is basically equal to PPCA or FA. Hence, users could utilize ZIFA as an immediate substitute with the advantage that it will consequently represent dropouts while remedial endeavors might be required with standard PCA. There procedure varies from methodologies, for example, the numerous variations of strong PCA, which mean to show corrupted perceptions. ZIFA regards dropouts as genuine perceptions, not exceptions, whose event properties have been described utilizing an observationally educated factual model.

In 2015, Esra, Hamparsum, and Sinan worked on a novel hybrid dimension reduction method for small high dimensional gene expression datasets with information intricacy principle for cancer classification. There study addressed the restrictions inside the setting of Probabilistic PCA (PPCA) by presenting and building up new and novel methodology utilizing most extreme entropy covariance matrix and its hybridized smoothed covariance estimators. To diminish the dimensionality of the data and pick the quantity of probabilistic PCs (PPCs) to be held, they further presented and created observed Akaike's information criterion (AIC), consistent Akaike’s information criterion (CAIC), and the information theoretic measure of complexity (ICOMP) rule of Bozdogan. Six openly accessible undersized benchmark informational collections were broke down to demonstrate the utility, adaptability, and flexibility of their methodology with hybridized smoothed covariance matrix estimators, which does not decline to play out the PPCA to diminish the measurement and to do regulated characterization of malignancy bunches in high measurements. Their proposed technique can be utilized to take care of new issues and difficulties present in the investigation of NGS information in bioinformatics and other biomedical applications.

In 2016, Wenyan, Xuewen and Jingjing worked on feature selection for cancer classification for disease utilizing microarray data expression. This paper used information on microarray gene expression level to decide marker genes that are pertinent to a sort of malignancy. They researched a separation-based element choice strategy for two-gather grouping issue. So as to choose marker genes, the Bhattacharyya separation is actualized to quantify the uniqueness in gene expression levels. They used SVM for classification with utilization of the selected marker genes. The execution of marker gene selection and classification are represented in both recreation studies and two genuine information analyses by proposing a new gene selection method for classification based on SVMs. In the proposed method, they firstly ranked every gene according to the importance of their Bhattacharyya distances between the two indicated classes. The optimal gene subset is chosen to accomplish the least misclassification rate in the developed SVMs following a forward selection algorithm. 10-fold cross-validation is connected to locate the optimal parameters for SVM with the final optimal gene subset. Subsequently, the classification model is trained and built. The classification model is evaluated by its prediction performance for testing set. The execution of the proposed B/SVM technique with that of SVM-RFE and SWKC/SVM gives normal misclassification rate (1.1%) and high normal recovery rate (95.7%).

In 2017, Nancy and VijayKumar, worked on Alzheimer's infection determination by utilizing dimensionality reduction based on KNN classification algorithm for analyzing and classifying the Alzheimer malady and mild cognitive mutilation are available in the datasets. Their study gave more precision rate, accuracy rate and sensitivity rate to give a better output. This paper proposed a narrative dimensionality reduction based KNN classification Algorithm dissected the Alzheimer's illness present in the datasets. With the algorithm, the dataset was separated into 3 classes; first class having the Alzheimer's disease (AD), second class was having the normal outcome, third class having the mild cognitive impairment. The information's were taken from the researcher’s data dictionary - Uniform Data Set (RDD-UDS).

The relative investigations between the current PNN classification procedures with the proposed KNN classification demonstrated that high measure of normal accuracy, sensitivity, specificity precision, recall, jaccard and dice coefficients furthermore diminish the information dimensionality and computational multifaceted nature. Their future work, stated that the feature extraction and classification algorithm will improve the classification performance.

In 2017, Usman Shazad, and Javed worked on PCA and Factor Analysis for dimensionality reduction of bio-informatics data, they utilized the dimensionality reduction model of bioinformatics information. These systems were applied on Leukemia dataset and the number of attributes was decreased. An investigation was exhibited on reducing the number of attributes using PCA and Factor Analysis. Leukemia data was used for the analyses. PCA was carried out on the dataset and 9 components were chosen out of the 500 components. The Factor Analysis was used to extract the critical features.

In 2017, Gökmen, Dincer, Selcuk, Vahap, Gozde, Izzet, and Ahmet, worked on a simulation study for the RNA-Seq data classification, they contrasted a few classifiers including PLDA renovation, NBLDA, single SVM, bagging SVM, CART, and random forest (RF). They analyzed the impact of a few parameters, for example, over-dispersion, sample size, number of genes and classes, differential expression rate, and the transform technique on model performances. A broad modeled study was conducted and the outcomes were contrasted using the consequences of two miRNA and two mRNA exploratory datasets. The outcomes uncovered that expanding the sample size, differential expression rate and transformation method on model presentation. RNA-Seq data classification requires cautious consideration when taking care of data over-scattering. They ended up that count-based classifier, the power changed PLDA and asclassifiers, vst or rlog changed RF and SVM classifiers might be a decent decision for classification.

In 2017, Chieh, Siddhartha, Hannah, and Ziv used neural network algorithm to reduce the dimensions of single cell RNASeq data containing a few new computational complexities. These incorporate inquiries concerning the best strategies for clustering scRNA-Seq data, how to recognize unique cells, and deciding the state or capacity of explicit cells dependent on their expression profile. To address these issues, they created and tested a technique based on neural network (NN) for the analysis and recovery of single cell RNASeq data. They showed different NN structures, some of which fuse prior biological learning, and utilized these to acquire a reduced dimension representation of the single cell expression data. They demonstrate that the NN technique enhances earlier strategies in the capacity to accurately group cells in analyses not utilized in the training and the capacity to effectively derive cell type or state by questioning a database of a huge number of single cell profiles. Such database queries (which can be performed utilizing a web server) will empower researchers to better characterize cells while investigating heterogeneous scRNASeq tests.

In 2017, Ian and Jorge reviewed recent ongoing advancements in PCA as a strategy for diminishing the dimensionality of RNA-Seq datasets, for expanding interpretability and yet limiting data misfortune by making new uncorrelated factors that progressively maximize variance. This study presented the essential thoughts of PCA, talking about what it can, can't do and after that depict a few variations of PCA and their application.

III MATERIALS AND METHODS

A. *Dataset Used for Analysis*

Colon cancer dataset was used for this experiment, it contains an expression of 2000 genes with highest minimal intensity across 62 tissues, derived from 40 tumor and 22 normal colon tissue samples. The gene expression was analyzed with an Affymetrix oligonucleotide array complementary to more than 6,500 human genes. The gene intensity has been derived from about 20 feature pairs that correspond to the gene on the DNA microarray chip by using a filtering process. Details for data collection methods and procedures are described (Alon, Barkai, Notterman, Gish, Ybarra, Mack, and Levine, 1999), and the data set is available from the website http://microarray.princeton.edu/oncology/.

MATLAB (Matrix Laboratory) is utilized to perform the experiment, due to its ease and beneficial programming environment for engineers, architects, scientists, researchers, among others. MATLAB is a multi-worldview numerical processing environment and exclusive programming language created by MathWorks. It permits framework controls, plotting of functions and information, execution of algorithms, production of User Interfaces, and interfacing with projects written in different languages, such as; C, C++, C#, Java, Fortran and Python (Bezanson, Karpinski, Shah, Edelman, 2012). The principle point of this study is the prediction of the RNA-Seq technology utilizing the MATLAB tool by utilizing the Malaria database. Table-1 demonstrates a concise description of the dataset.

B. *Experimental Methodology*

This study summarizes the proposed framework in figure-1 below. The fundamental idea is to predict machine learning task on high dimensional microarray data, for cells and genes into lower dimensional dataset. The plan is adjusted to fetch out important data in a given dataset by utilizing ANOVA and RFE feature selection methods and evaluate the performance of colon cancer microarray dataset on SVM classification algorithm.

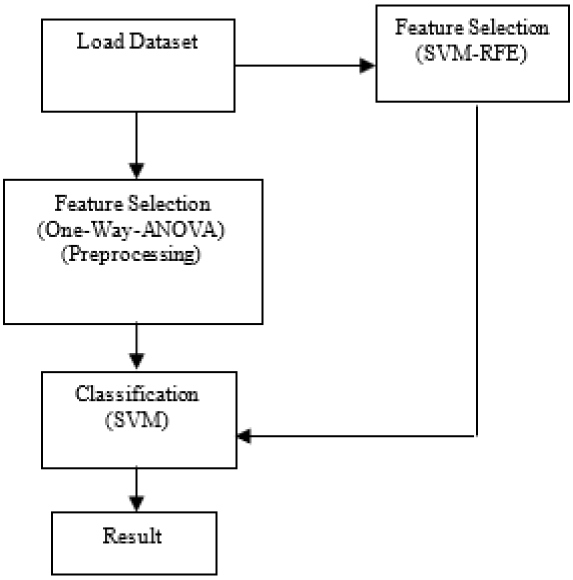

Load Dataset
Feature Selection (SVM-RFE)
Feature Selection (One-Way-ANOVA) (Preprocessing)
Classification (SVM)
Result

Fig 1: Proposed Framework.

Microarray data is the next generation sequencing technology to think about in transcriptome. It is utilized as an option to microarrays for gene expression analysis, without the need to earlier realize the RNA/DNA succession. RNA-seq offers progressively precise information and applications including identification of gene fusion, variations, alternative joining, post-transcriptional changes as well as analysis of small RNAs, such as; tRNA or miRNA profiles. A total image of the RNA/DNA substance can be gotten from low quantity biological samples. A few expository advances are basic for an effective portrayal and evaluation of the transcriptome. Bioinformatics tools are proposed for quality control, information handling, annotation, quantification and representation for translation and biological science investigation for understanding gene information.

C. *Analysis of Variance ANOVA*

ANOVA algorithm simplifies the value of intensity as a sum of components. ANOVA algorithm helps in normalization and gene-specific model.

The normalization eliminates properties due to total differences in intensity among diverse arrays. ANOVA normalization is trivial and basically deducts the mean of the log-transformed intensity from each array and refer to the distorted and normalized intensity values as Y (Gary, 2018). ANOVA test is used to compare the ‘multiple means’ values of the dataset, and visualize whether there exists any significant difference between mean values of multiple groups (classes). The statistic for ANOVA is called the F-statistic, which can be calculated using following steps (Mukesh, Nitish, Amitav and Santanu, 2015):

$$F\text{-}score = BMS/WMS \qquad (1)$$

The input to the algorithm is a matrix of the form N × M, where N is the total number of feature sets and M is the number of samples in the dataset.

D. *Recursive Feature Elimination RFE*

Guyon introduce RFE (Ding, 2015), RFE makes feature selection by iteratively training a set of data with the current set of features and eliminating the least significant feature indicated. In the linear case, the separating hyperplane (decision function) is $D\,(\vec{X}\,) = (\vec{W} \cdot \vec{X}\,) + b$. The feature with the smallest weight w^2 contributes the least to the resulting hyperplane and can be discarded. Due to the heavy computational cost of RFE, several variants have been introduced to speed up the algorithm. Instead of removing only one least important feature at every iteration, removing a big chunk of features in each iteration will speed up the process. The goal is to remove more features during each iteration, but not to eliminate the important features. (Shruti, and Mishra, 2015).

E. Classification

A few classification algorithms exist, for example, Logistic Regression, SVM, K-Nearest Neighbor, among others (Rimah, Dorra, and Noureddine, 2012). After reducing the dimensional complexity of data, the subsequent stage is the classification procedure. Classification is the fundamental goal; the analyzed data is classified. Two SVM kernels techniques were utilized: Polynomial Kernel and Gaussian Kernel. The results of the algorithms are analyzed and compared based on computational time, training time and performance metrics such as accuracy.

F. Support Vector Machine SVM

SVM is a learning machine algorithm presented by Vapnik in 1992 (Aydadenta and Adiwijaya, 2018). The algorithm works with the point of finding the best hyperplane that isolates between classes in the input space. SVM is a linear classifier; it is created to work with nonlinear problems by joining the kernel ideas in high-dimensional workspaces. In non-linear issues, SVM utilizes a kernel in training the data with the goal of spreading the dimension widely. When the dimensions are tweaked, SVM will look for the optimal hyperplane that can separate a class from different classes (Chang, and Lin, 2011). As indicated by the adoption of Aydadenta and Adiwijaya (2018), the procedure to locate the best hyperplane utilizing SVM is as follows:

i. Let $y_i \in \{y_{1,}y_2, \dots, y_n\}, where\ y_i\ is\ the\ p - attributes\ and\ target\ class\ z_i \in \{+1, -1\}$

ii. Assuming the classes +1 and -1 can be separated completely by hyperplane, as defined in equation 2 below:

$v.y + c = 0$ *(2)*

From equation (2), Equations (3) and (4) are gotten:

$v.y + c \geq +1$, *for class +1* *(3)*

$v.b + c \leq -1$, *for class -1* *(4)*

Where, *y* is the input data, *v* is the ordinary plane and *c* is the positive relative to the center field coordinates.

SVM intends to discover hyperplanes that maximizes margins between two classes. Expanding margins is a quadratic programming issue that is solved by finding the minimal point. The advantage of SVM is its capacity to manage wide assortment of classification problems in high dimensional data (Soofi and Awan, 2017).

Compared to other classification methods, SVM is outstanding, with its exceptional classification adequacy (Khan, Baharudin, Lee, and khan, 2010). SVM is grouped into linear and non-linear separable. SVM's has kernel functions that change data into a higher dimensional space to make it conceivable to perform seperations. Kernel functions are a class of algorithms for pattern analysis or recognition. Training vectors *xi* is mapped into higher dimensional space by the capacity Φ. SVM finds a linear seperating hyperplane with the maximal in this higher dimension space. $C > 0$ is the penalty parameter of the error term.

There are several SVM kernels that exist such as; the polynomial kernel, Radial basis function (RBF), linear kernel, Sigmoid, Gaussian kernel, String Kernels, among others. The decision of a Kernel relies upon the current issue at hand, since it relies upon what models are to be analyzed, a couple of kernel functions have been found to function admirably in for a wide assortment of applications (Bhavsar and Panchal, 2012). The prescribed kernel function for this study is the SVM-Polynomial Kernel and Gaussian Kernel.

SVM-Gaussian Kernel

Gaussian kernel (Devi and Devaraj, 2015) compare to a general smoothness supposition in all k-th order subordinates. Kernels coordinating a certain prior recurrence substance of the data can be developed to reflect earlier issues in learning. Each input vector $\underline{x}$ is mapped to an interminable dimensional vector including all degree polynomial extensions of *x's* components.

SVM Polynomial Kernel

For instance, a polynomial kernel model features conjunction up to the order of the polynomial. Radial basis functions permit circles in contrast with the linear kernel,which permits just selecting lines (or hyperplanes).

$K\left(y_a, y_j\right) = (\gamma y_a^S y_b + \ q)^e, \gamma > 0$ *(5)*

SVM-Linear Kernel Function

For instance, polynomial kernel is the least complex kernel function. It is given by the inner product *(a,b)* in addition to a discretionary constant K.

$K(y_a, y_b) = y_a^S y_b$ *(6)*

SVM-RBF Kernel Function

In SVM kernel functions, γ, *a*, and *b* are kernel parameters, RBF is the fundamental kernel function due to the nonlinearly maps tests in higher dimensional space unlike the linear kernel, it has less hyper parameters than the polynomial portion.

$K(y_a, y_b) = \exp(-\gamma||y_a, y_b||\ ^2), \gamma > 0$ *(7)*

XIV. Results And Discussions

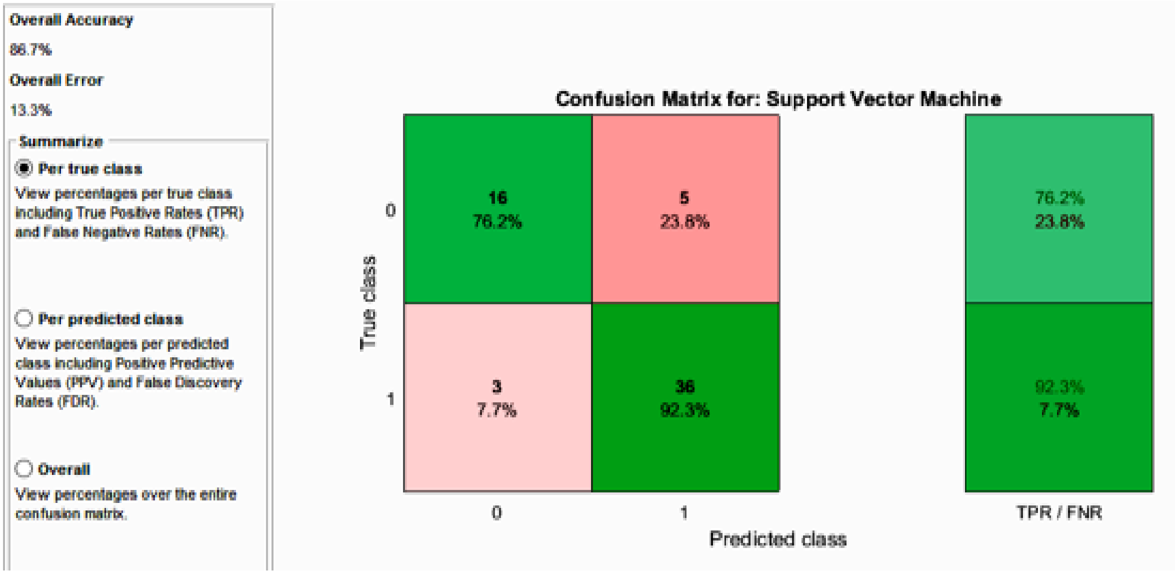
The colon cancer dataset extracted were classified, the classification results obtained show the features capability for classifying the colon's status. The average classification accuracy, which is using features with ANOVA and RFE are recorded in tabular form below. The proposed methodology was applied to the publicly available colon cancer database, the classification algorithm applies SVM kernel by utilizing MATLAB tools to implement the model.

Using ANOVA as a dimensionality reduction method, 416 features where fetched from the 2001 attributes of colon cancer dataset obtained from Alon, 2001. Using ANOVA, the output of the analysis is a statistically significant difference between group means. The significance value is 0.05 which is the mean length of time to complete the spreadsheet problem between the different courses taken.

ANOVA is appropriate when the model holds, have a single "treatment" with, say, k levels. "Treatment" may be interpreted in the loosest possible sense as any categorical explanatory variable. There is a population of interest for which there is a true quantitative outcome for each of the k levels of treatment. The selected features a processed for classification.

A supervised SVM kernel classifier methods, is among the most well-established and popular machine learning approaches in bioinformatics and genomics, 10-folds cross validation was used to evaluate the execution of the performance of the classification models, using 0.05 parameter holdout of data for training and 5% for testing to check the accuracy of the classifiers.

To each of the classifiers, a basic supervised learning assessment protocol is carried out. In particular, the training and testing stages are assessed as a 10-fold cross validation to eliminate the sampling bias. This protocol is implemented using MATLAB. The reported result of assessment is based on the computational time and performance metrics (Accuracy, Sensitivity, Specificity, F-score, Precision and Recall) (Nathan,Andi, Katelyn, and Dmitry, 2017).



Overall Accuracy

86.7%

Overall Error

13.3%

Summarize

● **Per true class**

View percentages per true class including True Positive Rates (TPR) and False Negative Rates (FNR).

○ **Per predicted class**

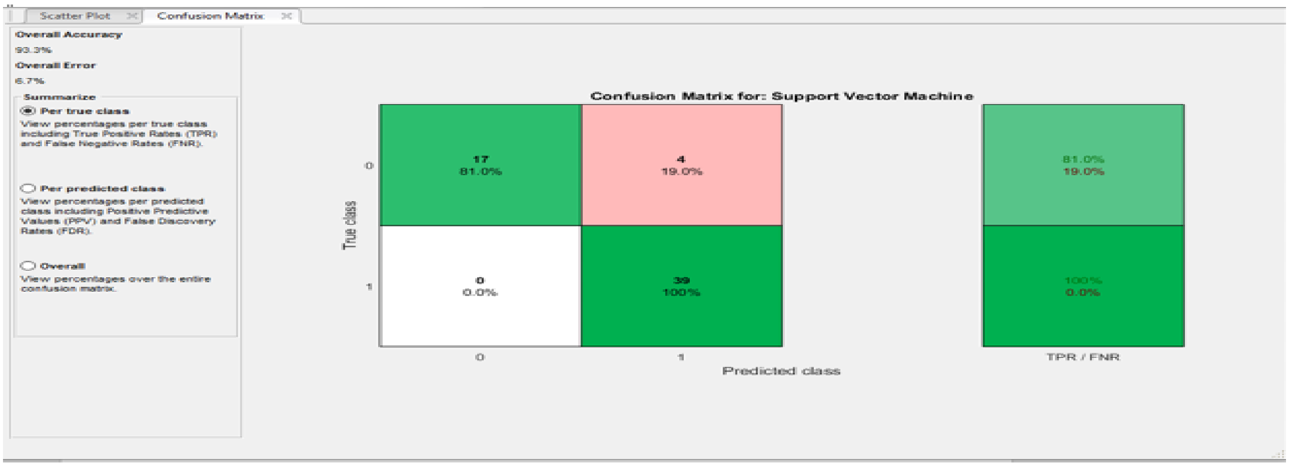
View percentages per predicted class including Positive Predictive Values (PPV) and False Discovery Rates (FDR).

○ **Overall**

View percentages over the entire confusion matrix.

Fig 2: Confusion Matrix and Performance Metrics for SVM-ANOVA

RFE-SVM algorithm was used to fetch out relevant data in the colon cancer dataset, 868 features were selected. SVM-RFE improve the computational performance of recursive feature elimination by eliminating chunks of features at a time with as little effect on the quality of the reduced feature set as possible. The RFE algorithm is implemented using an SVM to assist in identifying the least useful gene(s) to eliminate. Using SVM-RFE, the selected data was classified and accomplish 93.3% Accuracy



Overall Accuracy

93.3%

Overall Error

6.7%

Summarize

● **Per true class**

View percentages per true class including True Positive Rates (TPR) and False Negative Rates (FNR).

○ **Per predicted class**

View percentages per predicted class including Positive Predictive Values (PPV) and False Discovery Rates (FDR).

○ **Overall**

View percentages over the entire confusion matrix.

Fig 3: Confusion Matrix and Performance Metrics for SVM-RFE classification

In this study, data analysis of a well-known dataset colon cancer dataset by Alon, consisting of expression levels of 2000 genes describing 62 samples (40 tumor and 22 normal colon tissues, was analyzed using MATLAB tool. The dataset was used to compare the performance of the One-Way-ANOVA and SVM-RFE. The dataset was trained and tested. A different number of genes were selected by each of the algorithms, 416 and 868 respectively. The SVM was trained on the training data that was trimmed to the selected genes from each algorithm respectively.

The SVM model produced was evaluated by its performance to predict the class labels (since cross validation results on the training data tend to be optimistic).

Comparisons of the two algorithms in terms of prediction rate and time required are made. A comparison between ANOVA and SVM-RFE is also performed.

The performance of ANOVA was comparable to the SVM-RFE algorithms in terms of prediction accuracy rate (each achieving around 87% and 93% accuracy on the test data).

Experiment on the Alon colon cancer data sets also show that ANOVA has similar performance when compared with SVM-RFE with respect to accuracy, when comparing computational time, ANOVA is much faster than the SVM-RFE.

In general, SVM-RFE allows an enormous increase in the efficiency of the algorithm without a decrease of classification accuracy. Thus, the gene selection process is made much more practical in domains with a large number of features, such as gene expression data. This improvement is especially important as the number of samples available increases. The table 4.1 below shows the comparative analysis of ANOVA and SVM-RFE feature selection algorithm using SVM classification algorithm to improve the performance of Colon Cancer data in microarray technology.

Table 4.1 Comparative Analysis of the Classification of One-Way-ANOVA and SVM-RFE

Performance Metrics	**ANOVA Based**	**SVM-RFE Based**
Accuracy (%)	**86.70**	**93.33**
Sensitivity (%)	**92.30**	**100**
Specificity (%)	**77.27**	**80.95**
Precision (%)	**87.81**	**90.70**
Misclassification (%)	**13.12**	**6.67**
Time (Sec)	**23.1409**	**7.340**

The performance analysis of classification using Support Vector Machine on colon cancer dataset shows that, SVM-RFE feature selection technique method achieves necessary higher value in the datasets on performance parameters such as the accuracy, timing, sensitivity, specificity, and prediction when compared to the ANOVA feature selection method. When the dataset is of high dimensional, by application of dimensionality reduction, some valuable data are considered and the accuracy of an algorithm increases by removing unnecessary data. The feature selection algorithm using ANOVA for high dimensional datasets plays an important role, it improves the performance of feature extraction methods, and SVM-RFE also enhances the classification algorithm "SVM" performance, in terms of accuracy, sensitivity, specificity and precision.

V. Conclusion

In the past few years, remarkable works have been done on the innovation of microarray, improvement as far as the execution measurements and productivity that are extraordinarily influenced by exploratory plan, activity and the data analysis forms are in trends to enhance the performance. Cancer is a deadly insect comprising of various kinds. The significance of classification of malaria vector into gatherings has prompted numerous looks into. By examination, this study classifies a colon cancer data by using SVM on reduced dimensional data that employs RFE and ANOVA algorithms. The experiment accomplished a comparable result that shows that SVM-RFE outperforms ANOVA-SVM with 93%. Further studies should be conducted to improve performance of Machine Learning based methods by using more data and hybridized models.

References

T.L. Aaron, J.M. Davis, J.M, John,C.M. (2016). A Step-By-Step Workflow for Low-Level Analysis of Single-Cell RNA-Seq Data. F1000Research Vol. 1, No. 5, pp. 1-62. https://doi.org/10.12688/f1000research.9501.2

C. Ana, M. Pedro, T. Sonia, G. David, C. Alejandra, M. Andrew, W.S. Michał, J.G. Daniel, L.E. Laura, Z. Xuegong, and M. Ali. (2016). A survey of best practices for RNA-seq data analysis. Genome Biology. Vol. 17, No.13, pp. 1-19. DOI 10.1186/s13059-016-0881-8.

J.Z. Levin, M. Yassour, X. Adiconis, C. Nusbaum, D.A. Thompson, N. Friedman. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods. Vol. 7, pp. 709–715.

W. Dongfang, and G. Jin. (2018). VASC: Dimension Reduction and Visualization of Single-Cell RNA-Seq Data by Deep Variation Autoencoder. Genomics Proteomics Bioinformatics. Doi.org/10.1016/j.gpb.2018.08.03.

R. Bacher, and , C. Kendziorski. (2016). Design and Computational Analysis of Single-Cell RNA-Seq Experiments. Genome Biology. Vol 17. No. 63.

E. Pierson, and C. Yau. (2015). ZIFA: Dimensionality Reduction for Zero-Inflated Single-Cell Gene Expression Analysis. Genome Biology. Vol. 16. Pp. 241-257.

L. Chieh, J. Siddhartha, K. Hannah, and B. Ziv. (2017). Using neural networks for reducing the dimensions of single-cell RNA-Seq data. Nucleic Acids Research, Vol. 45, No. 17, pp. 1-11 doi: 10.1093/nar/gkx681.

K. Junhyong. (2012). Computational Analysis of RNA-Seq Data: From Quantification to High-Dimensional Analysis. University of Pennsylvania. Pp. 35-43.

B. Mariangela, O. Eric, A.D. William, W. Monica, A. Yaw, Z. Guofa, H. Joshua, L. Ming, X. Jiabao, G. Andrew, F. Joseph, and Y. Guiyun. (2015). RNA-seq analyses of changes in the Anopheles gambiae transcriptome associated with resistance to pyrethroids in Kenya: identification of candidate-resistance genes and candidate-resistance SNPs. Parasites and Vector. Vol. 8, No. 474, pp. 1-13 https://doi.org/10.1186/s13071-015-1083-z

https://figshare.com/articles/Additional_file_4_of_RNA-seq_analyses_of_changes_in_the_Anopheles_gambiae_transcriptome_associated_with_resistance_to_pyrethroids_in_Kenya_identification_of_candidate-resistance_genes_and_candidate-resistance_SNPs/4346279/1

J. *Bezanson, S. Karpinski, V. Shah, A. Edelman. (2012).*

.V. Keerthi, B. Surendiran, B. (2016). Dimensionality reduction using Principal Component Analysis for network intrusion detection. Perspectives in Science. Vol. 8, pp.510—512

H. Aydadenta, and Adiwijaya. (2018). On the classification techniques in data mining for microarray data classification. International Conference on Data and Information Science, Journal of Physics: Conf. Series Vol. 971. Pp. 1-10. doi :10.1088/1742-6596/971/1/012004

V. Sofie. (2017). A comparative review of dimensionality reduction methods for high-throughput single-cell transcriptomics. Master's dissertation submitted to Ghent University to obtain the degree of Master of Science in Biochemistry and Biotechnology. Major Bioinformatics and Systems Biology. pp.1-88.

Elavarasan and K. Mani. (2015). A Survey on Feature Extraction Techniques. International Journal of Innovative Research in Computer and Communication Engineering. Vol. 3, No. 1, pp.1-4.

C. Chang, C. and Lin. LIBSVM: A library for support vector machines. ACM TIST. Vol. 2, No.3, pp.27.

A.A. Soofi, and A. Awan. (2017). Classification Techniques in. Machine Learning: Applications and Issues. Journal of Basic and Applied Sciences, Vol. 13, pp. 459-465

A. Khan, B. Baharudin, L.H. Lee, K. khan. (2010). A Review of Machine Learning Algorithms for Text-Documents Classification. Journal of Advances in Information Technology, Vol. 1, No. 1. Pp. 1-17.

U. Alon, N. Barkai, D. Notterman, A. Gish, S. Ybarra, D. Mack, and A. Levine: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl myAcad Sci USA 8; 96(12): 6745-6750 (1999)

H. Bhavsar., and M.H. Panchal. (2012). A Review on Support Vector Machine for Data Classification. 3 International Journal of Advanced Research in Computer Engineering and Technology (IJARCET) Vol.1, No. 2, pp. 185-189.

A. Rimah., B.A. Dorra., and E. Noureddine. (2012). An Empirical Comparison of SVM and Some Supervised Learning Algorithms for Vowel recognition. International Journal of Intelligent Information Processing(IJIIP) Vol. 3. No. 1., pp. 1-5.

A.V. Devi, and D.V. Devaraj. (2015). Gene Expression Data Classification using Support Vector Machine and Mutual Information-based Gene Selection. Procedia Computer Science, Vol. 47, pp. 13 – 21.

A.C. Gary. (2018). Using ANOVA to Analyze Microarry Data. Biotechniques Future Science. Vol. 37, No. 2, pp. 1-5.

M. Balamurugan, Nancy, A., And Vijaykumar, S. (2017). Alzheimer's Disease Diagnosis by Using Dimensionality Reduction Based on KNN Classifier. *Biomedical & Pharmacology Journal* Vol. 10, No. 4, pp. 1823-1830.

P. Esra, B. Hamparsum, and C. Sinan. (2015). A Novel Hybrid Dimension Reduction Technique for Undersized High Dimensional Gene Expression Data Sets Using Information Complexity Criterion for Cancer Classification. Computational and Mathematical Methods in Medicine. Vol.1, pp. 1-14 http://dx.doi.org/10.1155/2015/370640

K. Mukesh, K.R. Nitish, S. Amitav, and K.R. Santanu. (2015). Feature Selection and Classification of Microarray Data Using MapReduce Based ANOVA and KNN. Procedia Computer Science. Vol. 54, pp. 301-310.

Z. Wenyan, L. Xuewen, and W. Jingjing. (2017). Feature Selection for Cancer Classification Using Microarray Gene Expression Data. Biostatistics and Biometrics journals. Vol.1, No.2, pp. 1-7.

A. Usman, A. Shazad, and F. Javed. (2017). Using PCA and Factor Analysis for Dimensionality Reduction of Bio-informatics Data. *(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 5, pp. 515-426*

Z. Gökmen, G. Dincer, K. Selcuk, E. Vahap, E.Z. Gozde, P.D. Izzet, and O. Ahmet. (2017). A comprehensive simulation study on classification of RNASeq Data. PLoS One Journal, Vol.12, No.8, pp. 1-24.

L. Chieh, J. Siddhartha, K. Hannah, and B.J. Ziv. (2017). Using neural networks for reducing the dimensions of singlecell RNASeq data. Nucleic Acids Research. Vol. 45, No.17, pp.1-21.

T.J. Ian, and C. Jorge. (2017). Principal component analysis: a review and recent developments. Philosophical Transaction A Mathematical Physical Engineering Science. Vol.374, pp. 1-21.

Y. Ding, and W. Dawn. (2015). BMC Bioinformatics. Improving the Performance of SVM-RFE to Select Genes in Microarray Data. Vol. 2, No. 12, pp. 1-11.

T.J. Nathan, D. Andi, J.H. Katelyn, and K. Dmitry. (2017). Biological classification with RNA-Seq data: Can alternative splicing enhance machine learning classifier**?.** bioRxiv . doi: http://dx.doi.org/10.1101/146340.

M. Shruti, and D. Mishra. (2015). SVM-BT-RFE: An improved gene selection framework using Bayesian T-test embedded in support vector machine (recursive feature elimination) algorithm. Karbala International Journal of Modern Science. Vol. 1, No. 2, pp. 86-96.

J. Divya, and S. Vijendra. (2018). Feature selection and classification systems for chronic disease prediction: A review. Egyptian Informatics Journal. https://doi.org/10.1016/j.eij.2018.03.002

Appraisal of Selected Principal Component Analysis-Based Methods for Face Recognition System's Accuracy

MADANDOLA Tajudeen Niyi
Department of Computer Sciences,
Kwara State College of education
Oro, Nigeria
nmadandola@yahoo.com
GBOLAGADE Kazeem Alagbe
Department of Computer Sciences,
Kwara State University,
Malete, Nigeria.
gkazy1@gmail.com
ISIAK Rafiu Mope
Department of Computer Sciences,
Kwara State University,
Malete, Nigeria.
abdulrafiu.isiaka@kwasu.edu.ng

ABSTRACT— A face recognition system is a technology proficient for identifying or verifying an individual from a digital image. Kernel Principal Component Analysis (KPCA) and Principal Component Analysis (PCA) are most frequently used face recognition system. KPCA is an example of non-linear and PCA is an example of linear appearance based face recognition system. All research work read prove that non-linear methods recognize better than linear appearance based method face recognition system. This study appraises the assertion on selected database and affirms that type of database employed has effect on the recognition accuracy not the methods only. Yale database was employed to implement the appraisal. Recognition index was used as performance metrics to determine the recognition accuracy of the two methods employed. MATLAB 2015a was used to implement the two algorithms. The implementation revealed that PCA has 71.11% recognition accuracy while KPCA also has 71.11% recognition accuracy. The two PCA-based algorithms have equal facial images recognition accuracy on the database employed despite that one of the algorithms is linear and the other one is nonlinear appearance based method.

Keywords— Recognition index, Principal Component Analysis, Yale Database, Implementation, Eigenvector.

I INTRODUCTION

The main challenge facing the security of the nation is inability to recognize the suspected criminals. Instituting the distinctiveness of a person is of principal significance in our environment [1],[2]. Employing biometrics system provided a better means of identification and verification of individuals. Biometric systems recognize persons based on behavioural and physiological qualities. It is found to be a strong, dependable, and suitable way of authentication. Some of the commonly employ biometric system are voice, deoxyribonucleic acid (DNA), retina, face and fingerprint. Biometric system has the merits of both high accuracy and low inappropriateness [3].

Face recognition system employs unique features of the face to confirm users. There are multiple techniques in which face recognition systems operate, some operate by comparing particular facial features from given image with faces within a database. The field that spread through face recognition comprises computer vision, neural network, pattern recognition and image processing [4]. The numerous face recognition system approaches can be grouped into appearance-based approach and model-based approach. Appearance based approach can be classified as linear or non-linear, while model-based approach can be 2D or 3D [5]. Linear appearance-based methods perform a linear dimension reduction such as PCA, Linear Discriminant Analysis (LDA) or Independent Component Analysis (ICA) while Non-linear appearances perform a non-linear dimension reduction [6]. Kernel PCA (KPCA) is an example of non-linear appearance algorithm [7]. PCA is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss [8].

The kind of face recognition algorithm adopted contributed seriously to its effectiveness and recognition accuracy of the system. Experimentally, non-linear PCA-based appearance algorithms have more recognition accuracy than linear appearance-based methods. Linear subspace analysis is an approximation of a non-linear manifold. The database setup used for the face recognition system is very important in recognition accuracy. Reference [9] stated that the type of database used for image acquisition is among the acknowledged obstacles hindering face recognition system accuracy.

The study affirmed the influence of database especially Yale database on recognition accuracy of nonlinear PCA-based using KPCA and linear PCA-based using PCA appearance face recognition system when MATLAB R2015a was used to implement the two algorithms. The Yale database comprised of 165 frontal images of 15 individuals with 11 subjects of each individual. The type of database sometimes referred to as Template employed in FRS is very important in recognition accuracy. The number of images available in the template, mode of normalization of the images and number of subjects per individual in the database counts [10].

The study presents experimental result that shown that both PCA and KPCA have equal recognition accuracy on Yale database despite that one of the algorithms is linear and the other one is nonlinear appearance based method.

II LITERATURE REVIEW

a. Dimensionality Reduction Algorithms

The dimensionality reduction algorithms perform two major roles in face recognition system. It is used for feature extraction and feature selection. Feature Extraction algorithms are grouped into Appearance based (Linear and nonlinear) and Model based (2D and 3D) Algorithms [11]. There are numerous Feature Extraction algorithms some of them are Discrete Cosine Transform (DCT), LDA, ICA, Active Shape Models (ASM), Neural Network based methods, Semi-supervised Discriminant Analysis, PCA and KPCA.

b. Principal Component Analysis

PCA is mathematical tool for accomplishing dimensionality reduction in face recognition system. It is used for reducing the size of data by implementing a linear mapping of the data to a lower-dimensional space in such a technique that the variance of the data in the low-dimensional representation is maximized. The covariance matrix of the data is created and the eigenvectors on this matrix are computed. The eigenvectors that match to the principal eigenvalues can now be used to recreate a large fraction of the variance of the main data. The eigenvectors may be viewed as representation of differences amid the faces. It was enjoyed in face representation and recognition in the early 90's [12]. The foremost reason of employing PCA for face recognition is to direct the urge 1-D vector of pixels constructed from 2-D facial image into the compact principal components of the feature space [13]. PCA is employed specifically in the cases where the data sets are too big and their reduction for feature extraction without losing the meaningful information is the case.

When x (mean) is obtained from the data. Permit $x_n x_m$ be the data matrix.

Such that, $x_1, ..., x_m$ are the image vectors and n is the number of pixels per image.

The KLT basis is gotten by solving the eigenvalue drawback where C_x is the covariance matrix of the data.

$$C_x = \phi \wedge \phi^T$$

$$C_x = \frac{1}{m}\sum_{i=1}^{m} xixi^T$$

$\phi = [\O_1, ...\O_n]$ is the eigenvector matrix of C_x. $\wedge$ is a diagonal matrix, the eigenvalues $\lambda_1, ..., \lambda_n$ of C_x are located on its main diagonal. λ_i is the variance of the data projected on $\O_i$

c. Kernel Principal Component Analysis

Principal component analysis can be engaged in a nonlinear manner by means of the kernel trick. The resultant method is able to make nonlinear mappings that maximize the variance in the data. The resulting technique is called Kernel PCA. Its actual procedure is to relate a non-linear mapping to the input (Ψ(x): $R^N \rightarrow R^L$) and then elucidate a linear PCA in the resulting feature subspace. The mapping of Ψ(x) is prepared implicitly using kernel functions

$$k(x_i , x_j) = (\Psi(x_i). \Psi(x_j))$$

such that n (input space) tally to dot-products in the higher dimensional feature space.

d. Related Works

Reference [14] experimental work titled "performance assessment of PCA and KPCA face recognition techniques" was carried out base on computational time using testing time and recognition accuracy on local database branded as TOAM database. The database comprises of 120 face images. The experiment revealed an average testing Time of 1.5475s for PCA and 67.0929s for KPCA indicating a longer Computational time for KPCA than PCA. It also revealed that PCA has 72.5% performance recognition accuracy while KPCA has 80.0% performance recognition accuracy showing that KPCA outstrips the PCA in terms of recognition accuracy.

In the study of [15] PCA algorithm with KPCA algorithm were compared using AT&T dataset. The comparison was done on recognition of accuracy, variation in facial expression, illumination changes, and computation time of each method. The error rate in training set and test data in AT&T database has shown that KPCA with less error produced better result. The error rate of PCA training and testing data calculated are 9.35% and 25.32% respectively, while KPCA training and testing data calculated are 7.90 and 13.56% respectively. The overall outcome revealed that KPCA for face recognition have better performance than PCA.

Reference [13] selected Principal Component Analysis (PCA), Binary Principal Component Analysis (BPCA) and Principal Component Analysis – Artificial Neural Network (PCA-ANN) for the experiment. A database totaling 400 images was used. One hundred images were used for testing while three hundred images were used for training the three face recognition systems. The implementation evaluation of the 3 PCA-based methods exposed that PCA – ANN method gave the best recognition rate of 94% with a trade-off in recognition time. Likewise, the recognition rates of PCA and B-PCA improved with decreasing number of eigenvectors but PCA-ANN recognition rate was negligible.

Likewise, [16] assessed the performance of Optimised PCA (OPCA) and Projection Combined PCA ((PC)2A) techniques based using recognition accuracy, total training time and average recognition time as performance metrics. A database consisting of 252 images was used for the experiment. MATLAB was used to implement both algorithms on a Pentium dual processor with 2.00GHz processor speed. The consequences of assessment between the two algorithms based on black faces showed that OPCA and (PC)2A gave recognition accuracies of between 96% to 64% and between 95% to 60% respectively. OPCA algorithm achieves better than (PC)2A when totally parameters considered are taken into matter.

VI. Methodology

The images were acquired from Yale database. The database is a standard database. Both PCA and KPCA were employed as dimensionality reduction algorithms. PCA and KPCA were implemented separately with MATLAB 2015a. PCA results were compared with KPCA results as shown in Fig. 1, using recognition index as performance metrics. The experiment was performed on each of the database. The two results were analyzed with column and pie chart.

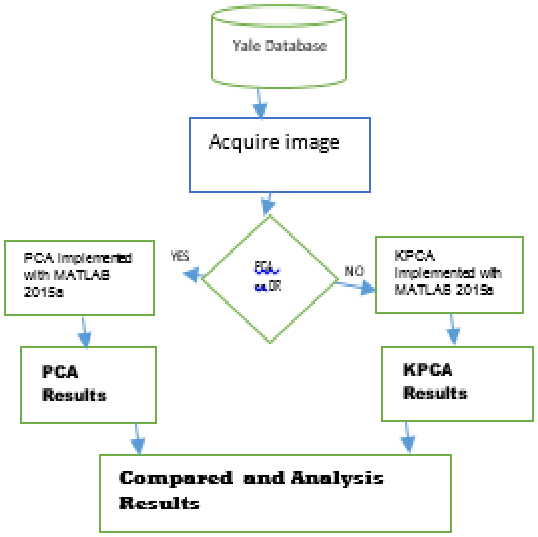

Yale Database
Acquire image
PCA
OR
YES
NO
PCA Implemented with MATLAB 2015a
KPCA Implemented with MATLAB 2015a
PCA Results
KPCA Results
Compared and Analysis Results

Fig. 1: Research Methodology

a. Database

Yale database that encompasses of frontal images of 165 images of 15 people were employed. One hundred and twenty (120) images were used for training while forty five (45) images were used for testing as shown in Fig. 2 and Table I.

Fig. 2: Some Images in the Yale database ([17]).

TABLE I. ANALYSIS OF THE IMAGE USED

Variables	Frequency
Number of persons	15
Number of sample per persons	11
Number of Total sample	165
Number of Training set	120
Number of Testing sample	45

System used for Implementation

MATLAB R2015a was used to implement the two algorithms on Intel(R) Celeron (R) CPU with 1.60GHz Processor speed. The experiment was conducted using Recognized index in Database as performance metrics to determine the recognition accuracy.

III RESULTS AND DISCUSSIONS

The Recognition accuracy of PCA and KPCA on the images was tested using a Recognition index as revealed in Table II.

TABLE II. RECOGNITION INDEX OF BOTH PCA AND KPCA TO DETERMINE RECOGNITION ACCURACY

IMAGE	PCA RECOGNIZED INDEX IN DATABASE	KPCA RECOGNIZED INDEX IN DATABASE	PCA IMAGE MATCHED	KPCA IMAGE MATCHED
1	2.jpg	2.jpg	YES	YES
2	5.jpg	5.jpg	YES	YES
3	8.jpg	8.jpg	YES	YES
4	10.jpg	10.jpg	YES	YES
5	64.jpg	109.jpg	NO	NO
6	17.jpg	17.jpg	YES	YES
7	19.jpg	19.jpg	YES	YES
8	23.jpg	23.jpg	YES	YES
9	25.jpg	25.jpg	YES	YES
10	28.jpg	28.jpg	YES	YES
11	31.jpg	31.jpg	YES	YES
12	51.jpg	35.jpg	YES	YES
13	37.jpg	37.jpg	YES	YES
14	42.jpg	40.jpg	YES	YES
15	43.jpg	43.jpg	YES	YES
16	48.jpg	48.jpg	YES	YES
17	50.jpg	50.jpg	YES	YES
18	30.jpg	52.jpg	YES	YES
19	55.jpg	55.jpg	YES	YES
20	48.jpg	24.jpg	NO	NO
21	28.jpg	62.jpg	YES	YES
22	41.jpg	110.jpg	YES	YES
23	44.jpg	69.jpg	YES	YES
24	6.jpg	6.jpg	NO	NO
25	75.jpg	75.jpg	YES	YES
26	78.jpg	78.jpg	YES	YES
27	30.jpg	30.jpg	NO	NO
28	17.jpg	56.jpg	NO	NO
29	41.jpg	59.jpg	YES	YES
30	30.jpg	90.jpg	YES	YES
31	30.jpg	93.jpg	YES	YES
32	41.jpg	94.jpg	YES	YES
33	30.jpg	61.jpg	NO	NO
34	8.jpg	101.jpg	YES	YES
35	16.jpg	105.jpg	YES	YES
36	78.jpg	106.jpg	YES	YES
37	48.jpg	24.jpg	NO	NO
38	28.jpg	113.jpg	YES	YES
39	10.jpg	10.jpg	YES	YES
40	45.jpg	45.jpg	NO	NO
41	10.jpg	10.jpg	NO	NO
42	74.jpg	74.jpg	NO	NO
43	30.jpg	30.jpg	NO	NO
44	42.jpg	40.jpg	NO	NO
45	35.jpg	36.jpg	NO	NO

TABLE III. SUMMARY OF RECOGNITION INDEX OF PCA AND KPCA RECOGNITION ACCURACY

	Database	
Variable	PCA IMAGE RECOGNIZED IN DATABASE	KPCA IMAGE RECOGNIZED IN DATABASE
Number of YES	32	32
Number of NO	13	13
Total	45	45
Percentage of Recognition Performance	71.11%	71.11%

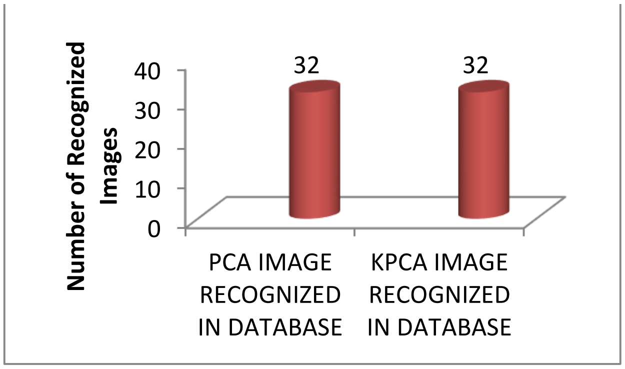


Fig. 3: Number of Recognized Images by both PCA and KPCA

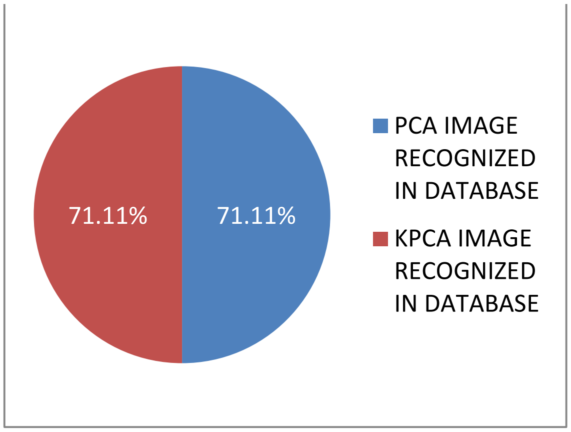
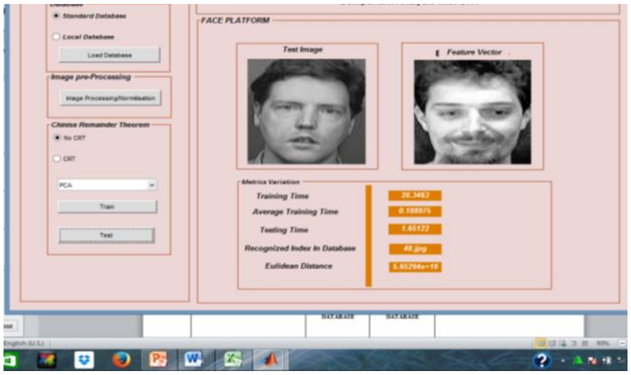


Fig. 4: Percentage of Recognition Accuracy for PCA and KPCA

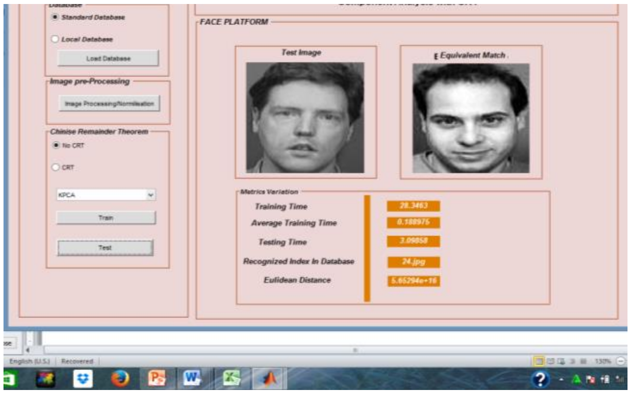
The implementation revealed that PCA and KPCA were able to recognised 32 images among the 45 images used as Testing samples as shown in Table III and Fig. 3. PCA has recognition accuracy of 71.11% likewise KPCA has recognition accuracy of 71.11% as shown in Table III and Fig. 4. Critical comparison of recognition index of Table II revealed that twenty-one of the images have different recognized index in the database such as PCA and KPCA having 48.jpg and 24.jpg recognition index for image 37 as shown in Fig. 5 and Fig. 6 respectively.



PCA

Training Time 28.3463
Average Training Time 0.188975
Testing Time 1.65122
Recognized Index In Database 48.jpg
Eulidean Distance 5.65294e+16

Fig. 5: PCA Recognized Index of 48.jpg using Yale Database of Image 37

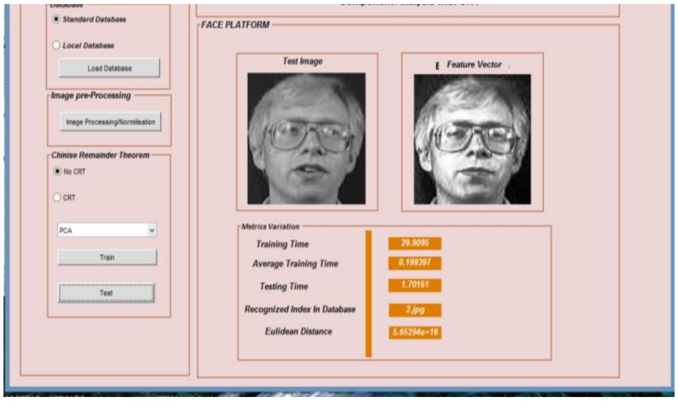


KPCA

Training Time 28.3463
Average Training Time 0.188975
Testing Time 3.09858
Recognized Index In Database 24.jpg
Eulidean Distance 5.65294e+16

Fig. 6: KPCA Recognized Index of 24.jpg using Yale Database of Image 37.

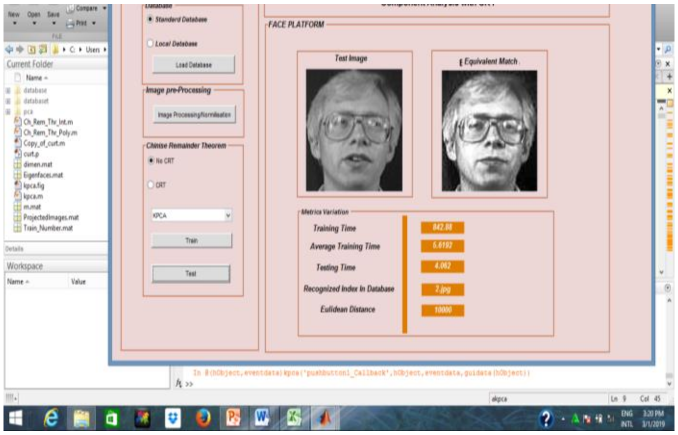
Despite the differences observed in some of the PCA and KPCA recognized index in the database, images that were able to be recognised by PCA was also recognized by KPCA and those not recognize by PCA was not recognize by KPCA. This was as a result of the type of database which had 11 face images of each of the 15 subjects. Also PCA and KPCA images 5, 20, 28, 33, 37, 44 and 45 have different recognition index but both did not recognize. Both Image 1 for PCA and KPCA have recognized index of 2.jpg and images were recognized as shown in Fig. 7 and Fig. 8.



PCA

Training Time 29.9095
Average Training Time 0.199397
Testing Time 1.70151
Recognized Index In Database 2.jpg
Eulidean Distance 5.65294e+16

Fig. 7: PCA Recognized Index of 2.jpg using Yale Database of Image 1



KPCA

Training Time 842.88
Average Training Time 5.6192
Testing Time 4.062
Recognized Index In Database 2.jpg
Eulidean Distance 10000

```
In @(hObject,eventdata)kpca('pushbutton1_Callback',hObject,eventdata,guidata(hObject))
fx >>
```

Fig. 8: KPCA Recognized Index of 2.jpg using Yale Database of Image 1

There is an equal recognition accuracy performance for both PCA and KPCA when implemented with Yale Database as a result of many number of sample per persons.

VI CONCLUSION

A face recognition system is a technology proficient of identifying or verifying an individual from a digital image. The procedure of face recognition is implemented in two stages, namely:

dimensionality reduction (feature extraction and selection) and classification of objects [18]. Prevalent recognition algorithms comprise of PCA using eigenfaces, LDA, Elastic Bunch Graph Matching using the Fisherface algorithm, the Hidden Markov Model, and Multilinear Subspace Learning using tensor representation. Some literatures were reviewed all research works read proved that nonlinear PCA-based methods recognized better than linear PCA-based method face recognition system.

The PCA and KPCA PCA-based face recognition algorithms were implemented using MATLAB 2015a. Data were acquired from Yale database. The experimental results revealed that PCA has 71.11% performance recognition accuracy while KPCA also has 71.11% performance recognition accuracy, indicating that KPCA and PCA have the same recognition accuracy, that was as a result of type of database employed to implement the experiment. The implication of the results obtained shown that the type of database employed has effect on the recognition accuracy of any face recognition techniques adopted.

Acknowledgment

Our utmost appreciation goes to the management of Kwara State University, Malete, Nigeria for providing enabling avenue to implement this study.

References

[1] A. Ross, “An Introduction to multibiometrics”. *15th European Signal Processing Conference (EUSIPCO)*, Poznan, Poland, 2007.

[2] T.N. Madandola and K. A. Gbolagade, “Reducing computtional time of Principal Component Analysis with Chinese Remainder Theorem”. *International Journal of Discrete Mathematics (DMATH).* 4(1) :1-7, 2019.

[3] E. O. Omidiora, “A prototype of knowledge-based system for black Face recognition using Principal Component Analysis and Fisher Discriminant Algorithms”, Unpublished Ph.D thesis, Computer Science and Engineering Department, Ladoke Akintola University of Technology, Ogbomoso, Nigeria, 2006.

[4] J. Sushma, S. B. Sarita and S. J. Rakesh, “3D face recognition and modelling system”, *Journal of Global Research in Computer Science*, 2(7): 30-37, 2011.

[5] X. Lu, “Image analysis for face recognition”. Personal notes, 2003.

[6] B. Moghaddam, “Principal manifolds and probabilistic subspaces for visual recognition”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 24(6):780–788, 2002.

[7] L. Torres, “Is there any hope for face recognition”*? In Proc. of the 5th International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS*, 21-23, 2004.

[8] I. T. Jolliffe and J. Cadima, “Principal component analysis: a review and recent Developments”. *Phil. Trans. R. Soc. A 374: 20150202. Retrieved from* http://dx.doi.org/10.1098/rsta.2015.0202 on 2016.

[9] T. M. Fagbola, S. O. Olabiyisi, F. I. Egbetola, and A. Oloyede, “Review of technical approaches to Face recognition in unconstrained scenes with varying pose and Illumination”, *FUOYE Journal of Engineering and Technology*, 2(1): 1-8. ISSN 2579-0625, 2017.

[10] T. N. Madandola, “Optimizing computational time of face recognition system using Chinese Remainder Theorem” Unpublished Ph.D thesis, Department of Computer Science, Kwara State Univetrsity, Malete, Nigeria, 2020.

[11] G. D. Guo, H. J. Zhang and S. Z. Li, "Pairwise face recognition". In *Proceedings of 8th IEEE International Conference on Computer Vision*. 2: i – xviii. Vancouver, Canada, 2001.

[12] M. Turk and A. Pentland, “Eigenfaces for recognition”, *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[13] J. O. Aluko, E. O. Omidiora, A. B. Adetunji and O. A. Odeniyi, “Performance evaluation of selected Principal Component Analysis-Based techniques for Face image recognition”, *International Journal of Scientific & Technology Research.* 4(01): 35-41, New York, Springer-Verlag, 2015.

[14] T. N. Madandola, K. A. Gbolagade and A. W. Yusuf-Asaju, “Performance assessment of Principal Component Analysis and Kernel Principal Component Analysis using TOAM database”, *Asian Journal of Research in Computer Science.* 3(2): 1-10, 2019.

[15] M. M. Ahmadinejad and E. Sherly, “A comparative study
on PCA and KPCA methods for Face recognition”, *International Journal of Science and Research (IJSR)*.5(6): 2319-7064, 2016.

[16] O. T. Adedeji, E. O. Omidiora, S. O. Olabiyisi and A. A. Adigun, “Performance evaluation of optimised PCA and Projection combined PCA methods in facial images”, *Journal of Computations & Modelling*, 2 (3): 17-29, 2012. 1792-8850 (online).

[17] The Yale database, Available: http://cvc.yale.edu/

[18] B. Max, “Artificial Intelligence in Theory and Practice”, IFIP 19th World Computer Congress, TC 12: 395. Santiago, Chile. Berlin: Springer Science+Business Media, 2006.

Enhancing Efficiency of Data Compression Techniques for Text Data using Chinese Remainder Theorem

Mohammed Babatunde Ibrahim and Kazeem Alagbe Gbolagade
Department of Computer Science
Kwara State University
Malete, Nigeria
imbamok@gmail.com kazeem.gbolagade@kwasu.edu.ng

ABSTRACT- Chinese Remainder Theorem (CRT) has been applied in various fields along with its applications in encryption, fault tolerant, error correction and detection, frequency estimation, coding theory as well as distributed data storage rather than in data compression. The traditional CRT is known to be highly sensitive to the errors in residues due to noises. The availability of texts in various applications has improved due to technological improvements having no impact on several of its operations on availability of sophisticated software tools that manipulate the text management. Compression ***is a known technique for storing large amount of data by dropping the number of character representations and pass on digital media. In this paper, CRT is applied through its residual attributes to the traditional data compression algorithms of Huffman coding and LZW for the purpose of enhancing its performance through simulation. Compression size, Compression time, Compression ratio and Space savings were used as performance metrics. The results indicate that LZW-CRT with 6.23% on the average performed better than other algorithms employed for compression size by 30.27%, 21.89% and 10.10% respectively for Huffman, LZW and Huffman-CRT, for compression time LZW-CRT had a value of 5.45s as against Huffman, LZW and Huffman-CRT having the following values 10.54s, 5.85s and 8.75s. Compression ratio for LZW-CRT is 10.68x better than Huffman, LZW and LZW-CRT with 2.60x, 3.10x and 7.79x while for space savings LZW-CRT outperformed other algorithms of Huffman, LZW and Huffman-CRT by 89.09% to 47.70%, 62.23% and 82.56% respectively. The results of this study performed better than some previous study highlighted in this work.***

Keywords: Text, Compression, Huffman coding, Lempel-Ziv-Welch (LZW), Chinese Remainder Theorem (CRT).

I. INTRODUCTION

The need for speed and storage space is the focus of recent computer technologies leading to the growth of an increased number of applications and important data giving rise to new methods of having efficient compression thereby aiding rise in greater performance [1]. So also, over the past decade, digital data transmission via the internet and cellular networks has brought about an unprecedented increase. This is due to the fact that data compression proposes a smarter approach in reducing the cost of communication through the effective usage of bandwidth [1] Data compression also known as source coding is the method of encoding information by the use of less bits than the original representation through the use of known encoding schemes. A technology known as compression is used for reducing the quantity of data used to represent any content without less difference from the quality from the original data. Compression also lessens the number of bits required to store and/or transmit digital media [2]. A data compression method is said to be lossless if the original data after decompression looks identical or a replica to the original data [3], while on the other hand, a compression method is said to be lossy if an approximate of the original data is generated. Data can be characters in text file, numbers that are samples of speech or image waveforms, or sequences of numbers that are generated by other processes. Most text files contain a lot of redundancies such as multiple occurrences of characters, words, phrases, sentences and even spaces. These redundancies are the structures which most compression algorithms tend to exploit to bring about reduction in the amount of data. Achieving a high amount of compression depends not only on the compression algorithm used but also on the types of data that is being compressed [4]. Text compression is a common scheme that reduces the number of bits needed to signify the data. This is capable of accumulating smaller storage spaces; enhance the speed of communication, and declining expenses incurred in hardware and network bandwidths setups [5]. The chief principle behind text compression is the simple removal of unnecessary characters, inserting specifically recurring characters with a string of characters and displacing a minor bit string for a repeatedly taking bit string [6]. Different compression algorithms are suited for different types of data, for example, MPEG (lossy) algorithm for video data compression, JPEG (lossy) algorithm is appropriate for picture data compression, while Run Length Encoding (RLE), Huffman and Lempel-Ziv-Welch (LZW) being lossless algorithms are suited for text data compression. In recent times, Chinese Remainder Theorem (CRT), a conversion algorithm in Residue Number System (RNS) have been established in fields of coding theory, phase unwrapping, frequency estimation and distributed data storage. This is certainly due to the fact; CRT is profound for isolating errors in residue caused by noise. The traditional CRT reconstructs a single integer for error-free co-prime and residues [7]. This paper aims to introduce Chinese Remainder Theorem (CRT) to enhance the performances of Huffman coding and Lempel-Ziv-Welch (LZW) algorithms for text compression.

II. Related Works

[8] utilized the inter-images correlations of external images to effectively compress images. The concept is based on cloud storage for image prediction scheme in which a semi-local approach is used to take advantage of the inter-image correlation. The reference image is divided into several planar regions using the local attributes that are matched and super-pixels. The photometric and geometric differences amongst the matched regions of current image and the reference image are resolved accordingly with classical video coding tools. Results showed significant rate-distortion efficiency as against high efficiency video coding techniques. [1] provided a comparison for text compression between Huffman and LZW algorithm using Chinese Remainder Theorem (CRT) where Huffman-CRT performed better than LZW-CRT in compression size by 55.64KB to 57.11KB, with a time of 44.61s to 55.39s Huffman-CRT still outperformed LZW-CRT. [9] presented an approach effective for short text compression for smart devices. Their study afforded a light-weight compression scheme where the storage required for compressing text is very low. Statistical context model for prediction of single symbols was used along with Arithmetic coding.

In [10], the notion of their compression algorithm was to describe an encryption technique to minimize all words in the dictionary so as to reduce every expression identified in glossary by substituting certain characters in the expressions via some distinct character, and retaining same characters so that the word is retrievable after compressing. In [11], the writers debated on the lossless techniques for text data compression of Huffman Coding, Shanon Fano coding and Run Length Encoding. Where the authors resolved in their study after undergoing assessment of these techniques and it was observed that the Huffman technique is the best for lossless data compression. [12] compared the performance between the algorithms of arithmetic coding and Huffman coding using the following metrics compression ratio, performance, and implementation. The algorithms were applied and tested with their results showing that results for compression ratio was higher for arithmetic coding than Huffman coding, while for effectiveness, Huffman coding performed superiorly to Arithmetic coding. Lastly, the authors also observed that it was easier to implement Huffman coding than the Arithmetic coding.

[13] considered for compression ratio the pre-transformation of images. The goal of their proposed approach is to preserve image recognition correctness making use of deep neural network for highly compressed images. It was observed that, recognition accuracy impacted on images for higher compression ratios using the conventional compression methods. Therefore, the proposed image pre-transformation minimizes the bitrate in order to keep hold of the recognition accuracy of the resulting images. Using ImageNet, it was observed that there was a 21.5% reduction on images. [14], carried out a research on algorithms used for text compression, the experimental results found out that the LZW algorithm performed the lowest grounded on the bits per compression (BPC). The LZW glossary/dictionary was revised by [15] as content addressable memory (CAM) based array utilizing smaller bits relative to the ASCII code. Information on tape or disk significantly contain lots of redundancy. The automatic usage of schemes used for compression has not been achieved due to multiple issues encountered, [16] provided a new principle that is not seen in most available commercial methods. [17] introduced Chinese Remainder Theorem (CRT), a subset of Residue Number System (RNS) in increasing the computational time for Huffman coding using compression time, compression ratio and space savings in measuring performance of the text compression schemes. The architecture of coding and decoding process was created using MATLAB 2015a. The results revealed significant improvement of compression time for CRT over Huffman coding alone by 1.55secs to 3.04secs. Again, the CRT based compression approach over-performed against Huffman coding alone by 38.55% to 19.08% for space savings. The trend is repeated for compression ratio, which is 1.63 to 1.24 for CRT and Huffman

coding only respectively. The results indicate that, using CRT consumes lesser execution time than when not in use with Huffman coding compression scheme.

A. **Chinese Remainder Theorem**

The basic operation of Chinese Remainder Theorem (CRT) which is a subset of Residue Number System is to generate a single integer through its residue modulo within moduli set [7]. The CRT is popular referred to as theorem of number theory that states that when the remainder of Euclidean division of an integer *n* by many integers, then, it is possible to certain exclusively the remainder of the division of *n* by taking the product of these integers having satisfied the condition that the divisor are pairwise coprime [18]. Given that $m_1, m_{2,} m_{3,} \dots, m_n$ are the pairwise relative prime positive numbers. And, the modular multiplicative inverse of integer P_i mod p_i is expressed as P_i^{-1} and must conform to Equation 1 [19]:

$$P_i . P_i^{-1} \equiv 1(mod\ p_i) \qquad (1)$$

where, i = {1,2,3, …, n}. For any given *n* positive integers, $b_1, b_{2,} b_{3,} \dots, b_n$, the CRT states that the pair of congruences are represented in Equation 2:

$$Y \equiv b_1 mod\ m_1, Y \equiv b_2 mod\ m_2, \dots, Y \equiv b_n mod\ m_n \qquad (2)$$

The exclusive solution $mod\ \partial_g = m_1\, m_2 \dots m_i = \prod_{i=1}^{n}(m_i)$. The solution realised from the key server can be expressed by Equation 3.

$$X \equiv b_1 + b_2 + \cdots + b_n \left(mod\ \partial_g\right) = \sum_{i=1}^{n} b_i \beta_i \gamma_i \left(mod\ \partial_g\right) \qquad (3)$$

where, $\beta_i = \frac{\partial_g}{m_i}$, and $\beta_i \gamma_i \equiv 1\ mod\ m_i$.

The residue independence, carry-free operation and parallelism attributes of the RNS have been intensively used in variety of areas, such as digital signal processing (DSP), digital filtering, digital communications, cryptography, error detection and correction [20,21]. The addition, subtraction and multiplication are dominant. And, division, comparison, overflow and sign detection are negligible. One key field of RNS-based applications is finite impulse response (FIR) filters. Likewise, digital image processing benefits from the RNS's features such as enhancing digital image processing applications [21].

B. Huffman Coding Algorithm

Huffman algorithm is the oldest and most widespread technique for data compression. It is an entropy encoding algorithm deployed for lossless data (text, audio, image and video) compression developed by David A. Huffman in 1952. It is based on building a full binary tree bottom up for the different symbols that are in the original file after calculating the frequency/probability for each symbol and put them in descending order. After that, it derives the code words for each symbol from the binary tree, giving short code words for symbols with large probabilities and longer code words for symbols with small probabilities. Huffman algorithm assigns every symbol to a leaf node, Root node and Branch node of a binary code tree. The tree structure results from combining the nodes step-by-step until all of them are embedded in a root tree. The algorithm always combines the two nodes providing the lowest frequency in a bottom up procedure. The new interior nodes get the sum of frequencies of both child nodes. This is known as Huffman code tree. The branches of the tree represent the binary values 0 and 1 according to the rules for common prefix-free code trees. The path from the root tree to the corresponding leaf node defines the particular code word [22].

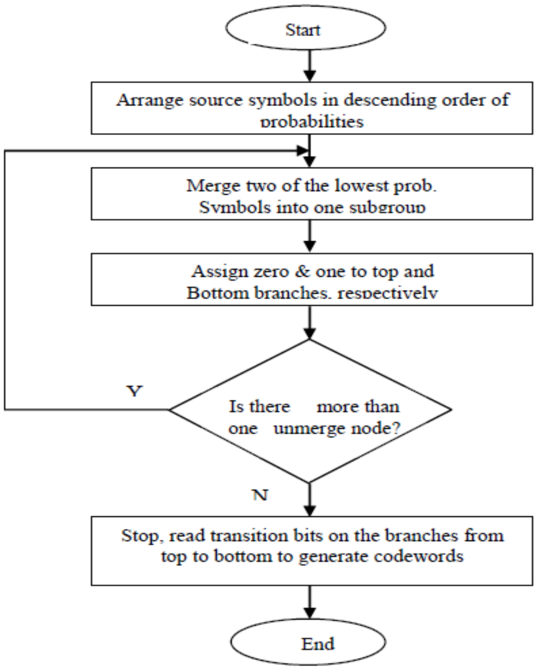

Start
Arrange source symbols in descending order of probabilities
Merge two of the lowest prob. Symbols into one subgroup
Assign zero & one to top and Bottom branches, respectively
Y
Is there more than one unmerge node?
N
Stop, read transition bits on the branches from top to bottom to generate codewords
End

Fig. 1: Algorithm of Huffman coding [12]

C. Lempel-Ziv-Welch Algorithm

The Lempel-Ziv-Welch (LZW) algorithm is one of the compression algorithms that capitalizes on the repetition or reoccurrence of words or phrases in a file for compression, reducing the number of bits or the execution time [23]. LZW algorithm is a universal lossless data compression algorithm created by Abraham Lempel, Jacob Ziv, and Terry Welch. It was published by Welch in 1984 as an improved implementation of the LZ78 algorithm published by Lempel and Ziv in 1978, which is simple to implement, and has the potential for very high throughput in hardware implementations. The scenario described by [16], encodes sequences of 8-bit data as fixed-length 12-bit codes. The codes from 0 to 255 represent 1-character sequences consisting of the corresponding 8-bit character, and the codes 256 through 4095 are created in a dictionary for sequences encountered in the data as it is encoded [14,16,24,25]. LZW (Lempel-Ziv-Welch) method maps strings of text characters into numeric code. To begin with, all characters that may occur in the string are assigned a code. For example, suppose that the string S = abaaabbbaaabaaabbb is to be compressed via LZW. Firstly, all occurring characters in string are assigned a code. In this case the occurring characters are “a” and “b” and are assigned 0 and 1 respectively. The mapping between character and their codes are stored in the dictionary. Each dictionary entry has two fields: key and code. The characters are stored in the key field while the numbers are stored in the code field. The LZW compressor repeatedly finds the longest prefix, p, of the unencoded part of S that is in the dictionary and outputs its code. If there is a next character c in S, then pc (pc is the prefix string p followed by the character c) is assigned the next code and inserted into the dictionary. This strategy is called the LZW rule [26].

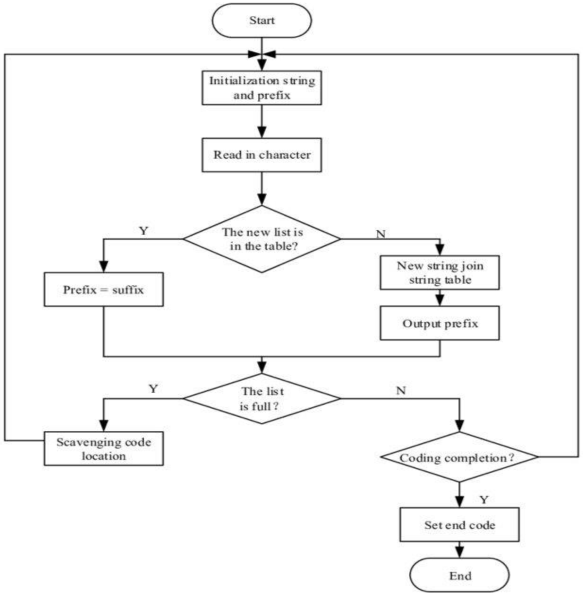


Fig. 2: The LZW Algorithm flow chart [27]

III. Methodology

In this lossless text compression implementation process, the process was coded from scratch using MATLAB R2015a programming language. The implementation was done in two stages. The first stage was the application of Huffman coding on the text data to extract relevant features from the text, while the second phase is the application of Lempel–Ziv–Welch (LZW) algorithm on the same text data. At the second stage, the Chinese Remainder Theorem algorithm was applied on all the algorithms to provide enhancements for the texts data. The algorithm used for the text compression is discussed given below:

The steps for performing Huffman-CRT text
compression scheme is presented
in algorithm 1 below;

Step 1: INPUT original text file

Similarly, the operational algorithm of proposed LZW-CRT text compression scheme is presented below

Step 1: Extract first byte from input STRING.
Step 2: Extract the next byte from input CHARACTER.
Step 3: Lookup in table for the STRING and CHARACTER stored up.
Step 4: Generate code for the STRING and update the lookup table.

Step 2: Run Huffman coding functions
Step 3: Extract symbols of the characters from input Text
Step 4: Create the probability of pixel symbols and organize in decreasing magnitude and smaller probabilities are combined.
Step 5: Concatenate the Huffman codeword ready for CRT
Step 6: Generate code for the STRING and update the lookup table.
Step 7: Apply CRT on the resulting STRING.
Step 8: The moduli set is chosen to obtain the best redundancy in data.
Step 9: The compressed text data is attained as final encoded values.
Step 10: The reconstructed text is obtained by applying decoding of Huffman and CRT
Step 11: Output is reconstructed text data.

Step 5: Output STRING same as CHARACTER.
Step 6: STRING = STRING and CHARACTER.
Step 7: Apply CRT on the resulting STRING.
Step 8: The moduli set is chosen to obtain the best redundancy in data.
Step 9: The compressed text data is attained as final encoded values.
Step 10: The reconstructed text is obtained by applying decoding of LZW and CRT.
Step 11: The final output is reconstructed text data.

Table 1. The CRT algorithm pseudocode

```
INPUT: y1, y2, ..., yn are integers in congruent equations and cp1, cp2, ..., cpn are relatively co-prime integers.
$x \equiv yi \bmod cpi$, for $I = 1, 2, \dots, n$.
OUTPUT: Solution of congruent equations (i.e. value of x)
Multiply the relatively co-prime numbers.
P = cp1 x cp2 x … x cpn
COMPUTE the constant Pi.
$\text{Pi} = \frac{p}{cpi}$, gcd (pi, cpi) = 1, for $i = 1, 2, \dots, n$.
Determine multiplicative inverse of Pi as $Pi^{-1}$ in the set $Z_{cpi}$
$Pi \times Pi^{-1} \equiv 1 \bmod cpi$, for $I = 1, 2, \dots, n$.
The OUTPUT of CRT operation is distinct value of x
$x = (\text{y1} \times P1 \times P2^{-1} + \text{y2} \times P2 \times P2^{-1} + \dots + \text{y1} \times Pn \times Pn^{-1}) \bmod P$
Return $x$
```

Performance Metrics

The Performance metrics are used in determining which techniques is better off according to some criteria. The nature of application determines the metrics to be used for compression algorithm performances [28]. The study intends to use the following performance metrics in analyzing the results. MATLAB 2015a was used in designing the system for simulation.

Compression Time (CT): This is the time taken to compress bits in data in a second (s).

Compression Ratio (CR): This is measured as the ratio between uncompressed size (US) of text and compressed size (CS) of text in relation to its bits. It is denoted by

$$CR = \frac{US}{CS}$$

Space Savings: This is the reduction in size relative to the original size. It is denoted by

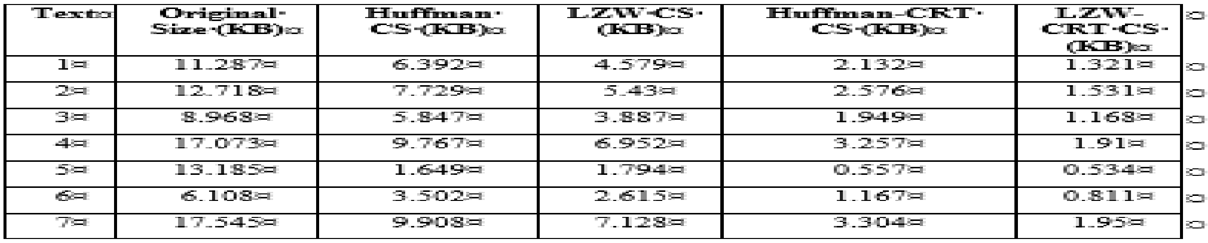
$$Space\ savings = \left(1 - \frac{CS}{US}\right) \times 100\%$$

US: Uncompressed (Original) Size

CS: Compressed Size

IV. Results

Compression Sizes computed for the Huffman coding, LZW and hybrid algorithm compression schemes using CRT enhancements for the various text files are shown in Table 4.1.



Text	Original Size (KB)	Huffman CS (KB)	LZW CS (KB)	Huffman-CRT CS (KB)	LZW-CRT CS (KB)
1	11.287	6.392	4.579	2.132	1.321
2	12.718	7.729	5.43	2.576	1.531
3	8.968	5.847	3.887	1.949	1.168
4	17.073	9.767	6.952	3.257	1.91
5	13.185	1.649	1.794	0.557	0.534
6	6.108	3.502	2.615	1.167	0.811
7	17.545	9.908	7.128	3.304	1.95

Table 4.1: The sizes of texts for Huffman coding, LZW compression schemes without and with CRT performed

From Table 4.1, the compression values of text files sizes for Huffman coding, LZW, Huffman-CRT and LZW-CRT are of different sizes. Figure 4.1 shows the variation in sizes between the compression algorithm. The compression values of the algorithms enhanced with CRT provided the best compression size procedures with LZW-CRT performing best when compared to other algorithms with and without CRT operation performed independently. After the compression, the resultant sizes of Huffman, LZW, Huffman-CRT, and LZW-CRT algorithms to the original text file are 30.27%, 21.89%, 10.10% and

6.23% respectively. The LZW-CRT enhancement procedures provided the best compression size as shown in Figure 4.1.

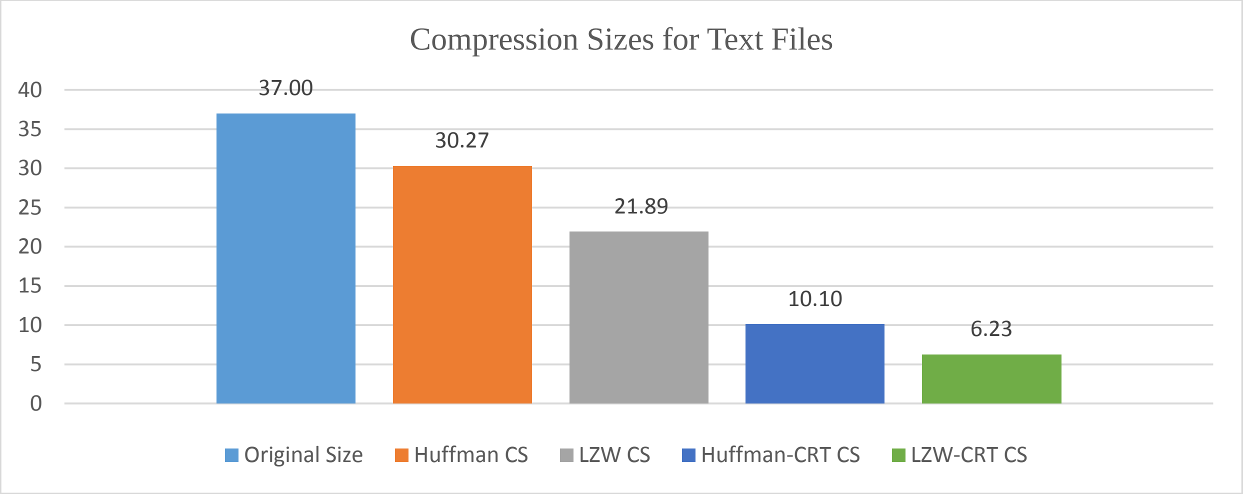


Fig.4.1: Compression Sizes for Text Files

Compression Time observed for the Huffman coding and LZW algorithms based on CRT enhancements for the distinct text files samples are presented in Table 4.2

Table 4.2: The compression time of texts for Huffman coding and LZW algorithms without and with CRT.

Text	**Huffman CT (s)**	**LZW CT (s)**	**Huffman-CRT CT (s)**	**LZW-CRT CT (s)**
1	9.3797	7.0187	9.1331	6.5383
2	18.7583	5.8027	10.076	7.1618
3	7.4023	4.3103	6.9223	3.4251
4	9.6238	4.8154	8.2003	3.972
5	8.6985	4.3741	7.9537	3.9009
6	7.6191	7.0626	6.8967	6.0261
7	12.2848	7.6006	12.0707	7.1003
Total	73.7665	40.9844	61.2528	38.1245
Average	10.53807	5.854914	8.7504	5.446357

From table 4.2, the values for compression time with the incorporation of CRT enhanced the time significantly especially for Huffman-CRT. Though, LZW and LZW-CRT performed better in isolation, on the average LZW-CRT performed better with 5.45s as against Huffman-CRT with 8.75s, Huffman with 10.54s and LZW having 5.85s respectively. The LZW-CRT enhancement provided better time as shown in figure 2 below.

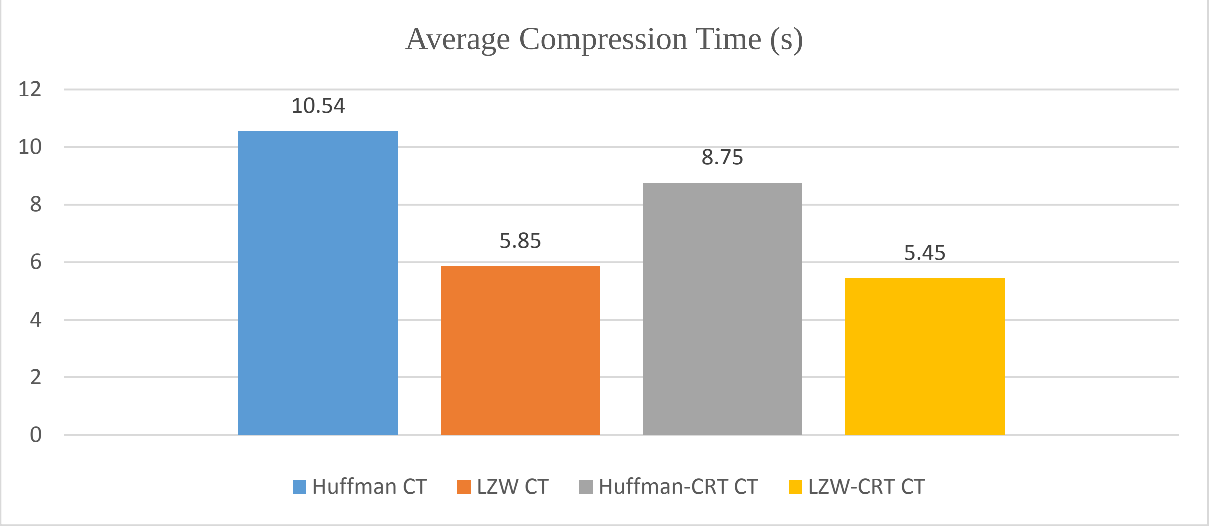


Fig. 4.2: Average compression time for Text files.

Compression Ratio calculated for the Huffman coding and LZW algorithms schemes enhanced with CRT for the different sample text files is presented in Table 4.3.

Table 4.3: The compression ratio of texts for Huffman coding and LZW algorithms without and with CRT.

Text	Huffman CR	LZW CR	Huffman-CRT CR	LZW-CRT CR
1	1.7658	2.4629	5.2941	8.5443
2	1.6455	2.3422	4.9371	8.307
3	1.5338	2.3072	4.6013	7.7681
4	1.748	2.4558	5.2419	8.9387
5	7.9958	7.3495	23.9292	24.691
6	1.7442	2.3358	5.2339	7.5314
7	1.7708	2.4614	5.3102	8.9974
Total	18.2039	21.7148	54.5477	74.7779
Average	2.600557	3.102114	7.792529	10.68256

In Table 4.3 and Figure 4.3, the compression ratio of text files for Huffman coding and LZW algorithms showed considerable improvements using CRT enhancement operations. In particular, the LZW algorithm with CRT was better for text files compression ratio than Huffman, LZW and Huffman-CRT accordingly. After successful compression operations based on without and with CRT, the Huffman, LZW, Huffman-CRT and LZW-CRT algorithms reduced the original data representations by 2.60x, 3.10x, 7.79x and 10.68x averagely in that order. Consequently, the LZW-CRT outperformed other compression algorithms understudy because of its capability to compress the original data representation as shown in Figure 4.3.

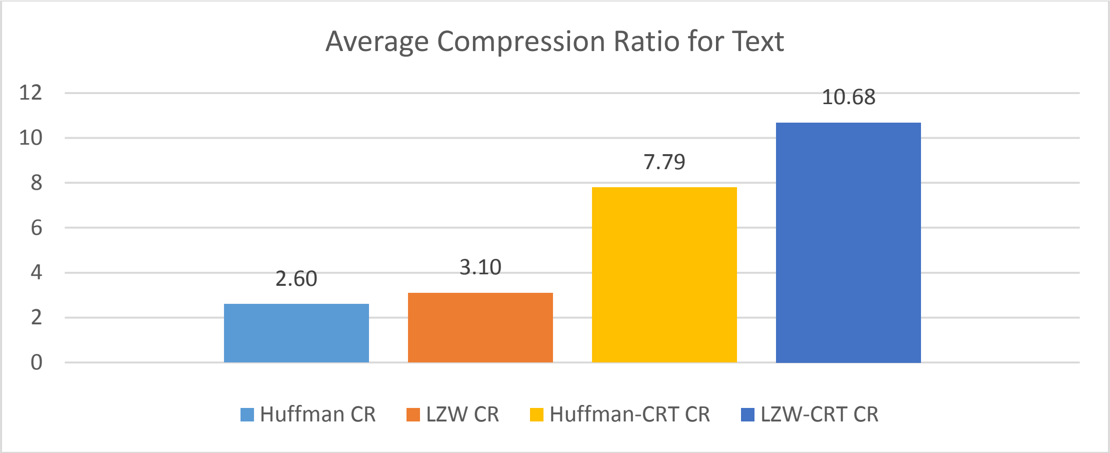


Fig. 4.3: Average Compression Ratio for text files

Space savings is used to determine the amount of redundancy removed after applying the compression algorithm of the Huffman coding and LZW algorithms schemes enhanced with CRT for the different sample text files are presented in Table 4.4.

Table 4.4: The space savings texts for Huffman coding and LZW algorithms without and with CRT.

Text	**Huffman SS (%)**	**LZW SS (%)**	**Huffman-CRT SS (%)**	**LZW-CRT SS (%)**
1	43.3685	59.4312	81.111	88.2963
2	39.2279	57.3046	79.7542	87.9619
3	34.8015	56.657	78.2672	86.9759
4	42.7927	59.2807	80.9231	88.8127
5	87.4934	86.3936	95.821	95.9499
6	42.6654	57.1873	80.8939	86.7223
7	43.5281	59.373	81.1684	88.8857
Total	333.8775	435.6274	577.9388	623.6047
Average	47.69679	62.23249	82.56269	89.08639

From Table 4.4 and Figure 4.4, the values for each compression algorithm for text files were stated. The average space savings for the CRT based compression procedure of LZW algorithm at 89.09% is the best against Huffman, LZW and Huffman-CRT at 47.70%, 62.23% and 82.56% respectively. The graphical representation of the various compression algorithms performances is shown in Figure 4.4.

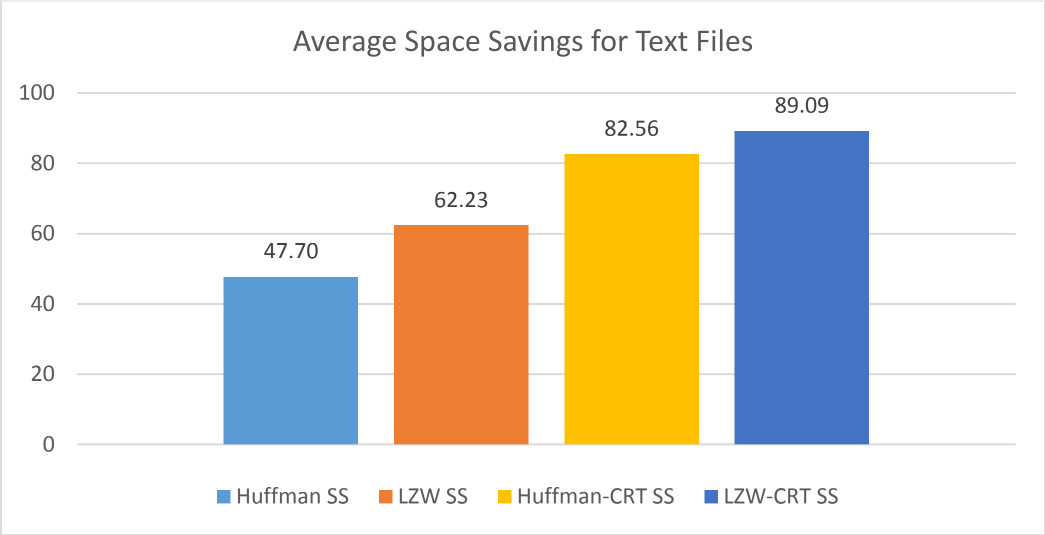


Fig. 4.4: Average Space savings for Text files.

Table 4.5: Summary of average values of the Algorithms

Evaluation Parameters	Huffman	LZW	Huffman-CRT	LZW-CRT
Compression Size (%)	30.27	21.89	10.10	6.23
Compression Time (s)	10.54	5.85	8.75	5.45
Compression Ratio (x)	2.60	3.10	7.79	10.68
Space Savings (%)	47.70	62.23	82.56	89.09

Table 4.6: Comparison with other related works

Author(s)	Original /Compressed Size (kb)	CT (s)	CR	Space saved (%)	Algorithm
Saravanan and Surender, 2013	43520/31747.5	-	0.40	60.23	Huffman coding
Sailunaz et al., 2014	454.75/402.33	-	1.13	19.67	Huffman coding
Alhassan et al., 2015	91/54	50.91		40.66	LZW-RNS
Ibrahim and Gbolagade (2019a)	10587/6041.3	44.61	3.70	-	Huffman coding
	10587/6200.3	55.39	2.37	-	LZW
Ibrahim and Gbolagade (2019b)	2276/1424.8	1.55	1.63	38.55	Huffman-CRT
Proposed approach	86.884/14.942	8.75	7.79	82.56	Huffman-CRT
	86.884/9.225	5.45	10.68	89.09	LZW-CRT

V. Conclusion

In this paper, a new text compression procedure involving the association of Huffman, LZW, Huffman-CRT and LZW-CRT was presented. After the computation and comparison with compression size, compression ratio and space savings for the algorithms, the results generated shows the amount of compression which can be obtained from each method. From the simulated results, it was observed that the most effective algorithm based on the LZW-CRT outperformed all other compression algorithms. For compression size, the LZW-CRT was best with 6.23% as against Huffman, LZW and LZW-CRT with 30.27%, 21.89% and 10.10% respectively. With compression time, LZW-CRT with 5.45s outperformed other used algorithms of Huffman, LZW, Huffman-CRT having the values 10.54s, 5.85s and 8.75s. For compression ratio, LZW-CRT was still better than other algorithms on the average of 10.68x, where Huffman, LZW and Huffman-CRT had 2.60x, 3.10x and 7.79x. The space savings of LZW-CRT was best with 89.09% as against Huffman, LZW and Huffman-CRT with 47.70%, 62.23% and 82.56%. It was discovered that the introduction of CRT being a reverse algorithm of RNS was able to enhance the traditional compression techniques due to its parallelism nature. The Future work will involve incorporating CRT or Mixed Radix Conversion (MRC) on its hybridization as well as other data compression techniques and on other forms of media data.

ACKNOWLEDGMENT

Our utmost appreciation goes to the management of Kwara State University, Malete, Nigeria for providing an enabling avenue to implement this study.

References

[1] M. B. Ibrahim and K. A. Gbolagade, “Performance Comparison of Huffman Coding and Lempel-Ziv-Welch Text Compression Algorithms with Chinese Remainder Theorem”. *University of Pitesti Scientific Bulletin: Electronics and Computers Science*, 19(2): 7-12, 2019a.

[2] M. Sharma, “Compression Using Huffman Coding”. *International Journal of Computer Science and Network Security (IJCSNS)*, 10(5): 133-141, 2010.

[3] W. E. Shannon and W. Weaver, “The Mathematical Theory of Communication”, University of Illinois Press, 1949.

[4] A. Moronfolu and D. Oluwade, “An Enhanced LZW Text Compression Algorithm”. *African. Journal of Computing and ICT (AJCICT)*, 2(2): 13-20, 2009.

[5] R. Radescu, “Transform Methods used in Lossless Compression of Text Files”, *Romanian Journal of Information Science and Technology*, *12*(1): 101-115, 2009.

[6] H. Kaur and B. Jindal, “Lossless Text Data Compression using Modified Huffman Coding-A Review”. *Proceedings of International Conference on Technologies for Sustainability-Engineering, Information Technology, Management and the Environment*, 1017-1025, 2015.

[7] H. Xiao, Y. Huang, Y. Ye and G. Xiao, “Robustness in Chinese Remainder Theorem for Multiple Numbers and Remainder Coding”. *IEEE Transaction on Signal Processing.* 1-16, 2018.

[8] J. Begaint, D. Thoreau, P. Guillotel and C. Guillemot, “Region-Based Prediction for Image Compression in the Cloud”. *IEEE Transactions on Image Processing*, 27(4):1835-1846, 2018.

[9] M. R. Islam and S. A. Ahson Rajon, “An Enhanced Scheme for Lossless Compression of Short Text for Resource Constrained Devices”, *International Conference on Computer and Information Technology (ICCIT)*, 2011.

[10] R. Franceschini and A. Mukherjee, “Data Compression using Encrypted Text*”. Proceedings of the Third Forum on Research and Technology Advances*, Digital Libraries, 1996.

[11] K. Rastogi and K. Sengar, “Analysis and Performance Comparison of Lossless Compression Techniques for Text Data”, *International Journal of Engineering Technology and Computer Research*, 2(1): 16-19, 2014.

[12] A. Shahbahrami, R. Bahrampour, M. S. Rostami and M. A. Mobarhan, “Evaluation of Huffman and Arithmetic Algorithms for Multimedia Compression Standards”. *International Journal of Computer Science, Engineering and Applications (IJCSEA)*, 1(4): 34-47, 2011.

[13] S. Suzuki, M. Takagi, K. Hayase, T. Onishi and A. Shimizu, “Image Pre-Transformation for Recognition-Aware Image Compression”, NTT Media Intelligence Laboratories, NTT Corporation, Japan. *2019 IEEE International Conference on Image Processing (ICIP)*, 2686–2690, 2019.

[14] S. Shammugasundaram and R. Lourdusamy, “A Comparative Study of Text Compression Algorithm”, *International Journal of Wisdom Based Computing*, 1(3): 68-76, 2011.

[15] S. Kaur and S. Verma, “Design and Implementation of LZW Data Compression Algorithm”. *International Journal of Information Sciences and Techniques (IJIST)*, 2(4): 71-81, 2012.

[16] T. A. Welch, “A technique for high performance data compression”. *IEEE Computer*, 17:8-20, 1984.

[17] M. B. Ibrahim and K. A. Gbolagade, “Performance Evaluation of Huffman Coding Algorithm Based Text Compression with Chinese Remainder Theorem”, Published in *International Journal of Information Processing and Communication*, 7(2):256-265, 2019.

[18] J. Zhang, J. Cui, H. Zhong, Z. Chen, L. Liu, “PA-CRT: Chinese Remainder Theorem Based Conditional Privacy-preserving Authentication Scheme in Vehicular Ad-hoc Networks”. *Journal of LATEX Class Files*, 14(8):1-14, 2017.

[19] X. Yan, Y. Lu, L. Liu, J. Liu and G. Yang, “Secret Data Fusion based on Chinese Remainder Theorem”, *Third (3rd) IEEE International Conference on Image, Vision and Computing*, 380-385, 2018.

[20] P. V. A. Mohan, “Residue Number System: Algorithms and Architectures”. Massachusetts: Springer, 2002.

[21] A. Omondi and B. Premkumar, “Residue Number System: Theory and Implementation”. Imperial College Press, London, 2007.

[22] D. A. Huffman, “A Method for the Construction of Minimum Redundancy Codes”, *Proceedings of Institute of Radio Engineers*, 40: 1098-1101, 1952.

[23] A. Alhassan, K. A. Gbolagade and E. K. Bankas, “A Novel and Efficient LZW-RNS Scheme for Enhanced Information Compression and Security”. *International Journal of Advanced Research in Computer Engineering & Technology*, *4*(11): 4015-4019, 2015.

[24] J. Amit, “Comparative Study of Dictionary Based Compression Algorithms on Text Data”. *International Journal of Computer Engineering and Applications*, 1(2):1-11, 2003.

[25] H. Jane and J. Trivedi, “A Survey on Different Compression Techniques Algorithm for Data Compression”, *International Journal of Advanced Research in Computer Science and Technology*, 2(3): 1-5, 2014.

[26] S. Sahni, “Data Structure, Algorithms and Applications in Java”, McGraw Hill Publishing Company, 437 – 450, 1999.

[27] G. Xiong, “Big Data Compression Technology Based on Internet of Vehicles”. *International Journal of Online and Biomedical Engineering*, 15(1):85–97, 2019. https://doi.org/10.3991/ijoe.v15i01.9773

[28] S. Pooja, “Lossless Data Compression Techniques and Comparison between the Algorithms”. *International Research Journal of Engineering and Technology (IRJET),* 2(2): 383–386, 2015

An Intelligent System for Student Placement Using Fuzzy Logic

Femi Temitope Johnson
Department of Computer Science
Federal University of Agriculture, Abeokuta
Ogun State, Nigeria.
femijohnson123@ hotmail.com

Olufunke Rebecca Vincent
Department of Computer Science
Federal University of Agriculture, Abeokuta
Ogun State, Nigeria.
vincentor@funaab.edu.ng

Abstract* - This present age is characterized with the usage of intelligent systems in almost every spectrum of human life and endeavour. Considerably, these intelligent systems have embedded knowledge and exhibit artificial intelligence which enable them perform special functions with utmost relevance within the shortest time duration, thus generating result similar to that of a natural (human expert). Intelligent systems deployed through innovations in Artificial Intelligence have greatly and positively impacted the human race in several areas of applications (i.e., agriculture, business, economy, business, medicine, warfare etc.). However, its influence can be more resounding when properly channelled towards education, as education is known to be the bedrock of knowledge. Many intelligent systems developed have classified students only based on their learning rates, one-time achievement test score or previous cognitive performance without consideration of the teachers and parents role in departmental selection. In this paper, the role of Artificial Intelligence was channelled towards education by developing a rule based intelligent system to aid school administrators in the placement of students (grade nine) into departments which suits their cognitive ability by combining the three educational assessment domains from students with both parent choice and teachers assessment in the departmental placement of students and later develop their interest in a chosen field or career. A high accuracy of 95.87 % generated by the fuzzy system depicts it as a very accurate and reliable tool for students' departmental placement.

Keywords - ***Intelligent systems, Fuzzy logic, Education, rules, student placements***

I. Introduction

The application of Artificial Intelligence (A.I) in education is, without doubt, creating a new dimension in the ways, manners, and approaches in which knowledge is being impacted and passed from one person to another. An indispensable tool for individual, societal and national development is education for without it, the development of any nation will be greatly affected. An individual must learn the accepted ways of doing things, culture and norms of the society he represents. This learning will also imbibe in him the rules, conduct, and expectations about future behaviour. The current and most innovative trend in 21st-century education is the application and utilization of Artificial Intelligence through the inclusion of smart learning which is highly dependent on the use of technology and related devices in smart education environments. With the introduction of smart learning and other related technologies in education, researches had revealed that it is a more efficient way through which learning can be enhanced in this present century.

However, for effective learning to take place several factors need to be put into consideration. [5] identified various factors hindering effective learning. If those factors are properly considered, they will greatly influence the extent to which learning is achieved. Another key factor for consideration as proposed in this paper is the system of education adopted by the learner's country of residence. Countries in the world have an adopted system of education they operate with. These systems of education may vary from one country to another based on their educational goals and the prospective level of development they envisage for their national development.

Educational system as summarized online by the educational glossary is "everything that goes into educating public-school students at the federal, state, or community levels." The term everything comprises laws, policies, and regulations, funding, infrastructures, human resources, teaching resources i.e. books, computers, teaching resources, and other learning materials as specified by the body solely responsible for its inclusion in different countries.

The body solely responsible for the formulation of Nigeria's educational system is the Federal Ministry of Education which is currently headed by the Minister of Education (Malam Adamu Adamu). The institution was established in the 1988 and saddled with the responsibility of using education as a tool for fostering the development of all Nigerian citizens to their full potentials. It had also formulated several policies to aid improve educational standards in the past. Some of these policies had a great significant effect on the country's education thus aiding national development while others do not see the limelight due to improper implementation. During the pre-independence days and shortly after Independence in 1960, the educational system adopted was similar to the British system (6-5-2 system) i.e. primary education takes 6 years to complete, 5 years for secondary education and about 2 years of higher-level / A Levels.

The introduction of the 6-3-3-4 system of education in Nigeria was introduced in 1973 to mirror the American system of education which allows a student to spend 6-years for primary education, three (3) years in junior secondary school, three (3) years of senior secondary education and four (4) years tertiary education. Likewise, the principal National Policy on Education was created and received in 1982 which prompted the presentation of different changes of various degrees of training. With the reception of the New National Policy on Education (2004), an increasingly huge methodology was conveyed in 2014 through the Universal Basic Education Board (UBE) into the instructive framework by transforming the current framework into another framework known as the 9-3-4 framework which adjusts to the Millennium Development Goals (MGDs) and the Education for All (EFA) approach activity.

It also accommodated some vocational skills in formal education thus enabling the Nigerian child to become self-reliant and self-dependent. In this paper, we present a novel approach to aid schools in making precise and accurate placement of grade nine students into two main departments (Science and Arts) within a short period of time thereby eliminating the challenges and problems posed by the traditional method of students choosing a department at an earlier stage of their senior secondary education or wavering between the departments during the period of secondary education which significantly has a negative effect on the students overall performance and the school at large. This paper is organized as follows: Section 1 gives a brief introduction on the application of artificial intelligence in education, the constituents of a nation's educational system and the goal of this paper. Section 2 presents a review of related works from which various computing techniques, algorithms and tools had been deployed in improving students' performance and classification. In section 3, the methodology adopted in carrying out the research was clearly explained in details and section 4 describes the implementation procedures and tool with analysis of the results obtained. Lastly section 5 gives a summary and conclusion of the research work.

A. OVERALL STRUCTURE OF EDUCATIONAL SYSTEM.

Basic Education Level

The Universal Basic Education, UBE, speculated a minimum of nine (9) years for the first schooling experience a child should undergo. During these years, elementary subjects and indigenous language are taught in the primary school for the first six-years allowing transitions from one grade to another until the sixth year (sixth grade) when the final primary school examination is written thus enabling the student to possess the ***Primary School Leaving Certificate*** on completion of the sixth grade. It is also further expected that the next three years will allow for the automatic and compulsory progression of students to junior secondary education (i.e. grades 7-9) with the introduction of other science-related and vocational subjects. At the end of grade 9, students are subjected to a state government administered examination known as Junior *School Certificate* Examination which will lead to the award of *Junior School Certificate* also known as *Basic Education Certificate* (*BEC*) for further progression into the next phase i.e. Secondary School Education.

Consequently, successful placement of students into respective departments that suits their overall ability had posed a challenge using the classical method. However, the introduction and adoption of an expert system for this fuzzy exercise will significantly improve and eliminate the associated problems with the classical method.

II RELATED WORK

[10] performed analysis on student prediction for campus placement adopting Data Mining Algorithm and K-Nearest neighbourhood technique for different cases of students. A total of nine hundred data sets were used, six hundred (600) data were used for training and the remaining 300 data for validating the model. The model showed an accuracy level of 92.67% and an execution time of 450(msec).

[21] in her research examined the influence of social media on undergraduate student's performance using Neuro-fuzzy modelling technique. This technique incorporated the use of statistical tool to analyse and test hypothesis. Result was also compared with the Neural Network model and a list of four(4) major statistical evaluation were used for result analysis but the Neuro-fuzzy model adopted proved more significant reliable with the minimum mean square error.

[14] presented a student placement prediction model adopting WEKA for mining the large collected set of data. They used different algorithms to predict and classify the data with the following result obtained: RBF network (65%), Bayes (70%), ID3 and J48 (71%), RT (73%). From the result obtained they concluded that the RT algorithm was more accurate for the prediction.

[1] modelled a prediction system for student placement using test results. A support vector machine, Artificial Neural Networks and Decision Tree Algorithm were deployed in the model resulting in an accuracy of 95%.

In addition, [15] work on predicting student placements using trained data from students on a fuzzy inference engine was able to efficiently predict and accurately analyse the numerous lists of students for placement without breakdown. The system classified two major groups of students with five (5) input data sets and more than five hundred (500) rules were embedded in the prediction system.

[12] opined in their research that fuzzy expert with cognitive mapping Approach is a more efficient model for developing Intelligent Tutoring Systems (ITS). Their model comprises of both an expert system and modules with different assessment techniques to determine the level of understanding and reception of concept taught and learnt by the students.

In [2] classification method for predicting and classification of secondary school students proved effective through the application of data mining techniques to classify students marks and grades in various subjects offered and implementation was performed using Decision Tree Algorithm.

Similarly, [11] collected over 200 data of students from previous academic performance over the years, applied a decision tree Algorithm on the collated data to predict students final Cumulative Grade Point Average(CGPA).

Dawod et al., (2017), developed a flexible fuzzy inference model to depict the pedagogical needs of students using teachers' fine-tuned processes. The tested developed model demonstrated effectiveness in learning the interaction among students, this aiding accurate proper placement with higher degree of accuracy.

III METHODOLOGY

The methods adopted in this paper is an interwoven concept which adopts fuzzy logic at each progressive stage to determine, rationalize ambiguity and provides solution to the fuzziness experienced during the procedural analysis. The stages are grouped into three: Data Requisition stage (DSR), Data Analysis Stage (DAS) and System Implementation Stage (SIS).

A. Data Requisition Stage (DRS)

This is the initial stage of the research process. This stage involves meeting with the domain expert i.e. the school registrar from which data, rules and conditions required for student placements are gotten. These data were further categorized into three sections as:

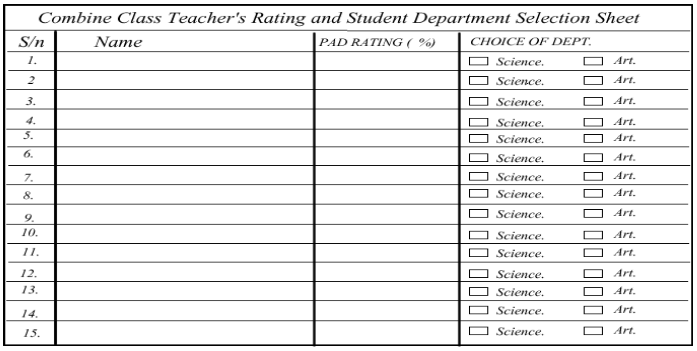
i. Cognitive Assessment Data (CAD)
ii. Psychomotor and Affective Data(PAD)
iii. Choice Selection Data(CSD)

List of subjects from which students are tested at both internal and external examinations are depicted in Table 1. Performance after examination are scored, graded and remarked with corresponding alphabets as shown in Table 1.

TABLE 1. GRADED STUDENTS SUBJECTS.

S/N	Subjects	Grade	Remark	Point
1.	ENGLISH	A	Excellent	5
2.	MATHEMATICS	B	Very Good	4
3.	BASIC SCIENCE (COMPUTER, PHE, INT.SCIENCE)	C	Good	3
4.	BASIC TECH	D	Pass	2
5.	YORUBA	E	Pass	1
6.	BUSINESS STUDIES	F	Fail	0
7.	FRENCH			
8.	HOME ECONOMICS			
9.	AGRIC			
10.	VOCATIONAL STUDIES (FINE ART, MUSIC)			

A sample form is also filled by the parents of concerned students to get data regarding their choice of department. Figure 1 is a template comprising of Student Departmental Choice selection and class teacher's assessment which forms the remaining constituent of the Choice Selection Data (CSD) and Psychomotor Assessment Data (PAD) respectively.



Combine Class Teacher's Rating and Student Department Selection Sheet

S/n	Name	PAD RATING (%)	CHOICE OF DEPT.
1.			□ Science. □ Art.
2			□ Science. □ Art.
3.			□ Science. □ Art.
4.			□ Science. □ Art.
5.			□ Science. □ Art.
6.			□ Science. □ Art.
7.			□ Science. □ Art.
8.			□ Science. □ Art.
9.			□ Science. □ Art.
10.			□ Science. □ Art.
11.			□ Science. □ Art.
12.			□ Science. □ Art.
13.			□ Science. □ Art.
14.			□ Science. □ Art.
15.			□ Science. □ Art.

Fig.1 Student Departmental Choice selection and Class Teacher's Assessment

B. Data Analysis Stage (DAS)

At this stage, the data collected were analysed to suit the research. It involves turning the fuzzy variables into mathematical models for representations.

1 Mathematical Modelling

The initial mathematical modelling is the Cognitive Assessment Data which deals with knowledge, understanding and comprehension of the students. The data for each student was derived by subjecting them (students) to two different types of examinations i.e. Internal and external examinations.

2 Cognitive Assessment Data Ratings

This examination is administered by the individual schools for students. It subjects the students to ten compulsory subjects from which performances are rated and graded as shown in Table 1. The mathematical equation representing the set containing the collection of subjects to which students are graded is shown in Eq. (1),

$$I(i.ex)_{\text{subject s}} = \left\{ \begin{matrix} \textit{English Studies, Mathematics, Basic Sc.,} \\ \textit{Basic Technology, Yoruba Lang., Frenc} \\ \textit{Business Studies, Home Econs, Agric Sci.} \\ \textit{Voc. studies} \end{matrix} \right\} \quad (1)$$

$$G(i.ex) = \begin{cases} pass, & i.ex \geq 50\% \\ fail, & i.ex < 50\% \end{cases} \quad (2)$$

A student who obtains score greater or equals to fifty per cent in any of the subject examined in the internal examination is assigned a pass grade and otherwise if the score is less than fifty per cent. The overall percentage in ten subjects is computed with Eq. (3),

$$OIex = \sum_{S=1}^{10} \left(\frac{G(i.ex)}{10} \right) \times 100\% \quad (3)$$

Placement of students in the next class is also dependent on students' performance in external examination assessment. Scores obtained in the same subject as reflected in Eq. (1) are graded as shown in Eq. (4) where S.Ex denotes scores obtained in external examination and *OE.ex*$_{(Grade)}$ represents the over-all external result grade obtainable by each student.

$$OE.ex_{(Grade)} = \begin{cases} distincion, 75\% < s.Ex \leq 75\% \\ merit, 46\% < s.Ex \leq 75\% \\ fail, 0\% \leq s.Ex \leq 45\% \end{cases} \quad (4)$$

3 Psychomotor and Affective Data

Educationist are of the opinion that cognitive domain isn't sufficiently accurate to determine if learning has taken place. As a result, the PAD is an essential attribute that must be considered as it denotes if positive changes and skills have been acquired by the students during the period of learning.

These data is supplied by the student's class teacher who over the time must have considered and noticed positive changes in the attitudinal behaviour of the students. The PAD rating is measurable in terms of percentage. In Eq. (5), the Psychomotor and Affective Data (PAD) is subject to the Class Teacher's Assessment (CT.Ass.)

$$PAD = f(CT\ ASS) = \begin{cases} Science, (60 \leq \ CT.Ass \ \leq 100)\% \\ \quad Art, (\ 0 \leq \ CT.Ass \ < 60)\% \end{cases} \quad (5)$$

4 Choice Selection Data (CSD)

The students and parents are also involved in the process as their contribution towards the departmental choice is taken into consideration. However, Parent's Choice (PCH) is attached an appreciable significant value than the Students Choice (SCH) except when the parent is indifferent to the choice of department. Both choices are depicted mathematically in Eq. (6)

$$PCH = \begin{cases} Science, (71 \leq pch \leq 100) \\ Art, (40 \leq pch \leq 70) \\ Ind, (\ 0 \leq pch \leq 39) \end{cases} \qquad SCH = \begin{cases} Science, (71 \leq sch \leq 100) \\ Art, (40 \leq sch \leq 70) \\ Ind, (\ 0 \leq sch \leq 39) \end{cases} \quad (6)$$

where PCH = Parents' choice values, SCH = Students' choice values

$$SCH \ > PCH \ \ (iff\ pch \leq 40) \ \ else$$

$$PCH \ > SCH \ \ \forall\, pch, sch \in (1 - 100)$$

3.3 Fuzzy Input Variables and Membership Functions

Every expert system requires input for which processing must be performed. The input specified for the systems have related membership functions as there exist a significant interaction among the variables.

$$S.iv = \{E.ex, I.ex, PAD, PCH, |SCH\} \quad (7)$$

$Given\ that: S.iv =$
$System\ input\ variables \quad CAD = E.ex +$
$I.ex\ \ and\ CSD = PCH + SCH$
$where\ CT.Ass \neq 0$

Since CAD comprises of two distinct variables as shown in Eq.(7), distinct ratings and membership functions exist as shown in equations (8 and 9) respectively. However, the other variables are assigned respective membership functions as shown in Eq. (10) and Eq. (11).

$$\mu_{CAD}(E.ex) = \{Distinction, Merit, Fail\ \} \quad (8)$$

$$\mu_{CAD}(i.ex) = \{Pass, Fail\ \} \quad (9)$$

$$\mu_{CSD}(PCH, SCH) = \{science\ , Art, Ind.\} \quad (10)$$

$$\mu_{PAD}(CT.Ass) = \{Science, Art, \ \ \} \quad (11)$$

D. Fuzzification and Defuzzification Process.

The system is fully in charge of the process. Each fuzzy variable is turned into linguistic type with assigned membership function. An inference engine interacts with the linguistic variables and the knowledgebase by mapping it to its associated rules, thus generating a fuzzy output which can be defuzzify for a better understanding by the user. Eq. (12) represents the triangular membership function used for defuzzifying the output where S, A and R denotes result to indicate if a student should be placed in either Science department, Art department, or repeats the present class.

$$\mu_{S.iv}(P.D; [R, A, S]) = \begin{cases} 1\ if\ P.D = S \\ \frac{PD-AS}{R-S} if\ P.D \ \in [S, R] \\ \frac{A-PD}{S-A}, if\ P.D \ \in [A, S] \\ 0\ if\ P.D \geq S \end{cases} \quad (12)$$

$$\sqrt{\sum_{i=1}^{n}} \vartheta^2 = \sqrt{\vartheta^2{}_1 + \vartheta^2{}_2 + \vartheta^2{}_3 + \cdots + \vartheta^2{}_n} \quad (13)$$

ϑ^2 = root mean square method used by the inference engine

In Eq.(13), $\vartheta^2{}_1, \vartheta^2{}_2, \vartheta^2{}_3 \ldots \vartheta^2{}_n$ are values of the different rules with the same conclusion in the knowledgebase. The output is derived by computing the aggregate of the value obtained in equation (13) and defuzzifying using the centroid model.

E. Fuzzy rule knowledge base

The fuzzy rules embedded in the knowledgebase was provided by an expert and arranged in a similar way as the Mamdani type. A set of IF-THEN control rules are matched or compared with the knowledge in the database to generate an output similar to that of an expert admission expert (registrar). The rules are depicted as shown in Table 3.

Table 3. Sample fuzzy rules for departmental Placement.

S/N	CAD		PAD	CSD		OUTPUT
	I.Ex	E.Ex		S.Ch	P.Ch	DEPT
1.	P	D	S	S	S	SCIENCE
2	P	D	A	A	A	ARTS
3.	P	M	A	I	I	ART
4.	F	F	A	I	S	REPEAT
5.	F	M	S	I	S	ART
6	P	M	A	A	I	ART
7	P	M	S	S	A	SCIENCE
8	P	M	S	A	S	SCIENCE
9	P	F	A	I	A	REPEAT
10	F	F	S	S	S	REPEAT

The tenth rule in the table 3 can be interpreted as thus:
If a student (Fails the Internal Examination) and (Fails the External Examination) and (PAD is Science inclined) and (Student wishes to be in Science Dept) and (Parents Choice is Sciences) THEN placed department is REPEAT present class. Other rules which form the knowledgebase can also be interpreted in the same manner.

F. Intelligent Student Placement System Algorithm

The algorithm depicted in Fig.(3) shows the systematic approach used by the sytem to genarate output. Inputs as specified in Eqs. (8, 9 10 and 11) are supplied by the user , processing is carried out by the system with generated results (i.e Student Department).

```
INPUT: collectedData
OUTPUT: Student Department
BEGIN
  while collectedData is not empty Do
    for each rule in $R^x$ of rulebase
      compare collectedData with $R^x$
      If collectedData = $R^x$
        Place student in Dept
      Else
        Update R
      Endif
    Endfor
  Endwhile
  For each Predicted $R^x$ do
    Case (switch)
    If (switch== a)
        StudentDept is Science
    If (switch== b)
        StudentDept is Art
    If (switch== c)
        StudentDept is Undetermined (Repeat)
    End Case
  End for
End
```

Fig3. Intelligent System Algorithm for Student Placement Algorithm

IV. IMPLEMENTATION AND RESULT ANALYSIS

This research paper requires series of data which were collected and analysed. The implementation and analysis procedure was divided into three phases which requires gathering of data i.e. (subject offered by students, allot-able grades and points) from schools in both internal and external examinations. The list of subjects were limited to ten (10) as some subjects were merged and embedded as one. Figure 4 shows the list of subjects and maximum allot-able score that can be obtained in the examination.

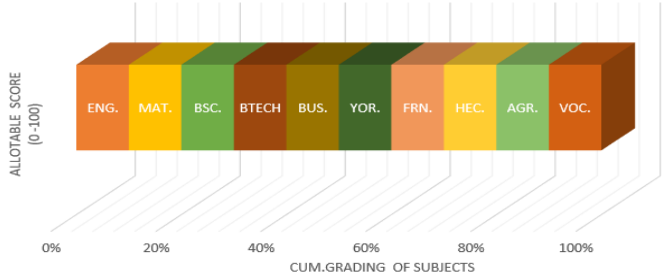


Fig 4. Subjects and allocable scores for Departmental placements.

Departmental choices collected from both students and parents were also instrumental in the development of the improved system. Choices of students and parents were sorted and graded according to the range of values shown in fig. (5). The data serves as input to the next stage of the implementation process.

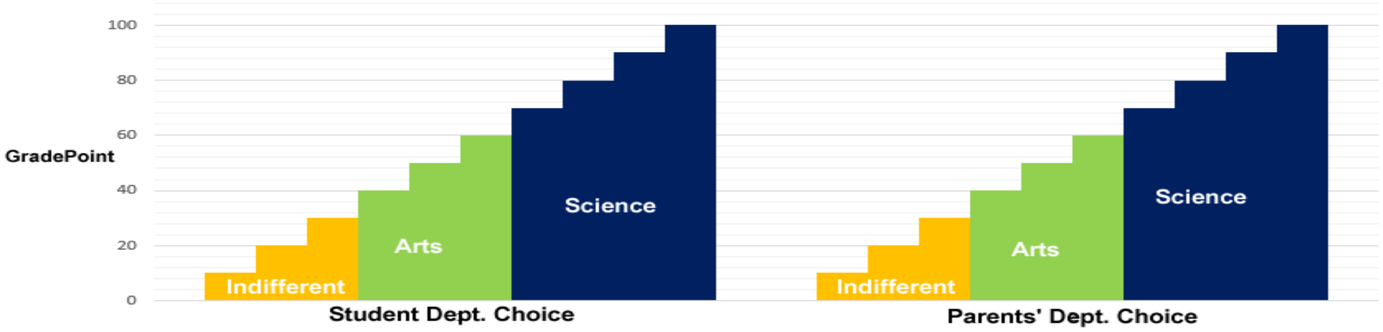
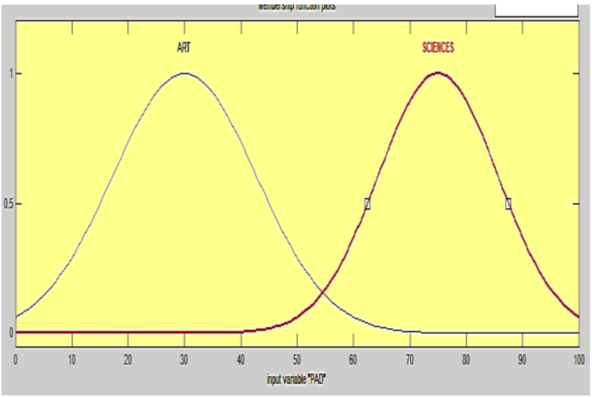
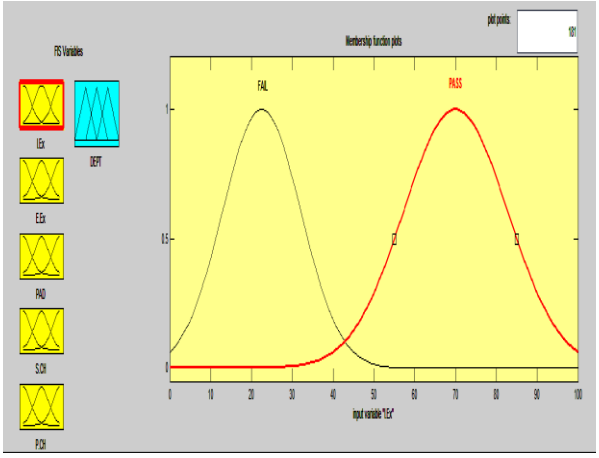


Fig 5. Analysis criteria for Departmental Choice selection.

A. Implementation and Result Analysis.

Simulation and implementation of the fuzzy based intelligent system was performed with the aid of MATLAB tool box. Marks and data obtained by 100 (grade nine) students of FCEMSS in both internal and external examinations were tested. For each student, Psychomotor Assessment performed by the teacher and internal examination scores obtained by the students were fuzzified using the Gaussian membership function as shown in fig. (6).





Triangular membership function was used to fuzzify other input variables (i.e. External Examination grades, Student Choice and Parent Departmental Choice). The Mamdami Fuzzy Decision Techniques combined with the IF THEN rules were adopted to determine the active output membership function. Output obtained (i.e. predicted department) was defuzzified by calculating the centroid thus generating values as shown in fig. (7).

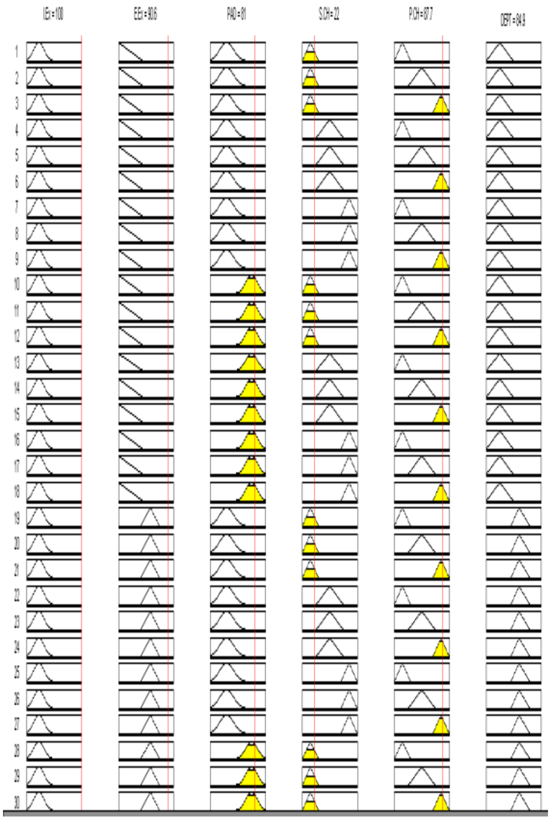


Fig 7. Surface viewer result interface

Figs. 8a and 8b are reflections of the relationship which exits between various input variables (i.e. Student Departmental Choice and Internal Examination grade). By research experimentation, it is also evident that similar correlation exists between the Internal Examination grade and Student Departmental Choice.

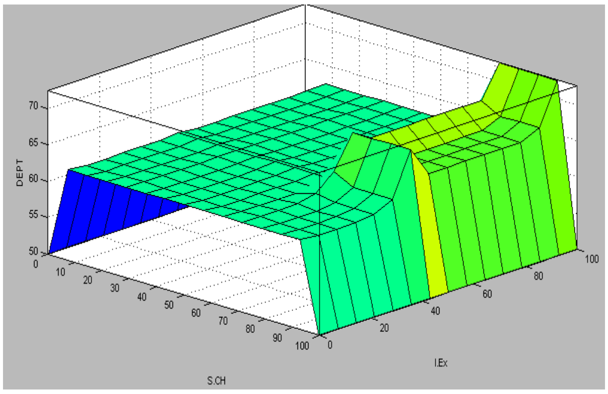


Fig. 8a. Plot of S.Ch against I.Ex

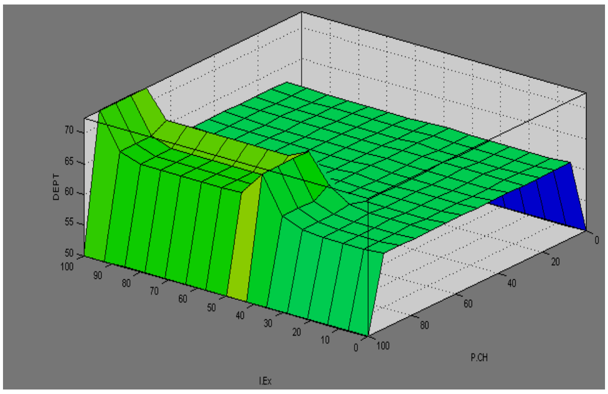


Fig. 8b. Plot of I.Ex against P.Ch

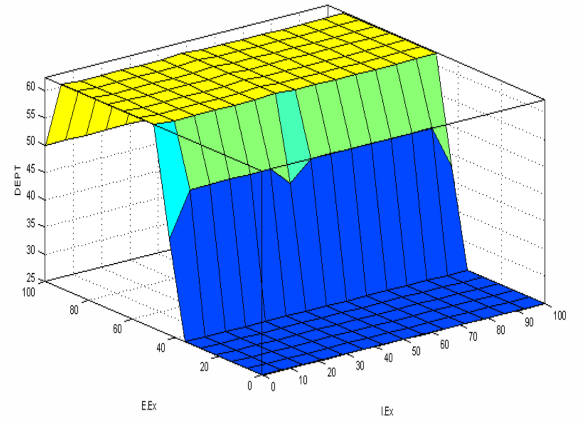


Fig. 8c. Plot of E.Ex against I.Ex

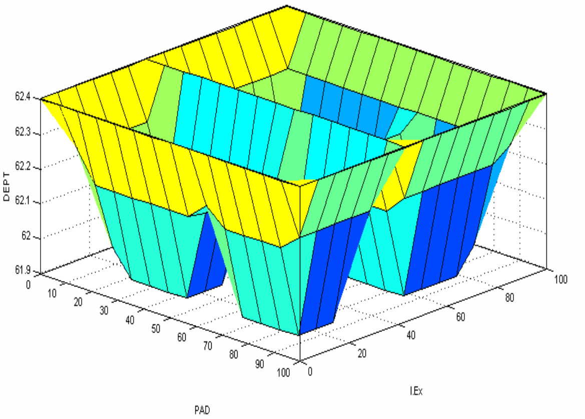


Fig. 8d. Plot of PAD against I.Ex

Grades of students obtained in External and Internal Examination were analysed to depict the type of correlation which existed between them. Similarly, assessment of Internal Examination Grade with Psychomotor Assessment for student departmental placement was performed and depicted in Figs 8c and 8d respectively.

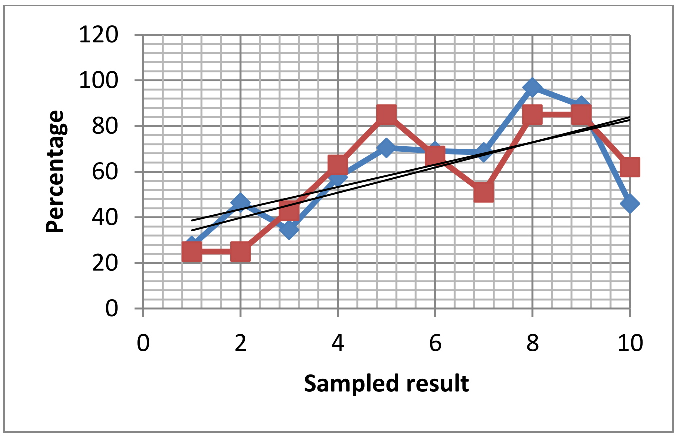


Fig.9 Graphical Correlation between Cognitive Assessment and Placement result

The correlation which exists between the Cognitive Assessment of students and Placement Result is depicted in Fig.9. Similarly, other input parameters and their corresponding results are as shown in Fig. 10.

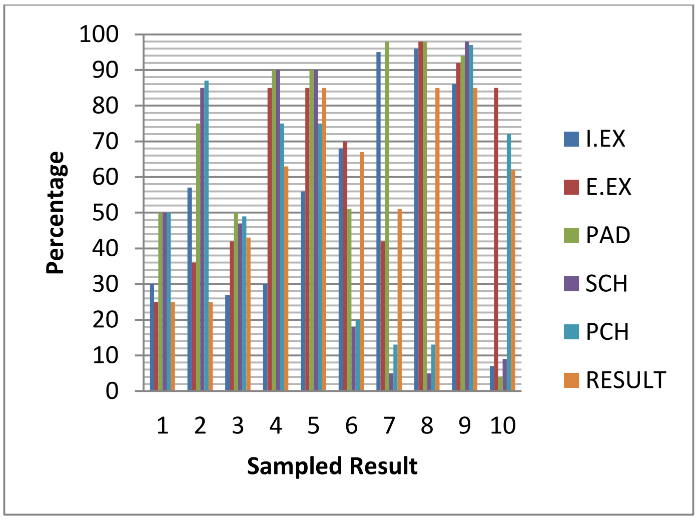


Fig.10 Analysed result of sampled parameters

The functionality and usability of the intelligent system was extremely significant as the output derived shows a remarkable result with an accuracy of 95.87%. Fig.11 is a formal comparison of the accuracy generated by the developed system with other known systems from literature reviewed.

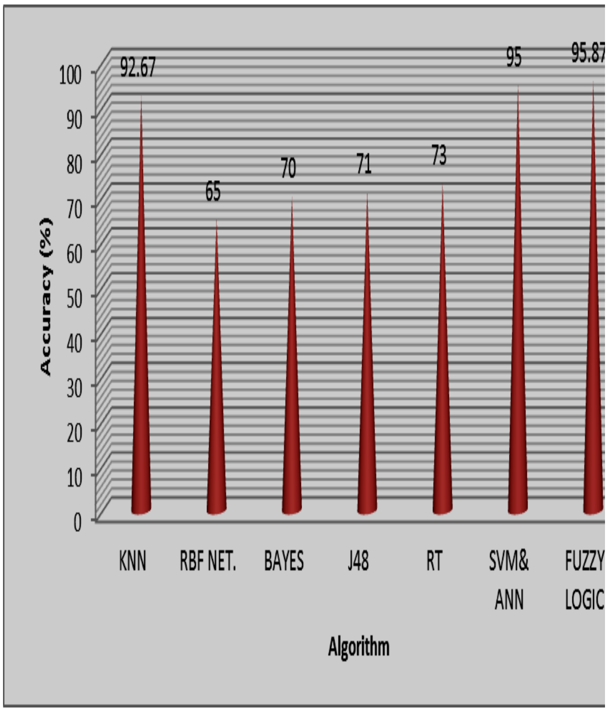
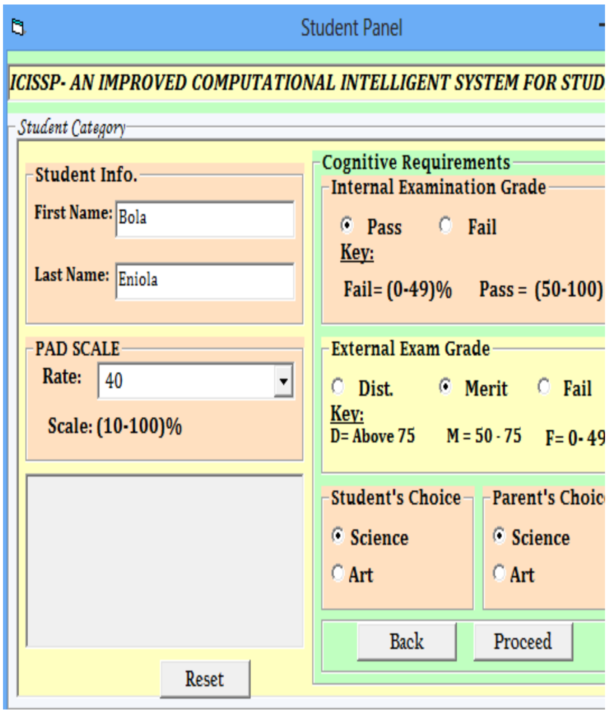


Fig. 11.Accuracy Evaluation and Comparison metric

In addition, technological acceptance of the system is also worth commending as it reduces the challenges faced by the school registrar in placement of students into departments. The developed system is an interactive one with user friendly output depicted in figs. 12a and b respectively.


Student Panel

ICISSP- AN IMPROVED COMPUTATIONAL INTELLIGENT SYSTEM FOR STUD

Student Category

Student Info.

First Name: Bola

Last Name: Eniola

PAD SCALE

Rate: 40

Scale: (10-100)%

Cognitive Requirements

Internal Examination Grade

● Pass ○ Fail

Key:

Fail= (0-49)% Pass = (50-100)

External Exam Grade

○ Dist. ● Merit ○ Fail

Key:

D= Above 75 M = 50 - 75 F= 0- 49

Student's Choice

● Science

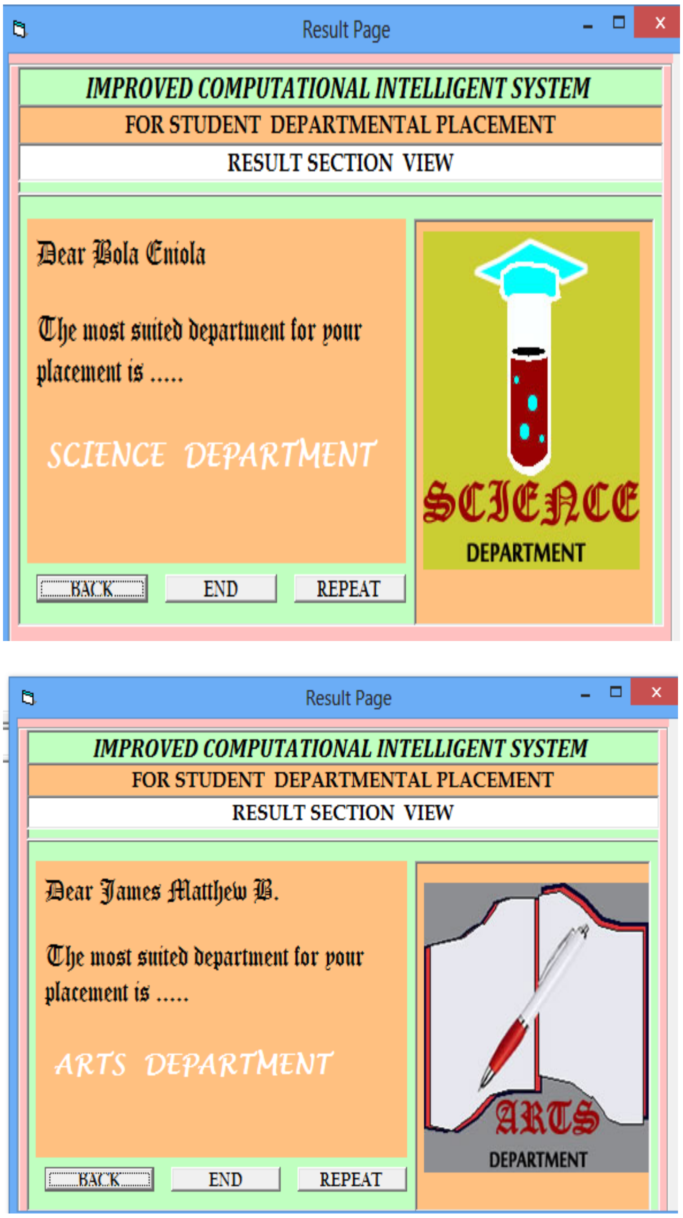
○ Art

Parent's Choic

● Science

○ Art

Fig 12a. Intelligent Placement System Data Requisition Interface


Result Page

IMPROVED COMPUTATIONAL INTELLIGENT SYSTEM

FOR STUDENT DEPARTMENTAL PLACEMENT

RESULT SECTION VIEW

Dear Bola Eniola

The most suited department for your placement is .....

SCIENCE DEPARTMENT

Result Page

IMPROVED COMPUTATIONAL INTELLIGENT SYSTEM

FOR STUDENT DEPARTMENTAL PLACEMENT

RESULT SECTION VIEW

Dear James Matthew B.

The most suited department for your placement is .....

ARTS DEPARTMENT

Fig. 12b. Intelligent System for Student Placement Result

Interfaces

V. CONCLUSION

This paper has adopted the use of Fuzzy logic to predict the most suited department for students. The prediction criteria adopted is a novel type combining Student choice, Parent choice Class teacher’s Assessment and school factors.

The algorithm describes the processes performed by the modelling engine in the selection and prediction of student department. Furthermore, the framework and architecture of the system presented depicts the variables deployed in the robust system. Output generated with an accuracy level of 95.87% by the system is a true replica of the action the expert (school registrar) would perform in order to place student into various department.

Conclusively, the data set for testing was limited but can be extended to an appreciable large number depending on the available number of students to be placed into departments. However, the fuzzy intelligent system for students' departmental placement has been noted to enhance and efficiently solve the rigorously task consuming mental power and enormous time of registrar during student departmental placement. Significantly noticeable in this research is the relationships which exist between the adopted variables and the computational time required for the system to generate the output for each student input.

REFERENCES

[1] B. Minaei-bidgoli, Ben Henson D.A. Kashy, G. Kortemeyer and W.F. Punch,(2003),"Predicting Student Performance: An Application of Data Mining Methods with An Educational Web-Based System", Proceedings of 33rd Annual Frontiers in Education, pp. 1-6,

[2] Bashir Khan, Malik Sikandar Hayat and Muhammad Daud Khattak,(2015) "Final Grade Prediction of Secondary School Student using Decision Tree", International Journal of Computer Applications, Vol. 115, No. 21, pp. 32-36

[3] Carmona, C., Castillo, G.,Millan, E., (2007)"Discovering Student Preferences in E-Learning", Proceedings of the International Workshop on Applying Data Mining in e-Learning, ADML-07, pp. 33-43.

[4] Chen, C., Hong, C., Chen, S., Liu, C., (2006), "Mining Formative Evaluation Rules Using Web-based Learning Portfolios for Web-based Learning Systems", Educational Technology & Society, Vol. 9, No. 3, pp. 69-87.

[5] Essays, UK. (November 2018), "Internal and External Factors of Effective Learning Education" Essay. Retrieved from https://www.ukessays.com/essays/education/internal-and-external-factors-of-effective-learning-education-essay.php?vref=1

[6] Esposito, F., Licchelli, O., Semeraro, G., (2004), "Discovering Student Models in e-learning Systems", Journal of Universal Computer Science, vol. 10, no. 1, p. 47-57.

[7] Hatzilygeroudis, I., Prentzas, J., (2004), "Using a hybrid rule- based approach in developing an intelligent tutoring system with knowledge acquisition and update capabilities", Expert Systems with Applications, Vol. 26, pp. 477–492

[8] Indriana Hidayah, Adhistya Erna Permanasari and Ning Ratwastuti,(2013),"Student Classification for Academic Performance Prediction using Neuro Fuzzy in a Conventional Classroom",Proceedings of IEEE Conference Information Technology and Electrical Engineering, pp.1-5.

[9] Jerry M. Mendel,(1995),"Fuzzy logic systems for engineering: A Tutorial", Proceedings of the IEEE, 83(3):345-377

[10] Mangasuli Sheetal B, Savita Bakare, (2016) Prediction of Campus Placement Using Data Mining Algorithm-Fuzzy logic and K nearest neighbour. International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 6, June 2016 , 309- DOI 10.17148/IJARCCE.2016.5666

[11] Mashael A. Al-Barrak and Muna Al-Razgan,(2016), "Predicting Students Final GPA Using Decision Trees: A Case Study", International Journal of Information and Education Technology, Vol. 6, No. 7, pp. 528-533.

[12] Mohammed Hussein, Negoita, M., Pritchard, D., (2004), "Using a virtual student model for testing intelligent tutoring systems", Interactive Technology and Smart Education, Vol. 1, pp.195–20.

[13] Nykanen, O., (2006), "Inducing Fuzzy Models for Student Classification", Educational Technology and Society, Vol. 9, No. 2, pp. 223-234.

[14] Ravi Tiwari and Awadhesh Kumar Sharma, (2015) "A Data Mining Model to Improve Placement", International Journal of Computer Applications , Volume 120 – No.12, (0975 – 8887)

[15] Ravi and Jayanthi (2017), "Student Prediction System For Placement Training Using Fuzzy Inference System. " ICTACT Journal of Soft Computing, Volume 7, Issue 3, pp 1-7.

[16] Romero, C., Romero, J.R., Luna, J.M. and Ventura, S., (2010). "Mining rare association rules from e- learning data". In Educational Data Mining 2010, pp. 123- 135

[17] Rubiano, S.M.M. and Garcia, J.A.D., (2015). "Formulation of a predictive model for academic performance based on students' academic and demographic data". In Frontiers in Education Conference (FIE), 2015. 32614 2015. IEEE (pp. 1-7).

[18] Rusli, N.M., Ibrahim, Z. and Janor, R.M., (2008). "Predicting students' academic achievement: Comparison between logistic regression, artificial neural network, and Neuro-fuzzy". In Information Technology, 2008. ITSim 2008. International Symposium Vol. 1, pp.1-6.

[19] Sanchez–Torrubia, M., Torres–Blanc, C., Krishnankutty, S., (2008), "Mamdani's Fuzzy Inference eMathTeacher: A Tutorial for Active Learning", WSEAS transactions on computers, Vol. 7, pp 76-87

[20] Sarker, Farhana, Thanassis Tiropanis, and Hugh C. Davis.(2013) "Exploring student predictive model that relies on institutional databases and open data instead of traditional questionnaires." In Proceedings of the 22nd International Conference on World Wide Web, pp. 413-418.

[21] Sofowoke Deborah (2017), "Influence of Social Media on Students Performance, Msc Computer Science Thesis, pp 22-25

Approximate Probability of Satisfactory State for Varying Number of Peers in P2P Live Streaming Networks using Poission Distribution

Dima, R. M.,[1] Aina S. K.,[2] Bashir, A. J.[1] and Yunus, A. A.[1]
Federal University Dutsin-ma,[1] and Federal University, Gashua[2]
Nigeria.

ASTRACT

Video has become an undoubtedly important medium for communications and entertainment for years. Some researchers tried to find solution of P2P challenges of video sharing such as impact of varying number of peers in the quality of video streaming using binomial distribution. In real sense, there is no Internet Protocol Television (IPTV) system with less than 30 channels, so computing a system with more than 30 channels using binomial distribution is close to impossible unless when evaluated with high performance computers. This study proposes an approximate probability expression that evaluates the satisfactory state of peers in P2P live streaming networks using Poisson distribution. This is because Poisson distribution can be used to evaluate N ≥ 30 and as far as N →∞ (infinity) approximately without the need of high performance computers and using the shortest possible time. Assumptions were stated and analytics were formed mathematically. Mathematical analytics were converted to the programmable format and was executed using java programming language. After successful compilation of the java program, results obtained for channel popularity and approximate probability of satisfactory state of varying number of peers in P2P live streaming networks considering 2000 peers in eighteen (18) different scenarios were presented. One of the eighteen scenarios where peers having high upload capacity (i.e. 1900 peers) are greater than peers having low upload capacity (i.e. 100 peers); thirty eight (38) channels achieved satisfactory state using Poisson distribution (approximate result) where forty (40) channels achieved satisfactory state using Binomial distribution (exact result).

Keywords:

1.0 INTRODUCTION

The Peer-to-Peer (P2P) paradigm depicts a fully-distributed, cooperative network design, where nodes collectively form a system without any supervision. Its advantages (although application-dependent in many cases) include robustness in failures, extensive resource-sharing, self-organization, load balancing, data persistence and anonymity (Tsoumakos and Roussopoulos, 2003).

The proliferation of portable devices such as mobile phones and tablets with multimedia capabilities and large storage capacities has projected video as a vital tool for dispensing visual information.

According to Thampi (2013), recent advances in digital technologies such as high-speed networking, media compression technologies and fast computer processing power, have made it possible to offer real-time multimedia services over the Internet. Real time multimedia, as the name implies, has timing constraints. For example, audio and video data must be played out continuously. If the data does not arrive in time, the play out process will pause, which is annoying to human ears and eyes. Real-time transport of live video or stored video is the predominant part of real-time multimedia.

There are three important means in which a streaming service may be offered over the Internet which are: caching and replication for the web based distribution for small amount of streaming media; use of a network (specifically designed for the distribution of streaming content) e.g. video On-demand Multimedia Streaming Networks are specialized in on-demand delivery of video and thirdly, live streaming systems allow clients to simultaneously watch a number of Television stations through the broadband internet connectivity available at their homes (Thampi, 2013).

1.1 Statement of the Problem

Aminu and Bashir (2017) in the Analysis of the probability of channel satisfactory state in P2P Live Streaming looked at user behavior and contribution of peers for a channel to attain satisfactory state considering a scenario where a system has equal number of high and low upload peers.

Li, *et al.,* (2011) undertook a research on measurement study in PPLive based on channel popularity, looked at channel popularity based on peer population on the channel, where they considered channel with high population as a channel with abundant resources, and less popular channel as channel with less resources.

In most papers, the authors focus mainly on the channel popularity based on number of peers, and their upload contribution. Variation of peers in bandwidth contribution is very important to see how the system will behave. In real sense, a channel cannot have equal number of high and low upload bandwidth; either high is greater than low or low is greater than high. Peer upload bandwidth plays a major role in P2P. Whenever a peer participates in a system to consume streams, its upload capacity is added to the overall system resources.

Since binomial and Poisson distributions are both theoretical and mathematical methods, binomial distribution provides an exact solution while Poisson distribution provides an approximate solution. This study derives an approximate expression using Poisson distribution to evaluate channel satisfactory state for varying number of peers.

1.2 Aim and Objectives

This study is aimed at deriving an expression that will calculate the approximate probability of satisfactory state for varying number of peers in P2P live streaming networks.

The objectives of the study are to:

(a) evaluate operational and theoretical studies on P2P live networks.

(b) derive an expression for approximate probability of satisfactory state of peers in P2P live streaming networks.

(c) analyze the effectiveness of the derived expression.

2.0 REVIEW OF LITERATURE

Research works that encompass operational and theoretical studies in P2P live streaming networks are reviewed in this section

A. Review of Operational Studies in P2P Live Streaming Networks

Push-to-Pull Peer-to-Peer live streaming was understudied by Locher *et. al.,*(2007). They presented and evaluated new neighbor selection and data allocation schemes for P2P live streaming. Their algorithms united low-latency push operations along an ordered overlay with pull operations' flexibility so as to be able to allocate data competently with minimal delay. They implemented the algorithms presented in their P2P live streaming system (Pulsar). They carried out broad simulations of their protocol and according to their emulations with up to 100,000 peers (using the real code base), their system scaled well such that network topology had a low diameter and guaranteed small round trip times due to the latency-aware preference of neighbors.

In Wu, *et al.,*(2008) a focus on multi-channel live P2P streaming was on servers where a thorough analysis was done on 400 GB and 7 months of run-time traces from UUSee (a commercial P2Pstreaming system in china). They discovered that existing capacities on streaming servers had been unable to keep up with the increasing demand forced by hundreds of channels. They proposed a novel online server capacity provisioning algorithm (Ration) that was able to dynamically predict the demand in each channel by using an array of dynamic learning techniques, and to proactively make provision of optimal server capacities across unlike channels.

Hammamia *et. al.,* (2014), proposed Hybrid Live P2P Streaming Protocol (HLPSP) in which live P2P streaming system was based on a mixture of overlay (tree and mesh topology). Broad simulations were conducted to evaluate and compare HLPSP alongside the improved edition of Coolstreaming system (DenaCast). Their simulation results showed through different scenarios that HLPSP performed better with regards to the set up delay, the end-to-end delay, the play-back delay and the data failure

The design of Contracts (a new incentive strategy that rewards contribution with quality of service by evolving the overlay topology in P2P live streaming systems) by Piatek *et. al.,* (2019) was motivated by the unique features of the P2P live streaming environment which limited the success of several widely-used incentive strategies based on balanced or joint exchange. Using measurements of tens of thousands of PPLive users, they showed that widely-used two-sided incentive approaches cannot be applied efficiently to the live streaming surroundings. Also, while using a modified PPLive client, it was shown that Contracts improves performance and strengthens contribution incentives related to existing strategies without controlling scalability.

B. Review of Theoretical Studies in P2P Live Streaming Networks

Agarwal and Rejaie (2005) presented the design and evaluation of an adaptive streaming device from multiple senders to a lone receiver in Peer-to-Peer (P2P) networks which they named P2P Adaptive Layered Streaming (PALS). They illustrated that their detailed simulation-based evaluations can effectively cope with several angles of dynamics in the system including: bandwidth variations, peer participation, and partially available content at different peers.

While exploring Large-Scale Peer-to-Peer Live Streaming Topologies, Wu, *et al.,*(2008) completed

a careful and in-depth study of the topological properties in large-scale live P2P streaming, as well as their evolutionary behavior over time. They sought to discover real-world P2P streaming topologies with respect to their graph theoretical metrics
which included the degree, clustering coefficient and reciprocity. Their findings were compared with results obtained from existing studies on topological properties of P2P file sharing applications. They observed that live P2P streaming sessions demonstrated outstanding scalability, a high point of reciprocity, a clustering occurrence in each ISP and a degree distribution that does not trail the power-law allocation.

The architecture and the design of P2P live streaming system schedulers, was researched by Christakidis *et. al.,*(2009). They analyzed P2P live streaming systems and obtained the vital parameters for their performance with respect to bandwidth consumption, set-uptime, equality and steadiness. The scheduler was developed to perform two separate but interdependent decisions (i.e. selecting the next neighbor and selecting the specific block to transmit to). The second which performed a receiver driven block selection process with a content dissemination optimization strategy achieved fast and competent dissemination of each block while significantly reducing duplicate block transmissions. They further observed high degrees of equality in upload bandwidth consumption among all nodes and a very steady and scalable system as it grew in numbers of the participating peers.

In their study titled: *"A P2P Video Delivery Network (P2P VDN)",* Nguyen, *et al.,*(2009) described a Peer-to-Peer Video Delivery Network (P2P-VDN) which provided both performance enhancement and scalability based on three architectural fundamentals - employing data distribution, data replacement and a path variety streaming protocol. They investigated the performance of their proposed architecture by comparing the non-network coding (Non-NC) and random network coding (RNC) scheme in assumption that a video publisher distributes either uncoded or coded packets to a number of peers which are from there, responsible for streaming the video to the client. Their simulation resulted in bandwidth saving of up to 60% over the long-established architecture in specific conditions.

A proficient application-layer handover scheme to support seamless mobility for P2P live streaming was proposed by Kim *et al.,* (2014). Going by their simulation experiments using QualNet network simulator, the P2P live streaming system with their proposed handover scheme improved the playback continuity drastically.

Efthymiopoulou and Efthymiopoulos (2015) conducted a study on "*QoS in Peer to Peer Live Streaming through Dynamic Bandwidth and Playback Rate Control".* They proposed a scalable and steady service organization architecture for a cloud assisted P2P live streaming system. They developed a systematic model and a hybrid control approach for a non-linear system that is able to dynamically allocate exact amount of bandwidth that is required from the cloud and at the same time dynamically adapts the playback rate to the obtainable bandwidth resources so as to guarantee the complete and on time stream sharing.

In a related study, Efthymiopoulos *et. al.,* (2015) working on *"Congestion Control for P2P Live Streaming*" proposed a P2P live streaming traffic aware congestion control protocol that is capable of managing sequential traffic heading to numerous network destinations, can competently utilize the accessible bandwidth, can correctly measure the inactive peer resources, avoids network congestion and is responsive to traditional TCP generated traffic. The proposed P2P congestion control was implemented, tested and evaluated via a sequence of real experiments powered across the BonFIRE infrastructure. Three sets of experiments were carried out in order to prove the properties of their proposed P2P congestion control. Firstly, its robustness to changes in the latency of the subordinate network path was demonstrated. Secondly, demonstration of its ability to dynamically adapt to very curt changes bandwidth of the bottleneck network point and thirdly, its coexistence to unrelated TCP traffic.

3.0 METHODOLOGY

This section provides some mathematical assumptions that are used to define the QoS parameters i.e. probability of channel popularity, derivation of approximate probability expression for achieving a good quality streaming in each channel with regards to it's kind of peer upload bandwidth contribution. The tools and techniques employed include Java programming language, MathType, MathCard and NetBeans IDE.

A. Basic Assumptions

Consider a P2P live system with multiple channels streamed to multiple peers. Let k denote the individual channel while K denote the total number of channels webcasted by a server, and N denote the total number of peers viewing each of the K-channels

Video originate from a server; denotes V_k (kbps) for the rate at which the server uploads the stream for the kth -channel, let R_k (kbps) denote the streaming rate for the kth -channel, i.e. the required bandwidth to view k-channel without pauses during playback. The video is then streamed to all participating peers.

Let U_i (kbps) denote the upload rate of peer i.e. assuming the relative popularity for k-channel to be ρ_k where:

$$\sum_{k=1}^{K} \rho_k = 1 \qquad \textit{equation 1}$$

then it means the ρ_k to the proportion of peers currently viewing k-channel. Let X_k denote a random variable representing the number of peers viewing channel k. since each of the N-peers view one of the k-channel, then:

$$\sum_{k=0}^{n} \mathrm{X}_k = N \qquad \textit{equation 2}$$

for k = 1, 2,..., K. Hence, the random variable $X_1, X_2, ... X_k$ are dependent random variables. The dependency of $X_1, X_2, ... X_k$ enables us to immediately arrive at the result that the number of viewers watching channel k possess a Binomial distribution

ρ_k is assumed to follow Zipf distribution, where

$$\rho_k = \frac{1}{\sum_{K=1}^{K} \frac{1}{K^Z}} \qquad \textit{equation 3}$$

Where $k \in K$ and Z=1

3.2 Analytics

Since N ≥ 30 takes longer time to calculate using **binomial** distribution; it is permitted to evaluate N ≥ 30 and N → ∞ using **Poisson** distribution.

In order to achieve satisfactory state using Poisson distribution, we have:

$$\Pi_{\text{approximate}} = \frac{v + uhxh + ulxl}{(xh + xl)} \geq R \qquad \textit{equation 4}$$

Table 3.1: Symbols and their definitions

Symbol	**Definition**
V	Server contribution
Xh	Number of high upload peers
Xl	Number of low upload peers
Uh	Upload capacity for high upload peers
Ul	Upload capacity for high upload peers
R	Channel streaming rate
Yh	Normalized variable for high upload peers
Yl	Normalized variable for low upload peers
Bh	Average number of high upload peers
Bl	Average number of low upload peers

From the *equation 4*, it follows that:

$$= v + uhxh + ulxl \geq (xh + xl)R$$

$$= v + uhxh + ulxl \geq xhR + xlR$$

$$= v + uhxh - xhR \geq xlR - ulxl$$

$$= uhxh - xhR \geq xlR - ulxl - v$$

$$=$$

$$(uh - R)xh \geq (R - ul)xl - v$$

$$= xh \geq \frac{(R - ul)xl - v}{(uh - R)}$$

$$= xh \geq \frac{(R - ul)xl}{(uh - R)} - \frac{v}{(uh - R)}$$

Where

$$c = \frac{(R - ul)}{(uh - R)}, V = \frac{v}{(uh - R)}$$

We therefore have:

$$xh \geq c * xl - V$$

$$\pi_{approximate} = P(xh \geq c * xl - V)$$

To calculate the ratio of average number of high and low upload peers denoted by C:

$$C = \frac{\beta h_k}{\beta l_k}, where \beta l_k \neq 0$$

Where

$$\beta h_k = nh * \rho_k$$

and

$$\beta l_k = nl * \rho_k$$

To derive $\Pi_{\text{approximate, let}}$ Yh and Yl be normalized random variables of xh and xl so that:

$$Yh = \frac{xh - \beta h_k}{\sqrt{\beta h_k}}$$

and

$$Yl = \frac{xl - \beta l_k}{\sqrt{\beta l_k}}$$

Making xh and xl subjects of relation in the above equation, it follows that, from:

$$\pi_{approximate} = P(xh \geq c * xl - V)$$

We can have:

$\Pi_{\text{approximate}}$ =

$$P((Yh\sqrt{\beta h_k} + \beta h_k) \geq c * (Yl\sqrt{\beta l_k} + \beta l_k) - V)$$

Dividing through by square root of average number of low upload peers, we have:

=

$$P(Yh\sqrt{\frac{\beta h_k}{\beta l_k}} + \frac{\beta h_k}{\sqrt{\beta l_k}} \geq cYl\sqrt{\frac{\beta l_k}{\beta l_k}} + c\frac{\beta l_k}{\sqrt{\beta l_k}} - \frac{V}{\sqrt{\beta l_k}})$$

=

$$P(Yh\sqrt{\frac{\beta h_k}{\beta l_k}} + \frac{\beta h_k}{\sqrt{\beta l_k}} \geq cYl + c\frac{\beta l_k}{\sqrt{\beta l_k}} - \frac{V}{\sqrt{\beta l_k}})$$

=

$$P(\frac{\beta h_k}{\sqrt{\beta l_k}} \geq cYl - Yh\sqrt{\frac{\beta h_k}{\beta l_k}} + c\frac{\beta l_k}{\sqrt{\beta l_k}} - \frac{V}{\sqrt{\beta l_k}})$$

Let $Y = cYl - Yh\sqrt{\frac{\beta h_k}{\beta l_k}}$ such that:

$$P((\frac{\beta h_k}{\sqrt{\beta l_k}} - c\frac{\beta l_k}{\sqrt{\beta l_k}}) + \frac{V}{\sqrt{\beta l_k}} \geq Y)$$

By rationalizing $(\frac{\beta h_k}{\sqrt{\beta l_k}} - c\frac{\beta l_k}{\sqrt{\beta l_k}})$, we have:

$$P((\frac{\beta h_k}{\beta l_k} - c)\sqrt{\beta l_k} + \frac{V}{\sqrt{\beta l_k}} \geq Y)$$

or

$$P(Y \leq (\frac{\beta h_k}{\beta l_k} - c)\sqrt{\beta l_k} + \frac{V}{\sqrt{\beta l_k}})$$

$$\pi_{approximate(k)} = \phi(\frac{(\frac{\beta h_k}{\beta l_k} - c)\sqrt{\beta l_k} + \frac{V}{\sqrt{\beta l_k}}}{\sqrt{\frac{\beta h_k}{\beta l_k} + c^2}})$$

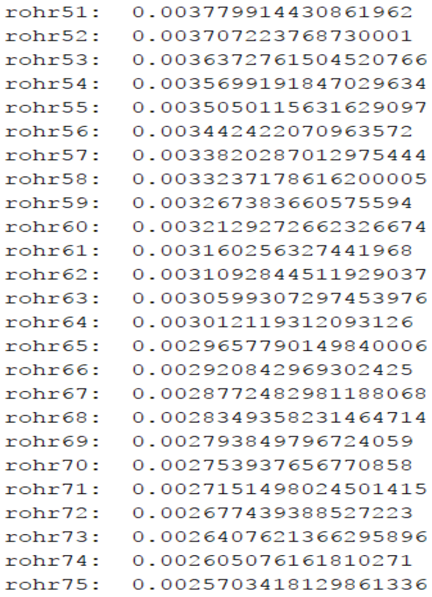
$$\pi_{approximate(k)} = \phi(\frac{(C - c)\beta l_k + \frac{V}{\sqrt{\beta l_k}}}{\sqrt{C + c^2}})$$

$$\pi_{approximate(k)} = \phi(\frac{(C - c)\beta l_k + V}{\sqrt{C + c^2} * \sqrt{\beta l_k}})$$

equation 5

4.0 Results and Discussion

Result on channel popularity i.e. which follows zipf

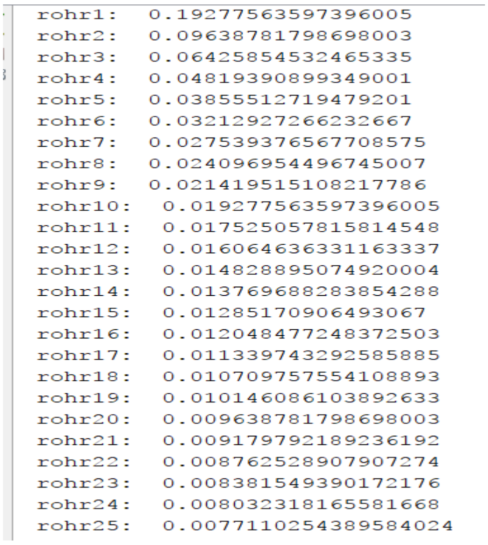


```
rohr51:  0.003779914430861962
rohr52:  0.003707223768730001
rohr53:  0.0036372761504520766
rohr54:  0.0035699191847029634
rohr55:  0.0035050115631629097
rohr56:  0.003442422070963572
rohr57:  0.0033820287012975444
rohr58:  0.0033237178616200005
rohr59:  0.003267383660575594
rohr60:  0.0032129272662326674
rohr61:  0.003160256327441968
rohr62:  0.0031092844511929037
rohr63:  0.0030599307297453976
rohr64:  0.003012119312093126
rohr65:  0.0029657790149840006
rohr66:  0.002920842969302425
rohr67:  0.0028772482981188068
rohr68:  0.0028349358231464714
rohr69:  0.002793849796724059
rohr70:  0.002753937656770858
rohr71:  0.0027151498024501415
rohr72:  0.002677439388527223
rohr73:  0.0026407621366295896
rohr74:  0.002605076161810271
rohr75:  0.0025703418129861336
```

distribution was presented. The result for the computation of approximate probability of achieving quality video streaming on each k-channel was computed.

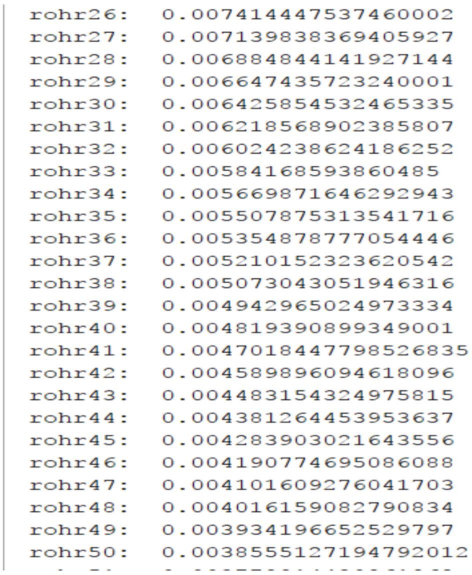
4.1 Result and Interpretation

The result of channel popularity (rohr(k)) which follows zipf distribution for k = 1, 2, …, K (100) is as follows:



```
rohr1:  0.19277563597396005
rohr2:  0.09638781798698003
rohr3:  0.06425854532465335
rohr4:  0.04819390899349001
rohr5:  0.03855512719479201
rohr6:  0.03212927266232667
rohr7:  0.027539376567708575
rohr8:  0.024096954496745007
rohr9:  0.021419515108217786
rohr10:  0.019277563597396005
rohr11:  0.017525057815814548
rohr12:  0.016064636331163337
rohr13:  0.014828895074920004
rohr14:  0.013769688283854288
rohr15:  0.01285170906493067
rohr16:  0.012048477248372503
rohr17:  0.011339743292585885
rohr18:  0.010709757554108893
rohr19:  0.010146086103892633
rohr20:  0.009638781798698003
rohr21:  0.009179792189236192
rohr22:  0.008762528907907274
rohr23:  0.008381549390172176
rohr24:  0.008032318165581668
rohr25:  0.0077110254389584024
```

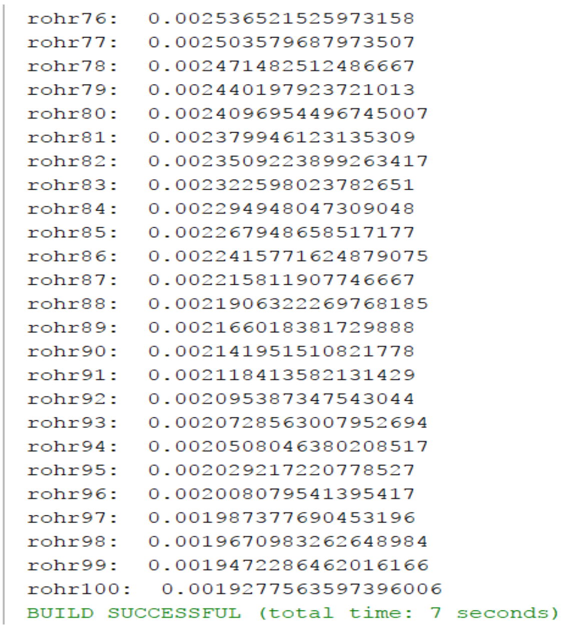
Figure 4.1: Values of rohr(k) from k = 1 to 25



```
rohr26:  0.007414447537460002
rohr27:  0.007139838369405927
rohr28:  0.006884844141927144
rohr29:  0.006647435723240001
rohr30:  0.006425854532465335
rohr31:  0.006218568902385807
rohr32:  0.006024238624186252
rohr33:  0.00584168593860485
rohr34:  0.005669871646292943
rohr35:  0.005507875313541716
rohr36:  0.005354878777054446
rohr37:  0.005210152323620542
rohr38:  0.005073043051946316
rohr39:  0.004942965024973334
rohr40:  0.004819390899349001
rohr41:  0.0047018447798526835
rohr42:  0.004589896094618096
rohr43:  0.004483154324975815
rohr44:  0.004381264453953637
rohr45:  0.004283903021643556
rohr46:  0.004190774695086088
rohr47:  0.004101609276041703
rohr48:  0.004016159082790834
rohr49:  0.003934196652529797
rohr50:  0.0038555127194792012
```

Figure 4.2: Values of rohr(k) from k = 26 to 50

Figure 4.3: Values of rohr(k) from k = 50 to 75



```
rohr76:  0.002536521525973158
rohr77:  0.002503579687973507
rohr78:  0.002471482512486667
rohr79:  0.002440197923721013
rohr80:  0.0024096954496745007
rohr81:  0.002379946123135309
rohr82:  0.0023509223899263417
rohr83:  0.002322598023782651
rohr84:  0.002294948047309048
rohr85:  0.002267948658517177
rohr86:  0.0022415771624879075
rohr87:  0.002215811907746667
rohr88:  0.0021906322269768185
rohr89:  0.002166018381729888
rohr90:  0.002141951510821778
rohr91:  0.002118413582131429
rohr92:  0.002095387347543044
rohr93:  0.0020728563007952694
rohr94:  0.0020508046380208517
rohr95:  0.002029217220778527
rohr96:  0.002008079541395417
rohr97:  0.001987377690453196
rohr98:  0.0019670983262648984
rohr99:  0.0019472286462016166
rohr100:  0.0019277563597396006
BUILD SUCCESSFUL (total time: 7 seconds)
```

Figure:4.4: Values of rohr(k) from k = 76 to 100

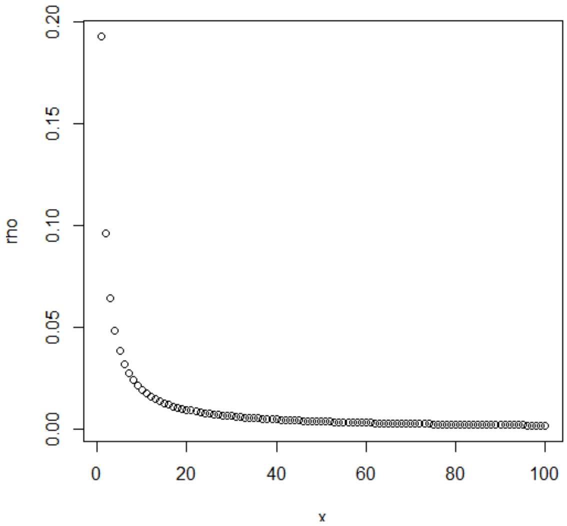


Figure 4.5: Graph of Rohr(k) from k = 1 to 100

Having N = 2000, eighteen (18) scenarios where there are varying number of high (nh) and low (nl) upload peers were evaluated and graphically illustrated in figure 4.5. In summary, the number of channels that achieved satisfactory state under varying number of peers are depicted in table 4.2.

Table 4.2: Channels that achieved satisfactory state

S/N	**number of high(nh)**	**number of low(nl)**	**Number of channels that achieved satisfactory state**
1.	1900	100	38
2.	1800	200	36
3.	1700	300	33
4.	1600	400	32
5.	1500	500	30
6.	1400	600	28
7.	1300	700	26
8.	1200	800	24
9.	1100	900	21
10.	900	1100	19
11.	800	1200	18
12.	700	1300	16
13.	600	1400	13
14.	500	1500	11
15.	400	1600	9
16.	300	1700	7
17.	200	1800	5
18.	100	1900	3

5.0 Conclusion

An expression was derived for the computation of a QoS parameter (i.e. channel popularity denoted by rohr(k)) of P2P Live Streaming and was evaluated as well. It was discovered that channels having low upload bandwidth peers hardly achieve satisfactory state approximately unlike channels possessing high upload bandwidth peers. Effect of peer population on channel popularity based on high and low upload bandwidth peers was carried out and a formula was obtained for the calculation of approximate probability of satisfactory state for varying number of peers in P2P live streaming networks using Poisson distribution. Unlike the computation carried out using binomial distribution which needs to be tested on high performance computers due to the amount of time it takes on normal PCs, the approximate solution was able to save us the stress in a negligible time compared to the exact solution (using binomial distribution) and the results obtained were efficient.

In conclusion, effect of varying number of peers on channel popularity based on high and low peer upload bandwidth capacity was carried out. A formula was derived in order to calculate channel popularity (rohr(k)) where k = 1, 2, …, K and K = 100 channels. A formula was further obtained for the calculation of approximate probability of satisfactory state for varying number of peers in P2P live streaming networks using Poisson distribution. The results gotten illustrated how faster and efficient the proposed distribution was compared to binomial distribution.

REFERENCES

Agarwal, V., & Rejaie, R. (2005). Adaptive Multi-Source Streaming in Heterogeneous Peer-to-Peer Networks. In *Multimedia Computing and Networking (MMCN), San Jose*

Aminu, A. and Bashir, A. J. (2017) Analysis of the Channel Satisfactory State in P2P Live Streaming Systems. *Bayero journal of* *pure and applied sciences*, 10(1): 20-27 ISSN 2006 – 699

Christakidis, A., Efthymiopoulos, N., Denazis, S., & Koufopavlou, O. (2009). On the architecture and the design of P2P live streaming system schedulers. *International conference on ultra-modern telecommunications ICUMT 2009.*

Efthymiopoulos, N., Christakidis, A., Efthymiopoulou, M., Corazza, L., Denazis, S. & Koufopavlou, O. (2015). Congestion Control for P2p Live Streaming. *International Journal of Peer to Peer Networks (IJP2P). 6*(2), 1–21.

Efthymiopoulou, M., & Efthymiopoulos, N. (2015). QoS in Peer to Peer Live Streaming through Dynamic Bandwidth and Playback Rate Control. In *the 4th International Conference on Communications, Computation, Networks and Technologies, 2015*

Hammamia, C., Jemili, I., Gazdar, A., & Belghith, A. (2014). Hybrid Live P2P Streaming Protocol. *Procedia - Procedia Computer Science, 32,* 158–165. https://doi.org/10.1016/j.procs.2014.05.410

Kim, E., Kim, S., & Lee, C. (2014). Supporting Seamless Mobility for P2P Live Streaming. https://doi.org/10.1155/2014/134391

Locher, T., Meier, R., Schmidt, S., & Wattenhofer, R. (2007). Push-to-Pull Peer-to-Peer Live Streaming. In *DISC 2007*. 388–402.

Nguyen, K., Nguyen, T., & Kovchegov, Y. (2009). A P2P VIDEO DELIVERY NETWORK (P2P-VDN). In *International Conference on Computer and Communication Networks, 2009, pp.1-7*

Piatek, M., Krishnamurthy, A., Ventaraman, A., Yang, R., Zhang, D., & Jaffe, A. (2010). Contracts : Practical Contribution Incentives for P2P Live Streaming. In *USENIX NSDI*

Thampi, S. M. (2013). A Review on P2P Video Streaming. *arXiV:1304.1235*

Tsoumakos, D. and Roussopoulos, N. (2003). A Comparison of Peer-to-Peer Search Methods. In *Proceedings of the WebDB'03, san Diego. pp. 61-66*

Wu, C., & Li, B. & Zhao, S. (2008). Multi-channel Live P2PStreaming : Refocusing on Servers. In *Proceedings of IEEE INFOCOM'08*

Wu, C., Li, B., & Zhao, S. (2008). Exploring Large-Scale Peer-to-Peer Live Streaming Topologies. *4*(3). https://doi.org/10.1145/1386109.1386112

Adapting Schools' Curriculum for the Fourth Industrial Revolution: A Nigerian Perspective.

Ijeoma Okoronkwo
ComputerProfessional Registrational Council of Nigeria.
ijayronk@yahoo.com

Ogochukwu Fidelia Nwosu
Department of Computer and RoboticsEducation, University of Nigeria, Nsukka.
fidelia.nwosu@unn.edu.ng

Isaiah Odinakachi Nwogbe
Department of Computer Science Spiritan University Nneochi.
odinakachi.nwogbe@spiritanuniversity.edu.ng

Chioma Chigozie-Okwum
Department of computer science, Spiritan University Nneochi.
chioma.chigozieokwum@spiritanuniversity.edu.ng

ABSTRACT— The fourth industrial revolution powers a digital economy that aims to achieve a fusion of all economies into one by interconnection and sharing. If Nigeria intends to key into the fourth industrial revolution there is need to make conscientious efforts geared at reducing barriers that hinder the growth and development of the fourth industrial revolution. Policy implementation issues, poor power supply, cyber security risks as well as widening skills gap were factors identified as challenges to adaptation of the fourth industrial revolution in Nigeria. The study identified that the fourth industrial revolution is powered by a set of novel and disruptive technologies like Internet of things, Robotics, Artificial Intelligence, and Virtual Reality. To this end there is need to grow the knowledge base and capacity of both professionals and citizens. This can only be achieved by adapting our current school curriculum to align with the requirements of the fourth industrial revolution if we aim to enjoy its attaining benefits.

Keywords: Industrial Revolution, Digital Economy, Curriculum, Skills, Technologies.

I. INTRODUCTION

There is a global trend and shift towards digitalization of world economies. Countries are gradually embracing digital economies in the wake of fourth industrial revolution. The fourth industrial revolution is disruptive in nature and invasive in operation, it has brought with it a change in the way society live, work and even interacts. It has changed industrial processes as well as brought a significant change in ways businesses are transacted. Technological revolutions have happened in human history, starting from the first to the current fourth revolution. A technological revolution is a period in which one or more technologies is replaced by another technology in a short amount of time. It is an era of accelerated technological progress characterized by new innovations whose rapid application and diffusion causes an abrupt change in society.

According to the World Economic forum, previous industrial revolutions liberated humankind fromanimal power, made mass production possible and brought digital capabilities to billions of people. This Fourth Industrial Revolution is, however, fundamentally different. It is characterized by a range of new technologies that are fusing the physical, digital and biological worlds, impacting all disciplines, economies and industries, and even challenging ideas about what it means to be human. The resulting shifts and disruptions mean that we live in a time of great promise and great peril. The world has the potential to connect billions more people to digital networks, dramatically improve the efficiency of organizations and even manage assets in ways that can help regenerate the natural environment, potentially undoing the damage of previous industrial revolutions [1].

The fourth industrial revolution brings with it a new era of economic disruption with uncertain socio-economic consequences for Africa and Nigeria in particular. According to [2], Africa has been left behind during the past industrial revolutions, and it does not appear that Africa has yet claimed the 21st century. [3] Suggests that Africa still lags behind in several indicators essential for a successful digital revolution. If African countries especially Nigeria hopes to benefit from the numerous advantages of the 4th industrial revolution there is need to mount strategies, polices and roadmaps towards achieving this goal.

This paper attempts to buttress how education, curriculum adaptation in particular contributes in preparing Nigeria and especially its workforce to adapt to the changing roles and approaches that has come with the advent of the industry 4.0.

II. LITERATURE REVIEW/RELATED WORKS

A. The fourth Industrial revolution; a conceptual overview

According to [4] the world stands on the brink of a technological revolution that will fundamentally alter the way society live, work, and relate to one another. In its scale, scope, and complexity, the transformation will be unlike anything humankind has experienced before. It is yet unknown just how it will unfold, but one thing is clear: the response to it must be integrated and comprehensive, involving all stakeholders of the global polity, from the public and private sectors to academia and civil society.

The fourth industrial revolution, a term coined by Klaus Schwab, founder and executive chairman of the World Economic Forum, describes a world where individuals move between digital domains and offline reality with the use of connected technology to enable and manage their lives [5]. The Fourth Industrial Revolution is characterized by a fusion of technologies that blurs the lines between the physical, digital, and biological spheres [6]. The Fourth Industrial Revolution dawned through the use of cyber-physical systems (CPSs), the Internet of Things (IOT), and services [7]. The fourth industrial revolution include smart factories, cyber-physical systems, self-organization, new systems in distribution and procurement, new systems in the development of products and services, adaptation to human needs, and corporate social responsibility [8].

The Fourth Industrial Revolution can be defined as the revolutionary change that occurs when IT proliferates in all industries, that is, the primary, secondary, and tertiary industries. In other words, it is a result of the horizontal expansion of IT. Therefore, the Fourth Industrial Revolution features the creative connection between technology and the market in all industries based on IT, that is, the creative and open combination of technology and the market through open innovation, or growth based on the open business model [9]. The catchphrase of the day is the Fourth Industrial Revolution or Industry 4.0. Industrialists have started talking about it since 1999, predicting human civilization that is geared by the Internet in the near future. True enough, it did not take long for the revolution to go full-fledged with many breakthroughs in technologies such as artificial intelligence, robotics, the internet of things, autonomous vehicles, 3D printing, the block chain, biotechnology and so on[10]. Now a Fourth Industrial Revolution is building on the Third, the digital revolution that has been occurring since the middle of the last century. It is characterized by a fusion of technologies that is blurring the lines between the physical, digital, and biological spheres. The fourth revolution is unfolding before our eyes. Its genesis is situated at the dawn of the third millennium with the emergence of the Internet. This is the fourth industrial revolution rooted in a new technological phenomenon digitalization rather than in the emergence of a new type of energy. This digitalization enables us to build a new virtual world from which we can steer the physical world. The fourth industrial revolution is the current and developing environment in which disruptive technologies and trends such as the Internet of Things (IoT), robotics, virtual reality (VR) and artificial intelligence (AI) are changing the way society live and work [11].

2.1 Evolution of Industrial revolutions

The First Industrial Revolution took place in the 18th and 19th centuries, it involved a change from mostly agrarian societies to greater industrialization as a consequence of the steam engine and other technological developments. The first industrial revolution used water and steam power to mechanize production. The Second was driven by electricity and involved expansion of industries and mass production as well as technological advances, it used electric power to create mass production. The Third industrial revolution sometimes called the digital revolution, involved the development of computers and IT (information technology) since the middle of the 20th century. It used electronics and information technology to automate production. The Fourth Industrial Revolution is building on the Third, the digital revolution that has been occurring since the middle of the last century. It is characterized by a fusion of technologies that is blurring the lines between the physical, digital, and biological spheres.

B. Agents of Change in the Fourth Industrial revolution

Key players fuelling the fourth industrial revolution include the following disruptive technologies and trends like the Internet of Things, Robotics, Virtual Reality and Artificial Intelligence. Disruptive technology refers to any enhanced or completely new technology that replaces and disrupts an existing technology, rendering it obsolete. It is designed to succeed similar technology that is already in use. For example, cloud computing serves as a disruptive technology for in-house servers and software solutions [12].

1. Internet of Things is the concept of basically connecting any device with an on and off switch to the Internet (and/or to each other) [13]. This includes everything from cell-phones, coffee makers, washing machines, headphones, lamps, wearable devices and almost anything else you can think of. This also applies to components of machines, for example a jet engine of an airplane or the drill of an oil rig.

2. Robotics is a branch of engineering that involves the conception, design, manufacture, and operation of robots. This field overlaps with electronics, computer science, artificial intelligence, mechatronics, nanotechnology and bioengineering [14].

3. Virtual Reality (VR), is the use of computer technology to create a simulated environment.

Unlike traditional user interfaces, VR places the user inside an experience [15]. Instead of viewing a screen in front of them, users are immersed and able to interact with 3D worlds. By simulating as many senses as possible, such as vision, hearing, touch, even smell, the computer is transformed into a gatekeeper to this artificial world. The only limits to near-real VR experiences are the availability of content and cheap computing power.

Artificial intelligence is the branch of computer science concerned with making computers behave like humans [16]. The term was coined in 1956 by John McCarthy at the Massachusetts Institute of Technology.

C. Benefits of Industry 4.0 in Nigeria.

The 4th industrial revolution will fuel the next wave of sporadic development in Nigeria, the impact of this inevitable tide is getting clearer as a rising sun. The mode of trade is being changed from the conventional mode of buying and selling within a brick-and-mortar store to ecommerce. Start-ups like Jumia, Konga, Mall for Africa, Olx, Jiji, and others have revolutionized trade. In the process, new jobs have been created and many old ones have disappeared as a result of automated processes. Furthermore, Automated manufacturing which requires very few humans as seen in the Dangote-Sinotruk initiative a joint venture that aims to locally produce 10, 000 commercial vehicles annually with very few employees are all indication that Nigeria is not entirely idling away from the trend but the efforts are minimal and not commensurate with our counterparts the world over. With industry 4.0, every aspect of the productionchain, like the turning over machine, filling machine, inspection line, debugging workshops to finished products is to be handled by high-tech devices. If Nigeria keys into the industry 4.0 the better chances we stand to become and industrialized economy as opposed to our developing economy status.

D. Challenges Facing Industry 4.0 in Nigeria

Notwithstanding the numerous benefits accruing to implementation of the fourth industrial revolution, there seems to be increasing challenges hindering full implementation of the digital economy in Nigeria. These challenges include but are not limited to:

Policy Implementation issues:

Policies may encourage application development providing innovative technological solutions with relevant local content. Policies will deal with ICT diffusion, and ICT literacy, and awareness of the benefits of ICT, the creation of new economic and social opportunities for poverty eradication, job creation and empowerment. Policies should address how public-private partnership (PPP) initiatives can be effective. It is particularly important for PPP initiatives to provide, support and use the information infrastructure, to encourage the deployment and use of ICTs within the economy and society. The right environment for the private sector should promote fair competition, opening up new markets, global opportunities and the delivery of high quality products and services. ICT policies in Nigeria need to address high cost of ICT infrastructure, and also address expensive and poor quality internet access provided by internet service providers in the nation. ICTs cannot develop in Nigeria if subscribers cannot afford infrastructure, devices and internet access. The more people who have access to ICTs the more the likelihood of Nigeria joining leagues of developed nations in harnessing the massive and unique benefits of ICTs

Poor Power supply

The poor electricity supply in Nigeria is proving a major impediment to the operation and growth of information and communication technologies in Nigeria. The Nigerian Electricity Supply Industry Statistics (NESI) revealed that the country lost an estimated N1.202 billion ($283,081) on September 5, 2017 due to several constraints inhibiting the power sector. Access to power expands the number and variety of business and job opportunities available. The fourth industrial revolution is a digital economy where every component is powered by electricity to produce its seamless benefits. A lack of a consistent access to reliable power costs businesses and the economy as a whole. Even with access to energy, unreliable power makes operating a business even more challenging than usual.

Electricity provides business owners with access to online information and resources, while power provides business owners with information that is critical to operating their business successfully, whether that information is about local or national markets, new economic policies or tax regulations. This allows small business owners in rural areas to engage with the wider business community and learn best practices from other individuals working in the same industry. It is clear that by investing in energy infrastructure, governments can help both small and large firms simultaneously, while also helping to alleviate poverty. Nigerians seem to have adapted to the poor power and epileptic power supply in the country as owners of business have resorted to alternative forms of getting electricity to power their businesses. If Nigeria seeks to rapidly develop and deploy ICT usage in the wake of the digital age and the fourth industrial revolution there is need to fix the deficiencies in the power sector.

Cyber Security Issues

The federal government has estimated the annual cost of cybercrime in Nigeria to be about 0.08% of the country’s Gross Domestic Products (GDP), which represents about N127 billion . 2019 saw a rise in the number of sophisticated phishing attacks; these occurred on multiple Nigerian financial institutions and utility companies. We also efforts by the regulatory bodies in setting up committees responsible for implementing and monitoring the cybercrime act. Cyberattacks would continue to grow and only the informed and prepared would survive with minimal losses. In 2017, cyber threats and countermeasures took the following dimensions, namely, rise and fall of cyber ponzi schemes, increase in ransomwares, increase in cloud based attacks, Rise in IoT (Internet of Things) compromises, distributed denial of service attacks, phishing attacks and identity theft. As technology and the internet continue to evolve, the world is rapidly becoming a global village, with almost everything running on the cyber space affecting most aspects of human lives, enabling growth, dismantling barriers to commerce and allowing people across the globe to communicate, collaborate and exchange ideas. But hackers are becoming more sophisticated by the day. This places the burden of securing IT infrastructure and users on us IT professionals hence the need to be vigilant and prompt in responding to incidents of cyber-attacks as well as proactive in ensuring that cyber-attacks are mitigated against in all its entirety. Nigerian law enforcement agencies as it were lack the cutting edge expertise to match that of hackers and cybercriminals who are sophisticated and keep upping their games hence always staying ahead of security personnel. There is an urgent need for massive investment in training of cyber security experts (ethical hackers, digital forensic experts, etc), if Nigeria intends to combat this security issues on the cyber space.

Widening Skills Gap

There exists a rapidly widening skills gap between the academia and industry practitioners in the ICT sector in Nigeria. Graduates are found inadequate to fit into these challenging industry roles. These challenge results to the importation of manpower and infrastructure. The current computing curriculum in Nigeria is grossly inadequate. A computing profession is one characterized with rapid evolution. Nothing prepares the graduate from Nigeria tertiary institution for they are to be faced from the industry after graduation. Common practice as well is that companies now have to retrain employees to suit into industry roles after employment, hence incurring higher overhead and running cost.

IV. Global Efforts towards aligning towards the fourth industrial revolution

New schools and new views on teaching are springing up around the world to help prepare the next generation for a rapidly changing employment landscape. This overhaul of teaching and education methods is much needed and not only because of the breath taking pace of change being ushered in by digital technologies, AI and data but because the methods have outlived their use and essence. Critical efforts and strategies have been adopted by several countries in preparation for the fourth industrial revolution. This is evident in the report as contained in a report which analysed New Models of Education for the Fourth Industrial Revolution. The report looked at how education needs to change with the change in industrial processes and procedures. The report identified eight "critical characteristics in learning content and experiences" and highlighted 5 schools, systems and initiatives around the world that are leading the way, they include [17]:

1. Child’s play in China

Anji Play was established in the Zhejiang Province of China in 2002. It follows an early childhood curriculum that fosters learning entirely through child-led play. Its core belief is that any setting can become a learning environment, with a minimum of 90 minutes every day set aside for outdoor play, using equipment such as ladders, buckets and climbing cubes. Key to the model’s success is that it makes use of low- to no-cost items, ensuring it is accessible to low-income families. Initially, 14,000 children in Zhejiang were enrolled. It has since been spread to over 100 public schools in more than 34 provinces in China. There are now Anji Play pilots in the US, Europe and Africa, too.

2. Finland’s budding entrepreneurs

Finland routinely ranks high for the quality of its education system, which is regarded as one of the best in the world. Founded in 1958, South Tapiola High School is one of the best schools in the country. As well as following the Finnish national curriculum, it adds a special focus on teaching collaboration through entrepreneurship, active citizenship and social awareness with real-world applications. The school’s Young Entrepreneurship Programme gives students the opportunity to work in groups to create a business of their own, then enter their ideas in national competitions.

3. Growing green leaders in Indonesia

Shaping the green leaders of the future is central to the Green School, which opened in Bali in 2008. Its 800-strong student body is comprised of 3- to 18-year-olds. The school now has plans to expand into New Zealand, South Africa and Mexico by 2021. Maintaining a sustainable school environment is one of the school’s main activities, and in the 2017-2018 school year students produced over 150 kilogrammes of food every month. In 2018, it joined forces with Sunseap, Singapore’s largest clean energy provider, to help the school with its goal of becoming completely off the grid.

4. Refugees in Kenya get connected

In 2015, Belgian teacher, activist and entrepreneur Koen Timmers set up a crowdfunding campaign after speaking to an outreach worker in the Kakuma refugee camp in Kenya. He sent more than 20 laptops (including his own), solar panels and internet equipment to the camp to connect volunteer teachers with refugee children. There are now 350 teachers across six continents offering remotely taught courses in English, mathematics and science to children in the camp. The Kakuma model is now expanding through a network of Innovation Lab Schools to Tanzania, Uganda, Nigeria, Morocco, Argentina, South Africa, Brazil and Arctic Canada.

5. Tech immersion in Viet Nam

TEKY is the first STEAM (science, technology, engineering, art and mathematics) academy in Viet Nam for children ages 6 to 18 years old. Founded in 2017, it has since established 16 centres in five cities across the country. Through partnerships with 30 schools across Viet Nam, the academy is able to deliver nine- to 18-month-long technology courses, as well as a coding camp for the holiday periods. TEKY teaches modules on programming, robotics, web design, multimedia communications and animation, with students spending about 80% of their learning time interacting with technology.

V. CURRICULUM ADAPTATIONS AS A KEY DETERMINANT TOWARDS ACHIEVING A DIGITAL ECONOMY; THE NIGERIAN EXPERIENCE.

If Nigeria would benefit adequately from the prospects and benefits of a digital economy there is need to raise skills personnel with higher mental capacity. That is why we must invest aggressively in 21st century education. If Nigeria wants to benefit from the dividends of the fourth industrial revolution we must teach our children to code and write programs from a very early age so that they can develop technology solutions for the challenges that are specific to our environment and beyond. This means emphasis on Science, Technology, Engineering and Mathematics, First, if technology is the core resource of the information age, programming and coding literacy becomes the currency of trade in that world. A computer is not a toy; every child in school today should have access to one. The Internet is not just for Facebook, WhatsApp, Instagram and Snapchat; it is the platform for exponential knowledge and information that will help our young people to develop real world problem-solving skills.

The system of education in Nigeria is inadequate to prepare us for the technological tsunami associated with the 4th industrial revolution. Apart from the emphasis on a theoretical mode of learning divorced from application to real-life problems, it doesn't belong in this age. Hence, a remodelling of the base of our educational framework is in order. Government has to be at the forefront of the process of building the next generation's tech-savviness.

Regulatory agencies such as National University Commission (NUC), National Board for Technical Education (NBTE), National Council for Colleges of Education (NCCE), are making efforts to ensure compliance and strict adherence but more need to done. Revision of computing curriculum should be done in line with global trends and best practices. Nigeria computing graduates need to be taught the right and most current technologies to position for effective global competition with their peers in other climes. The need for the academia and faculty members to be mandated to annually engage in skill upgrade is also a way regulatory bodies can assure that lecturers have the basic skills needed to impact on their students the required skills to fit into industry. Groups like the Academia in Information Technology and interest group of the Nigeria Computer Society are making conscientious efforts geared at ensuring its members in addition to acquiring required degrees, are also exposed to cutting edge technologies. These will ensure closure of the skills gap between academia, quality of graduates and expertise required by the industry. Cisco with its Networking Academics Safari, an IBM skills academy are some public private collaborative efforts, geared at covering IT skills gap in Nigeria. Training of the academia and faculty members in research and development through local and international scholarship programmes, is one effort made by the government at growing content. However, there is need for more collaboration and investment in human capital development and enhancement of skills and technical knowhow of faculty members. Monitoring of educational programmes should not be scheduled, frequent and unannounced visitation to schools for compliance evaluation and adherence level by regulating agencies in the educational system will reveal a lot of decadence and window dressing in IT education in Nigeria.

The government is making conscientious efforts but more need to be done in areas of overhaul of computing curriculum from the lowest levels to the tertiary and graduate levels. The initiative of the federal government to establish the ministry of digital economy lays credence to the fact that the government is on top of their game, but more need to done. The "Data science Nigeria" group recently launched a new text that aims at improving teaching and learning of artificial intelligence from the cradle. This is a very laudable stance. Worth mentioning is the efforts of the Lagos State Government growing technology education of its citizens. Through its CodeLagos programme, the state made effort at training 1 million Lagosians with coding skills by 2019. Through Public-Private

Partnership, it was able to carry along many private sector entities and so points to a fundamental lesson for survival in the tech age, which revolves around partnerships.

V. CONCLUSION

An educational framework that produces people ill-fitted for employment or who cannot contribute to the growth of a digital economy is just a clog in the wheels of progress of the 4th industrial revolution. Hence, a remodelling of the base of our educational framework is in order. Another point of action is the creation of an environment conducive for new and old businesses to flourish. A good way to prepare for the 4th industrial revolution tsunami is to invest massively in science, technology, engineering and mathematics (STEM) education. Beyond this, government has to be at the forefront of the process of building the next generation's tech-savviness. For Nigeria to partake fully in the next industrial revolution, which is the fourth, local content development, must be a priority.

References

[1] World Economic Forum. The fourth industrial revolution.Available@https://www.weforum.org/aout/the-fourth-industrial-revolution-by-laus-schwab.

[2] Landry Signé. Africa's Role in the Fourth Industrial Revolution: Riding the World's Biggest Wave of Disruptive Innovation. Forthcoming. See the summary online: landrysigne.com.

[3] World Bank, Can Africa Claim the 21st Century? (Washington, D.C.: World Bank, 2000).

[4] Klaus Schwab. (2016). The Fourth Industrial Revolution: what it means, how to respond. Available @ https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/.

[5]. Xu, Min & David, Jeanne & Kim, Suk. (2018). The Fourth Industrial Revolution: Opportunities and Challenges. International Journal of Financial Research. 9. 90. 10.5430/ijfr.v9n2p90.

[6]. Schwab, K. The Fourth Industrial Revolution; Crown Business: New York, NY, USA, 2017.

[7] Jazdi, N. Cyber physical systems in the context of Industry 4.0. In Proceedings of the 2014 IEEE International Conference Automation, Quality and Testing, Robotics (AQTR), Cluj-Napoca, Romania, 22–24 May 2014; pp. 1–4.

[8]. Lasi, H.; Fettke, P.; Kemper, H.-G.; Feld, T.; Hoffmann, M. Industry 4.0. Bus. Inf. Syst. Eng. 2014, 6, 239–242. [CrossRef]

[9]. Yun, J.J. Business Model Design Compass: Open Innovation Funnel to Schumpeterian New Combination Business Model Developing Circle; Springer: Cham, Switzerland, 2017.

[10]. Mark Thonhill. The 4th industrial revolution and software systems. Available @ www.statti.co.uk. 2017

[11]. Techtarget (n.d). Fourth Industrial Revolution. Available @ www.techtarget.com.

[12] Technopedia. Disruptive Technology. Available @ www.technopedia.com. 2018.

[13] Jacob Morgan. What is the fourth industrial revolution? Forbes, available @ www.forbes.com. 2016.

[14] Techtarget. Robotics. Available @ www.whatis.techtarget.com. 2018.

[15] Brian Jackson, (2015). What is virtual reality? AR blog. Available @ www.marxentlabs.com.

[16] Vangie Beal (n.d), Artificial Intelligence. Available @ www.webopedia.com.

[17] Sean Fleming. 2020. How can we prepare students for the Fourth Industrial Revolution? 5 lessons from innovative schools around the world. Available @ https://www.weforum.org/agenda/2020/02/schools-of-the-future-report-2020-education-changing-world/

An Evolutionary Computing Model for Cancerous Genome Detection

Ikenna L Ezea[1], Nneka Ernestina Richard-Nnabu[2]
[1]Mathematics/Computer Science/Statistics/Informatics
Alex Ekwueme Federal University, Ndufu-Alike
ezeaikenna@yahoo.com
[2]Mathematics/Computer Science/Statistics/Informatics
Alex Ekwueme Federal University, Ndufu-Alike
ernestina.richard-nnabu@funai.edu.ng

Abstract - Scientists have been able to develop an electronic system that can sample over 3.2 billion letters in an individual's genome and make reasonable observations that can improve a person's medical condition, but this area has not fully been explored as several genetic related diseases like cancer is still predominant and can only be treated through surgery. This paper used experimental data from Genome Informatics Research Lab to build a model that will compare the genome of cancerous individuals in a family with non-cancerous family members so as to detect the portion of the genome that is responsible for the cancerous growth. The result of the study shows that after 33 generations the genetic algorithm was able to detect the portion of the DNA that is responsible for the disease. The main significance of the research is that it will enhance cancer treatment based on pharmacogenomics and gene transplant.

Keywords: Evolutionary Computing, Genetic Algorithm, Genome Sequencing, Genetic Disorder, Deoxyribonucleic Acid (DNA)

I. INTRODUCTION

Every living organism is composed of cells which generate the tissues and the organs needed for human development and growth. The cells are the fundamental building blocks of life. The cell determines the physical appearance of human beings because of the presence of Deoxyribonucleic Acid (DNA) in the nucleus. During cell reproduction (cell division) the nucleotides or molecular base pairs: adenine (A), cytosine (C), guanine (G), and thymine (T) that make up the human DNA undergo some mutations which alters their underlying sequence, for example instead of having a base pair of AT and GC which is the evolutionary transcript of life, mutation could interchange the pairs making it difficult to produce the ideal protein needed for normal body maintenance and growth [1]. This mutation error if not corrected may result in cancerous growth in the body. Over the years scientists have made significant progress on performing whole genome sequencing to identify some underlying deformities in the genome but this field still remains a very fertile research area as the 3.2 billion molecular base pairs that make up the human DNA have not fully been understood. Most of the research has been based on identifying the exact gene that is responsible for erroneous protein formation, but the main question is what sequence of nucleotides needs to be present in two reproductive mates that might lead to mutations that may be very difficult to handle internally?

In other to address this question this research aims at building a genetic algorithm model that will perform genome sequencing on family members that has cancerous growth and compare them with other family members with no cancerous growth and check if a pattern can be found that will help in identifying the genome strings that are responsible for the disparity in the mutation and protein formation. To achieve this, this paper will use datasets from [2] for testing of the model. The main significance of this research is that it will enhance cancer treatment based on pharmacogenomics and gene transplant.

II. LITERATURE REVIEW

Research in genome sequencing had been inhibited by technology but with the advent of modern computers with high processing speed and Artificial Intelligence (AI) capabilities the end is closer than it seems [3]. Several researchers have approached genome sequencing using different Artificial Intelligence tools of which Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNN) have demonstrated to be the best in performance when dealing with genome sequencing [4], however genetic algorithm has been the most preferred due to its generalization ability.

In the study conducted by [5], Genetic Algorithm was used for early detection of pancreatic cancer by finding minimal set of genetic biomarkers that can be used for establishing diagnosis. Their result shows that pancreatic cancer can be predicted with high accuracy using a minimum group of genes.

Gayathri [6] used a combination of Genetic Algorithm and Neural Network for early detection and prediction of cancer using a real-world cancer datasets. The paper showed that cancer can be predicted in less time without medical or laboratory tests.

Fatima and Mohammad [7] proposed a genetic algorithm model for the prediction of oral cancer based on probabilistic cost function. Their approach is based on the abnormality of the x-ray and bronchitis of the individual. Their result showed an improved reduction in time needed to carryout normal medical prognoses.

Diaz, Pinon, and Solano [8] used genetic algorithm for the selection of features used by support vector machine and artificial neural network for lung cancer status classification. Their result showed a prediction accuracy of 95.87% for support vector machine and 93.66% for artificial neural network.

The literatures have shown the efforts and progress made by researchers on the application of genetic algorithm on genome sequence for cancer detection, however some inherited DNA mutation can actually predispose an individual to high risk of cancer. Certain types of cancer are predominant in some family, so this work uses genetic algorithm approach to detect the exact genes in a family line that undergoes such cancer causing mutations.

III. METHODS

This research is an experimental research that was conducted using genetic algorithm and python 3 programming language. It uses a reference genome dataset and a simulation of some family genome to carry out genome sequencing on a set of family members with some cancerous genes so that the mutation pattern can be traced down to the parent that contributed the genome to the offspring. The result of the family genome sequencing is cross matched with the reference genome to detect some variation in the sequenced genomes. The framework of the genome datasets used for this research is as shown in figure. 1.

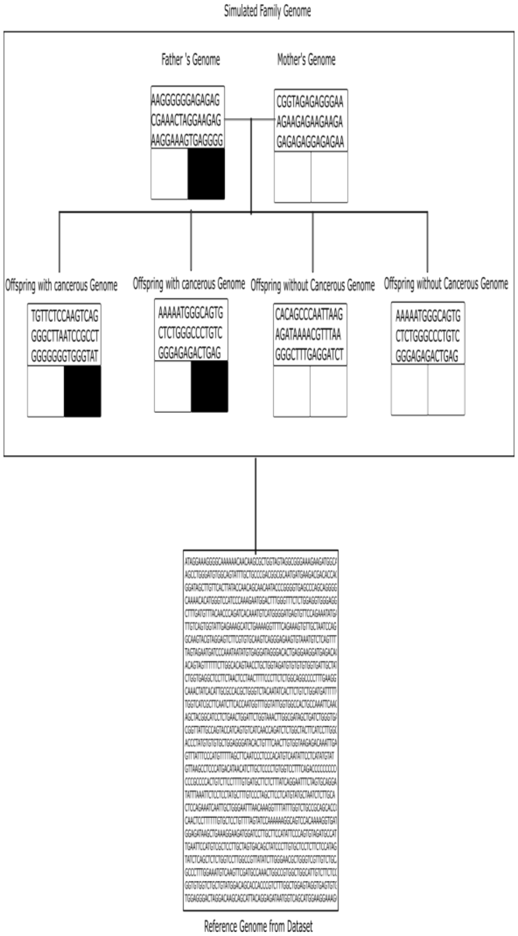


Figure 1: Architectural Framework of the index and reference genome

Figure 1 is composed of the simulated family genome and the reference genome which is from the dataset in [2]. The simulated family genome is a hierarchical structure consisting of the parents at the upper layer and the offspring at the bottom layer. The father's genome has a black and white nucleotide sequence. The black sequence shows the presence of cancerous genes whereas the white color sequence shows the healthy genes. The mother on the other hand has two healthy genes. Thus, because of the father's disposition to cancer two of his offspring have inherited the bad gene from the father and the good gene from the mother; as a result they are equally predisposed to cancer. This is in contrast to the healthy children who were lucky to inherit the good gene from both parents. On the other hand the second path of the framework is the reference genome that is used as a guide for determining errors in the family genome sequence. The steps used for determining the genome pattern of the family offspring is as shown below;

1. Pick the offspring with cancer and sequence the genome

2. Match the genome with the reference genome to determine errors in the genome
3. Compare the erroneous part with the parents genome to know who contributed the gene to the children
4. Sequence other family members genome
5. Compare the genome with other family members
6. Repeat step 1 to 5 for offspring without cancer
7. Display result
8. Stop

A. Genetic Algorithm

Genetic Algorithm is one of the optimization techniques used in Machine Learning [9]. It is widely used in scientific, engineering, stock, medical application, etc. Genetic Algorithm is used in this research to determine the sequence of mutation pattern that is capable of causing cancer in an individual. Figure 3 shows the flow of activities in the Genetic Algorithm.

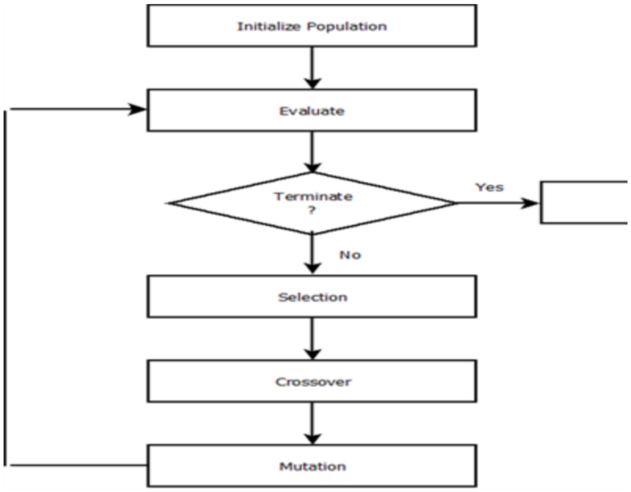


Figure 2: A general genetic algorithm process

Step 1: Genetic algorithm starts the process of reproduction by first getting a random number of individual from the population sample. This is to ensure that every aspect of the population is taking into consideration. The population consists of the reference and simulated family chromosomes. Examples are:

ATTC CGGA TTTG, CTGG GCTA AAAG, TCCC TTAG CCAA, GAAA AATG GGCC

Step 2: in this step the fitness value of every individual in the population is checked to know the number of individuals that will participate in the reproduction process. At this point the fitness of each individual is taken and evaluated against the fitness of the overall population to find the actual fitness of an individual. To calculate the fitness the source string and the target string are compared and the numbers of character they have in common are computed against every individual of the population.

Step 3: At this step the algorithm will evaluate the candidate solution against its target solution. If it has been able to get the required solution then it will terminate, else it will proceed to the next step.

Step 4: If the termination condition is not met then the algorithm will go through a selection phase of which individuals are selected based on their fitness score. An individual with the highest fitness score is usually selected. To select the fitness score the algorithm uses roulette wheel selection method, see equation 1.

$$P_i = \frac{f_i}{\sum_{j=1}^{n} f_j} \quad (1)$$

i. fi = fitness for string *i* in population,
ii. pi = probability of string *i* being selected.
iii. n = no of individuals in the population,
iv. n*pi = the expected count

Table 1 shows the distribution of the population according to their fitness scores. Based on the fitness count the individual with the highest expected count will be selected for reproduction while the one with the lowest fitness score may disappear or become the best candidate for mutation, see table 2 and figure 3 and 4.

TABLE 1: SAMPLE DISTRIBUTION OF THE POPULATION ACCORDING TO THEIR FITNESS SCORES

String No	Initial Population (Source String)	Target String	Fitness	Pi	Expected count
1	ATTC CGGA TTTG	ATTC AAAA GGTG	7	0.412	1.648
2	CTGG GCTA AAAG	ATTC AAAA GGTG	3	0.176	0.704
3	TCCC TTAG CCAA	ATTC AAAA GGTG	2	0.118	0.472
4	GAAA AATG GGCC	ATTC AAAA GGTG	4	0.235	0.940
Sum			17	1.000	3.996
Average			4.25	0.250	0.999
Max	3		7	0.235	1.648

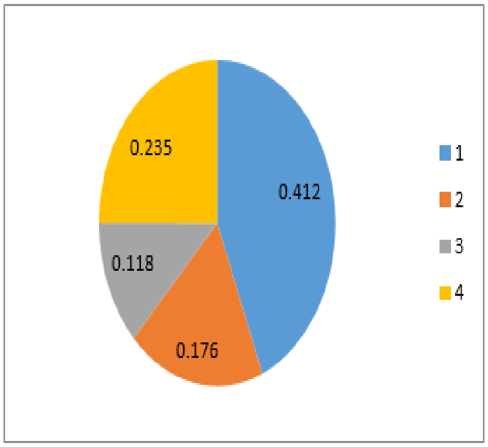


Figure 3: Pie chart to show the eligibility of an individual based on its fitness score

Step 5: This stage is where new individuals are created for the next generation based on the result of the crossover applied on the population pool. The algorithm uses one point crossover, see table 2.

TABLE 2: THE DISTRIBUTION OF THE POPULATION AFTER CROSSOVER

String No	Mating Pool	Crossover point	Offspring after crossover	Fitness	Pi	Expected count
Target String: ATTC AAAA GGTG						
1	ATTC CGGA TTTG	8	ATTC CGGA AAAG	6	0.286	1.143
2	CTGG GCTA AAAG	8	CTGG GCTA TTTG	4	0.190	0.762
3	ATTC CGGA TTTG	4	ATTC AATG GGCC	8	0.381	1.524
4	GAAA AATG GGCC	4	GAAA CGGA TTTG	3	0.141	0.571
Sum				21	1	4
Average				5.25	0.25	1
Max				8	0.381	1.524

Step 6: After the crossover has been completed the offspring with the list expected count are taking for mutation and this can be seen in table 3 and figure 4. The mutation approach used on this implemenation is called scramble mutation.

String no	Offspring Before Mutation	Offspring After Mutation	Fitness	Pi	Expected count
1	ATTC CGGA AAAG	ATTC CGGA AAAG	6	0.273	1.092
2	CTGG GCTA TTTG	CTGT ATCG GTTG	5	0.227	0.908
3	ATTC AATG GGCC	ATTC AATG GGCC	8	0.364	1.456
4	GAAA CGGA TTTG	GAAT AGGC ATTG	3	0.136	0.544
Sum			22	1	4
Average		ATTC AAAA GGTG	2	0.25	1
Max			8	0.364	1.456

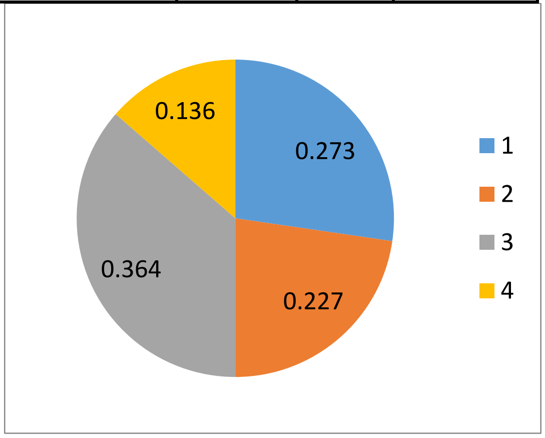


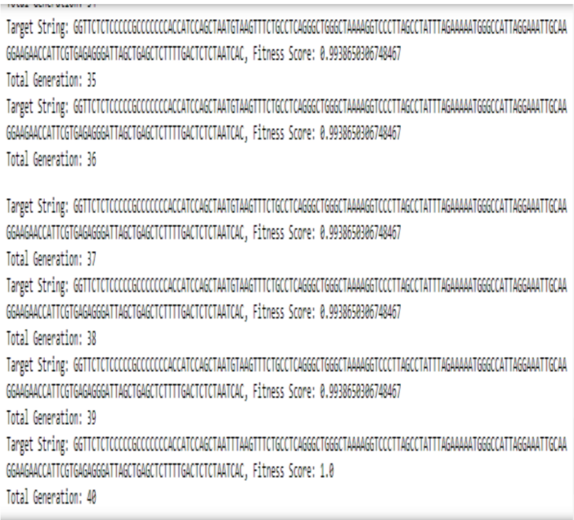
Figure 4: Pie chart to show the eligibility of an individual based on its fitness score after mutation

Step 7: After the algorithm has met its termination condition by getting the candidate solution that satisfies the search criterial, it will get out of the loop and terminate, else it repeats all the steps again.

IV RESULTS AND DISCUSSION

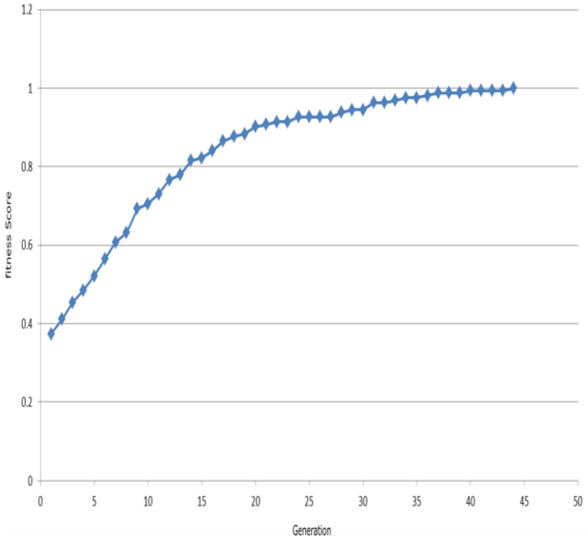
This section shows the result of the genetic algorithm implementation using python programming language. The result of the implementation can be shown in figure 5. It comprises the target string, the fitness score and the number of generations. The target string is the chromosome of the cancerous gene which is being compared with the rest of the genes of other family members to know if there is any detectable pattern. As the algorithm evolves through the generations the fitness score tends to converge to 1.0 meaning that it reaches 1.0 when all the conditions has been satisfied and that the string can be found in the population of offspring in the family. In some cases the algorithm may not converge if the string cannot be found in the population.

The result shows how the algorithm converges to the optimum solution as the generation evolves. The effectiveness of the algorithm is dependent on how accurate and fast it searches the population in other to detect some useful patterns that will guide decision on the exact genome that is causing illness in an individual. Figure 6 shows how the fitness score increases from 0 to 1.0 as the generation increases. In most cases the algorithm converges very fast depending on the mutation value and the size of the population. The graph in figure 6 may evolve forever if the pattern present in the target string cannot be found in the population under consideration.



```
Target String: GGTTCTCTCCCCCGCCCCCCCACCATCCAGCTAATGTAAGTTTCTGCCTCAGGGCTGGGCTAAAAGGTCCCTTAGCCTATTTAGAAAAATGGGCCATTAGGAAATTGCAA
GGAAGAACCATTCGTGAGAGGGATTAGCTGAGCTCTTTTGACTCTCTAATCAC, Fitness Score: 0.9938650306748467
Total Generation: 35
Target String: GGTTCTCTCCCCCGCCCCCCCACCATCCAGCTAATGTAAGTTTCTGCCTCAGGGCTGGGCTAAAAGGTCCCTTAGCCTATTTAGAAAAATGGGCCATTAGGAAATTGCAA
GGAAGAACCATTCGTGAGAGGGATTAGCTGAGCTCTTTTGACTCTCTAATCAC, Fitness Score: 0.9938650306748467
Total Generation: 36

Target String: GGTTCTCTCCCCCGCCCCCCCACCATCCAGCTAATGTAAGTTTCTGCCTCAGGGCTGGGCTAAAAGGTCCCTTAGCCTATTTAGAAAAATGGGCCATTAGGAAATTGCAA
GGAAGAACCATTCGTGAGAGGGATTAGCTGAGCTCTTTTGACTCTCTAATCAC, Fitness Score: 0.9938650306748467
Total Generation: 37
Target String: GGTTCTCTCCCCCGCCCCCCCACCATCCAGCTAATGTAAGTTTCTGCCTCAGGGCTGGGCTAAAAGGTCCCTTAGCCTATTTAGAAAAATGGGCCATTAGGAAATTGCAA
GGAAGAACCATTCGTGAGAGGGATTAGCTGAGCTCTTTTGACTCTCTAATCAC, Fitness Score: 0.9938650306748467
Total Generation: 38
Target String: GGTTCTCTCCCCCGCCCCCCCACCATCCAGCTAATGTAAGTTTCTGCCTCAGGGCTGGGCTAAAAGGTCCCTTAGCCTATTTAGAAAAATGGGCCATTAGGAAATTGCAA
GGAAGAACCATTCGTGAGAGGGATTAGCTGAGCTCTTTTGACTCTCTAATCAC, Fitness Score: 0.9938650306748467
Total Generation: 39
Target String: GGTTCTCTCCCCCGCCCCCCCACCATCCAGCTAATTTAAGTTTCTGCCTCAGGGCTGGGCTAAAAGGTCCCTTAGCCTATTTAGAAAAATGGGCCATTAGGAAATTGCAA
GGAAGAACCATTCGTGAGAGGGATTAGCTGAGCTCTTTTGACTCTCTAATCAC, Fitness Score: 1.0
Total Generation: 40
```

Figure 5: Sample output of the programme showing the target string and the sequence of generations



V. CONCLUSION

This paper has been able to prove that inherited cancerous genes in offspring can be traced back to the parent that contributed the gene to the offspring using genetic algorithm. The major significance of this research is that with good understanding of human genome medicine can be administered based on an individual's genetic makeup rather than generalized drug administration. This research work could not explore all the genome for effective medical diagnoses and treatment, so it is our intension to explore the genome deeper in future work so as to detect more patterns that will help in medical diagnoses and treatment of diseases.

REFERENCES

[1] S. E. DeWeerdt, (2003, Jan. 15). What is a Genome? Genome News Network [Online]. Available: http://www.genomenewsnetwork.org/resources/whats_a_genome/Chp1_1_1.shtml.

[2] M. Burset and R. Guigo, (2020, Jan. 9). Evaluation of gene structure prediction programs [Online]. Available: https://genome.crg.cat/datasets/genomics96/#SEQS

[3] R. L. Haupt and S. E. Haupt, Practical Genetic Algorithms, New York: WileyInterscience, 1998.

[4] E. D'Agaro, "Artificial intelligence used in genome analysis studies", The EuroBiotech Journal, pp. 78-88, 2018

[5] C. Moschopoulo, D. Popovic, A. Sifrim, G. Beligiannis, B. Moor, and Y. Moreau, "A genetic algorithm for pancreatic cancer diagnosis", *Engineering Applications of Neural Networks,*

Communications in Computer and Information Science, Springer, Berlin, Heidelberg, 2013.

[6] R. Gayathri, (2017, August). "Genetic Algorithm Based Model for Early Detection of Cancer", International Journal for Modern Trends in Science and Technology, vol. 3, issue 8. pp. 28-31.

[7] N. Fatima, N. and S. Mohammad, (2018, August). "Paper on Genetic Algorithm for Detection of Oral Cancer," International Journal of Advanced Research in Computer and Communication Engineering, vol. 7, issue 8. pp. 102-107.

[8] J. M. Diaz, R. C. Pinon and G. Solano, "Lung Cancer Classification Using Genetic Algorithm to Optimize Prediction Models", The 5th International Conference on Information, Intelligence, Systems and Applications, Chania, Greece: IEEE, 2014.

[9] D. E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley, 1989.

Enhanced CAPTCHA Based Authentication using Mathematical Operation

Olanrewaju Oyenike Mary[1] and Abdulwasiu Adebayo Abdulhafeez[2]
[1] Department Computer Science, Federal University Dutsinma.
[2]Computer Science Department, Federal University Dutsinma.
oolanrewaju@fudutsinma.edu.ng ibnwasar@gmail.com.

Abstract

Despite the fact that many research works have been done on the way to securing documents on websites, attackers are day by day finding means of detecting a way to break all the security methods. One of the ways of these prevention is Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA) which is used to protect websites such as E- mail providers, social networks, and online business from unauthorized user. This research work reviewed major text CAPTCHA in existence, and proposed an Enhanced CAPTCHA based authentication using Mathematical operation to provide solutions to most common problem of the existing ones such as inlegibility of character and easy penetration by smart robots. The proposed system based the authentication on random mathematics expression with alteration in the operator for the user to respond to the challenge by supplying the correct answer. The implementation was developed using Visual Basic 2010 and result were presented in this paper.

Keywords: security,website, CAPTCHA, mathematical operation

I. INTRODUCTION

Security is an essential domain and it plays vital role in protecting one's resources so as to keep them from hackers, theft or the likes. It has been noticed that technology advances every day because many new things are being designed to the extent that machines are now designed to predict or to serve as a Robot which can penetrate into user's account and manipulate things. Many security methods have been design to curb the problem of insecurity and to protect documents and one of the security ways of protecting users document is using CAPTCHA. The term CAPTCHA was coined in 2003 by Luis Von Ahn , Manuel Blum, Nicholas J. Hopper, and John Langford [15]. The most common type of CAPTCHA (displayed as Version 1.0) was first invented in 1997 by two groups working in parallel. This form of CAPTCHA requires that the user type the letters of a distorted image, sometimes with the addition of an obscured sequence of letters or digits that appears on the screen because the test is administered by a computer [9], in contrast to the standard Turing test that is administered by a human, therefore, a CAPTCHA is sometimes described as a reverse Turing test [8]. This user identification procedure has received many criticisms, especially from people with disabilities, but also from other people who feel that their everyday work is slowed down by distorted words that are difficult to read. It takes the average person approximately 10 seconds to solve a typical CAPTCHA [3]. There are different kinds of CAPTCHA, one of them is text-based CAPTCHA , where user would be requested to supply the text displayed in pictures which computer cannot understand but made easy for human. This research deals with Mathematical operation CAPTCHA based on text to solve the insecurity problem in other to reduce the guessing attacks and increase the simplicity of CAPTCHA password. The proposed algorithm is built on CAPTCHA and mathematical operation. Mathematical operation combines two numbers(operands) with either addition or subtraction between them and Visual Basic 2010 was used for the implementation.

II. LITERATURE REVIEW

CAPTCHA as a program can generate and grade tests that most humans can pass but computer programmed robots should not pass [15]. Such a program can be used to differentiate humans from computers. In general, CAPTCHA methods can be divided into five groups: Text-based CAPTCHAs, Image-based CAPTCHAs, Audio-based CAPTCHAs, Motion-based CAPTCHAs, and Hybrid CAPTCHAs [3].

A text-based CAPTCHA is a distorted image of a sequence of characters on which different types of degradations, background clutters and color mixtures are applied to make it more challenging for attackers. Text-based CAPTCHAs is introduced in six sub-groups [10]

a) CAPTCHAS with "English Words" as their CAPTCHA text:

i. Gimpy: Gimpy is one of the most famous CAPTCHAs which are primarily based on distorted text. Gimpy picks seven words from a dictionary; then renders a distorted image containing those words. It finally presents them to its users and asks them to type three words of the image to gain entry to the service [12]. The sample is presented in Figure 1.

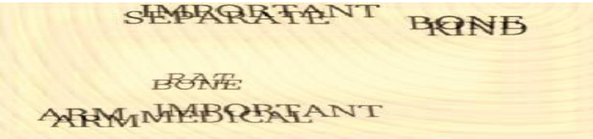


Figure 1: Gimpy CAPTCHA [12]

ii. EZ-Gimpy: In this CAPTCHA, a word is chosen from a dictionary. In the next step, the word is rendered to an image using various fonts and different types of distortions such as black or white lines, background grids and gradients, blurring and pixel noise are added. Then, the user is asked to type the word.[16]. The sample picture is depicted in Figure 2.

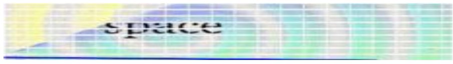


Figure 2: Sample of EZ-Gimpy[16].

iii. PessimalPrint: PessimalPrint Concentrates on degradations, such as adding noise to or blurring the images to defeat OCR techniques; the designers of this CAPTCHA argue that under the conditions of inferior image quality, there is a significant gap in pattern recognition ability between humans and machines. In this CAPTCHA, a word is randomly selected from a fixed list containing 5-to-8-letters of English words [1]. The feature of PessimalPrint is in Figure 3.

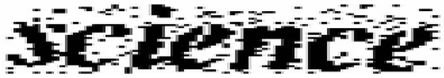


Figure 3: Sample of PessimalPrint **[1]**

b) CAPTCHAs with "Random Strings": Using English words in some current CAPTCHAs makes them vulnerable to dictionary attacks. The solution for this issue is exploiting random strings instead of words. This technique is utilized by MSN CAPTCHA, Yahoo,Ticketmaster, Google, etc.

i. Hotmail or MSN CAPTCHAs : In this CAPTCHA used in the Hotmail email service registration, it selects eight English characters (upper case letters and digits); then, after applying local and global wraping, it renders the characters with dark blue color on a light gray background [4].

ii. Yahoo! CAPTCHA (Yahoo version2): Starting in August 2004, Yahoo! introduced its second generation CAPTCHA. Its characteristics include using a string of characters instead of English words, using both letters and digits, and having connected lines and arcs as clutter, examples are Ticketmaster, Google/Gmail [1]. The image in figure 4 shows example of CAPTCHA used in hotmail.

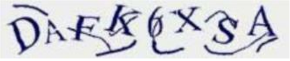
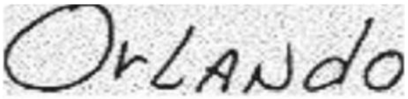


Figure 4: **Yahoo! CAPTCHA (Yahoo version2)** [1].

c). CAPTCHAS Based on Handwritten Text:

While most current text-based CAPTCHAs use machine printed text, which makes them vulnerable to pattern recognition attacks, there are CAPTCHAs that use handwritten text in their challenges. An example is Handwritten CAPTCHA which is based on distorted handwritten text [14]. The picture in figure 5 is an



handwritten CAPTCHA

Figure 5: Handwritten CAPTCHA [14].

d). CAPTCHAS Based on Linguistic Knowledge: Some current CAPTCHA systems combine an OCR (optical character recognition) problem with linguistic knowledge in order to strengthen their tests. Examples of such CAPTCHAs include semCAPTCHA, odd-words-out, number-puzzle-text CAPTCHA, SSCAPTCHA and text-domain CAPTCHA [11] Figure 6 is an example of semCAPTCHA.

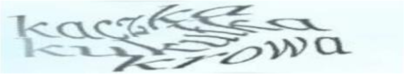


Figure 6 : Sample semCAPTCHA [11].

e). Non-English CAPTCHAS: Besides English CAPTCHAs, some CAPTCHAs have been developed in other languages. One reason for producing localized CAPTCHAs is attacks against current English CAPTCHAs. Another reason is that people are more comfortable with solving tests in their own languages. An example is Arabic CAPTCHA [11]. Figure 7 depict Arabic CAPTCHA.

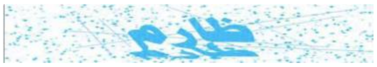


Figure 7: Arabic CAPTCHA. [11]

All the above CAPTCHA were developed so as to make user documents secured, where some were blurred, some were written in users language and some deviated from usinlg dictionary words just for the CAPTCHA test to be robust, however, it was discovered that some text CAPTCHA above are not clearly seen because the developer intend to make them secure and in CAPTCHA, consideration must be given to usability, therefore, this research is also deviating from using dictionary words thereby imploring the use of numbersfrom 0 to 9 and operators (addition and subtraction) and moreso, mathematical numbers seems to be a general language that every computer user can operate with, and lastly scattered lines are used on the challenge so that it will be difficult for machine to solve.

3.0 METHODOLOGY

To develop an enhanced CAPTCHA that aim at securing the account of web users the following procedures were considered and implemented using Visual Basic (2010).

CAPTCHA's codes are chosen digit numbers (0-9 for this experiment) with a mathematical operator (addition (+) or subtraction (-) are implemented), Multiple random functions are used to generate code of numbers in each task so as to avoid susceptibility to attack. The numbers are varied in every refresh (and it consist of two numbers with operatorin between them, The numeric image is displayed using distorted lines for CAPTCHA to be difficult for malicious Software and addition and subtraction are used randomly to prevent prediction by robot.

The algorithm below shows how the CAPTCHA system is designed:

1. CREATE RANDOM VALUE FOR NUMERIC CODE
2. MAKE A STATEMENT OF THE NUMBERS AT RANDOM AND ADD SCATTERED LINES
3. USE THE SECURE NUMBER STORED
4. SUPPLY THE ANSWER
5. IF THE CORRECT ANSWER IS SUPPLIED THEN IT MOVES TO THE NEXT
6. DISPLAY CORRECT! YOU CAN PROCEED
7. IF THE ANSWER SUPPLIED IS WRONG THEN IT DISPLAYS INCORRECT, TRY AGAIN IF YOU ARE NOT A ROBOT.

The Flowchart For The Proposed Method

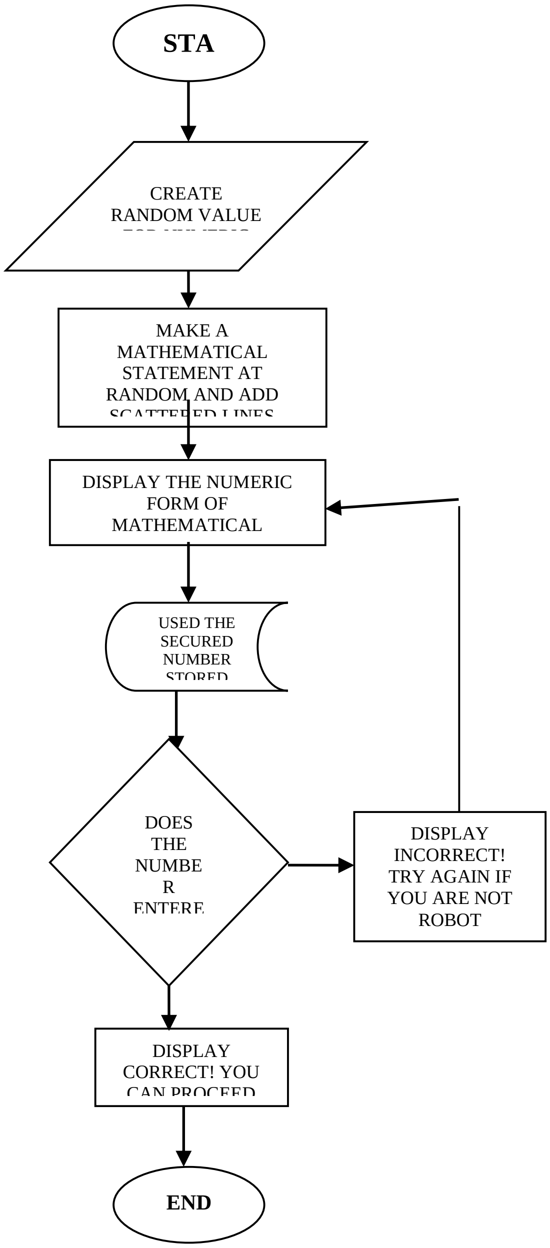
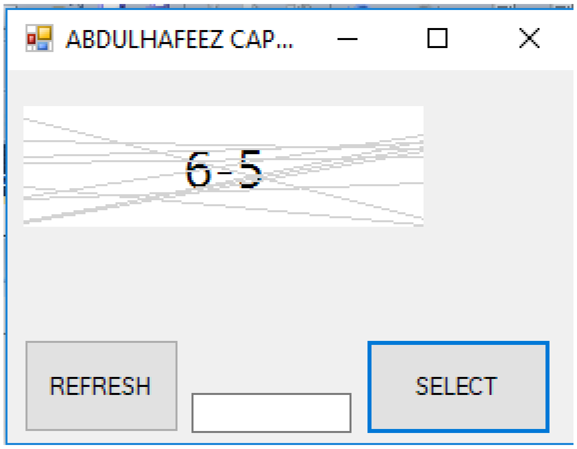

STA
CREATE RANDOM VALUE
MAKE A MATHEMATICAL STATEMENT AT RANDOM AND ADD SCATTERED LINES
DISPLAY THE NUMERIC FORM OF MATHEMATICAL
USED THE SECURED NUMBER STORED
DOES THE NUMBE R ENTERE
DISPLAY INCORRECT! TRY AGAIN IF YOU ARE NOT ROBOT
DISPLAY CORRECT! YOU CAN PROCEED
END

Figure 8: SYSTEM FLOWCHART



4.0 IMPLEMENTATION AND RESULT

VB.Net 2010 was used for design of the CAPTCHA because it consist of several parts such as Application Programming Interface (API) class libraries that contains all built in functions that make the windows platform to function the way it does, with adequate programming tools.It is also platform independent because it can easily be moved to different computer system.

The following are the results:

a. Graphical User Interface Of CAPTCHA for Addition

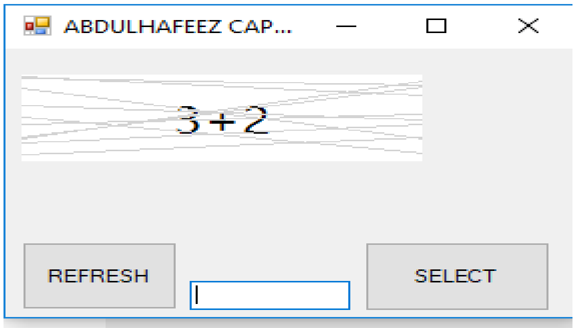
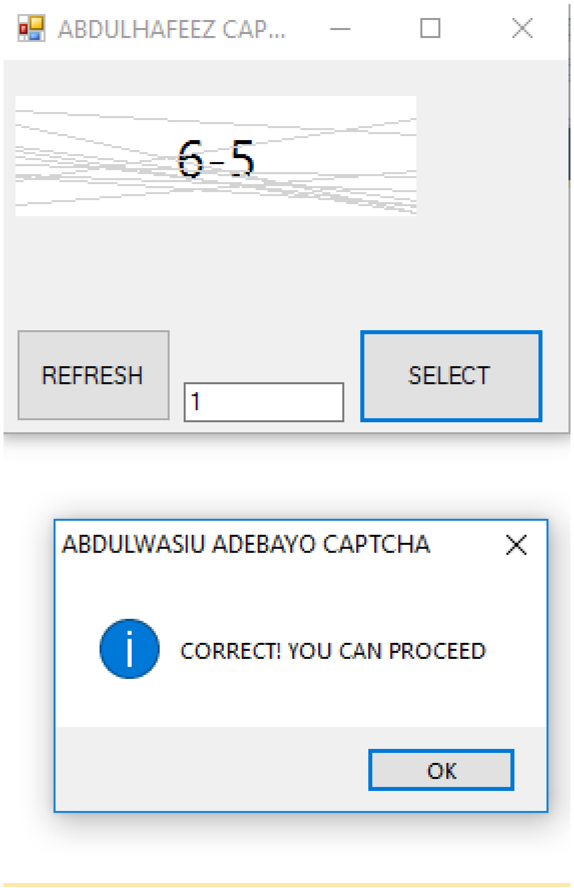


Figure 9: Addition operation CAPTCHA

Figure 9 displays the interface for addition of operands with scattered line on it, the user is expected to enter the correct answer into the text box where the cursor is blinking.And from this interface, the Refresh Botton is to refresh the challenge by changing the expression while select botton is used to submit the answer entered.

b. Graphical User Interface For Subtraction



1

CORRECT! YOU CAN PROCEED

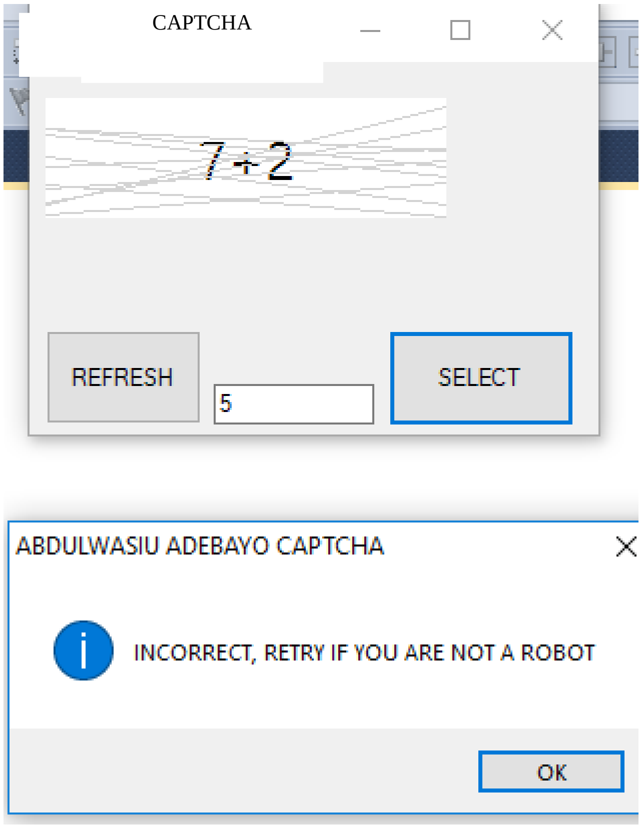
Figure 10**:**Subtraction operation CAPTCHA

Figure 10 displays the interface for subtraction of operands with scattered line on it, the user is expected to enter the correct answer into the text box where the cursor is blinking.

c. CAPTCHA with Wrong user response
Figure 11**:**CAPTCHA with wrong answer.

When a user enters wrong answer for CAPTCHA challenge as shown above, then it displays "incorect, retry if you are not a robot" as shown in figure 11

d. CAPTCHA withCorrect Response



5

INCORRECT, RETRY IF YOU ARE NOT A ROBOT

Figure 12: CAPTCHA with Correct Response
When the challenge of CAPTCHA system is displayed, and the user enters the appropriate answer, then it responded by displaying "correct! You can proceed", and then the user can view or make use of the account as shown in figure 12.

5.0 DISCUSSION

From the above results, figure 9 is the interface which displays Mathematical CAPTCHA in form of "3 + 2" with lines being scattered on the numbers to prevent robot from solving the problem, basically it is for addition operation. "Refresh" was thenused to change the question while "select" was used to submit the answer written inside the text box. Figure 10 is the interface for Subtraction which displays "6-5" together with other features as described in addition.

Figure 11 shows the outcome when the answer entered is wrong, this will not allow the user to be able to proceed to login or to next task, and it will displays "incorrect! Retry if you are not a robot".Figure 12 shows that the answer entered is correct, then it displays “correct! You can now proceed”.This indicates that user is not a robot but the authenticated user.

5.1 CONCLUSION

This research is one of the contributions towards improving security against attacks experienced by automated programs.CAPTCHAs or *Reverse Turing tests* are used by programs (or machines) to differentiate humans and machines apart. The race between makers and breakers of CAPTCHAs is at a juncture where the Text CAPTCHAs proposed today are not answerable even by humans due to their complexity or the effect added to them which make the usability difficult.This work proposed Enhanced Captcha Based Authentication Using Mathematical Operation, which falls between making the CAPTCHA system easily used by human and added some effects to make it difficult for machine to solve.

Many researchers have done a lot of research works on a way to secure documents on websites, while attackers are day by day finding means of detecting a way to break all the security methods. This research is a stepping stone in the complex field of using CAPTCHA as security and my hope is that the research presented in this paper will help other researchers continue to examine CAPTCHA challenges and improve security to prevent SPAM. The future work should exploit the use of image to display mathematics CAPTCHA instead of text because image are more robust than text.

REFERENCES

1. Baird, H., and Chew, M. (2003). Baffletext: A human interactive proof. 10th document recognition and retrieval conference. http://doi.org/10.1117/12.479682
2. Bin, B., Zhu, J., Ning Xu, (2014). Captcha as graphical passwords- a new security primitive based on hard AI problems. 891-904. http://doi.org/10.1109/TIFS.2014.2312547
3. Bursztein, E., Bethard, S., Fabry, C., Mitchell, C. And Jurafsky, D. (2010). How good are humans at solving CAPTCHA? a large scale evaluation. proceedings of the 2010 IEEE symposium on Security and Privacy, (pp. 399-413). http://doi.org/10.1109/SP.2010.31
4. Chellapilla, K., Larson, K., Simard, P., and Czerwinski, M., (2005) "Designing human friendly human interaction proofs (HIPs)," in *ACM Conference on Human Factors in Computing Systems(CHI 05)*, 2005, pp. 711-720.
5. Coates, A., Baird, H. and Fateman, R. (2003) "PessimalPrint: A reverse Turing test," *International Journal on Document Analysis and Recognition,* vol. 5, pp. 158-163, 2003.
6. Yan, J. "Bot, cyborg and automated turing test," in *Security Protocols Workshop,* 2006, pp. 190-197. DOI http://doi.org/10.1007/s10032-002-0089-1
7. Converse, T. (2005). "CAPTCHA generation as a web service," Human Interactive Proofs, . In T. Converse, "CAPTCHA generation as a web service," Human Interactive Proofs, (pp. 82-96). DOI http://doi.org/10.1007/11427896_6
8. Darko Brodic, Alessia Amelio, Nadeem Ahmad, and Syed Khuram Shahzad,(2017) “Usability Analysis of the Image and Interactive CAPTCHA via Prediction of the Response Time”, Springer International Publishing AG pp. 252–265. Retrieved from http://www.springerlink.com. //doi.org/10.1007/978-3-319-69456-6_21
9. Ibrahim F., Ilker Y., Yucel B.S., (2008) “*Designing Captcha Algorithm: Splitting And Rotating The Images Against Ocrs*”Third 2008 International Conference on Convergence and Hybrid Information Technology 978-0-7695-3407-7/08 $25.00 © 2008 IEEE DOI 10.1109/ICCIT.2008.195 596
10. Kaur, I., & Hornof, A.J. (2005). "Is it human or computer? Defending e-commerce with CAPTCHAs,". In C. P. Kaur, "Is it human or computer? Defending e-commerce with CAPTCHAs," (pp. 43-49). IT professionals.
11. Lupkowski, P. and Urbanski,(2008) M. "SemCAPTCHA—user-friendly alternative for OCR-based CAPTCHA systems," in *International Multiconference on Computer Science and Information Technology (IMCSIT 2008)*, 2008, pp. 325-329
12. Mir AmanSheheryar, Pradeep Kumar Mishra and Ashok Kumar Sahoo, (2016), “A review on CAPTCHA generation and evaluation techniques”,ARPN Journal of Engineering and Applied Sciences, VOL. 11, NO. 9, MAY 2016 ISSN 1819-6608,Asian Research Publishing Network (ARPN). Retrieved from www.arpnjournals.com
13. Pawel, L., Mariusz, U., (2008). SemCAPTCHA—user-friendly alternative for OCR-based. international multiconference on computer science and information technology, (pp. 325-329). http://doi.org/10.1109/MCSIT.2008.4747260
14. Rusu, A., Thomas, A., and Govindaraju, V. (2010) "Generation and use of handwritten CAPTCHAs. international journal on document analysis and recognition , 49-64.http://doi.org/10.1007/s10032-009-0102-z
15. Manuel, V., Nicholas, B., John L., (2003). CAPTCHA: Using Hard AI Problems for Security (PDF), EURPCRYPT. international conference on the Theory and Applications of Cryptographic Techniques.http://doi.org/10.1007/3-540-39200-9_18
16. Von, L., Blum, M. and Langford, J. "Telling humans and computers apart automatically," *Communications of the ACM,* vol. 47, pp. 56-60, 2004.

APPENDIX

```
PublicClassForm1

Private cAnswer AsString = Nothing

Function genQuestion() AsString
Dim cOperators AsString() = {"+", "-"}
Start:
Dim p1 AsInteger = NewRandom().Next(1, 9)
Dim p2 AsInteger = NewRandom().Next(1, 9)
If p1 = p2 ThenGoTo start
Dim cOperator AsString = cOperators(NewRandom().Next(0,
cOperators.Length))
SelectCase cOperator
Case"+"
        cAnswer = p1 + p2
If cAnswer <= 0 ThenGoTo start
Case"-"
        cAnswer = p1 - p2
If cAnswer <= 0 ThenGoTo start
EndSelect
ReturnString.Format("{0}{1}{2}", p1, cOperator, p2)
```

```
EndFunction


PrivateSub generatelines(ByVal G AsGraphics)
IfNot G IsNothingThen
Dim R AsNewRandom()
Dim linebrush AsNewSolidBrush(Color.LightGray)
For i% = 0 To 9
            G.DrawLines(NewPen(linebrush, R.Next(1, 2)),
NewPoint() {NewPoint(0, R.Next(0, 60)), NewPoint(200, R.Next(0,
60))})
Next
EndIf
EndSub


PrivateFunction generateimage() AsImage
Dim B AsNewBitmap(200, 60)
Using G AsGraphics = Graphics.FromImage(B)
With G
            .Clear(Color.White)
            .DrawString(genQuestion(), NewFont("segoe UI", 20),
Brushes.Black, NewRectangle(0, 0, 200, 60), NewStringFormat()
With {.Alignment = StringAlignment.Center, .LineAlignment =
StringAlignment.Center})
EndWith
         generatelines(G)
EndUsing
Return B
EndFunction



PrivateSub Button1_Click(ByVal sender As System.Object, ByVal e
As System.EventArgs) Handles Button1.Click
SelectCase TextBox1.Text
CaseIs = cAnswer
MessageBox.Show("CORRECT! YOU CAN PROCEED", "
CAPTCHA", MessageBoxButtons.OK, MessageBoxIcon.Asterisk)
            PictureBox1.Image = generateimage()
            TextBox1.Clear()
Case Else
MessageBox.Show("INCORRECT, RETRY IF YOU ARE NOT A
ROBOT", " CAPTCHA", MessageBoxButtons.OK,
MessageBoxIcon.Asterisk)
            PictureBox1.Image = generateimage()
            TextBox1.Clear()
EndSelect
       TextBox1.Clear()

EndSub

Button2_Click(ByVal sender As System.Object, ByVal e As
System.EventArgs) Handles Button2.Click
       PictureBox1.Image = generateimage()
EndSub


PrivateSub Form1_Load(ByVal sender As System.Object, ByVal e
As System.EventArgs) HandlesMyBase.Load
       PictureBox1.Image = generateimage()
EndSub
EndClass
```

A Predictive Model for Student Performance in Examination Using Supervised Machine Learning Algorithm

Bukola Taibat Adebiyi
bek2705@yahoo.com
Department of Computer Science
Federal University of Agriculture, Abeokuta
Ogun State, Nigeria

Olufunke Rebecca Vincent
vincent.rebecca@gmail.com
Department of Computer Science
Federal University of Agriculture, Abeokuta
Ogun State, Nigeria

ABSTRACT* - Student academic performance has to do with the extent to which students attained short and long term educational goal. Several models have been proposed by different researcher in evaluating the performance of student and predicting the student performance. These models have not considered all performance parameters. This study presents an improved model for student's performance using a supervised machine learning approach. The study takes into consideration performance attributes which are classwork, assignment, attendance, punctuality, and participation in tutorials. These attributes formed the basis for which the models for this work is built. The research concentrated on predicting the actual score using supervised machine learning algorithms (RandomTree and RandomForest). The randomtree outperformed the randomforest with a higher correlation of than those in literature. Furthermore, the students' performance was classified into three classes (low, average and high) using Hoeffding, J48, LMT, RandomForest and RandomTree. Both RandomTree and RandomForest had the highest precision, recall and fmeasure.

Keywords: supervised learning, machine learning, predictive, performance, classifier.

I INTRODUCTION

End of semester examination is one of the major ways students' academic performance are usually measured. There have been several research works carried out to predict students' academic performance which focused on resource allocation, mode of study, student background among others. The aim of this research work is to predict students' academic performance through comparative analysis of examination scores to predict academic performance.

The academic performance of students in tertiary institution is the most important yardstick to measure the quality of college students. It is the vital criteria taking into cognizance by college or university in order to monitor the quality of teaching and learning and for college to assess and select students. Tertiary institutions today seek to attract prospective student through setting a track record of excellent performance amongst her graduates. At this present time, most tertiary institutions are encountering difficulties in attracting prospective students due to high competitiveness in the educational terrain. These has made the study of students' performance a subject matter that cannot be jeopardized in promoting the student developments and the betterment of academic quality in tertiary institutions. Nevertheless, the performance of students is affected by various element in a complex manner, and the student socio-economic background and their historical academic performance may potentially affect their academic performance. Obviously, most existing research works focused on analysing and predicting students' performance in a relatively simple problem formulation and adopting statistical techniques.

To handle these limitations, machine learning technique has in recent times been adopted in data science applications in the analyses of complex relationships. It has the capability of learning automatically without being programmed in detail. An Artificial Neural Network (ANN) model, which has established a history in computing and data science, is fast increasing in term of popularity and wide applications. Neural Network extends the capability of analysing complicated amount of data sets that are not easily to be simplified through the conventional statistical techniques. It has also the ability to implicitly detect non-linear relationships between dependent and independent variables. Neural Network has been gaining popularity and has proven a great success in the application of classifications, forecasting pattern recognitions and prediction in the areas of healthcare, stock markets, climatic condition, etc.

However, the use of Neural Network is not very popular in the educational research. This can be due to the complexity of the modelled network, the obstacle for a modelled Neural Network system to provide an appropriate explanation (its black-box nature), its proneness to over-fitting, and the period required for training the neural network. The educational data and examination scores of the student of one of the Federal College of Education in Nigeria will be studied to using a supervised machine learning technique model with suitable configuration settings to accurately predict and classify the students' performance.

Overall, this paper presents an exploratory modelling and analyses of students' performance through the data collected from one of the Federal College of Education in Nigeria. The machine learning model serves as a reliable educational quality tool that evaluates the students' performance in tertiary institution, addressing those discrepancies and consequently continual improvement in the quality of education.

II RELATED WORK

It has been proven that Neural Network has a splendid prediction method for different applications such as speech recognition, image processing and pattern recognition application. Neural network is a series of algorithms that strives to recognize underlying relationship in a set of data through a process that imitate the way the human brain operates. Neural Network has also been used for predicting performance of studies especially using multi-layer Neural Network and feature selection played a major role in the network performance [1], [2]. For instance, in [2] presentation, it considers personal and academic careers data related, 810 students registered in first year of health care professional courses. Machine learning Neural Network used 49 variables input factors and 34 nodes in the hidden layer predicting student dropout rate. The prediction model categorized students as 1, 2, and 3 meaning regular student, irregular student, and student at risk of abandonment, respectively. The chosen characteristics were captured from students' academic performance and students' demographics but did not capture other important characteristics like physiological, economical etc. that will have a significant influence on overall student performance.

[3] - [7] have delved into various ways to predict students' performances at various stages of their study. This comprised using different types of prediction variables in addition to using different mathematical prediction models such as Random Forest (RF), k-Nearest Neighbours (kNN), Naïve-Bayes (NB), Bayesian, Decision Tree (DT), and Support Vector Machine (SVM).

[8] developed a data mining predictive model for students' performance to establish the factors which are causative of poor performance of students in higher secondary examination in Tamil Nadu. They used 772 students' dataset collected from regular students and school offices for the prediction. Chi-Square Automatic Interaction Detection (CHAID) DT (decision tree) algorithm was used for prediction. A set of decision rules were formulated using this tree, which was for predicting student grades. The algorithm achieved an overall prediction accuracy of 44.69%.

Another work [9] proposed a model for the prediction of students' performance using data mining approaches with some few characteristics referred as student's behavioural features. This model was evaluated with three classifiers which are; Artificial Neural Network, Decision tree and Naïve Bayesian. Bagging, Random Forest, and Boosting were ensemble methods used in order to enhance the performance of the classifiers. The model achieved up to 22.1% more in accuracy as compared to when the behavioural features were not considered. The model improved to 25.8% accuracy on using the ensemble methods.

[10] proposed a multi-dimensional conceptual framework which looked at six interrelated domains:

Domain	**Factors Considered**
Cognitive	examination marks and presentation skills
Demography	age, ethnicity and gender
Economic	income, parents' financial status
Institutional	course of study, learning environment, support
personality	learning style, motivation
psychological	ability, accomplishment, interest

The researchers emphasized on the combination of the various factors selected from all these domains to support each other in order to predict the performance students.

In proposing a model for predicting the performance of students', [11] classified students into a binary class (successful / unsuccessful). The proposed model was constructed under the CRISP-DM (Cross Industry Standard Process for Data Mining) research approach. The classification algorithms (OneR, J48, MLP and IBK) were applied on the given dataset. The results show that the highest accuracy was achieved by the MPL model (73.59%) for identification of successful while other three models perform better for the identification of unsuccessful students.

[12] proposed a model to predict students' performance by overcoming the problem of class unevenness. Two algorithms used in this study were Naïve Bays and Decision tree. A cost-sensitive method. Metacost approach was used to manage this problem. After this was done, the highest accuracy was gotten by naive bays was 85%. The collection of data at the end of academic period was not feasible because no one can get benefit at that time. A student academic performance prediction model was proposed in this study [13]. The classifiers used are ANN, J48, Reptree, Decision Stump and NB with three categories of attribute were evaluated in this study. The J48 classifier outperformed the others with the high accuracy 90.51%.

[14] presented a model for predicting students' failure, this model identified the students who might be at-risk of failing. The model had four output classes (Average, Risk, below Average and Above Average) which were generated based on the CGPA of the students. Six different classifiers were adopted on the dataset for this work. The ID3 had the highest accuracy of 79.23%. However, the model was unable to work out for class imbalance problem.

[15] proposed a multilevel classification model to resolving multiclass classification problem in prediction of students' performance. The goal of the study was to increase the accuracy of the model as well as the that of the various classifier in general. The model comprises of two levels. Firstly, in order to deal with distribution problem in the pre-processing phase, a re-sampling technique was carried out on the dataset. In this level, the following four different classifiers were applied on the dataset which are IBK, MLP,

NB, J48. The results were compared and evaluated. The j48 outperformed the other classifiers thereby gaining the highest accuracy and was therefore selected for the next level. In the next level, outliers were selected by comparing the results of predicted with the actual results in previous level. The J48 classifier having been selected from the previous level was applied on the filtered dataset and the results were compared with the other classifiers. The J48 classifier had an accuracy above 90% for overall model and also for the various classifier classes prediction.

III METHODOLOGY

Through various search on the discussion with experts on student performance in examination, some factors that were considered as influence on the performance of a student in examination were identified. These factors were formed the input variables. On the other hand, the output variable was the possible score. Secondary dataset from Federal college of education Akoka. Machine learning algorithm were employed, decision tree was used as well as function algorithm in WEKA. The goal was to predict student score in examination using various parameters/attributes.

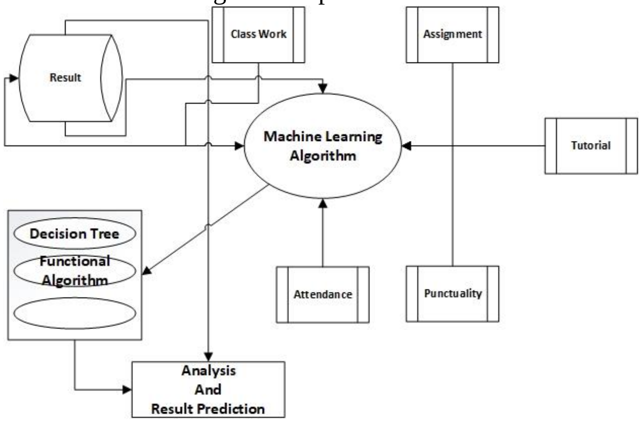


Fig 1. Proposed model for student score prediction

The model showed the parameters considered as determinants for the performance of students in examination. There are several parameters for predicting the performance of a student in an examination. However, for this study, the following parameters were considered; students' performance in class work, outcome of assignment of a student, attendance to classes, punctuality to class and participation to tutorial classes in preparation for examination.

Data collection

Students' performance data used in this study was collected from FCE, Akoka, Lagos. The data consists of student performance for the 2018/2019 academic session. It includes results of 3 departments with various courses taken for the session.

Data manipulation

The collected data was worked upon in order to extract usable features selected for the purpose of this work. The mid semester assessment usually was broken into class activity which was graded, assignment and a continuous assessment test which is written and graded. During each class, student attendance register was taken, this was converted into percentage to get the actual attendance rate of a particular student for the session under study. On the attendance register, the time a student arrives at class were also noted. This formed the rate of punctuality of the student for the course taken.

This study considered the following features as a determining factor for the performance of students' in examination; attendance to class was viewed as a major determinant of a student success factor, sequel to that, the study also assumed that being present in class is not enough but when does the student arrive in class. It was assumed that a student who was present in class before the commencement of the class has a more psychological stability all through the class than those who come later when the class is on or towards the tail end of the class. Participation in class assessment and activities was viewed as another major determinant of the students' performance in an examination. This was conceived from the first two features highlighted. Only a student who is present in class will participate in class assessment or activities for that class. This was taking note of as the performance of each class activities or assessment were summed up for each student. This formed 10% of the students' performance for each course taken during the course of the session. Assignments were given as considered necessary by the instructor. This also formed 10% of the performance for the semester. The study further considered tutorials organized among students prior to examination so as to know whether or not it has any impact on the performance of the student in the examination. This was put in form of percentage (given by 1-%absent in tutorials).

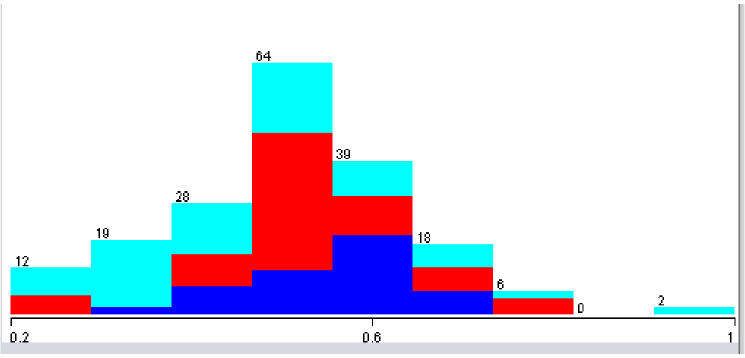


Fig 2: Attribute selection chart

Datasets used

The datasets used for this study were grouped into two categories. The academic category and non-academic category. The academic category includes students' score sheet containing the score in examination and continuous assessment which was broken down into class activity, assignment and written test. However, for this work the written test was discarded as it did not contribute to the performance of students in the examination. The non-academic record were the various attendance registers which are attendance registers, punctuality records and tutorial attendance record.

Adopted techniques

For this work, decision tree algorithms were used in WEKA.

4. Results

For this work, five classifiers J48, LMT, Hoeffding, Random Forest and Random Tree have been evaluated using 10folds cross validation technique. In this technique the dataset split into ten equal subsets of the datasets; nine of these subsets are used for training the model, while the remaining one is exempted for use in testing the model. The process is done iteratively for ten iteration, the final result is the estimation of the average error rate on test carried out.

Evaluation Measures

In this work, five common different measures were used for the evaluation of the classification quality. Details are as follows:

Correctly Classified Instances (CCI): this is the ratio of correctly classified instances and the total classified instances. It is also referred to as the accuracy.

Incorrectly Classified Instances (ICI): this is the ratio of the instances that were not incorrectly classified.

Precision: this is the percentage of accurate classified instances with respect to the classified instances.

Recall: this is the ratio of correctly classified instances and the total number of all instances (usually the recall value is almost equal as CCI).

F-Measure: this is gotten from the values of the recall and precision. It is given by the ratio of twice the precision and summation of precision and recall.

IV RESULT ANALYSIS

In the first experiment, two classifications were carried out on the dataset using the decision tree classifier algorithms (random tree and random forest). From the result presented in the figure below.

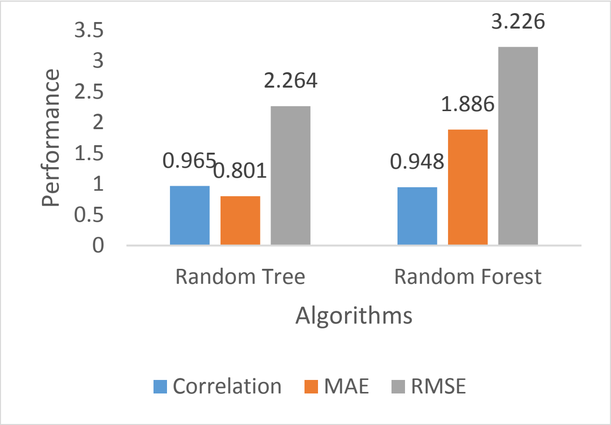


Fig 3: Performance of decision tree in score prediction

In predicting the performance of a student in an examination using the model, the random tree outperformed the random forest with a higher correlation coefficient of 0.965 against 0.948. The mean absolute errors are 0.801 and 1.886 respectively. The root mean square errors are 2.264 and 3.226 respectively. From the dataset, the random forest showed a better performance in predicting numeric values which could not be done by the bayes classifier.

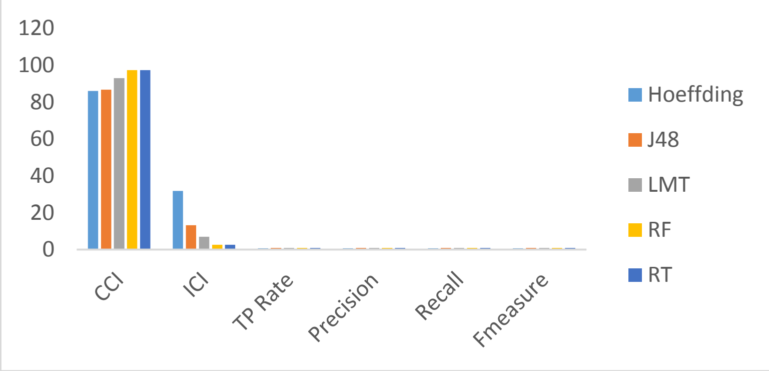


Fig 4: performance of selected classifiers on nominal data prediction

In the second experiment, five classifiers (J48, LMT, hoeffding, random tree and random forest) were performed on the datasets. In this experiment, the predicted class is nominal unlike the classification class in the first experiment which adopted numeric dataset for the class. The evaluation measures used are the correct classified instances (accuracy), incorrect classified instances, precision, recall and fmeasure. The result showed that the decision tree and random tree with the accuracy of 97.34 respectively outperformed the other algorithms which had 86.09, 86.702 and 93.085 respectively (i.e. Heoffding, J48 and LMT).

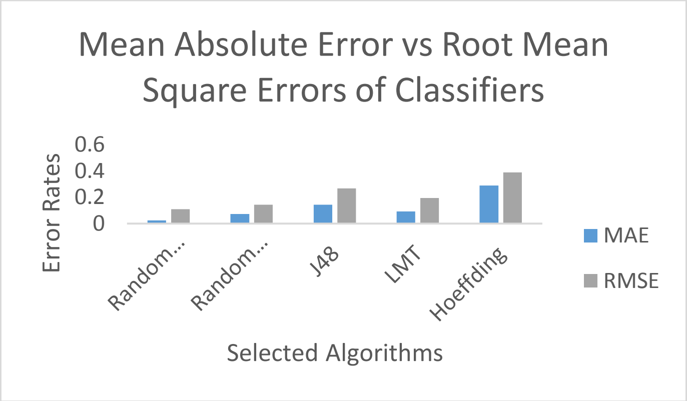


Fig 5: Error values of the selected classifiers.

The result presented in fig 4.3 above shows the error levels of the various classifiers used for the experiments. Random Tree had the lowest mean absolute error and root mean square errors while hoeffding had the highest error level. This showed that the Hoeffding classifier had the lowest performance for the classification of the dataset used for this work.

Attribute Ranking

In conducting this research work, there was need to determine the most determining attribute for the corresponding output. Of the algorithms used, attention was given to the trees generated by the decision trees. The various algorithms that generated a decision tree considered are the random tree, J48 and the LMT respectively. From the various decision tree, the topmost node is tutorial attribute. Which gave a view that the tutorial is the most dominating attribute in predicting the performance of a student.

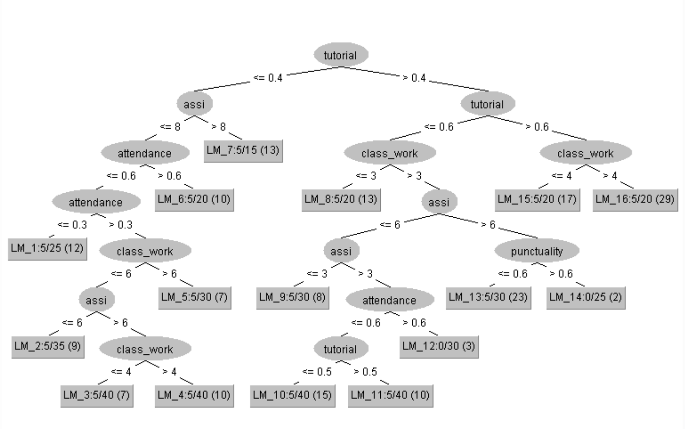


Fig 6: Decision tree generated by LMT classifier

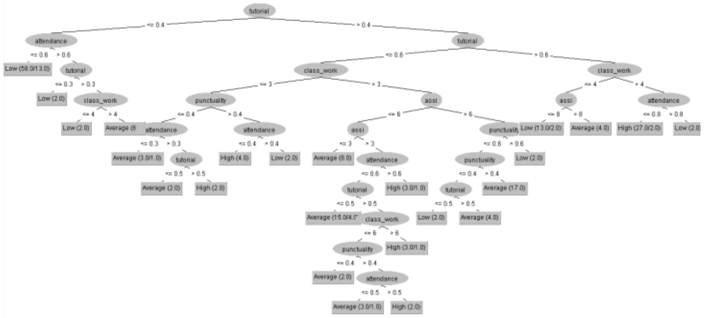
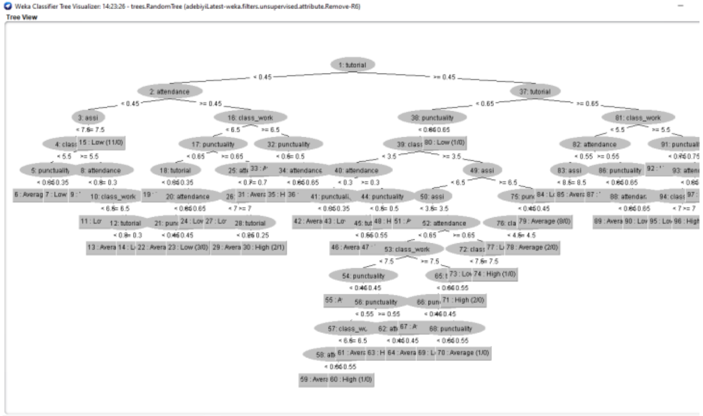


Fig 7: Decision tree generated by J48 classifier


Weka Classifier Tree Visualizer: 14:23:26 - trees.RandomTree (adebiyi.latest-weka.filters.unsupervised.attribute.Remove-R6)

Tree View

Fig 8: Decision tree generated by Random tree classifier

An attribute ranking was performed using the relief F attribute evaluation. It returned the tutorial as the highest determining attribute next to it was punctuality and then attendance. This was the case for predicting the nominal values which are (high, average and low) which was done in experiment 2.

In predicting the numeric class which was in experiment 1, the relief F attribute evaluation was used to rank the attributes. It returned the tutorial attribute as the highest determining attribute followed by attendance attribute and the class work attributes.

V CONCLUSIONS

An all-inclusive model is required in predicting the performance of student in any academic institution. The need to identify factors which are very key in enhancing students' performance is therefore a task worth taking cognizance of. This work considered a model for predicting the performance of students using both nominal and numeric output based on supervised learning approach. The classifier used are the J48, RMT, Random Tree, Random Forest and Hoeffding. The classifiers used in predicting the numeric class are random tree and random forest other classifiers performed poorly while Bayes classifiers could not handle numeric classification.

This work only considered the dataset of a department in the selected institution. Subsequent work will enhance the models in more departments to verify the claims the model was able to arrive at.

REFERENCES [1]. Cerny, P. A., & Proximity, M. A. (2001). Data mining and neural networks from a commercial perspective. In ORSNZ Conference Twenty Naught One, pp., 1–10.

[2]. Siri, A. (2015). Predicting students' dropout at university using artificial neural networks. Italian Journal of Sociology of Education, 7(2).

[3]. Bekele, R., & McPherson, M. (2011). A Bayesian performance prediction model for mathematics education: A prototypical approach for effective group composition. British Journal of Educational Technology, 42(3), 395–416.

[4]. Koutina, M., & Kermanidis, K. L. (2011). Predicting postgraduate students' performance using machine learning techniques. In Artificial intelligence applications and innovations (pp. 159–168). Berlin, Heidelberg: Springer.

[5]. Yadav, S.K., Bharadwaj, B., & Pal, S. (2012). Mining education data to predict student's retention: A comparative study. arXiv preprint arXiv:1203.2987.

[6]. Shahiri, A. M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques. Procedia Computer Science, 72, 414–422.

[7]. Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). Predicting student dropout in higher education. arXiv preprint arXiv:1606.06364.

[8]. Ramaswami, M., & Bhaskaran, R. (2010). A CHAID based performance prediction model in educational data mining. arXiv preprint arXiv:1002.1144.

[9]. Elaf Abu Amrieh, Thair Hamtini, and Ibrahim Aljarah. 2016. Mining educational data to predict student's academic performance using ensemble methods. International Journal of Database Theory and Application 9(8), 119-136.

[10].Adejo, O., & Connolly, T. (2017). An integrated system framework for predicting students' performance in higher education institution. International Journal of Computer Science and Information Technology (IJCSIT), 9(3), 149–157.

[11].Dorina Kababchieva. (2012). Student Performance Prediction using Data Mining Classification Algorithms. International Journal of Computer Science and Management Research, vol. 1.

[12].Camilo Ernesto Lopez Guarín, Elizabeth León Guzmán, and Fabio A. González. (2015). A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining IEEE Transactions, vol. 10(3)

[13].Sana Akhai and Ruchi Karia. (2017). Automated Performance Evaluation System.IJARIIT,3(2).

[14].Bilal Mehboob, Rao Muzamal Liaqat andNazar Abbas Saqib. (2016). Predicting Student Performance and Risk Analysis by Using Data Mining Approach. International Journal of Computer Science and Information Security (IJCSIS), vol. 14, No 7.

[15].Mrinal Pandey and S.Tarun. (2014). A Multi-Level Classification Model Pertaining to the Students' Academic Performance Prediction. International Journal of Advances in Engineering & Technology,4**:** 1329-1341.

A Cryptosystem Using Two Layersof Security – RNS and DNA Cryptography

Logunleko, Abolore Muhamin
Department of Computer Science
Gateway ICT Polytechnic,
Saapade, Ogun State, Nigeria

Gbolagade, Kazeem Alagbe
Department of Computer Science
Kwara State University,
Malete, Ilorin, Nigeria

Lawal, Olanrewaju Olaide
Dept. of Computer Engineering
Moshood Abiola Polytechnic,
Abeokuta, Ogun State, Nigeria.

Logunleko, Kolawole Bariu
Dept. of Computer Sci. & Statistics
DS Adegbenro ICT Polytechnic,
Eruku-Itori, Ogun state, Nigeria

Isiaka, R.M
Department of Computer Science
Kwara State University,
Malete, Ilorin, Nigeria

Oyekunle, Olurotimi Olufunso
Department of Computer Science
Gateway ICT Polytechnic,
Saapade, Ogun State, Nigeria

abolore.logunleko@gaposa.edu.ng, logunleko.kolawole@dsadegbenropoly.edu.ng, kazeem.gbolagade@kwasu.edu.ng, abdulrafiu.isiaka@kwasu.edu.ng, lawal.olanrewaju@mapoly.edu.ng, olurotimi.oyekunle@gaposa.edu.ng

Abstract— Residue Number System RNS, plays a major role everywhere including, error detection and correction, modularity, cloud computing, digital signal processing and data security. Many new encryption methods are being developed every day and a lot of research is being made towards finding a good cryptographic algorithm. Though modular arithmetic algorithms such as* Rivest, Shamir and Adleman, *RSA have been providing satisfactory level of security, but still there is need for adding more cryptography schemes. In this paper, we propose a RNS based algorithm which acts as an additional layer around the Deoxyribo Nucleic Acid, DNA algorithm. This research merges some features of the two algorithms to make a hybrid algorithm and the result revealed that it was found to be better cryptographic algorithm and less vulnerable to attacks. The algorithms were developed and implemented using python programming language. This enhanced data security using RNS and DNA sequence can be applied to secure SMS communication system and email system. Hence, security of private and confidential data via SMS and email could be adequately guaranteed.

Keywords— ***Cryptography, DNA Sequence, Chinese Reminder Theorem CRT, Residue Number System RNS, Cipher, Hybrid Algorithm***

I. INTRODUCTION

In recent times, much research works have been done on Residue Number System RNS, A lot of researches in computer science are enthusiastic to go through residue numbering system because of its characteristics such as, error detection and correction, modularity, and embedded parallelism [7]. Residue Number System is a non-weighted number system which is very much different from the weighted number system such as binary number systems or decimal number systems [3]. A Residue Number System (RNS) represents a large integer using a set of smaller integers, so that computation may be performed more efficiently. RNS speeds up arithmetic operations by dividing them into smaller parallel operations. [3][7]

A Residue Number System is characterized by a moduli set$\{\mathrm{m}_1\ \mathrm{m}_2\ \mathrm{m}_3\ \dots\dots\dots.\mathrm{m}_\mathrm{l}\}$, where the modulo, m_i,(i = 1, 2, …L), are pair wise relatively prime [4] [5]. Any integer X in the dynamic range, $\mathrm{M} = \mathrm{m}_1\ \mathrm{m}_2\ \mathrm{m}_3\ \dots\dots\dots.\mathrm{m}_\mathrm{l}$ is represented by an L-tuple $\left(\mathrm{x}_1\ \mathrm{x}_2\ \mathrm{x}_3\ \dots\dots\dots.\mathrm{x}_{\mathrm{l}-1}\ \mathrm{x}_\mathrm{l}\quad\right)$ where, x_i is the residue of X in modulo m_i for i = 1, 2, ., L.

An integer X is represented by an L-tuple where, x_i is a nonnegative integer satisfying. Thus,

$X = m_i\, q_i + x_i$ and $0 \leq x_i<m_i$.

The residues can be represented as:

$$x = |X|_m \tag{1.0}$$

Residue arithmetic operations like addition, subtraction, and multiplication are inherently carry-free, i.e., each digit of the result is a function of only one digit from each operand, thereby independent of all other digits. This feature helps in considerable enhancement on the processing speed, which is the major criterion in cryptography.
Cryptography provides techniques for keeping information secret, for determining that information has not been tampered and for determining who authored pieces of information [8]. In cryptography, encryption is the process of encoding a message or information in such a way that only authorized parties can access it and those who are not authorized cannot. Encryption does not itself prevent interference, but denies the intelligible content to a prospective interceptor. In an encryption scheme, the intended information or message, referred to as plaintext, is encrypted using

an encryption algorithm (a cipher) generating cipher text that can be read only if decrypted[13][14]. For technical reasons, an encryption scheme usually uses a pseudo-random encryption key generated by an algorithm. It is in principle possible to decrypt the message without possessing the key, but, for a well-designed encryption scheme, considerable computational resources and skills are required. An authorized recipient can easily decrypt the message with the key provided by the originator to recipients but not to unauthorized users. This approach has been extended by Lipton (1999) to solve another NP-complete problem, which is the satisfaction problem. In the following researchers, scientists find that the vast parallelism, exceptional energy efficiency and extraordinary information density are inherent in DNA molecules. DNA computing provides a parallel processing capability with molecular level, introducing a fire-new data structure and calculating method. It can simultaneously attack different parts of the computing problem put forward challenges and opportunities to traditional information security technology. For example, in 1995, Boneh et al. demonstrated an approach to break the Data Encryption Standard (DES) by using DNA computing methods. In 1999, Clelland et al. achieved an approach to steganography by hiding secret messages encoded as DNA strands among a multitude of random DNA. DNA cryptography is a newborn cryptographic field emerged with the research of DNA computing, in which DNA is used as information carrier and the modern biological technology is used as implementation tool. The vast parallelism and extraordinary information density inherent in DNA molecules are explored for cryptographic purposes such as encryption, authentication, signature, and so on.[10][11][12]

Generally DNA is represented using a set of bases which can be presented in sequential manner. There are four such bases namely Adenine, Guanine, Cytosine, Thymine and they are represented as A, G, C and T respectively [12][13][14]. A message consisting of a set of characters can be represented as a sequence of DNA bases A, G, C and T. The DNA structure was created by Watson and Crick in the year 1953 [15]. DNA Sequencing technology is the order in which the four bases are arranged and even a single change in the position of a base in a DNA sequence forms a totally different DNA [6]. DNA has two strands and there exist linkages between one strand to another. Thus it is said to have a Double helix structure. The bonds are formed between the bases in such a way that thymine always bonds with adenine and guanine bonds with cytosine. Apart from the bases in the DNA structure, it also has phosphate and sugar molecules. The DNA structure can be considered as a ladder like structure in which the phosphate and sugar molecule would be on the side and the basis would act as the rungs. One of the simplest ways to convert messaging to DNA form used by representing each character in its binary form and just mapping two bits to a DNA base (eg: 00 to A). There are various modern schemes of cryptography which ensures four securities attributes, namely; availability, confidentiality, integrity and authenticity. But there are speed, memory, compression and computational problems with these schemes. To overcome these problems, this study was carried out. The study aimed to integrate Residue Number System, RNS and Chinese Remainder Theory, CRT into Deoxyribo Nucleic Acid, DNA sequence to form a hybrid cryptosystem with high speed, high memory, and compressional and computational data encryption scheme.

The rest of the paper is arranged as follows: Section II provides some related works. Section III focuses on the proposed work. Section IV presents the results and discussions and section V concludes the paper.

II. RELATED WORK

Karimi and Haider (2017) [15] proposed an encryption and decryption algorithm based upon biological operations which take place in DNA molecule. The DNA operations such as transcription, replication, annealing, marking and mutation are used. The algorithm generates a set of keys using the user's password as an input. The user generated password ensures random key generation. First the password is converted into binary, and then the bits taken pairwise are encoded to nucleotides as follows 00-Adenine, 01-Guanine, 11-Cytosine and 10-Thymine. If the length of the data is not divisible by 3 (codon length) or if length of data is less than 60, the data is extended by DNA replication. Next DNA annealing is done to get double stranded DNA. Next the DNA is converted to mRNA by replacing Thymine (T) with Uracil (U). Next mutation of mRNA is done. Both nonsense and missense mutation is performed on the DNA strand. Next the mRNA is split into subparts depending on occurrence of the stop codons UAG, UAA and UGA. This results in generation of subkeys. The number of subkeys generated is random as it depends on the number of stop codons in the mRNA. The subkeys are converted into binary notation. Each subkey is grouped into 8-bit blocks. The 1st 8-bit block of input data is left shift 1 time and subkey1 is XOR with it. 2nd 8-bit block of input data is left shift 2 times and subkey2 is XOR with it. This is repeated for all the 8- bit blocks of input data to get final result. The encryption process is applied in reverse order to decrypt the message as it is a symmetric algorithm.

Zhang et al (2017) [16] revealed a solution to the generation of random keys required by one time pad encryption scheme and secure transmission. They propose the use of DNA molecule for generation and storage of the keys. They generate the keys from the DNA of the organism. This ensures its randomness. The secret key is then securely transported through a bacteria using recombinant DNA technology. The algorithm can be implemented in the biological DNA and bacteria with the current improvements in technology.

Kaur Karandeep (2016) [17] proposed a layered algorithm combining DNA and RSA cryptographic techniques. The DNA encryption is done with respect to a reference DNA strand from a genetic database which acts as a secret key. The DNA strand is converted to decimal values based on the sequencing of reference strand. The scheme is proposed for providing security in cloud infrastructure.

Saha and Haque (2017) [18] revealed an encryption algorithm based on DNA cryptography. They have used a dynamic mapping for encoding to DNA bases. They have also used operations such as Roll in encoding and data and key arrangement to improve its security.

Kalsi et. al. (2018) [19] discussed the concept of DNA deep learning cryptography to hide the ciphertext using deep learning and DNA cryptography techniques. They have also proposed method to generate keys using natural selection.

Aparna et. al. (2017) [20] discussed an audio steganography method which is encrypted using a combination of DNA cryptography and AES encryption schemes. Piracy detection of movie files is one of the applications for which their work can be used.

III. PROPOSED WORK

In the proposed paper, we have combined various efficient techniques used in the existing DNA cryptography algorithms and also combined them with RNS algorithm; Decimal to Residue Conversion and Chinese Reminder Theorem(CRT) which resulted in a RNS-DNA Hybrid Cryptography system. The input message is converted to decimal form and then RNS form. Then the output of the RNS form is converted to binary form and then DNA form. Thus this DNA sequence is the final cipher text which will be used in the communication.

A. Decimal to Residue Conversion

The process of conversion from conventional representation to RNS representation is called *Forward Conversion*. [9] Conceptually, this process can be done by dividing the given conventional number by all the moduli and finding the remainders of the divisions. This is the most direct way that can be applied to any general moduli-set. This is represented mathematically as:

$$r = |X|_m \qquad (1.1)$$

Where,

r = the reminder

X = decimal number

m = modulus

B. Reverse Conversion from RNS to Binary Representation

Reverse Conversion Algorithms are all based on either Chinese Remainder Theorem (CRT) or Mixed-Radix Conversion (MRC). The MRC is an inherently sequential approach. On the other hand, the CRT can be implemented in parallel. The main drawback of the CRT based R/B reverse converter, is the need of a large modulo adder in the last stage. The reverse conversion is one of the most difficult RNS operations and has been a major, if not the major, limiting factor to a wider use of RNS [1] [2]. In general, the realization of a Very Large Scale Integration VLSI implementation of R/B converters is still complex and costly. Here, we derive the mathematical foundations of the CRT and then we present possible implementations of these methods in reverse conversion.

Chinese Remainder Theorem

The statement of the Chinese Remainder Theorem (CRT) is as follows [1][2][3][4]:

Given a set of pair-wise relatively prime moduli $\{m_{1,} \ , m_{2\,,} \, m_3 \ , \dots \dots m_n\}$ and a residue representation $\{r_{1,} \ , r_{2\,,} \, r_3 \ , \dots \dots r_n\}$ in that system of some number X, i.e. $r_i = |X|_{m_i}$, that number and its residues are related by the equation:

$$|X|_M = \left|\sum_{i=1}^{n} r_i \left|M_i^{-1}\right|_{m_i} M_i\right|_M$$

(1.2)

Where is the product of the m_i's, and $M_i = M/m_i$.

C. Algorithm Process

Encryption Algorithm Process

Supply a plain text, s
s = "a0a1a2...an"
Get ASCII Number for each Character
asc = array[]
asci = ASCII(si)

```
Convert each ASCII Number (Decimal Number) to Residue Number r using Modulo Set
Convert each Residue Number r to 8-bits binary string and add padding where necessary
Binary = Binary(r)
Merge all Binary Strings accordingly:
Split the merged string to each of 2-bits Binary String:
Split2 = array[]
Split20 = "s0s1s2…s5"
Split21 = "s6s7s8…s11"
Split22 = "s12s13s14…s17"
… … …
… … …
Split2n = "snx2snx2+1snx2+2…snx2+5"
Convert each 2-bits Binary string to DNA Sequence:
split2n = DNA Sequence (split2n)
```

Decryption Algorithm Process

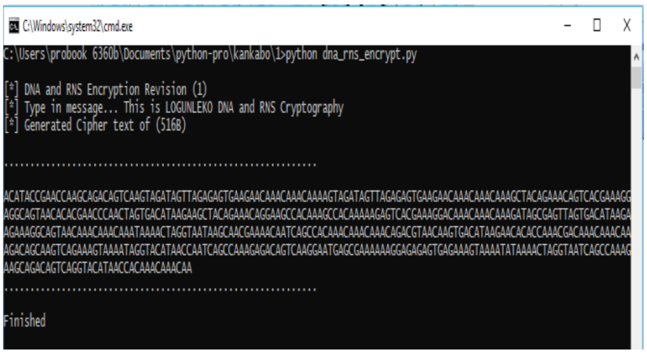
```
Convert C into Binary String B by replacing or substitute the DNA Nucleotides with their corresponding Binary Number
 Split B into 8-bits Binary String:
 Convert each of the 8-bits Binary String above to a Decimal Number D:
 D = D ( 8- bits Binary String)
 Pick the nth Three parts of D and apply Reverse Conversion or Backward Conversion Using Chinese Remainder Theorem CRT with Modulo Set :
 iD = CRT(D3n, D3n+1, D3n+2)
 Substitute each value from iD by the Corresponding Character in ASCII Table
 asci = ASCII(si)
 Combining S into one string gives the required Plaintext
```

IV. RESULTS AND DISCUSSIONS

The proposed system was developed using Python Programing Language, DNA Technology, and RNS. The system was able to encrypt and decrypt both intelligible and unintelligible messages respectively which were shown in figure1 and figure 2 respectively.

a) Encrypt/Encode Page

Figure 1 shows the encrypted message page. This page shows the plain text that was encrypted to a cipher text using RNS-DNA Hybrid Cryptography System.

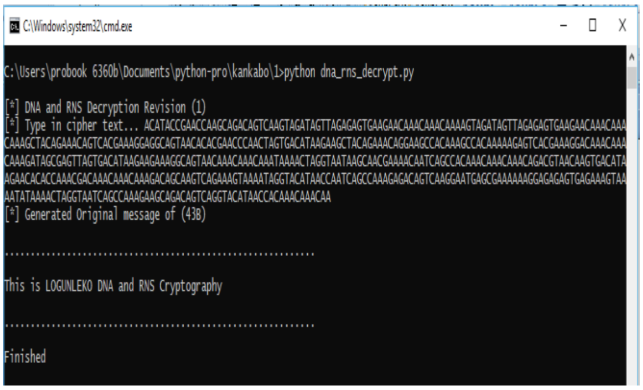


```
C:\Users\probook 6360b\Documents\python-pro\kankabo\1>python dna_rns_encrypt.py

[*] DNA and RNS Encryption Revision (1)
[*] Type in message... This is LOGUNLEKO DNA and RNS Cryptography
[*] Generated Cipher text of (516B)


..........................................................

ACATACCGAACCAAGCAGACAGTCAAGTAGATAGTTAGAGAGTGAAGAACAAACAAACAAAAGTAGATAGTTAGAGAGTGAAGAACAAACAAACAAAGCTACAGAAACAGTCACGAAAGG
AGGCAGTAACACACGAACCCAACTAGTGACATAAGAAGCTACAGAAACAGGAAGCCACAAAGCCACAAAAAGAGTCACGAAAGGACAAACAAACAAAGATAGCGAGTTAGTGACATAAGA
AGAAAGGCAGTAACAAACAAACAAATAAAACTAGGTAATAAGCAACGAAAACAATCAGCCACAAACAAACAAACAGACGTAACAAGTGACATAAGAACACACCAAACGACAAACAAACAA
AGACAGCAAGTCAGAAAGTAAAATAGGTACATAACCAATCAGCCAAAGAGACAGTCAAGGAATGAGCGAAAAAAGGAGAGAGTGAGAAAGTAAAATATAAAACTAGGTAATCAGCCAAAG
AAGCAGACAGTCAGGTACATAACCACAAACAAACAA
..........................................................

Finished
```

Figure 1: Encrypted Message

b) Decrypt/Decode Page

As shown in figure 2, this page shows the cipher text that was decrypted to a plain text using RNS-DNA Hybrid Cryptography system.



```
C:\Users\probook 6360b\Documents\python-pro\kankabo\1>python dna_rns_decrypt.py

[*] DNA and RNS Decryption Revision (1)
[*] Type in cipher text... ACATACCGAACCAAGCAGACAGTCAAGTAGATAGTTAGAGAGTGAAGAACAAACAAACAAAAGTAGATAGTTAGAGAGTGAAGAACAAACAAA
CAAAGCTACAGAAACAGTCACGAAAGGAGGCAGTAACACACGAACCCAACTAGTGACATAAGAAGCTACAGAAACAGGAAGCCACAAAGCCACAAAAAGAGTCACGAAAGGACAAACAAA
CAAAGATAGCGAGTTAGTGACATAAGAAGAAAGGCAGTAACAAACAAACAAATAAAACTAGGTAATAAGCAACGAAAACAATCAGCCACAAACAAACAAACAGACGTAACAAGTGACATA
AGAACACACCAAACGACAAACAAACAAAGACAGCAAGTCAGAAAGTAAAATAGGTACATAACCAATCAGCCAAAGAGACAGTCAAGGAATGAGCGAAAAAAGGAGAGAGTGAGAAAGTAA
AATATAAAACTAGGTAATCAGCCAAAGAAGCAGACAGTCAGGTACATAACCACAAACAAACAA
[*] Generated Original message of (43B)


..........................................................

This is LOGUNLEKO DNA and RNS Cryptography

..........................................................

Finished
```

Figure 2: Decrypted Message

V. CONCLUSION

The proposed work uses a hybrid scheme of RNS Algorithms and DNA processes which have shown the desired and good results. Thus the algorithm is proved to be cryptographically secured and it is suitable for applications where more than one layer of security is required. The RNS layer compresses the decimal number system to smaller values to hasten the conversion. Meanwhile, DNA layer adds confusion to the data so that it will be arduous for intruder to decrypt the message. In future, this algorithm can be improved further by adding more DNA processes such as converting DNA sequence to amino acids which will add to the security of the system.

REFERENCES

[1] Omondi, A. & Premkumar, B., (2007). *Residue Number System: Theory and Implementation.* Imperial College Press 2007, ISBN 978-1-86094-866-4.

[2] Neha, S., (2008). An overview of Residue Number System. *National Seminar on Devices, Circuits & Communication Organized by Department of ECE,* B.I.T, Mesra, Ranchi- 835 215

[3] Sharoun, A.O, (2013). Residue Number System (RNS). *Poznan University of Technology Academic Journals, Zawia University, Libya.* pp 265-270.

[4] Garner, H., (1959). The residue number system. *IEEE. Trans. Electron. Comput.,* 8: 140-147.

[5] Parhami, B., (2001). Computer Arithmetic: Algorithms and Hardware Designs. 1st Edn., Oxford University Press, Oxford, UK., ISBN: 0-19-512583-5.

[6] Abbasy, M. R. and Shanmugam, B. (2011). Enabling Data Hiding for Resource Sharing in Cloud Computing Environments Based on DNA Sequences. *IEEE World Congress on Services (SERVICES), Washington DC,* pp. 385-390.

[7] Mansour, B., Andraws, S., Mazin, A., And Baha, R., (2016). A Binary To Residue Conversion Using New Proposed Non-Coprime Moduli Set", Signal & Image Processing. *An International Journal (SIPIJ)* Vol.7, No.3, June 2016. DOI : 10.5121/Sipij.2016.7301

[8] Saheed, Y.K and Gbolagade, K.A, (2017). Efficient RSA Cryptosystem Decryption Based on Chinese Remainder Theorem and Strong Prime. *Annals. Computer Science Series,* 15th Tome 2nd Fasc, Vol 15.

[9] Aremu, I.A., Gbolagade, K.A., (2017). An overview of Residue Number System. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)* Volume 6, Issue 10, October 2017, ISSN: 2278 – 1323 1618.

[10] Vinay, J. and Sonam, B., (2016). Various Cryptographic Techniques: An Overview. *International Journal of Engineering Science and Computing,* Volume 6 Issue No. 11

[11] Beenish, A., Kazi, S.M., Alamgir, H., Keshav, D., (2010). Review on the Advancements of DNA Cryptography, arXiv:1010.0186v1 [cs.CR] 1.

[12] Shanmugasundaram, G., Thiyagarajan, P., and Pavithr, S. (2015). A Novel DNA Encryption System Using Cellular Automata. *International Journal of Security, Privacy and Trust Management (IJSPTM)* Vol 4, No 3/4, DOI: 10.5121/ijsptm.2015.4404

[13] Sridevi, R., and Karthika, S., (2015). Secured Image Transfer Through DNA Cryptography Using Symmetric Cryptographic Algorithm. *International Journal of Engineering Research and science & Technology*. Vol. 4, No. 2, May 2015, ISSN 2319-5991.

[14] Ahsan, O. and Muhammad, I. F., (2015). *DNA Cryptography Algorithms and Applications,* 14 MS-EE-015 and 14-MS-EE-113.

[15] Karimi, M. and Haider, W. (2017). Cryptography using DNA Nucleotides. *International Journal of Computer Application,* Vol.168, pp.16-18.

[16] Zhang, Y., (2017). DNA based random key generation and management for OTP encryption. *BioSystems,* Vol.159, pp.51– 63.

[17] Kaur, K., (2016). "A Double Layer Encryption Algorithm based on DNA and RSA for Security on Cloud. *International Research Journal of Engineering and Technology,* Vol.03, No.03, pp.1742-1745.

[18] Haque, R. and Saha, R. (2017). A novel Rolling based DNA Cryptography. *Journal of Bioinformatics and Genomics,* Vol.1, No.3, pp.1-6.

[19] Kalsi, Shruti, HarleenKaur, and Victor Chang,(2018). DNA Cryptography and Deep Learning using Genetic Algorithm with NW International Journal of Pure and Applied Mathematics Special Issue algorithm for Key Generation. *Journal of medical systems,* Vol. 42, No.1, pp.17.

[20] Aparna, A. A, Juvin, V.C.B and Kodakara, S. C. E. T. (2017). Video Piracy Detection Based on Audio Steganography, AES and DNA Cryptography. *International Journal of Engineering Science*, Vol. 7, No.3, pp. 5487-5489.

The Effects of Cyber Bullying on The Academic Performance of Students in Nigerian Tertiary Institutions

Chioma Chigozie-Okwum
Department of Computer Science,
University Nneochi.
chiomaokwum@gmail.com

Peter Ezeanyeji
Department of computer science,
Anambra State university, Uli.
peter.ezeanyeji@coou.edu.ng.

Ijeoma Okoronkwo
Computer Professionals Spiritan
Registration Council of Nigeria.
Ijayronk@yahoo.com

.

***ABSTRACT*— The growing rate of internet usage and online presence have brought a massive growth in cyber bullying. The study attempted to identify the effects of cyber bullying on the academic achievements of students in Nigerian. The study adopted a quantitative research methodology where a survey was carried out. The population of the study was undergraduate students in Nigeria, while 100 undergraduates were purposively sampled from the Imo state university Owerri. Structured questionnaires were used as instrument for data collected and was validated by measurement and evaluation experts. Results collected were analyzed using descriptive statistics made up of charts and frequency tables. Results showed that majority of the sampled respondents have been bullied at least once since they started using the internet and the forms of cyber bullying included trolling defamation of character, among others. Factors identified as promoters of cyber bullying include but not limited to lack of prosecution of offenders and poor awareness. The study identified abysmal drop in student grades as the most dangerous effect of cyber bullying on the academic achievement of students and recommends the government security agencies to take prosecution and punishment of cyber bullies seriously.**

Keywords: Cyber Bullying, Academic, Achievements, Tertiary Institutions, Technology.

I INTRODUCTION

The growth of the internet has many advantages as it affects live and work positively, saving time and easing out energy used in carrying out tasks. However proliferation of internet and online activities have promoted some social issues top among which is cyber bullying. Cyber bullying in schools are causing unimaginable problems to students, parents and to educational institutions. People can easily attack and bully people they know and strangers as well as using anonymous platform available to them. Unfortunately little has

been done to help those victims who are continuously trapped in the name of modernization

and digitalization [1]. This continuous neglect and lack of prosecution of bullies have led to serious incidents of cyber bullying of children and adults and these have had tragic consequences. The problem with cyber bullying is that it is very difficult to detect the culprit, as they hide under false identities covering their trails as they perpetrate their crimes.

Researchers agree that the stressful impact of bullying is escalating with the rapid expansion of technological advancement [2]. Many students who are victims of cyber bullies suffer silently; they are reluctant to complain to the authorities or their parents owing to the social stigma attached to victims of bullies [3]. As such, parents, teachers and institutional authorities have to be on the lookout for such occurrences.

Under great emotional stress, victims of cyberbullying are unable to concentrate on their studies, and thus their academic progress is adversely affected [4]. Since the victims are often hurt psychologically, the depressive effect of cyberbullying prevents students from excelling in their studies [5].

The aim of this study is to identify the effects of cyber bullying on the academic achievements of students in tertiary institutions in Nigeria. To achieve this aim the study attempts to satisfy these specific objectives:

1. actors contributing to the proliferation of cyber bullying in Nigerian tertiary institutions
2. Identify the effects of cyber bullying on the academic achievements of students in Nigerian tertiary institutions.

The study attempted to provide answers to the following research questions:

1. What are the factors promoting the proliferation of cyber bullying in Nigerian tertiary institutions?
2. What are the effects of cyber bullying on the academic achievement of students in Nigerian tertiary institutions.

II RELATED LITERATURE

2.1 Conceptual framework

Cyberbullying is defined as the electronic posting of mean-spirited messages about a person (such as a student) often done anonymously [6]. [7] define cyberbullying as online exchanges with the intention to harm the recipient. Cyberbullying involves the use of information and communication technologies such as e-mail, cell-phone and pager text messages, instant messaging, defamatory personal web sites, blogs, online games and defamatory online personal polling web sites, to support deliberate, repeated, and hostile behaviour by an individual or group that is intended to harm others." Characteristics like anonymity, accessibility to electronic communication, and rapid audience spread, result in a limitless number of individuals that can be affected by cyberbullying, [8].

According to [9], the most frequent and common media within which cyberbullying can occur are:

a. Electronic mail (email): a method of exchanging digital messages from an author to one or more recipients.
b. Instant messaging: a type of online chat that offers real-time text transmission between two parties.
c. Chat rooms: a real-time online interaction with strangers with a shared interest or other similar connection.
d. Text messaging (SMS): the act of composing and sending a brief electronic message between two or more mobile phones.
e. Social networking sites: a platform to build social networks or social relations among people who share interests, activities, backgrounds or real-life connections.
f. Web sites: a platform that provides service for personal, commercial, or government purpose.

According to [10] the following are some symptoms of victims of cyber bullying:
1. Anxious, insecure, unhappy and have low self-esteem
2. Cautious, sensitive, quiet, withdrawn and shy
3. Depressed and engaged in suicidal ideation much more often than their peers
4. Do not have a single good friend to discuss problems.
5. Often physically weaker than their peers in the school.

The following are some characteristics of cyber bullies:
1. often involved in other antisocial activities such as drug use.
2. Impulsive and easily angered
3. Strong need to dominate other students.
4. Show little empathy toward students who are victimized
5. Often physically stronger than other students in the class.
6. Often defiant and aggressive, including to parents and teachers.

Cyberbullying is a modern form of traditional or offline bullying and similar motives that instigated traditional bullying are now manifested online through text messages or social networking sites. Cyberbullying appears in various forms. [11] has provided the following list: sending threatening messages, spreading rumours, attacking someone verbally, intentionally excluding someone from the group, pretending to be someone else, publicizing unflattering pictures of a person, circulating sexually suggestive pictures and sharing confidential information online.

Different studies suggest that undergraduate students' use of the Internet is more significant and frequent than any other demographic group. A 2014 survey of 1006 participants in the U.S. conducted by the Pew Research Centre revealed that 97% of young adults aged from 18 to 29 years use the Internet, email, or access the Internet via a mobile device. Among them, 91% were college students [9].

2.2 Theoretical framework

Theories highlighting cyber bullying and its effects on society are discussed below:

Karl Marx in his **conflict theory** identifies levels and strata among teens. The levels identified are the popular kids at the top level and the losers at the bottom of the ladder. This strata introduces conflict among the popular kids and the losers. When we look at this from the perspective of cyber bullying, the conflict theory would say that conflict happens between these social levels and can cause cyber bullying. Teens will do what they need to do in order to increase their status. The actual cyber bullying that takes place signifies a social conflict that is unsolved and power that is unbalanced. Conflicts seem to arise when status and power are unevenly distributed between groups or cliques. The cyber bullies that are in peer repress the poor teens that are at the bottom of the ladder and they do this to maintain the status quo, basically so they stay on top. It doesn't seem to matter who they hurt along the way.

Labelling Theory as developed by Howard Becker is based on the fact that people's negative behaviours are deviant only because society labels them thus. This simply means that the labels teens are given impact their own and other's perceptions of them. This makes their behaviour deviant according to Howard. This theory also believes that a teen is not bad because of their actions, but rather negative behaviours are developed because people negatively judge them thus. Nobody is born bad, but social aspects of their peers constantly labelling them as bad influence their behaviours. This theory buttresses how the self-identity and behaviour of individuals may be determined or influenced by the terms used to describe or classify them. Terms like “loser, nerd, geek, slut etc.” used to label bullied individuals influence their behaviours. Once teens are given this label, they tend to continue to live up to the name. In line with cyber bullying, bullies at times try to become what they are labelled and in this attempt begin to bully others to satisfy themselves.

Social Disorganization Theory states that crime and deviant behaviours are more likely to happen in a social institution that is unable to control groups of people. According to the social disorganization theory of social problems, the rapid changes in the media and technology have disrupted the norms in society. Facebook, Instagram, cell phones and other social networks have developed so quickly and have basically taken over. Teens have access to the social media all day while they are at school, so cyber bullying is taking place in this setting all day long. This theory would say that cyber bullying is an issue at the school level and the cyber bullying is a sign of the disorder of the behaviours and attitudes at a larger level.

Emile Durkheim, describes the **anomie theory** as the condition in which society has not provided much, if any moral guidance to individuals. He see it basically as a mismatch between a teen’s actions and the social norms of the society and that teens are a product of their environment. If we look at the anomie theory from the perspective of our school, this makes sense. Cyber bullies torment their victims and this makes it even harder for the victim to fit in to the school environment. This can lead to disastrous consequences. Who are these teens that are cyber bullying? Where are the parents when this is happening? Who is teaching these teens that it is morally not ok to cyber bully? Somewhere along the line, these teens have not been morally guided by our society, community and perhaps family. This has led to disastrous impacts for those that are being bullied.

2.3 Empirical studies

Research results from previous empirical studies reviewed are provided below:

[12,13,4] all suggested in their findings that cyberbullied victims generally manifest psychological problems such as depression, loneliness, low self-esteem, school phobias and social anxiety.

[14] discovered that victims of cyber bullying often suffer psychosocial di*ffi*culties.

[5] investigated the emotional and physiological effects of cyber bullying on the university students. The results of the study indicated that a significant number of the respondents 35 (13%) had suffered emotionally due to cyber bullying. Furthermore, 300 (85%) of the respondents indicated that in their views cyber bullying causes emotional and psychological stress. Further, majority of the respondents 255 (70%) agreed that cyber bullying adversely affects students’ academic performance. Results further designate that 60 (16.6%) of the respondents specified that they had bullied someone inside the university at least 2 or 3 times a month while 4 (1.1%) of the participants said that they had bullied someone outside the university at least 2 or 3 times a month. It is interesting to note that majority of the students 75 (20.8%) signify that they have heard bullying taking place inside the university.

[15] in their research *fi*ndings submitted that cyberbullying causes emotional and physiological damage to defenceless victims.

[16] identified behaviour problems such as, drinking alcohol smoking, depression, and low commitment to academics as effects of cyber bullying on students.

[9] investigated the in*fl*uence of cyberbullying on the academic, social, and emotional development of undergraduate students. The study revealed that 57% of the students had experienced cyberbullying at least once or twice through different types of media. Correlation analyses were conducted and con*fi*rmed signi*fi*cant relationships between cyberbullying, mainly through instant messaging, and the academic, social and emotional development of undergraduate students. Instant messaging (IM) was found to be the most common means of cyberbullying among the students. The study concluded that although cyberbullying existence has been proven, studies of cyberbullying among undergraduate students have not been fully developed. The results of the study indicated that cyberbullying has an in*fl*uence on the academic, social, and emotional development of undergraduate students.

[17] observed that more than half of the participants in their study experienced a wide variety of cyberbullying, sexual offence being the highest. They were negatively affected both emotionally and academically to the extent that some thought of suicide. It was also observed that even though students in this rural high school have access to the latest cyber technology, they are not equipped to prevent or cope with its negative effects; hence, they suffer in solitude. The study recommends an anti-cyberbullying policy being established, as well as counselling at school, and advises stakeholders who intend to expand e-learning at schools to include cyber safety and supportive mechanisms in their programmes for successful implementation.

III METHODOLOGY

The study adopted a quantitative research methodology, where a survey was carried out in the month of October 2019. The population of the survey were undergraduate students in Nigerian tertiary institutions. 120 respondents were purposively sampled at the Imo state university Owerri. The instrument for data collection was a questionnaire which was validated by 2 measurement and evaluation experts at the faculty of education of the Imo State University. Out of the 120 questionnaires distributed 100 were properly filled out and returned hence making the sample size 100. Out of the 100 sampled respondents 62 were female while 38 were male undergraduates. The results obtained thereof were analysed using descriptive statistics.

IV RESULTS AND DISCUSSION

The result of the survey are presented and discussed below.

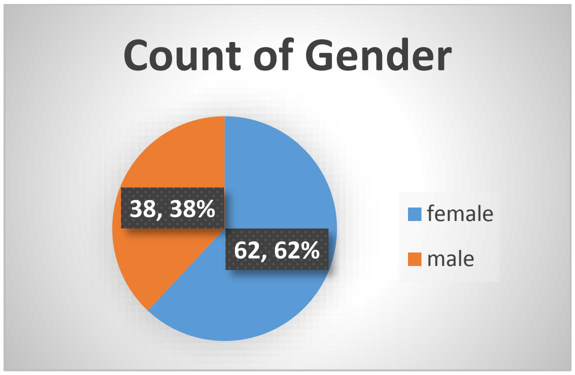


Figure 1: Gender Distribution of respondents

The results as presented in figure 1 above shows that 62% of the sampled respondents were female while 38% were male.

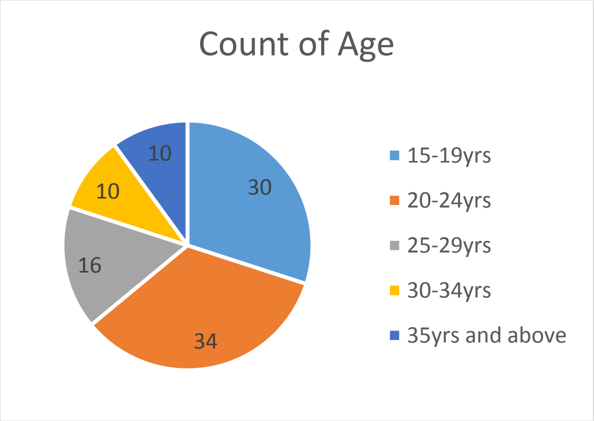


Figure 2: Age Distribution of respondents

The age distribution of the respondents as shown in figure 2 above show that 30 of the respondents were within the ages of 15-19years of age, 34 respondents fell between the 20-24years age range, 16 respondents were of the 25-29 years age range, 10 respondents were aged between 30-34years and 10 respondents were aged 35years and above.

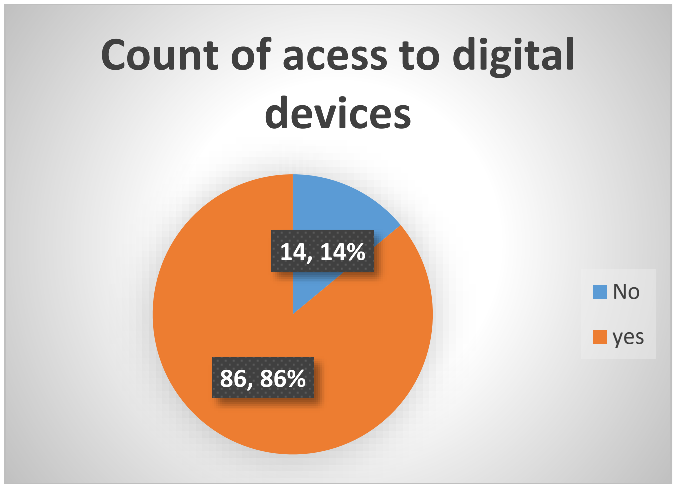


Figure 3: Respondent's access to digital devices

The results as shown in the figure 3 above shows that out of the 100 sampled respondents, 86 had access to digital devices while a minute 14% (14) did not have access to digital devices. The sampled respondents owned smart phones, tablets and even computers.

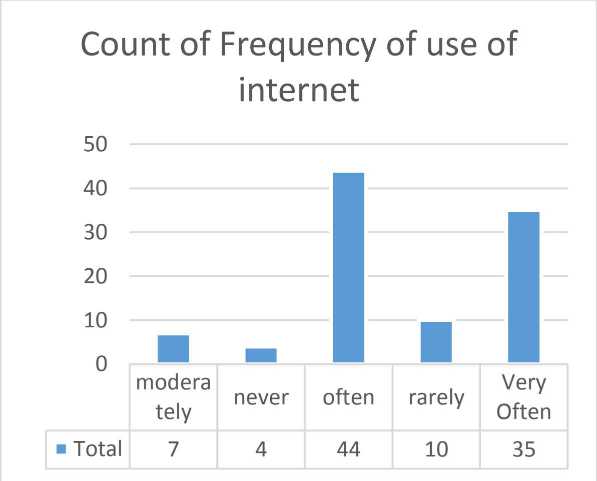


Figure 4: Frequency of use of the internet by respondents.

The survey results showed that 35 respondents visited the internet very often, 44 of the respondents often visited the internet, 7 respondents were neutral and reported the moderately visited the internet, 10 respondents rarely visited the internet and 4 respondents reported they never visited the internet.

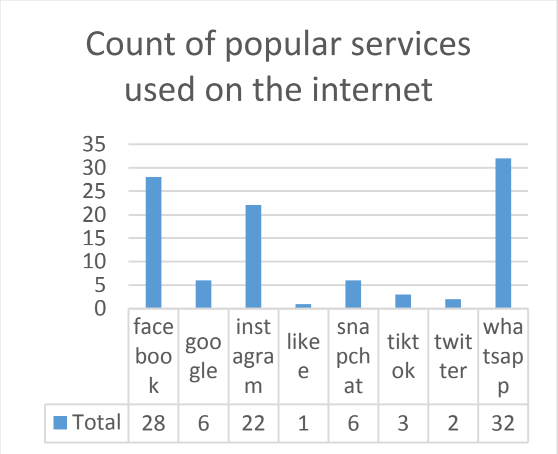


Figure 5: Popular services used on the internet by respondents

The results from figure 5 above shows that 32 of the respondents reported they used Whatsapp the most of all internet services, 28 respondents often used Facebook, 22 respondents used Instagram the most , while 6 respondents used google search engine the most. Another 6 respondents often used snapchat, 3 respondents used Tik Tik video posting app more, and 2 respondents were always on twitter while 1 respondent was always on Likee.

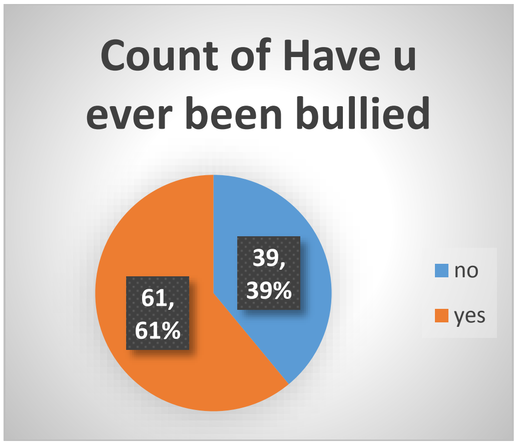


Figure 6:Respondent count on being bullied

The survey results presented in figure 6 above, shows that 61% of the sampled respondents have been bullied online either at least once since they started using the internet, while 39% of the respondents reported they have not directed being bullied before albeit they have witnessed friends, family and even random strangers being bullied.

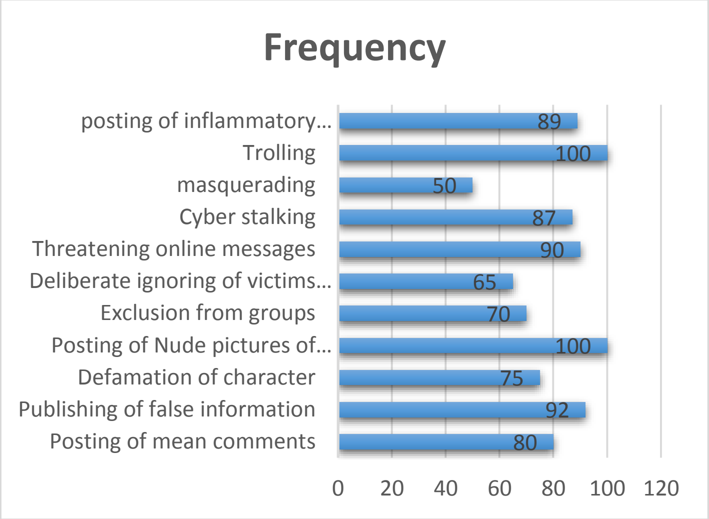


Figure 7: Forms of Cyber Bullying

The sampled respondents identified various forms they have either being bullied online or witnessed others being cyber bullied. These variants include Trolling (100%), where total strangers make demeaning and hurtful comments on posts made, posting of nude pictures of victims (100%), Publishing of false information about victims(92%), Cyber stalking (87%). Cyber stalking is not just a dangerous approach used by bullies but can also lead to physical crimes against victims. Other forms of cyber bullying identified included posting of inflammatory comments about victims (89%), these inflammatory comments are capable of sparking off series of negative events against the victim. Posting of threatening online messages (90%), posting of mean and insensitive comments (80%), defamation of character of victims (75%) and masquerading or hiding under fake aliases to taunt a victim (50%) were also identified in the study. The study also identified exclusion of victims from groups they belonged to as a way of cyber bullying and finally deliberately ignoring victim's comments to make them feel inferior as another way victims were bullied online.

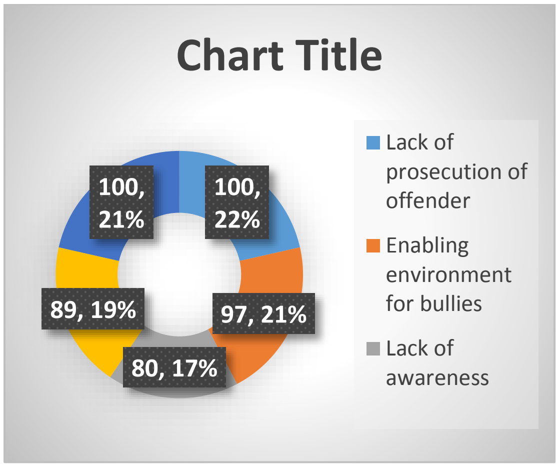


Figure 8: Factors Promoting Cyber bullying

The results as seen in figure 8 above identifies the following factors as being the promoters of cyber bullying especially in tertiary institutions in Nigeria. Factors identified include Lack of prosecution and punishment of offenders, shaming of the victims by the society hence boosting the ego of the bullies, enabling environment for the bullies as they are encouraged by others sharing and spreading their posts to the detriment of the victims. This is in line with the Social Disorganization Theory states that crime and deviant behaviours are more likely to happen in a social institution that is unable to control groups of people. Our society does not have control and hardly punish offenders hence the proliferation of this malaise. Other factors identified included carefree attitude of the society as no one cares about the feeling of the victims and finally lack of awareness; many do not know what constitutes cyber bullying, and are not even aware they are committing crimes against humanity and society and hence continue in that regard. This finding supports the anomie theory, as people do not even have the moral knowledge that they are breaking the law hence the growth of the crime.

Table 1:Effects of Cyber Bullying on the academic achievement of students

Effects of Cyber Bullying on Victims	Frequency
Total Abandonment of school	60
Skipping Classes	85
Loss of attention during lectures	98
Abysmal drop in grades	100
Repetition of classes	90
Overall Depression	92
Suicidal Thoughts	87
Alienation from Society	70

The study identified the direct effects cyber bullying have on the academic achievements of students top on the list included abysmal drop in student grades which is the lethal effect of cyber bullying on students. Other effects identified included, skipping of classes, loss of attention during classes, repetition of classes, overall depression, suicidal thoughts and total alienation from society.

V CONCLUSION/RECOMMENDATIONS

Cyber bullying is as old as the internet itself and it keeps evolving and morphing to more complex forms. The danger in cyber bullying is the anonymity of the crime as usually the bully can operate under anonymity to evade prosecution. The study shows that a greater percentage of the sampled respondents have been bullied, and the forms adopted by the bullies ranges from trolling to defamation of character among others. The greatest enabler of cyber bullying is the fact that the bullies are never punished to serve as deterrent to other like them and this leads to drastic consequences; coupled with the lack of awareness as to what constitutes cyber bullying. The greatest consequence and effect of cyber bullying on the academic achievement of students is a drastic drop in their grades which is caused by depression, skipping classes and even attempts to avoid everyone. Cyber bullying is growing with growth in access to the internet and if not checked could pose permanent damages and challenges to students and society at large.

The study hence recommends:

1. The law enforcement agencies should mount surveillance on cyber bullies and respond promptly when reported.
2. Cyber bullies should be openly prosecuted and punished to serve as deterrent to other.
3. Society should show more restraint in judging victims of cyber bullying and should desist from promoting bullying.
4. Sensitization campaigns should be launched to enlighten the society on what constitutes cyber bullying and what punishment it attracts.
5. Application developers should incorporate modules to detect and block cyber bullying.

REFERENCES

[1]. Watson, Scott E. J.; Vannini, N. Woods, Sarah; D., Kerstin; Sapouna, Maria; Enz, Sibylle; Schneider, W.; Wolke, D.; Hall, L.; Paiva, A.; Andre, E.; Aylett, R., 2010). Inter-Cultural Differences in Response to a Computer-Based Anti-Bullying Intervention, Educational Research, v52 n1 p61-80.

[2]. Patchin, Justin W.; Hinduja, Sameer, (2011). Traditional and Non-traditional Bullying among Youth: A Test of General Strain Theory. Youth & Society, v43 n2 p727-751 Jun 2011.

[3]. Susan., D.; Butler, C., W.; Emmison, M. (2011). "Have You Talked with a Teacher Yet?" How Helpline Counsellors Support Young Callers Being Bullied at School, Children & Society, v25 n4 p328-339.

[4]. Akcil, S., 2018. Cyberbullying-Victimization, Acculturative Stress, and Depression Among International College Students. Doctoral dissertation. Kent State University.

[5]. Qais Faryadi. Cyber Bullying and Academic Performance. International Journal of Computational Engineering Research ISSN: 2250-3005 / Vol 1, Issue 1 / Dec. 2011 Available Online through www.ijceronline.com.

[6]. Merriam-Webster, 2017. On-line Dictionary. https://www.merriam-webster.com/dictionary/cyberbullying.

[7]. Faucher, C., Jackson, M. & Cassidy, W., 2014, 'Cyber bullying among university students: Gendered experiences, impacts and perspectives', Education Research International 2014, Article ID 698545, 10 p. https://doi.org/10.1155/2014/698545.

[8]. Belsey, B., 2006. Cyber Bullying: an Emerging Threat to "Always on" Generation. From. http://www.cyberbullying.ca/pdf/Cyberbullying_Article_by_Bill_Belsey.pdf.

[9]. Yehuda Peled. Cyberbullying and its influence on academic, social, and emotional development of undergraduate

students. Heliyon 5 (2019) e01393. doi: 10.1016/j.heliyon.2019. e01393.

[10]. Olweus, D. (1996). Bullying In Schools: Facts and Intervention, Research Centre for Health Promotion, University of Bergen, Norway.

[11]. National Crime Prevention Council, 2011, Cyber bully statistics, viewed n.d., from http://www.bullyingstatistics.org/content/cyber-bully-statistics.html.

[12]. Grene, M.B., 2003. Counselling and climate change as treatment modalities for bullying in school. Int. J. Adv. Couns. 25 (4), 293e302.

[13]. Juvonen, J., Graham, S., Shuster, M.A., 2003. Bullying among young adolescents: the strong, the weak, and the troubled. Paediatrics 112 (6), 1231e1237.

[14]. Ybarra, M.L., Mitchell, K.J., 2007. Prevalence and frequency of internet harassment instigation: implications for adolescent health. J. Adolesc. Health 41, 189e195.

[15]. Akbulut, Y., Eristi, B., 2011. Cyberbullying and victimization among Turkish university students. Australas. J. Educ. Technol. 27 (7), 1155e1170.

[16]. Selkie, E.M., Kota, R., Chan, Y.F., Moreno, M., 2015. Cyberbullying, depression, and problem alcohol use in female college students: a multisite study. Cyber psychol. Behav. Soc. Netw. 18 (2), 79e86.

[17]. Farhangpour, P., Maluleke, C. & Mutshaeni, H.N., 2019, ‘Emotional and academic effects of cyberbullying on students in a rural high school in the Limpopo province, South Africa’, South African Journal of Information Management 21(1), a925. https://doi.org/ 10.4102/sajim.v21i1.925.

Comparative Analysis of Machine Learning Classifiers for Detecting Malware in Portable Executable

Faden David Nanven and Morufu Olalere
Department of Cyber Security Science
Federal University of Technology
Minna, Nigeria
nanvenfaden@gmail.com lerejide@futminna.edu.ng.

ABSTRACT
Over the years, malware vendors have evolved from using non intelligent malwares which are easily identifiable to intelligent malwares by employing polymorphism and metamorphism in malware behavior paving the way for evasive malware techniques ranging from environmental awareness, confusing automated tools, timing based evasion and obfuscated internal data. Modern malware detection techniques use machine learning algorithms mostly classifiers to detect malware signatures and malware behavior. Some of the machines learning algorithms are not effective in detecting malware behavior whereas some are. This research is a comparative analysis of commonly used machine learning classifiers ranging from Decision Tree, Random Forest and Bayesian Network. The training dataset comprises of 138,047 Portable Executable (PE) header file record samples which was divided into: 41,323 clean files containing exe and dll file samples and 96,724 malware file samples. This research evaluates the above-mentioned machine learning classifiers in relation to their performance using Machine Learning performance Metrics. Based on th experimental results of selected classifiers, the overall best performance was achieved by Gradient Boosting Classifier with accuracy of 98.5% and a Recall of 0.98 followed by both Random Forest and Decision Tree classifiers respectively.

Index Terms— Malware, Classifiers, Obfuscation, Polymorphism, Metamorphism.

I. Introduction

Malware is shorthand for malicious software. It was first introduced in the early 1970s when the creeper virus was introduced. Over the years we have seen multiple variants of malware running into well over 500 thousand malware variants which are all harmful to individuals and organizations that use the internet through electronic devices such as Portable executables and handheld devices. According to[1] the policy of bring your device (BYOD) has had a major impact on the security to organizational networks owing to the fact that BOYD policies paves the way for employees to come along with devices of their choice; chiefly PE's to enable the employee be efficient at work. Such policies has made PE's one of the major ways malware can be distributed within a secure network owing to the fact that non trusted devices are given acces to plug into an organizational network even when infected by malware which might be hidden to the user of the PE. Different techniques have been explored to mitigate and detect malware chiefly the use of machine learning algorithms mostly classifiers to evaluate the accuracy of a classification model.

Analysis of malware files are carried out in two ways either through static or dynamic techniques which are then classified into different malware families. Machine learning algorithms are used to predict and classify signatures based on features extracted from malware program code[2]. The features extracted from static malware analysis may range from byte sequence n grams, operational code (opcode) and syntactic library calls where function calls are checked to ascertain the libraries accessed by the functions[3]. Malware authors began to develop new ways to stealth the payload of a prevent Analysis of malware files are carried out in two ways either through static or dynamic techniques which are then classified into different malware families. Machine learning algorithms are used to predict and classify signatures based on features extracted from malware program code[2]. The features extracted from static malware analysis may range from byte sequence n grams, operational code (opcode) and syntactic library calls where function calls are checked to ascertain the libraries accessed by the functions[3]. Malware authors began to develop new ways to stealth the payload of a prevent ransomware malware are not so effective due to obfuscation of ransomwares. This obfuscated techniques used by malware authors to bypass static analysis paved way for the introduction and development of dynamic analysis of malware. In dynamic analysis of a malicious code, the behavior of the malware is monitored as it is executed in a sandbox which is a controlled environment[5][6]. The natural behavior of a malware can be observed without requiring the Portable Executable to be disassembled. This technique is more effective against evasive/obfuscated malware because it reveals the malwares running pattern before and after payload exposing the obfuscated behavior naturally[7].

Multiple research work has been carried out in the detection of malware using machine learning algorithms using classifiers the common ones being Decision Trees, Random Forest, Naïve Bayesian and Support Vector Machine (SVM) which are all classification algorithms. One major problem with some of the approaches used in detecting malware is the inability to detect evasive/obfuscated malware variants[8] but before training a dataset to detect unseen malware variants that a comparison of classification algorithms to ascertain the performance of each machine learning classifier should be carried out hence the motivation for this research work. We intend train several machine learning classification algorithms comparing their performance.

The remaining part of this paper is organized as follows: In section II, related works on malware classification with different machine learning classifiers are discussed. Methodology is discussed in section III. Results and analysis are given in section IV. Final conclusion of this work is discussed in section V.

II. RELATED WORK

Academic works have carried out comparative analysis of machine learning algorithms in detecting malware. The idea was to find the best algorithm with best accuracy and the least false positive rate either through a combination of several algorithms or by iterating through results using a single machine learning algorithm.

The comparative analysis by [9] was to find the best combination of machine learning algorithms to get the lowest false positive rate when classifying malware, their work was based on One Side Class (OSC) perceptrons algorithm which can detect malware samples with a low false positive rate, they achieved this through computing a set of malware features for every binary file in the training dataset which were trained with the OSC perceptrons algorithm using a dataset of clean and infected malware files furthermore cross validation was applied to the dataset to obtain the correct parameter values. They used a database of well over 2 million records where training features where extracted using both static and dynamic malware analysis techniques. The static features extracted includes file geometry, type of packer, type of compiler and executable flags while the dynamic features extracted during execution in a sandbox includes behavior such as if the Portable Executable (PE) clones itself on the disk, if it seeks permission to connect to the internet, or if it uses the concept of stealth to include itself in some system processes. The result provided the best detection rate although they couldn't get low false positive rates concluding that the OSC perceptron algorithm is best used with a method of false positive filtering. The above approach might not be feasible in detecting metamorphic malware based on the extracted features used for training. This research work will compare machine learning algorithms based on the lowest false positive rate using features from a simple metamorphic malware for our training set.

Another exhaustive survey on comparing heuristic malware detection methods was given by [10] where the researcher gave a description of three major malware detection methods commonly used namely signature based detection, behavioral based detection and heuristic based detection. The researcher was able to give advantages and disadvantages of each detection method and proffered reasons why heuristic malware detection technique is the most proffered detection method adopted by researchers against metamorphic malware. For example in a signature based detection which deals with identifiable patterns extracted from several malware files which are matched against known malware patterns already categorized into malware families will give low false positive rates making it the most preferred method adopted by antivirus merchants, the downside of this method is the inability to detect polymorphic and metamorphic malware variants. The second method was developed to mitigate the disadvantage of signature based malware detection; this method builds on the weakness of the signature based method through the concept of static and dynamic analysis where the behavior of the executable file is safely inspected by the use of sandboxes and Virtual Machines (VM's) so as to ascertain the hidden behavior of the file by allowing the file to execute in its natural environment[11]. It is worthy of note that machine learning algorithms are trained in any of the three ways namely Supervised Learning, Unsupervised Learning and Semi supervised Learning.

According to [12] supervised learning is the process of using the concept of classification where a machine learning algorithm known as classifiers map input features of a malware dataset to output labels which are already known. When the process is to map input features to a continuous output label it is known as regression. Accurate output is usually achieved from the training data; the end product is to learn a function that accurately approximates the relationship between input and output malware features.

This research adopts supervised learning to carry out a comparative analysis of the major classifiers used by previous research work based on the malware dataset. The major classifiers used in this research work for training and testing the malware dataset are Naïve Bayes Classifier, Random Forest, Aritificial Neural Networks (ANN), Support Vector Machine (SVM), Decision Trees, Multi-layer Perceptron, Quadratic discriminant, Linear Discriminant and Stochastic Gradient.

A. *Portable Executables (PE)*

Portable Executables is a file format used by Windows Operating System (OS) for executables, DLLs, executables, Object code, and .FON Font files. The PE file format contains the information necessary for the Windows OS loader to manage the enclosed executable code. Detection of malicious PE file from non malicious window files (benign) is key as PE file format is the widely used file format in Windows OS noting that the rate of computer systems that use the windows OS is on a global scale[11][12].

The filename extension and internet media file for PE's are

- Filename Extension: .acm, .ax, .cpl, .dll, .drv, .efi, .exe, .mui, .ocx, .scr, .sys, .tsp
- Internet Media File: application/vnd.microsoft.portable-executable

The Portable Executable (PE) file header contains the information that provides information regarding other data about the executable file. It is made up of a DOS stub, a signature, the architecture of the file's code, a time stamp, a pointer, and various flags. A very important feature in the PE header is the **Characteristics Header** and it contains flags that indicate the attributes of the file. Different flag values of Characteristics fields have information on different file characteristic. Another important feature is the Dynamic Link Library **(DLL) characteristics** and it contains information about the DLL behaviors which are mostly used by the linker and loader in the windows Operating System [15]. Other detection strategies on android application includes [29], [30], [31], and [32].

In this research, raw features are directly extracted from PE header field, and the selection of these header's fields are made on the basis of importance and relevance during feature selection phase before the training and testing phases.

B. *Machine Learning Classifier Algorithms*

Machine Learning classifiers are algorithms which map the input data to a specific category which allows the machine to learn from examples and experience without being explicitly programmed. Machine learning tools are provided in Python through a library named as scikit-learn, which can be conveniently applied on a malware dataset using a Python environment know as Jupyter Notebook. There are different types of classifiers; a classifier is an algorithm that maps the input data to a specific category. We shall now discuss several classifiers:

a) Naïve Bayes Classifier: Bayes theorem is based on the probability of an event based on past knowledge of particular conditions that might be similar or related to that event; Naïve Bayes classifiers are based on applying Bayes theorem with the assumption that features of measurement are independent of each other. It is a family of algorithms where all of the algorithms have a common principle in which every pair of features being classified is independent of each other. It works by predicting family probabilities for each class of feature such as the probability that a given data point belongs to a particular class. The class with the highest probability is seen as the most likely class. It is known to work very well with natural language processing problems.[16] made a research on a data mining framework which automatically detects malicious binaries. After feature selection was carried out on a data set consisting of of 4,266 programs broken down into 3,265 malicious binaries and 1,001 clean files; every example in the set was labeled either malicious or benign by the commercial virus scanner. The researcher compared signature based methods and several Classifier Algorithms such as RIPPER, Naive Bayes and Multi naïve Bayes on the extracted features to get the most accurate algorithm with least False Positives (FP) rate, for example the RIPPER algorithm was used on the dataset; which is a rule-based learner that builds a set of rules to identify the classes of either positive examples and negative examples while minimizing the amount of error.

b) Support Vector Machine (SVM): Support Vector Machine are learning models under supervised learning models that analyze data used for classification and regression analysis problems. SVM finds out the line in a hyper plane separating two defined classes. Support Vectors are simply the coordinates of individual observation. A good research on how Support Vector Machines are used was done by [17] in this research Support Vector Machines were used alongside decision trees to categorize malware. Support vector Machine was applied to the training dataset features to minimize the classification errors on a set of randomly selected samples to attain the best classification performance to detect malware evolution and zero-day attacks. Another good use of SVM to detect malware through real time classification was carried out by [18] where lexical features of malware were examined through empirical analysis of already identified malware URLs;96.95% accuracy was obtained.

c) Decision Trees: Decision Trees are excellent for helping a researcher to choose between several courses of action. They provide options and describe the possible outcomes of choosing those options. Decision trees are mostly used in operations research to provide decision analysis to aid in discovering a strategy that will attain to a prescribed outcome. **Decision Trees in machine learning** are used as classifiers in classification and regression under supervised learning. The basic algorithm used in decision trees is known as the ID3 algorithm. The ID3 algorithm builds decision trees using a top-down, greedy approach. Decision Tree shows the **correlation between several features** and **non-linearity between the features**. A Decision Tree is easy to understand, requires very little data cleaning and no constraints on the data type [19]**.**

d) Random Forest: **Random Forest algorithm can also be used for both classification and regression kind of problems. The Random Forest** Algorithm works by creating a forest given a number of trees. The **higher the number** of trees in the forest **the higher the accuracy of the** results. When a Random Forest classifier is fed the training dataset with targets and features, the Random Forest classifier will come up with some **set of rules** that are used to perform prediction on the test dataset. The underlying principle of Random Forest classifier is the principle that a group of weak learners can come together to form a strong learner thereby making the Random Forest classifier to be able to classify large amounts of dataset with high accuracy . Random Forests do not overfit because of the law of large numbers by introducing the right kind of randomness it makes them accurate classifiers [19].

e) K-Nearest Neighbor (k-NN): *k*-NN is a type of instance-based learning, or lazy learning, where the function is approximated locally and all computation is deferred until classification. It is used for both classification and regression predictive problems. Its major advantage is the ease to interpret output, calculation time and prediction power. The goal is usually to find the k influence in the algorithm. The k-NN algorithm always assumes that similar things exist within the same area of focus and they are usually close to each other which means similar things are near to each other.[20].

f) Gradient Boosting: Boosting is a process of enhancing weak leaning models into strong learning models therefore Gradient Boosting classifier is a machine learning classifier that joins several weak learning models together to create strong predictive models and it mostly uses Decision trees. The process of boosting involves fitting every new tree into a modified version of the original malware dataset. Agnihotri (2018) used Gradient boosting classifier to detect ransomware through a static analysis of the ransomware PE file. Extraction of the static attributes was first carried out to obtain numerical values for the attributes which were used as inputs to the gradient boosting classifier to predict if the given sample is malicious or not. The performance metrics used was the false positive rate to grade the performance of the classifier; 0.3 percent false positive rate was obtained. Furthermore Pham *et al.* (2018) did a research on Static PE Malware Detection Using Gradient Boosting Decision Trees Algorithm stating that the problem with gradient boosting is the training time and also the ability to predict using imbalanced data makes the performance metrics somewhat inaccurate. They were able to reduce the training time by selective feature extraction and obtained a detection rate of 99.394 percent and a false positive rate of 1 percent.

Feature Selection Analysis

Feature selection plays a vital role in training datasets with classifiers to categorize or detect novel malware variants most especially the new metamorphic malware variants. According to [23] the type of feature selected affects the ability to detect accurately the metamorphic malware variants proving by his research work that unnecessary and redundant PE features when selected may decrease the detection rate of metamorphic malware variants. The researcher also proved that feature selection phase in malware detection plays a vital role in the whole detection process and can efficiently reduce the redundant and unnecessary features in the malware dataset; this in turn will reduce the false positive rate for a malware detection model using classifiers. According to [24] importance should be given to some part of the malware dataset with the goal to extract the significant sequences of malware opcodes in the dataset, the researcher used the dissimilarity of these significant sequence of malware opcodes to the benign files to select the significant sequence because all parts of a malware dataset feature are not representative of the malicious nature of the malware. As we have seen above all the reviewed work used n-grams byte sequence, opcodes, API/System calls feature selection for training the various classifiers used in malware detection which are effective in detecting known and unknown metamorphic malware variants but not so effective in detecting metamorphic malware that inject themselves inside a white listed software so as to disguise itself as trusted system process. A very good example is the powershell.exe virus that masks itself inside a Windows system powershell tool and executes malicious powershell commands traditional malware datasets feature selection may miss the obfuscation hidden in the Powershell during feature selection due to the fact that Powershell is white listed software under the windows operating system. [25] in their research on detecting Malicious PowerShell Commands using Deep Neural Networks expanded on the wide gap between the lack of research on automatic detection of malicious PowerShell commands and the high cases of PowerShell based malicious cyber exploits. This point was later developed upon by (Bohannon, 2017) who showed that recent approach is effective in detecting metamorphic malware but not so effective in detecting PowerShell attack which is fast becoming an ever increasing trend due to the fact that PowerShell attack is evasive and nearly impossible to detect because command line arguments and PowerShell events are not logged and monitored therefore most malware samples do not capture command line arguments/Powershell commands.

III. METHODOLOGY

A. Experimentation

This study was conducted using a dataset which consists of 138,047 PE header file records samples which was divided into: 41,323 clean files containing exe and dll file samples and 96,724 malware file samples. The research used the above file samples to compare and obtain the most efficient using machine learning performance metrics. This research follows techniques from [27] and [26] and applied machine learning algorithms so as to ascertain the most accurate algorithm to use in detecting malware. Some of the dependencies include, windows 10 operating system, python 3 server, Anaconda; which is a free and open source python programming language platform for scientific computing (data science, machine learning, data processing and predictive analytics), that aims to simplify package management and deployment, PEfile; which is an independent module to parse and work with PE files, Pandas; which is a software library for the python programming language for data manipulation and analysis, Jupyter Notebook; an open source web application that allows you to create and share documents that contain live code, equations and visualizations used mostly for machine learning, modeling, simulation and data cleaning which will be used as our GUI and IDE for running tests on the malware datasets.

Statistical tests used to evaluate the performance of the machine learning classifiers are: Accuracy, Precision, Recall, F-1 score, Area Under the Curve (AUC) which will be discussed shortly. The choice of Machine learning algorithms used to train the dataset was motivated by the literature review owning to the fact that most of the algorithms we shall compare where selected based on the frequency of appearance in the literature review.

For the feature selection, the legitimate files were separated from the malware files to obtain two variables; one containing the benign files and the other containing the malware files. Optimization of the dataset was carried out by applying a baseline threshold variance after which we applied Pearsons Correlation. For analysis of the baseline variance, baseline values need to be accurate in this case all the 96,724 malware files are accurate. Calculating the threshold will be based upon the features for the Portable executable header files for the each corresponding malware file by removing all low variance features. Features with a training set variance lower than the set threshold will be removed The idea is when a feature doesn't have a variation much within itself; it generally has very little predictive power (low variance) and it will be removed by the application of sklearn ensemble library on the malware dataset. We got 56 features that define the malware datasets as either malicious or benign. It is worthy of note that some of the 56 features that defines the samples are more important than other features in defining the sample as legitimate or malware. The application of Pearson Correlation will further improve accuracy by finding the linear correlation between two variables in this case the features optimized by variation. It finds the mutual relationship or connection between the headers for the benign variable and the malware variable. We selected features which have a correlation above 0.5 (factoring the absolute value) with the output variable;

- A value closer to 0 implies weaker downhill correlation (exact 0 implying no correlation)
- A value closer to 1 implies stronger uphill positive correlation
- A value closer to -1 implies stronger downhill negative correlation

The application of sklearn feature selection library to the dataset will apply Pearson Correlation to further improve the accuracy during the training and testing phase with the machine learning. We got 13 features which the machine considers as important features after the application of the sklearn feature selection library. We intend to test and train our data with cross validation of 20 percent which means our classifier kept 20 percent of the samples to use as test samples.

B. Statistical Tests

To prove that the classifier correctly predicted a malware, Evaluation metrics are used to determine the prediction accuracy of a classifier. This research work will use the following evaluation metrics explained below:

Accuracy: provides general information about how many samples are misclassified. Accuracy is calculated as the sum of correct predictions divided by the total number of predictions.

$$ACC = \frac{TP + TN}{TP + FP + FN + TN}$$

Precision Metrics: This is used when the aim is to limit the number of false positives measured from highest to lowest which means the classifier with the highest precision score has the best precision.

$$Precision = \frac{TP}{TP + FP}$$

Recall Metrics: This is the direct opposite of the Precision Metrics. It is mostly used when the aim of the research is to reduce the number of false negatives; it measures the predictive ability of the classifier to find all true samples/benign labels.

$$Recall = \frac{TP}{TP + FN}$$

F1-Score: This optimizes both precision and recall metrics. It is the numerical mean average between a set of positive variables; in this case it is the mean between precision and recall.

$$F1\ Score = \frac{2 * (Recall * Precision)}{Recall + Precision}$$

C. RESULT AND ANALYSIS

As earlier stated we test and trained our data with a cross validation of k = 20 and got the following result on table. I. All metrics are in percentages except the AUC metrics.

TABLE I
RESULTS

Classifiers	Accuracy	Precision	Recall	F1_Score	AUC
Gradient Boosting	98.5	99.2	98.5	0.988	0.99
Random Forest	89.3	89.8	89.4	0.894	0.54
Naïve Bayes	34.8	34.2	34.5	0.344	0.13
KNeighbor	56.2	56.2	56.6	0.564	0.20
MLP	88.1	88.3	88.1	0.882	0.53
Decision Tree	88.2	88.3	88.1	0.882	0.52
Quadratic Discriminant	85.3	85.1	85.3	0.852	0.48
Linear Disciminant	74.2	74.6	74.4	0.745	0.39
Logistic Reg.	23.8	23.8	23.4	0.236	0.11
Stochastic GD.	87.9	87.7	87.9	0.878	0.51

We shall now discuss the above results according to the standard evaluation metrics**:**

a) Accuracy: Provides general information about how many samples are misclassified; this implies the classifier with the lowest percentage score has a high level of misclassification. The Gradient Boosting classifier had a low level of misclassification on the malware dataset; out of the 96,724 malware instances, The Gradient Boosting classifier correctly predicted 98,273 malware instances. The Random Forest and Decision Tree had the second highest and third highest accuracy respectively. The classifier with the least score was Logistic Regression with an accuracy of 23.8 percent. Fig. 1 shows a bar chart for the Accuracy metrics.

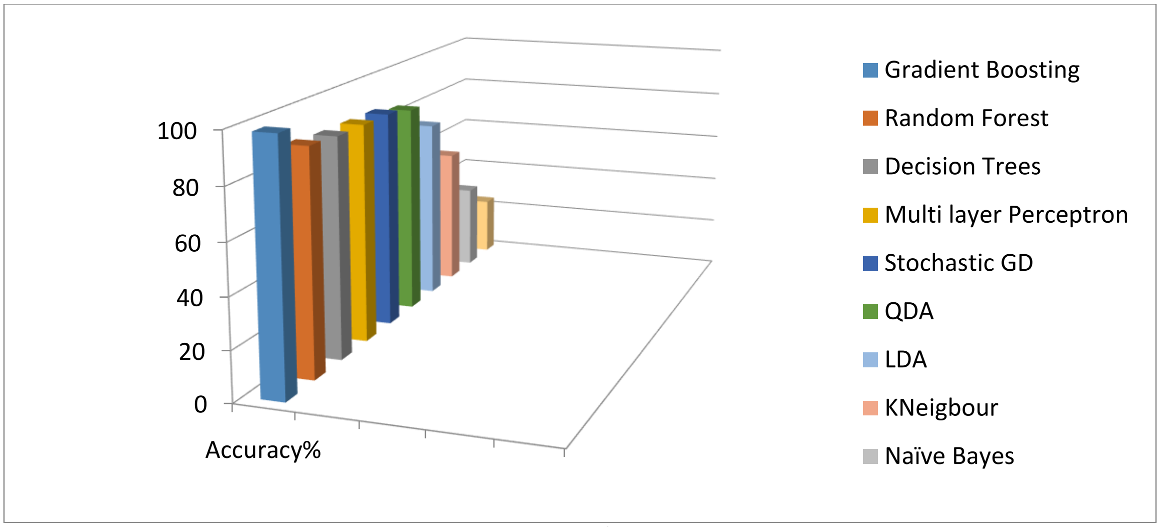


Fig. 1. Result Comparison in terms of Accuracy

b) Precision Metrics: The predictive performance of the classifiers in terms of Precision metrics is presented in figure 2. Precision metrics is used when we want to limit the number of false positives measured by from highest to lowest which means the classifier with the highest precision score has the best precision and can be used in cases where false positive reduction is the aim. The Gradient Boosting Classifier precision predictive output was at 97.2 percent this means the out of 96,724 malware instances the Gradient Boosting Classifier predicted 2709 instances as false positives which is relatively good. The Random forest classifier had the second highest Precision metrics while Logistic regression had the lowest. The Decision Trees and Multi-level Perceptron (MLP) classifiers had the same precision of 88.3 percent which is not so different from the Random Forest value of 89.8 percent implying that the Random Forest, Decision Tree and MLP classifiers can be used interchangeably in scenarios that need to reduce the false positive rate. Fig. 2 is a bar chart for the precision metrics of the different classifiers.

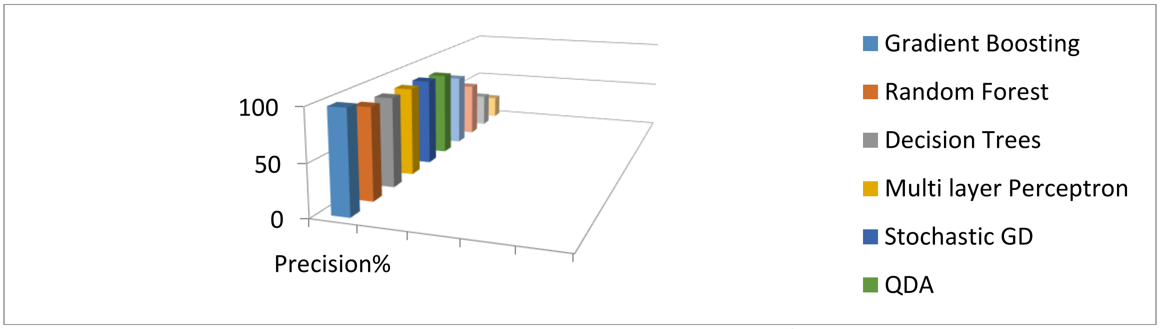


Fig. 2. Classifier Precision Metrics

c) Recall Metrics: This is the direct opposite of the Precision Metrics. It is mostly used when the aim of the research is to reduce the number of false negatives and it is also called sensitivity or true positive rate. From our results, Naïve Bayes and logistic regression got the lowest recall points in our research whereas all other classifiers had relatively high recall points. The Gradient Boosting classifier predicted 40165 to be benign/true samples out of 41323 benign/true samples having the highest Recall among the other classifiers. Fig 3 shows the Recall metrics graph for the various classifiers.

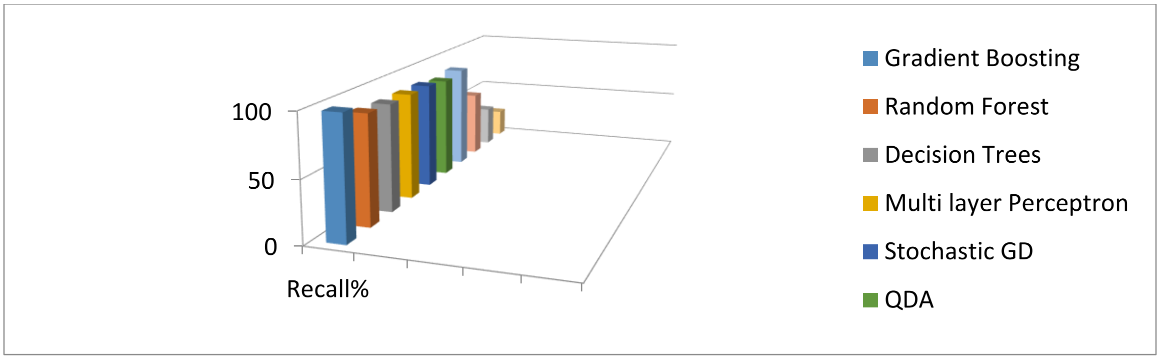


Fig. 3. Classifier Recall Metrics

d) F1_Score: Fig. 4 shows the f1_score metrics of the various trained classifiers. The F1_Score optimizes both precision and recall metrics. The F1-score will give the harmonic mean between the precision metrics and the recall metrics to understand how accurate the classifier is in predicting malware instances. From our results Naïve Bayes still had the lowest F1_score metrics followed by both Logistic Regression and Stochastic Gradient. The Gradient Boosting classifier still had the highest F1_score of approximately 0.99 which is very close to 1.

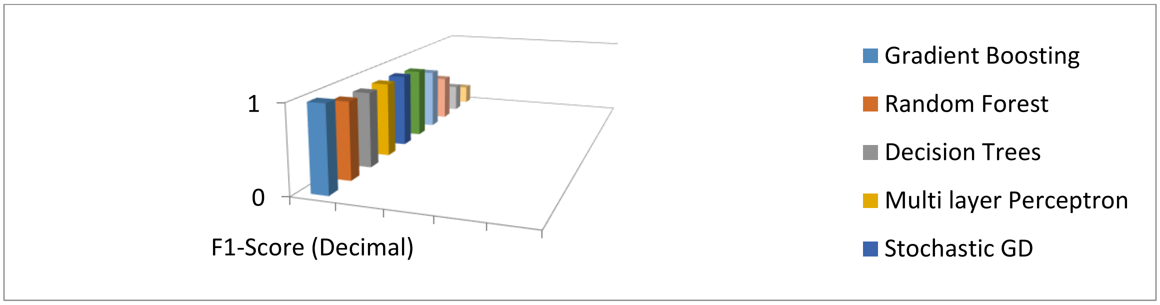


Fig. 4. Classifier F1_Score Metrics

e) Area Under Curve (AUC): AUC is the possible thresholds that are considered during the Receiver Operating Characteristics (ROC) curve plot that is underneath the curve; the ROC is a visual way to check the performance of a binary classifier. Fig. 5 shows the AUC metrics of the classifiers. The AUC is one of the most important evaluation metrics for checking the predictive performance of a classifier because a perfect classifier has an AUC near to one (1) meaning it has a good measure of classification and a poor classifier has an AUC near to zero (0). The Gradient Boosting, Random Forest, KNeigbour, and Decision Tree classifiers are the best performing classifiers according to the AUC score.

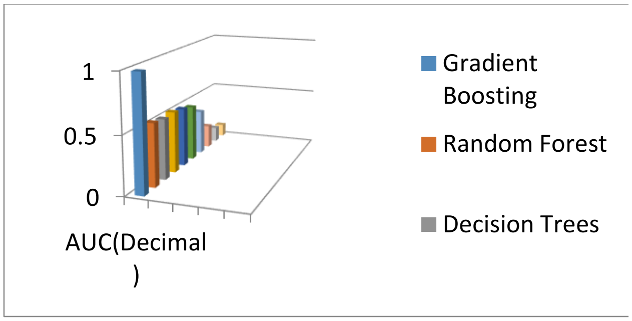


Fig. 5. Classifier AUC Metrics

V. CONCLUSION AND RECOMMENDATION

This research trained ten different classifiers with dataset of 138,047 PE header file records samples which was divided into: 41,323 clean files containing exe and dll file samples and 96,724 malware file samples. The Gradient Boosting classifier achieved the best results with 98.5% accuracy and an F1-Score of 0.988 followed by the Random Forest classifier although it was closely followed by the Decision Tree Classifier and the Multilayer Perceptron. The Logistic Regression and Naïve Bayes classifiers performed poorly compared to the other classifiers due to misclassification. Support Vector Machine couldn't be trained with this research's dataset because the dataset couldn't be fitted into this algorithm and also the time constraint in training the SVM classifier. The classifiers presented above reveal the detection rate of various machine learning classifiers. In addition, this research's experiments reveal that, while Gradient Boosting provides the best detection rate with the best AUC followed by Random Forest and Decision Tree respectively. The predictive ability of both classifiers can be improved with a method for false negative reduction and further feature analysis on the malware dataset.

REFERENCES

[1] M. Olalere, M. T. Abdullah, R. Mahmod, and A. Abdullah, "A Review of Bring Your Own Device on Security Issues," *SAGE Open*, vol. 5, no. 2, p. 215824401558037, 2015.

[2] Z. Bazrafshan, H. Hashemi, S. Mehdi, H. Fard, and A. Hamzeh, "A Survey on Heuristic Malware Detection Techniques," no. May, 2013.

[3] M. Christodorescu, "Static Analysis of Executables to Detect Malicious Patterns ∗."

[4] A. M. Maigida, S. M. Abdulhamid, M. Olalere, J. K. Alhassan, H. Chiroma, and E. G. Dada, "Systematic literature review and metadata analysis of ransomware attacks and detection mechanisms," *J. Reliab. Intell. Environ.*, vol. 5, no. 2, pp. 67–89, 2019.

[5] N. Kaur, "A Complete Dynamic Malware Analysis," vol. 135, no. 4, pp. 20–25, 2016.

[6] H. S. Anderson, B. Filar, and P. Roth, "Evading Machine Learning Malware Detection," 2017.

[7] I. Firdausi, C. Lim, and A. Erwin, "ANALYSIS OF MACHINE LEARNING TECHNIQUES USED IN BEHAVIOR-BASED MALWARE DETECTION," no. May 2014, 2010.

[8] J. J. Barriga and S. G. Yoo, "Malware Detection and Evasion with Machine Learning Techniques : A Survey Malware Detection and Evasion with Machine Learning Techniques : A Survey," no. September 2017, 2019.

[9] C. Vatamanu, D. Cosovan, and H. Luchian, "of Malware Detection Techniques Using Machine Learning Methods," vol. 9, no. 5, pp. 1157–1164, 2015.

[10] Y. Ye, "A Survey on Malware Detection Using Data Mining Techniques," vol. 50, no. 3, 2017.

[11] D. Carlin, P. O. Kane, and S. Sezer, "A cost analysis of machine learning using dynamic runtime opcodes for malware detection R," vol. 85, pp. 138–155, 2019.

[12] D. Ucci, L. Aniello, and R. Baldoni, "Survey of Machine Learning Techniques for Malware Analysis," *Comput. Secur.*, 2018.

[13] A. Kumar, K. S. Kuppusamy, and G. Aghila, "feature set feature set," *J. King Saud Univ. - Comput. Inf. Sci.*, 2017.

[14] T. Wang, "Detecting Unknown Malicious Executables Using Portable Executable Headers Detecting Unknown Malicious Executables Using Portable Executable Headers," no. November, 2016.

[15] A. Kumar, K. S. Kuppusamy, and G. Aghila, "A learning model to detect maliciousness of portable executable using integrated feature set," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 31, no. 2, pp. 252–265, 2019.

[16] M. G. Schultz, E. Eskin, and S. J. Stolfo, "Data Mining Methods for Detection of New Malicious Executables."

[17] Z. Chen, M. Roussopoulos, Z. Liang, Y. Zhang, Z. Chen, and A. Delis, "The Journal of Systems and Software Malware characteristics and threats on the internet ecosystem," *J. Syst. Softw.*, vol. 85, no. 7, pp. 1650–1672, 2012.

[18] M. Olalere, M. T. Abdullah, R. Mahmod, and A. Abdullah, "Identification and Evaluation of Discriminative Lexical Features of Malware URL for Real-Time Classification," *Proc. - 6th Int. Conf. Comput. Commun. Eng. Innov. Technol. to Serve Humanit. ICCCE 2016*, pp. 90–95, 2016.

[19] S. Ranveer and S. Hiray, "SVM Based Effective Malware Detection System," vol. 6, no. 4, pp. 3361–3365, 2015.

[20] M. Ahmadi and G. Giacinto, "Novel Feature Extraction , Selection and Fusion for Effective Malware Family Novel Feature Extraction , Selection and Fusion for Effective Malware Family Classification," no. March, 2016.

[21] N. Agnihotri, "Ransomware Classifier using Extreme Gradient Boosting," vol. 9, no. 2, pp. 45–47, 2018.

[22] H. Pham, T. D. Le, and T. N. Vu, *Static PE Malware Detection Using Gradient*. Springer International Publishing, 2018.

[23] Q. Jiang, "A Feature Selection Method for Malware Detection," no. June, pp. 890–895, 2011.

[24] V. Derhami, S. Hashemi, S. Mehdi, and H. Fard, "Proposing an approach to detect metamorphic mal- ware based on Hidden Markov Model," no. September 2016, 2015.

[25] D. Hendler, S. Kels, and A. Rubin, "Detecting Malicious PowerShell Commands using Deep Neural Networks arXiv : 1804 . 04177v2 [cs . CR] 14 Apr 2018."

[26] D. Bohannon and D. Bohannon, "Techniques on how to detect Invoke obfuscation 2018"

[27] M. Swamynathan, "swamynathan2017.pdf." p. 374, 2017.

[28] O. S Adebayo, M. A. Mabayoje, A. Mishra, and O. Osho, "Malware Detection, Supportive Software Agents and Its Classification Schemes," International Journal of Network Security & Its Applications (IJNSA), Vol.4 (6), pp. 33 – 49, 2012.

[29] Adebayo, Olawale Surajudeen, Aziz, Normaziah Abdul. Static Code Analysis of Permission-based Features for Android Malware Classification Using Apriori Algorithm with Particle Swarm Optimization. Journal of Information Assurance and Security, 10 (4), 2015, page 152 – 163.

[30] A.Shabtai, Y. Fledel, & Y. Elovici, "Automated Static Code Analysis for Classifying Android Applications using Machine Learning," International Conference on Computational Intelligence and Security (CIS), 2010.

[31] Andrew Walenstein, Luke Deshotels, and Arun Lakhotia "Program Structure-Based Feature Selection for Android Malware Analysis" MOBISEC 2012, LNICST 107, pp. 51–52, 2012. Institute for Computer Sciences, Social Informatics and Telecommunications Engineering.

[32] Suhas Holla, and Mahima M Katti "Android based Mobile Application and Its Security". International Journal of Computer Trends and Technology, 3 (3) Pp. 486 – 490, 2013. ISSN 2231 – 2801.

Post-Quantum Cryptographic Algorithm: A systematic review of round-2 candidates.

A.C. Onuora
Department of Computer Science
Akanu Ibiam Federal Polytechnic Unwana, Afikpo
Ebonyi State, Nigeria
aconuora@akanuibiampoly.edu.ng

C. E. Madubuike
Department of Computer Science
Akanu Ibiam Federal Polytechnic Unwana, Afikpo
Ebonyi State, Nigeria
cemadubuike@akanuibiampoly.edu.ng

A. O. Otiko
Department of Computer Science
Cross River State University of Technology, Calabar
Cross River State, Nigeria
otikotony@gmail.com

J. N. Nworie
Department of Computer Science
Akanu Ibiam Federal Polytechnic Unwana, Afikpo
Ebonyi State, Nigeria
praisejoyuh@gmail.com

***Abstrac*— The rise of the new paradigm (Quantum computing) in the recent years have created a major security challenge to classical and widely used primitive cryptography schemes such as ECC (Elliptic Curve Cryptography) and RSA (Rivest-Shamir-Adleman) Algorithm. These classical computing algorithms depend on the problems of discrete logarithm and integer factorization respectively. Recent advancements in quantum computing have made encryption schemes more vulnerable since they are weak to some quantum attacks, like Shor's Algorithm and Grove's Algorithm. Therefore the call for a new set of algorithms known as Post-Quantum cryptography (PQC) that would not be vulnerable to quantum attacks is imminent. NIST haven selected some candidates in the second round of Post-Quantum Cryptographic algorithms standardization project. This work's goal is to review these algorithms under there types. A rigorous survey on each Post-Quantum Cryptography schemes and their underlying properties will be x-rayed while recommending areas for research in this new security paradigm.**

Keywords*— *KEM, Cryptography, Digital Signature, Quantum Computer, Security

VII. INTRODUCTION

Cryptography is the science of hiding information meant for a recipient so that intruders cannot understand the message even when it is intercepted. Only the intended recipient can unravel the message using a key. Post Quantum cryptography is the application of existing cryptographic algorithms or the design of new algorithms that are quantum proof.

The computer system used in our daily computing activities ses the bit system: a "0" or a "1". In other words, Classical computers perform its operations using the binary position of the bit. This implies that very task carried out by our classical computers takes either of these two states. A single state such as on or off, up or down, 1 or 0 is called a bit. Qubit is what the quantum computer uses. They don't use the two-state position of 0s and 1s like the classical computers. Quantum computers use a four-state superimposition to encode data. The four states of the quantum computer qubits are represented as ions, atoms, photons or electrons on devices for processing data. A quantum computer will probably be more powerful than even the most powerful supercomputer because of its ability to represent data in multiple states.. (Zentachain, 2019)

A lot of research has been on-going on quantum computers (machines that will use quantum phenomena to bring solutions to hard mathematical problems that normal classical computers won't solve). These researches have open up week points in our present day classical computers that if eventually, the quantum computer is mass produced today, a lot of system will remain vulnerable to different attacks. Researches on quantum computer have proven that our present cryptographic system will not be capable of protecting data when the breakthrough is finally made. Digital communication will be adversely compromised. The integrity and confidentiality of users both online and offline will be jeopardized. This is the reason NIST highlighted the goal of post-quantum cryptographic system as building an enhanced or new cryptographic system that will secure both classical and quantum computers. They will seamlessly integrate with already existing network and communication protocols. (NIST, 2019).

VIII. Quantum Computing

The marriage of quantum theory in physics with computer science is know nas quantum computing (Mingsheng, 2010). In quantum computing, series of bits (qubits) are used to represent information. Unlike a normal bit, qubits can take the state of both 0 and 1 at the same time. When these qubits are extended, it gives rise to the great computational power of quantum computer.. (Horowitz, et.al, 2019).

Quantum computers cannot be likened to a group of parallel-arranged classical computers. Researchers have thought that an n-bit quantum computer can computer can solve NP-complete problems faster due to the fact that qubits can occupy both states (0s and 1s) at the same time. A systematic transformation of qubits are performed on the intended information so that at the end of computation, the algorithm will not destroy the needed information, this makes a successful quantum algorithm. The implication is that no quantum algorithm that can solve all Np-Complete problems at the same time. Therefore, in other to use quantum algorithms to solve hard classical problems, one most exploit the specific structure of the problem at hand. (Zentachain, 2019).

Renty (2019) went ahead to highlight the algorithms that necessitated the need for improved quantum security on the present day primitive cryptography. To achieve this improvement, Post-quantum cryptography was born. The algorithms as highlighted by Renty (2019) are:

a. *Shor's Algorithm:*

Among the best-known and oldest quantum algorithms is Shor's algorithm which can effectively factorize integers and solve discrete logarithm problems. Come the day we're able to construct a quantum computer powerful enough to handle very large numbers using Shor's algorithm, all public-key cryptography based on these problems will be rendered immediately obsolete.

b. *Grover's algorithm:*

Grover's algorithm is another well-known quantum algorithm, which enables the inversion of functions. This algorithm can be applied to search effectively for items in an unordered list or unstructured database, something that can be likened to looking for a needle in a haystack. Grover's algorithm thus makes it possible to accelerate a search for a symmetric encryption key but does not call into question the principles of secret-key cryptography.

IX. Post-Quantum Cryptography and Types

Post-quantum cryptography is the building of cryptosystems that can secure both classical computers and quantum computers should incase an intruder possess it. NIST initiated a process for standardizing post-quantum algorithms. The submissions included cryptosystems from the following types of cryptosystem. (NIST, 2019).

The major types of post-quantum cryptosystem are highlighted below;

a. *Lattice-Based Cryptosystem:*

Used for Key Encapsulation Mechanism (KEM) / Encryption, Due to the flexibility of Lattices, they are the most studied. They can be used for digital signature, key exchange and for fully hormomorphic encryption. The underlying foundation for lattice cryptosystem is basic linear algebra notwithstanding the complex math involved for optimization and security proofs. The simple algebra is shown below.

$$
\begin{aligned}
a_{0,0}x_0 + a_{0,1}x_1 + \cdots + a_{0,n}x_n &= y_0 \\
a_{1,0}x_0 + a_{1,1}x_1 + \cdots + a_{1,n}x_n &= y_1 \\
&\vdots \\
a_{n,0}x_0 + a_{n,1}x_1 + \cdots + a_{n,n}x_n &= y_n
\end{aligned}
$$

You can solve for x in a classic linear algebra problem that can be solved quickly using Gaussian elimination method. A mystery function is another way to solve this,

$$f_{\mathbf{x}}(\mathbf{a}) = a_0x_0 + \cdots + a_nx_n$$

Where "*a*" is a vector, we see the result of *ax*, and x is unknown. When this function is queried enough times we are able to learn f in a short amount of time (by solving the system of equations above). The linear algebra problem above can be modeled to a machine learning problem.

If a small noise is introduced to our algebraic function, add an error term *e* to the product of *x* and *a* then the whole thing is reduced to a modulo *a* (medium-sized) prime q. Then our noisy mystery function looks like

$$f_{\mathbf{x}}(\mathbf{a}) = a_0x_0 + \cdots + a_nx_n + \epsilon \mod q$$

It is mathematically extremely difficult to learn this noisy function. Each loop in the Gaussian elimination method makes use of the noisy function, this increase the error term. The error term continues to increase as we continue in the loop until it swallows all information about the noisy function. This approach is what is called Learning With Errors (LWE) Problem in Lattice Cryptographic system. (Research Institute, 2019)

Learning With Errors problem (LWE) cryptosystems are call lattices because finding the shortest vector when proofing that LWE is NP-hard is called a lattice. Let's spare ourselves the scope of knowing what a lattice is mathematically. A lattice can be likened to a tiling of n-dimensional space in a nutshell.

.

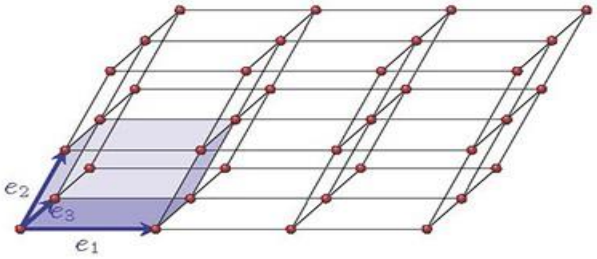


Fig1: Diagram of a lattice (Research Institute, 2019)

Some of the lattice-based cryptosystems that mad the second round of NIST PQC standard are Kyber, Saber, FrodoKEM, Dilithium.

b. Code-Based Cryptosystem

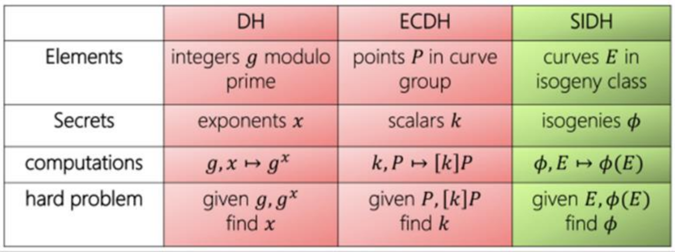
Code-based cryptosystem is an encryption / key encapsulation mechanism (KEM). During communication, error in form of bit flips occurs. As the errors occur, the ability to withstand certain number of bit flips at the expense of message compactness is known as Error-correcting codes. For instance, 0 can be corrected to 000 and 1 to 111 for single bits. In this way, if a recipient gets 101, one will understand that 111 was actually sent. In like manner if the recipient gets 001, one will understand that 000 was originally sent by taking majority of the vote. This cannot actually work where the error happened on two bits such as 111 received as 001. The recipient will take the majority vote and arrive at 0 instead of 1. (Research Institute, 2019).

Linear code are one of the most-popular error-correcting codes and are represented by k x n matrices, where k stands for the length original message while n stand for the length of encoded message. It is generally difficult to decode messages without knowing the corresponding linear code. This hardness is what gives strength to the McEliece public key cryptosystem. (Research Institute, 2019). McEliece uses a conservative approach for key encapsulation and public key encryption. Large keys are trademarks of this crypto-algorithm and researcher trying to use small keys for this cryptosystem and not water the underlying security.. (Basu, Soni, Nabee, & Karri, n.d.)

c. Isogeny-based (Supersingular) Cryptosystem

Used for Key Encapsulation Mechanism (KEM) / Encryption. Isogeny-based cryptography uses the approach of walking through a sequence of elliptic curve instead of points on the elliptic curve.

Table 1: The evolution of Isogeny from Diffie-Hellman. (Costello, 2017)



	DH	ECDH	SIDH
Elements	integers g modulo prime	points P in curve group	curves E in isogeny class
Secrets	exponents x	scalars k	isogenies ϕ
computations	$g, x \mapsto g^x$	$k, P \mapsto [k]P$	$\phi, E \mapsto \phi(E)$
hard problem	given g, g^x find x	given $P, [k]P$ find k	given $E, \phi(E)$ find ϕ

In Supersingular Isogeny Diffie-Hellman (SIDH) scheme, secret keys are obtained from chain of isogenies and public keys are obtained from curves. For instance Alice and Bob combine this information, they acquire curves that are different, but have the same j-invariant. It’s not so important for the purposes of cryptography what a j-invariant is, but rather that it is a number that can easily be computed by both Alice and Bob once they’ve completed the key exchange. The Isogeny-based cryptography uses only 330 bytes of public keys, the smallest among all post-quantum crypto-system competitors. They also possess perfect forward secrecy unlike other cryptosystems. (Research Institute, 2019)

d. Hash-Based Signature

Used for Digital signatures. The Merkle tree is used in the hash-based cryptosystem to reduce space. Hashes cannot be used to construct KEM or public key encryption schemes. Hashes are mainly for signatures based schemes. Furthermore, they are extremely fast due to the fact that the cryptosystem requires only the computation of hash functions. This cryptosystem are extremely strong also. There has not been any confirmation that current widely used hash functions like SHA3 or BLAKE2 are vulnerable to attacks, hash-based crpto-system are secure. (Research Institute, 2019). Examples of these algorithms are SPHINCS+ and Picnic. Both are NIST second –round candidates.

e. Multivariate-Based Cryptosystem

Used for Digital signatures and Public-Key schemes. Multivariate cryptography used a set of polynomial known as multivariate polynomial (MVP) in several variables, of small degree over a small finite field. With the polynomial, verification of digital signature is very fast and short. It also has good performance and flexibility to its name. (Research Institute, 2019).

Some algorithms under this cryptosystem are Rainbow (Unbalanced Oil and Vinegar), MQDSS, LUOV etc. Rainbow is another multivariate signature scheme that is built as a quantum-resistant digital signature scheme. (Research Institute, 2019).

Table2: Complete list of post-quantum cryptographic algorithm (NIST Second Candidates)

S/No	**PQC Algorithm**	**Type**	**Mechanism**
1	BIKE	Code-Based	KEM
2	Classic McEliece	Code-Based	KEM
3	CRYSTALS-	Lattice-Based	KEM

	KYBER		
4	FrodoKEM	Lattice-Based	KEM
5	HQC (Hamming Quasi-Cyclic)	Code-Based	KEM
6	LAC	Lattice-based	KEM
7	LEDAcrypt	Code-Based	KEM
8	NewHope	Ring-LWE	KEM
9	NTRU	Lattice-Based	KEM
10	NTRU Prime	Lattice-Based	KEM
11	NTS-KEM	Code-Based	KEM
12	ROLLO	Code-Based	KEM
13	Round5	Lattice-Based	KEM
14	Rank Quasi-Cyclic (RQC)	Code-Based	KEM
15	SABER	Lattice-Based	KEM
16	SIKE	Isogeny-based	KEM
17	Three Bears	Lattice-Based	KEM
18	CRYSTALS-DILITHIUM	Signatures Scheme	lattice-based signature
19	FALCON	Signatures Scheme	lattice-based signature
20	GeMSS	Signatures Scheme	Multivariate-based
21	LUOV	Signatures Scheme	Multivariate-based
22	MQDSS	Signatures Scheme	Multivariate-based
23	Picnic	Signatures Scheme	hash functions and block ciphers
24	qTESLA	Signatures Scheme	Code-based
26	Rainbow	Signatures Scheme	Multivariate-based
27	SPHINCS+	Signatures Scheme	Hash-based

X. REVIEW OF SOME WORK DONE ON POST-QUANTUM CRYPTOGRAPHY

Rijneveld (2019) researched on various group of post-quantum cryptography algorithm. Firstly, He examined hash-based digital signature schemes, a thorough discussion of historical constructions, leading up to the recent XMSS, SPHINCS, and SPHINCS+ schemes. He further discussed its scheme design, and described several implementations, in particular on embedded platforms. a non-standard approach towards designing signature schemes based on the MQ problem was employed by him and MQDSS and SOFIA schemes were introduced. Secondly, Lattice-based KEM was also examined. Lattice-based key-encapsulation mechanisms with a focus on NTRU was optimized and implemented.

Gyurik (2018) in his research developed a new cryptography from the RSA primitive cryptographic algorithm. He worked a Post-Quantum RSA variation by Bernstein et al that was designed to withstand supposed quantum attacks better than RSA. He used large-scale quantum computers to investigate secuirity of Post-Quantum RSA. He used element of relatively low multiplicative order to speed up Shor's algorithm then the unification of unification of Lenstra's elliptic curve factorization method (ECM) and Shor's order-finding algorithm. Furthermore, he showcased how to use certain number of processors to speed up Shor's algorithm for classical or quantum operations in parallel.

Nejatollahi et al. (2019) demonstrated that latticed based cryptography algorithms can be implemented in softwares, hardwares and both softwares / hardwares. The lattice-based cryptography has schemes for Public key encryption, digital signature and key exchange. He highlighted the scheme for public key encryption as NTRU or LWE (variety of variants exits such as RLWE, MLWE, ILWE and MPLWE. He surveyed the implementation of this post-quantum cryptography (Lattice-based) algorithm on diversity of computing platforms.

Endignoux (2017) focused on hash based scheme using hash functions. His work relied on Preimage and collision resistance with regards to their security with hash function properties. Stateless hash-based signatures was efficiently improved and made Stateful. He further presented a cryptanalysis of the subset-resilience problem, showing new attacks and proposing fixes.

Yang (2019) asserted that multivariate scheme's security is based on the Problem MQ: where m multivariate quadratic polynomials p(1), . . . , p(m), find a vector w = (w1, . . . , wn) such that p(1)(w) = . . . = p(m)(w) = 0.

XI. CONCLUSION

The latest advances in the technology of quantum computing have form a new challenge for modern cryptography. There are needs to find the novel ways of ensuring information security and its main properties -confidentiality, integrity, authentication and repudiation are achieved. More work should be done by scientists and researcher to unveil ways of improving these algorithms. We should make sure that tat post-quantum cryptosystem are ready before the mass production of large-scale quantum computers. With this is ready, our network and data will continue to be safe in the computing sphere.

XII. Future Research Areas

Post-Quantum Cryptography (PQC) is still in its infant stage and a lot of areas are begging for research and according to Ott & Peikert (2019), they include;

a. PQC migration: Research that addresses the application of candidate algorithms to specific contexts and how migration within any given cryptographic usage domain can be realized in a secure way. Deploying this algorithms on several platforms like Web, Mobile IOT, VPN and Trusted computing architectures

B. Cryptographic agility: Cryptographic agility addresses the important problem of future-proofing our global cryptographic infrastructure in a flexible and robust manner. Research are needed in the area of Implementation Agility, Compliance Agility, Security Strength Agility etc

C. Other Areas are in Policy making, Process and people. Areas of emerging trends like Blockchain PQC, Password-authenticated Key Agreement (PAKE), Secure Multi-Party Computation (MPC) and more

References

Basu, K., Soni, D., Nabee, M., & Karri, R. (n.d.). NIST Post-Quantum Cryptography-A Hardware Evaluation Study. Retrieved from https://eprint.iacr.org/2019/047.pdf

Bernstein, D. J., Heninger, N., Lou, P., & Valenta, L. (2017, April 19). Cryptology ePrint Archive: Report 2017/351 - Post-quantum RSA. Retrieved from http://eprint.iacr.org/2017/351

Costello, C. (2017). An introduction to supersingular isogeny-based cryptography. Retrieved from https://ecc2017.cs.ru.nl/slides/ecc2017school-costello.pdf

Endignoux, G. (2017). Design and implementation of a post-quantum hash-based cryptographic signature scheme (Unpublished master's thesis). Ecole Polytechnique of Palaiseau, Paris, France.

Gyurik, C. (2018). Quantum algorithms for factoring and Post-Quantum RSA (Master's thesis, University of Amsterdam, Amsterdam, Neitherland). Retrieved from https://esc.fnwi.uva.nl/thesis/centraal/files/f550606688.pdf

Horowitz, M. A., Aspuru-Gujzik, A., Awschalom, D. D., Blakley, B., Boneh, D., Coppersmith, S. N., … Vazirani, U. V. (2019). Quantum computing's implications for cryptography. In Quantum computing: Progress and prospects (pp. 95-112). Retrieved from https://www.nap.edu/read/25196/chapter/6

Nejatollahi, H., Dutt, N., Ray, S., Regazzoni, F., Banerjee, I., & Cammarota, R. (2019). Post-Quantum Lattice-Based Cryptography Implementations. ACM Computing Surveys, 51(6), 1-41. doi:10.1145/3292548

Ott, D., & Peikert, C. (2019, November 11). Identifying Research Challenges in Post Quantum Cryptography Migration and Cryptographic Agility. Retrieved from https://cra.org/ccc/wp-content/uploads/sites/2/2018/11/ CCC-Identifying-Research-Challenges-in-PQC-Workshop-Report.pdf

Renty, D. (2019, July 9). Status of post-quantum cryptography. Retrieved from https://www.riskinsight-wavestone.com/en/author/david-renty/

Research Institute (2019, July 24). A Guide to Post-Quantum Cryptography. Retrieved from https://medium.com/ hackernoon/a-guide-to-post-quantum-cryptography-d785a70ea04b

Yang, B. (2019, March 19). Multivariate Quadratic Public-Key Cryptography In the NIST Competition. Retrieved from https://www.maths.ox.ac.uk/system/files/attachments /MQ%20Public-Key%20Crypto%20in%20the%20NIST %20competition.pdf

Zentachain. (2019, October 27). Quantum Computers & Encryption. Retrieved from https://medium.com/ @zentachain/quantum-computers-encryption-b3d407da5099

Parallel Smith Waterman Algorithm Based RNS Accelerator for DNA Sequencing.

Olatunbosun Lukumon Olawale
ICT. Department of Computer Science
Federal University of Agriculture,
Abeokuta, Nigeria
Tel: +2348029290875
Email: Latunbosunol@funaab.edu.ng.

Gbolagade Kazeem. Alagbe
ICT.Department of Computer Science
Kwara state University,
Malete. Nigeria.
Tel::+2348136273074
Email: Kazeem.gbolagade@Kwasu.edu.ng

ABSRACT-Computing in DNA sequences requires effective algorithms to perform and accelerate sequence alignment activities. The Choice of Smith-Waterman (SW) exhibiting a highly robust and efficient parallel computing system development for biological Gene, sequence is distinctly used in addressing some of the lapses handle by other traditional software equivalents. Understand deep knowledge transfer about exiting approach for gene sequencing and alignment using Smith-waterman with emphases on it strength, weaknesses and improvement. However, the unique SWA characteristic exhibiting the most computationally intensive algorithm based on residue number system (RNS) makes it more viable, necessary to implement effective sequence alignment operation with hardware acceleration methods for practical applications, maximizing the inherent RNS arithmetic potential advantages.

Keywords - DNA, Sequence Alignment, Smith Waterman Algorithm, RNS, Moduli set. Hardware Accelerator, LCS.

I. INTRODUCTION

Genome sequencing problems are part of the main issues for researchers to develop an optimized system model that could facilitate the optimum processing and efficiency without affecting the performance of memory and time. This study is oriented towards developing such type of system while taking into consideration of the dynamic programming approach called a Smith Waterman based on RNS algorithm. The Smith Waterman (SW) algorithm described a method based on dynamic programming (DP)**]** in 1981and basically for local sequence alignment where common regions in DNA sequences that share same similarity characteristics were identified. [1][11] A human genome contains approximately 3 trillion DNA base pairs. In order to discover which amino acids are produced by each part of a DNA sequence, it is necessary to find the similarity between two sequences. This is done by finding the minimum string edit distance between the two sequences and the process is known as sequence alignment [20][21]. The predominant factor which has been optimized with its optimum possibility is Smith Waterman algorithm which functions in a unique parallel alignment rather being in conventional serial approaches. Smith Waterman algorithm is a dynamic programming approach that could accomplish the higher rate sequencing with parallel scheme. The dynamic programming approach uses a table or matrix to preserve values and avoid re computation at the expense of time and high computational cost. Sequence alignment algorithms detect similar or identical parts between two sequences called the query sequence and the reference sequence. [3].The global and local alignments are the most prevalent kinds of sequence alignment. In global alignment, problem finds the superior counterpart between the whole sequences. While in local alignment, algorithms must find the superior counterpart between parts of the of the sequence [4], When obtaining the local alignment, a matrix Hi;j is used to keep track of the degree of similarity between the two sequences to be aligned (A_i and B_j) [2][17].Each element of the matrix $H_{i,j}$ is calculated according to the following equation:

$$\mathbf{H}_{i,j} = \mathrm{Max} \begin{cases} 0 & \\ Hi-1, j-1+Si,j & \text{Diagonal entry} \\ Hi-1, j-d & \text{Upper entry} \quad \ldots\ldots\ldots \\ Hi, j-1-d & \text{Left entry} \end{cases} \qquad [1]$$

Where**:**

H is the matrix value of the essential cell with H (i; j) is the maximum similarity score between the two sequences. S is the score of the cell $S_{i;j}$ which is the similarity score of comparing sequence A_i to sequence B_j and **d** is the gap alignment and penalty for a mismatch. i, j describe row and column Diagonal, Upper and Left entries are the matrices entry position relative to the current $_{H(i;j)}$ calculation

Procedure of the Algorithm

- Initialization of matrix considering the two sequences A and B.
- Matrix filling with the suitable scores. The two sequences are set in a matrix form by means of **A+ 1** column and **B+1** row with the values in the first row and first column set to zero.
- The Trace back Matrix.

1. Data Representation in DNA Sequencing

Deoxyribonucleic Acid is a sequence of string code in which almost all genetic information is encoded using four key chemicals, adenine, thymine, guanine and cytosine (abbreviated as A, T, G and C) [21].[23] An example of a genome sequence is shown in Figure1:

**CCTTCATCTAGGAGTTGAGAAGGGTAGATAAGA
TTCTTGGATACTAGGTATTTAAGAACTTTCTCAG
ATGAAAGGAAGCTGGGAACAAAGTAAGAAAGAA
TACCTTTTAGGATTCACAAAATTATGAGAAGTCA
GCCACATACGGTAGGTCAGCTTTTTAATGTATTT
GCTCCTTTTCTTATTCT**

Figure 1: Sub-sequences of DNA for Western Gorilla

Published under licence by IOP Publishing Ltd

II. BACKGROUND

Deoxyribonucleic acid DNA contains the "Genetic instruction" for the development of function of living things (Albert et al 2014). Motifs are essential pattern for understanding the function of genes and human diseases and so essential for identifying transcriptional regulatory elements and factor binding sites.[15] [16] Motif are refers to as planted (I, d) where I represents motif length and d represents maximum number of mutation allowed in the motif. Motif can be used to determined the evolutionary and functional relationships of the genes and vary in length, positions, redundancy, orientation and bases. Locating and aligning these short sequences i.e. motifs or signals is a fundamental problem in molecular biology and computer science with important applications such as residue number system, RNS. Knowledge-base, KB. Drug design, DD. forensic, DNA analysis, agricultural Biotechnology, AB. (Xionger et al, 2006).The application of RNS have been used to speedup linear processing, to achieve high-speed and low-power VLSI implementations for the multiplication of accumulate operation which is also employed in other areas of Bioinformatics, linear signal and image processing [11].However, the non positional nature of the RNS prevents its usage to implement the division and, in general, non linear processing. Magnitude comparison, which is a fundamental operation to support this type of processing, is difficult to implement in RNS, being an important topic of research in the last few years [5], [2] and [13].

Many attempt has been made by deferent researchers to overcome the problem associated with SWA, a reconfigurable accelerator for Smith-Waterman algorithm is adopted and is presented from [11], where in the accelerator, a modified equation is projected to develop mapping efficiency of a processing which provides more than 330 speedup[12][16] when compared to a standard desktop platform with the 4GB memory and 2.8GHz Xeon processor and it has a 50% progress on the peak performance of a traditional implementation without the two special techniques.[1],[5].Due to the complexity of SW algorithm, there is a need for a methodology that could reduce the computation time while delivering accurate results. The total time that will be needed to alignment two strings of DNA using SWA based RNS architecture will be improved exponentially more than its software equivalent implementation [3],[11].Therefore RNS is a good platform to implement the SWA, since it has a very high prospect of improving the overall computational cost and the hardware foot prints of the algorithm.[12]; [16] ;[17]

A. Space, Time and Optimal Parallel Sequencing.

SWA is the most accurate sequence alignment algorithm available and the most computationally expensive, especially for long sequences of protein. It guarantees exact matches between sequences, at the cost of long processing time. .(Altschul et al, 1990) and (Lipman and Pearson, 1985). For an example, the time and space complexity of this algorithm for comparing two sequences is O(nm), where m and n are the lengths of the two sequences being compared. The complexity in the real world applications is **O (knm),** where k represents the exponential growth of the size in genetic databases. Thus the total time complexity of the SWA is O(M + N) + O(MN) + O(MN) = O(MN). The total footprint of the SWA is also O(MN), as it fills a single matrix size MN. In order to reduce the O(MN) complexity of the matrix fill stage, multiple entries of the H(i, j) are calculated in parallel **[**20];[21].It compares segments of all possible lengths and optimizes the similarity measure. One motivation for local alignment is the difficulty of obtaining correct alignments in regions of low similarity between distantly related biological sequences, because mutations have added too much noise over evolutionary time to allow for a meaningful comparison of those regions. Local alignment avoids such regions altogether and focuses on those with a positive score. Another motivation for using local alignments is that there is a reliable statistical model for optimal local alignments.

However, SWA is fairly demanding of time and memory resources; in order to align two sequences of lengths **m** and **n**, **O(kmn) time** and space are required. As a result, it has largely been replaced in practical use by the BLAST algorithm; which is not guaranteed to find optimal alignments. These limitations therefore call for the hardware acceleration of SWA algorithm using the inherent arithmetic advantages of RNS, in order to explore the full potentials that SWA has to offer to the DNA computing.[6], [10], [18].

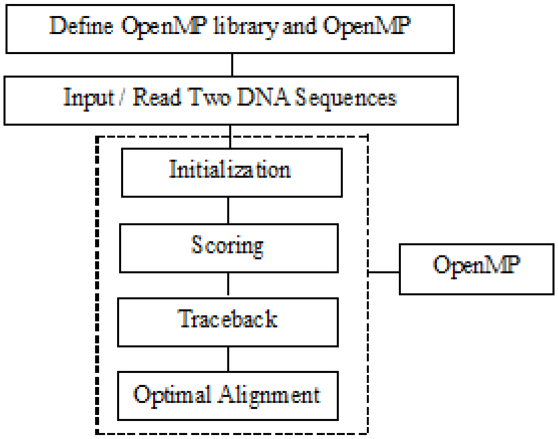

Define OpenMP library and OpenMP
Input / Read Two DNA Sequences
Initialization
Scoring
Traceback
Optimal Alignment
OpenMP

Figure 2: Parallelization of Smith program that run based on the Smith-Waterman algorithm.SWA.

As illustration in Figure 2, the implementation of parallel programming to Smith program is based on Smith Waterman algorithm that uses Dynamic Programming to find the best local alignment between any two given DNA sequences.

B. Methods for sequence alignment.

1. The Matrix Initialization:

The matrix is first initialized with Ho;j = 0 and Hi;o = 0, for all **i** and **j** (initialization step**).** Next is matrix fill step where all entries in the matrix are carried out using Equation 1.The last step is the trace back step, where the scores in the matrix are traced back to inspect for optimal local alignment. The trace back starts at the cell with the highest score in the matrix and continues up to the cell, where the score falls down to a predefined minimum threshold. For the trace back to commence, the algorithm requires to find the cell with the maximum value, which is done by traversing the entire matrix. E.g. SWA is used to compute the optimal local alignment of two sequences:

Such that:

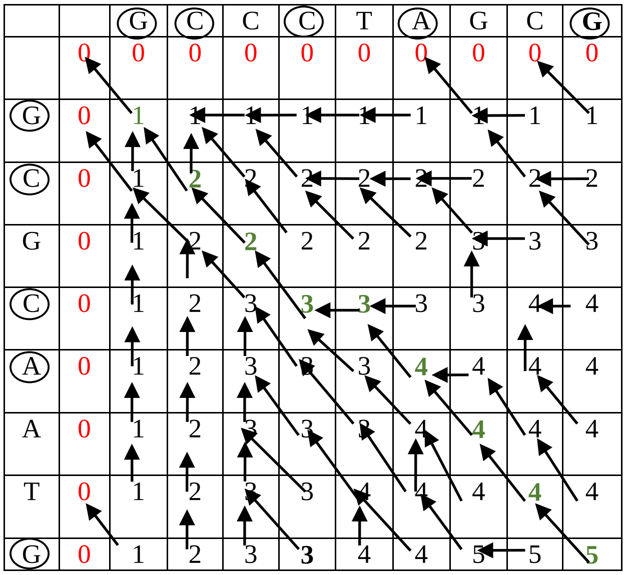
$$S_{i,j} = \begin{cases} +2 \text{ if } (A_i = B_j) \\ -1 \text{ else} \end{cases} \quad \textbf{[2]}$$

Where d = 2

Table **1** illustrates the calculation of the DP matrix H and the trace back path shown in **bold digits**. The best score found in the matrix is **5** and the corresponding optimal local alignment is

A : G C C C T A G C G
B : G C G C A A T - G

Consider the sequences: A: B with the proposed algorithm the table will begin to fill from the position (1, 1), the first entry in the first row is initialise in zeros.



		G	C	C	C	T	A	G	C	G
	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1
C	0	1	**2**	2	2	2	2	2	2	2
G	0	1	2	**2**	2	2	2	3	3	3
C	0	1	2	3	**3**	**3**	3	3	4	4
A	0	1	2	3	3	3	**4**	4	4	4
A	0	1	2	3	3	3	4	**4**	4	4
T	0	1	2	3	3	4	4	4	**4**	4
G	0	1	2	3	**3**	4	4	5	5	**5**

Table 1. The SWA Base DP Matrix and the Trace Back Path

2. Matrix filling

The first residue in both the sequence is ‘G’ and the matching and mismatching score is added to the neighboring value which is located diagonally.

Assumed scoring schemas

If the nucleotide in both sequences A and B are the same then a non exceptional user’s defined match score is assumed **(Si, j)** as + i. For i = - n,….,–3,-2,-1,0,+1,+2,+3,…+ n . It added to the diagonally located cell of the current cell such as **i, j** position. Suppose the residues are not same, the mismatch score is taken as -i. This score is added to the diagonally positioned cell of the existing cell. The gap penalty score is taken as - **n** and this score is added to the left and above positioned cells of the current cell where a negative value is assigned to the gap penalty and mismatch.

The score schema equation is shown in equation 3 as follows:

$$H11 = \text{Max}[H0,0 + S1,1, H1,0 + d, H0,1 + d, 0] \ldots\ldots \quad \textbf{[3]}$$
$$= \text{Max}\ [5*1] + [1*-2] + [3*-1]$$
$$= \text{Max}\ [5 + -2 + -3] = [0]$$

3. Trace Backing the Sequence for an Optimal Alignment

The maximum score in the matrix is **5**. So the trace back begin from the position which has the highest value, pointing back with the pointers, consequently find out the possible predecessor, then go to next predecessor and continue until it reach the score 0 the optimal alignment of smith program. From the trace back process, it can be seen that when an arrow skips a row, it is a gap in seq.A, and when it skips a column, it is a gap in seq.B.

The Smith Waterman Algorithm Pseudo code depicting the above matrix dynamic programming and the trace back path is shown below:

```
1 Declare an nxm similarity matrix;

2 Initialize the top row (i = 0) and left column (j = 0) with 0;

⌈→ 3 for i = 1; i < length (Sequence); i++ do
|
| ⌈→4 for j = 1; j < length (Sequence); j++ do
| | 5 H(i,j) = max{0;H(i-1; j-1)+S(i;j);H(i-1; j)-d; H(i; j-1)-d};
| |
| ⌊→6 end
⌊→ 7 end
8  Save index of term that contributed to the calculated value in H(i,j);
9  Find maximum value in nxm matrix;
10 Using saved indices in 8, trace back to find 0 encountered;
```

The Pseudo code implementation of the Smith Waterman Algorithm

As a result local alignment longest common subsequence (LCS) of the two DNA sequences, A and B is obtained with the resulting alignment as: **G, C, C, A, G.**

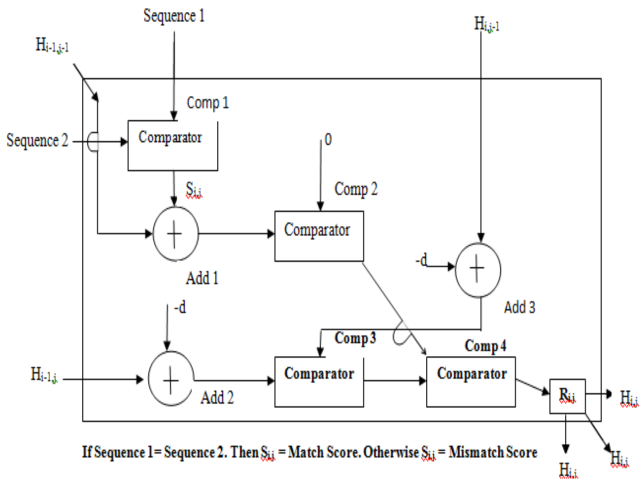


Fig. 3. Block diagram description of a basic cell for computing Hi;j values of Eq.1.

III. Implementation Using Traditional Acceleration Approach.

Figure 3 shows a block diagram of a basic cell for computing elements of the **Hi;j** matrix according to a traditional acceleration approach. **Comp1** is a comparator that compares the two input sequences and outputs the corresponding value of **Si;j**, base on the values of the match and mismatch scores, such that **Si;j** = match score, if the corresponding characters of Sequence1 and Sequence2 are equal, otherwise **Si;j** = mismatch score. Add1 is an adder that adds the diagonal element **Hi¡1; j¡1** and the value of **Si;j** . **Comp2** is a comparator that compares the output of the Add1 with a constant value 0 and outputs the greater of the two numbers. Add2 is an adder that adds the left element **Hi¡1;j** and **-d**, where **d** is the gap penalty. Add3 is an adder that adds the upper element **Hi;j¡1** and -d. **Comp3** compares the outputs of Add2 and Add3 and outputs the greater of the two numbers. **Comp4** compares the outputs of Comp2 and Comp3 and results the greater of the two numbers. The output of Comp4 is the corresponding Hi;j value, which is stored in register **Ri;j** .The matrix is initialized with the value zero. The gap penalty is assumed to have a value zero and a simple scoring scheme is assumed, such that **Si;j = 2**, if there is a match otherwise **Si;j = 0**.

The paper is organized as follows: Section 2 provides an overview of the background work as well as the merits and limitation of smith water man algorithm. Section3 highlighted solution to the limitations of SWA with the adoption of data conversion and moduli selection, matrix partitioning and RNS-SWA base processor taking the advantages of RNS arithmetic operations, and software implementation of the SWA

Section 4 shows the hardware description of the function (fill matrix) of the SWA and the profiling results for the software implementation of the S W algorithm and Section 5 gives a brief conclusion.

A. Data Conversion and Moduli selection.

Data conversion and moduli selection are two most indispensable issues for a successful RNS-SWA based processor realization [9],[10],[17] the forms and the number of moduli selected determine the speed of the sequence alignments, the dynamic range, and the hardware complexity of the resulting RNS architecture. The magnitude of the largest modulus dictates the speed of the arithmetic operations. In order to be able to use moduli set with smaller dynamic range; matrix partitioning has to be used based on the fact that the comparison of two long strings can be done in a divide-and-conquer fashion.

B. Restricted moduli selection

These are moduli set base on the powers of two and power of two related moduli. This class of moduli eliminates the need for ROM in building RNS-SWA base data converter, Prem kumar et al, 1995 [3],[6],[10],[12]. While the unrestricted moduli sets counterpart entails prime numbers chosen in sequence until the desire dynamic range M is obtained and this might not support sometimes simple conversion and RNS arithmetic computation. The solution for realizing all arithmetic operations are base on ROM so as to speed up the execution. The cost of implementing ROM based RNS data converter is generally very high. Hence, there is need for restricted moduli selection.

(Abdallah and Skavantzos, 1995, Prem kumar 1995, Parhami, 2000, Wang et al, 2003). [24] With the restricted moduli sets, the basic building blocks such as multipliers, adders, binary, **binary-to-RNS** converter and **RNS-to-binary** converters can be easily realized using logic gates.[8].The proposed algorithm takes similar advantage of both the characteristics of the two pairs of conjugate moduli sets ($2^n -1$, $2^n +1$, $2^{n+1} -1$, $2^{n+1} +1$).

IV. Hardware Acceleration implementation of the Matrix Fill Step.

In this section, we present the proposed novel method of using RNS to address the computational challenge associated with the SWA. This method exploits the potential arithmetic advantages and the modular nature of RNS to accelerate the SWA. The next paragraph accompanied is the architectural organization of the acceleration logic. The acceleration logic of the SWA implementation is made up of three major building blocks:.

A. The RNS-SWA Based Forward Converter

In this section, we describe the Residue Number System Based Smith Waterman Algorithm forward Converter (RNSFC), The Binary/Decimal to **RNS** Conversion stage as one of the components of the hardware implementation of the RNS SWA based architecture. This unit performs the conversion of the SWA inputs into their residue equivalents otherwise termed RNS Forward Converter. The Memory less Forward Converter is implemented using restricted moduli set m = $2^{n-1}{}_{+}$ 3, 2^{n-1} + 4, Where n = 5, M = 380 with (2n-1) Bits having dynamic range restricted with sub matrices element value within the range of **{-190,+189**}.The sizes, the numbers and the length of sub matrix string compared is dictated and determined by the unique decimal numbers in the RNS system that is applicable in combinational logic where the generation of the residue values is greatly simplified.

Implementation is carried out by partitioning the binary number into blocks and then concurrently carrying out modular exponentiation on all the partitions. 32 bit residue is implemented with the speed of the residue computation further increased using 5 times as many multiplexer enabling modular exponentiation to be performed in one clock cycle.

A 9 bit decimal number D is partitioned into two eight bits as x and y. With X→ X_0, X_1, X_2, X_3, X_4. X_5, X_6, X_7 as the high order bits of the binary representation of D and Y→ Y_0,Y_1,Y_2,Y_3,Y_4 Y_5,Y_6,Y_7.as the low order 2n-1 bits. The high order 8bit is added to the low order 8bit by a parallel adder (PA) The sum from PA1, identity called P1, P2, P3, P4,P5 P6, P7,P8, forms an operand for a second stage of addition emanated from X_8, the carry-out{C_{out}}and the sum from PA1 logic is executed between PA1 and PA2. D mod 19 i.e. {$\lvert X \rvert 2^{n-1}{}_{+}$ 3} representation is the sum without the carry-out {C_{out}**}.**D mod 20 i.e. {$\lvert X \rvert 2^{n-1}$ + 4} is the eight bits Y.

B. The SWA Based RNS Processor

The **RNS** based arithmetic operations stage. This is also termed the RNS SWA base microprocessor i.e. .RNS Processor stage. The logic in the control unit controls the sequencing of these additions exploiting the inherent potential properties of RNS to do carry-free Arithmetic without partial product. These binary/decimal values are converted into residues numbers by the Binary to RNS Converter (BRC), called **"**RNS forward Conversion**"** [9] [10].The residues produced using two sets of eight bit-sliced 2-to-1 multiplexers, two modulus 19 parallel adders, one modulus 20 parallel adder and a control unit are then used to execute carry free addition, borrow free subtraction by the two RNS processors in accordance with **Equation 1**.The Sequential process of these arithmetic logic operations is tabulated as follow:

	S/N	Sequential Process	Logical Component of Arithmetic Operation Process	Components
	Step1	Diagonal Addition	H(i-1,j1+S(i,j)	H(i-1,j-1; S(i, j),
	Step2	Upper Addition	H(i-1,j)+ (- d)	H(i-1,j);(-d).
	Step3	Left Addition	H(i,j-1) + (- d)	H(i, j-1);(-d)

Table2: Sequential process of arithmetic logic operations.

Each of the residue processors does concurrent data processing, independent of each other, and thereby speeding up the arithmetic operation involves in the SWA calculation as shown in the **Figure.4.**

C. The SWA Based RNS Reverse Comparator Implementation

The **RNS** magnitude comparison stage the RNS-SWA reverse comparator performs reverse conversion of the residue results of the arithmetic operation by the RNS processor to twos complement (M) binary representation and compares them with zero and with each other. The decimal values corresponding to the four values {H(i-1, j-1) + S(i, j), H(i -1, j) –d, H(i, j-1) –d and 0} being compared are read into two different registers in various clock cycles and then compared by a binary comparator. The maximum value for the matrix score assignment is output to H (i,j) as shown in Fig 4 which ends the comparisons process. Two sequential processes (Diagonal versus Left) Addition are compared, yielding to the maximum summation of the three values. These three stages are implemented on a PLD system employing the inherent arithmetic properties of **RNS**

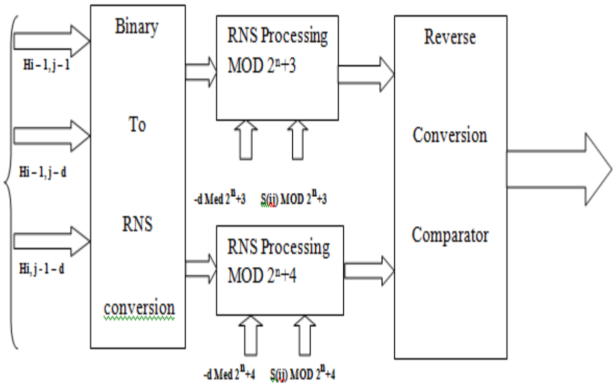


Figure 4: The Proposed RNS-SWA Accelerator Architecture

V. Performance Evaluation of the Accelerator Implementation.

The code was run using (6.4 GHz) processor, with the time period of the clock is

1
= 0.15625ns.
6.4 GHz

No of Clock cycle = Clock Ticks * 64

The actual times consumed by fill matrix functions [12] [17] = 5.23 ms
100 = 0.05232 *ms* → 52.32 µs
Total Simulation delay = 0.0146 µs
The % runtime improvement is calculated using the equation below:

$$=\left\{\frac{\frac{1}{\text{Hardware time\_fill matrix}} - \frac{1}{\text{2 time}}}{\frac{1}{\text{fill matrix 2 time}}}\right\}*100\% = \left\{\frac{\frac{1}{\text{Hardware\_time}} - \frac{1}{\text{Software\_Runtime}}}{\frac{1}{\text{Software\_Runtime}}}\right\}*100\% \quad [20] \quad \text{eq 4}$$

$$=\left\{\frac{\frac{1}{\text{Hardware\_time (A)}} - \frac{1}{\text{Hardware\_time (B)}}}{\frac{1}{\text{Hardware\_time (B)}}}\right\} \quad \left\{\frac{\frac{1}{12.012 \times 10^{-9}} - \frac{1}{14.6 \times 10^{-9}}}{\frac{1}{14.6 \times 10^{-9}}}\right\}*100\% \quad \text{Eq 5}$$

= 82.273%

The runtime improvement design is achieved by substituting the software runtime value and the propagation delay of the proposed accelerator from [Eq.4] into [Eq. 5][12][17] thus:

Percentage Runtime Ratio:

$= (14.6 \times 10^{-9} / 12.012 \times 10^{-9})* 100$

=121.5451

Hardware Runtime Improvement

= 82.272%

The performance of the proposed improvement was evaluated in terms of speed and hardware cost. The timing simulation of the proposed accelerator shows:

The total delay **= 12.012ns**
Clock speed = **185.53 MHz**

Function	**Related Scheme [12][17]**	**Proposed Scheme**	**(%) Ratio Deference**
The time period of the clock **[MHz]**	0.3120	0.15625	49.9190
No of Clock cycle	16769280	19210426	12.7000
Total times consumed by Fill matrix functions µs]	52.3200	65.5300	20.1587
Total Simulation delay **[ns]**	14.6010	12.0120	21.5534
Hardware Runtime Improvement[%]	**72.3300**	**82.2720**	**12.0840**

Table 3: Performance Evaluation Accelerator Implementation

The percentage runtime improvement of a hardware accelerator implementation of the fill matrix is achieved relatively as shown in eq 4 above.

Usage by Processor	**% Time Used**
Revision Name	RNS-SWA PROCESSOR
3 processors	0.0%
2 processors	20.0%
1 processor	100.0%
Number detected on machine	3 [100]%
Maximum used	2 [50]%
Maximum allowed	2 [50]%
Total PLLs	0/6[0%]
Total combinational functions	82.273
Total pins	9 / 329 (3 %)
Total logic element	2 / 18,752< [1%]
Total memory bits	0 / 421,318 (0 %
Total ALUTs	193 / 12,460 [1%]
Average used	1.33

Table 4: Parallel compilation report

Table 3 shows the summary report of the final compilation and the performance evaluation accelerator implementation and. table 4 shows the parallel compilation where 2 out of 18,752 total logic elements within the device are used and a negligible number of the logic cell 193 / 12,460 (1%) within the device are also used when implemented on EP2S19F484C7 device (Cyclone 11).

Testing Result: Time Consumed

Number of cells	1*SCM (ms)	Run/cell (ms)	Number of cells	1*SCM (ms)	Run /cell (ms)	Number of cells	1*SCM (ms)	Run/ cell (ms)
2*2	0.119	0.03	24*24	13.992	0.024	46*46	51.319	0.024
4*4	0.395	0.025	26*26	16.389	0.024	48*48	55.842	0.024
6*6	0.909	0.025	28*28	17.974	0.023	50*50	60.599	0.024
8*8	1.607	0.025	30*30	21.845	0.024	52*52	65.488	0.024
10*10	2.487	0.025	32*32	23.457	0.023	56*54	70.607	0.024
12*12	3.550	0.025	34*34	28.031	0.024	56*56	75.974	0.024
14*14	4.795	0.024	36*36	31.401	0.024	58*58	81.537	0.024
16*16	6.224	0.024	38*38	35.004	0.024	60*60	87.268	0.024
18*18	7.888	0.024	40*40	36.625	0.023	62*62	93.123	0.024
20*20	9.738	0.024	42*42	42.780	0.024	64*64	99.210	0.024
22*22	11.772	0.024	44*44	46.923	0.024			

Table 5: The SWA performance in custom instruction through Cyclone IV board.

The table above shows the identical length that is being tested in this paper is at ranges 1 to 64 base pair shows the performance of custom instruction through Cyclone 4 board. The time taken for each cell is average with 0.024 ms/cell. Increasing for the full run time is reduced by 3.319 to 1.065 with average runtime from 0.02 to 0.03.

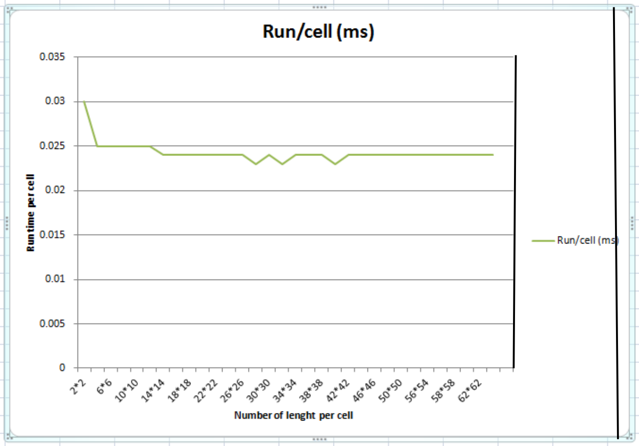


Figure 5: SWA Operating Characteristics Curve Performance in Cyclone IV board.

Fig.5. above illustrates the weakness in SWA, as the length per cell increases, the run time per cell fluctuate which later maintain constant. This calls for the inclusion with enhanced RNS implementation accelerator.

VI. CONCLUSION

In this paper, we have discussed Smith Waterman algorithm SWA which presents a kind of dynamic programming approach for identifying an enhanced local sequencing alignments of biological gene pairs. The local alignment reduces the running time and increases accuracy of the sequence matching within two sequences. Performing a Smith-Waterman search task is both time consuming and computer power intensive. The improvement of parallel sequencing based RNS reduces greatly the system memory and time. RNS design comparator with two moduli sets having a dynamic range of (2n -1) bits were used. The percentage speed gained in our work is 82.273 % faster than [17].

REFERENCES

[1]Premkumar,B.(1995) “An RNS to binary converter in a three moduli set with common factors,” IEEE Trans. on circuits and systems - II analog and digital signal proc., vol. 42, no. 4, pp. 98 – 301.

[2] Arslan A.N.(2005) "Multiple Sequence Alignment Containing a Sequence of Regular Expressions "Computational Intelligence in Bioinformatics and Computational Biology, Proceedings of the IEEE Symposium on. 14-15 pp.1-7.

[3] Brian Hang; Wai Yang; (2009) “A Parallel Implementation of Smith-Waterman Sequence Comparison Algorithm”; December 6, 2002.China),.

[4] Gish W. Miller W. Myers E. W. Altschul, S. F. and D. J. Lipman.(1990) A basic local alignment search tool. J. Mol. Biol., 215:403–410,.

[5] K .Gbolagade and S. Cotofana, Nov (2008) “A residue to binary converter for the moduli set (2n+2; 2n+1; 2n),”

[6]Gbolagade.K and Cotofana.S,(2009) “An O(n) Residue Number System to Mixed Radix Conversion,” (To appear) in proceeding of the IEEE International Symposium on Circuits and Systems (ISCAS 2009), (Taiwan).

[9]Gbolagade.K and Cotofana,S. .(2008) “Generalized matrix method for efficient residue to decimal conversion,” No. 3 in Proceedings of 19th IEEE Asia Pacific Conference on Circuits and Systems Macao, China), pp. 1414 – 1417,

[10]Gbolagade.K.andCotofana,S.((2008) “Residue number operands to decimal conversion for 3 - moduli set,” August,.

[1]Hasan,L.Z.AlArsandVassiliadis.S.(2007) Hardware Acceleration of Sequence Alignment Algorithms-An Overview, Proceedings of International Conference on Design & Technology of Integrated Systems in Nanoscale Era (DTIS’07): 96–101,

[12] Laiq Hasan, Zaid Al-Ars. Performance improvement of the Smith-Waterman. Delft University of Technology Computer Engineering Laboratory Mekelweg 4, 2628 CD Delft, The Netherlands.

[13] Giegerich..R. (2000) A systematic approach to dynamic programming in Bioinformatics, vol. 16:665–677.

[15]Wang, .Y.(1998)“New Chinese remainder theorems,” vol. 1 of in proc. 32nd Asilomer Conf. signals, systems computing, pp. pp. 165 – 171..

[16]Wang,Y..Song,X. Aboulhamid,M. and Shen:,H. (2002)“Adder based residue to binnary number converter for 2n-1;2n;2n+1,” IEEE Tans. on Signal processing, vol. 50, no. 7, .

[17] Hasan, L., and Al-Ars, Z.,(2007) November 29–30,, “Performance Improvement of the Smith–Waterman algorithm,” Annual workshop on circuits, systems and signal processing (ProRISC). Veldhoven,The Netherlands.

[18] Soderstrand, M. A., Jenkins, W. K., Jullien, G.A., and Taylor, F. J., (1986), Residue Number System Arithmetic: Modern Applications in Digital Signal Processing, New York: IEEE Press.

[19] Conway, R., and Nelson, J., (2004), “Improved RNS FIR Filter Architectures. IEEE Trans. on Circuits and System – II,” Express briefs, Vol. 51, No. 1, pp.26 – 28**.**

[20] Smith, T.F., and Waterman, M. S., (1981),“Identification of common molecular sub sequences, “journal of molecular biology, vol. 147, pp 195– 197.

[21] Altschul, S. F., et al., (1990), “A Basic Local Alignment Search Tools,” In Journal of Molecular Biology, vol 215, pp. 403 – 410.

[22] Proceedings of the 13th International Workshop on Field Programmable Logic and Applications.

[23] S.K. Moore. (2000). Understanding the human genome (vol. 11: pp, 34 – 35). IEEE Press.

Towards an Awareness Model to Caution and Mitigate Privacy and Security Invasion on Social Networking Sites

[1]D. Du Plessis
[1]School of Information and Communication Technology, Nelson Mandela Metropolitan University (NMMU), Port Elizabeth, South Africa
divanduplessis@rocketmail.com

[2]G. Thomas
[2]Department of computer Science University of Jos, Plateau State Nigeria
thomasg@unijos.edu.ng

Abstract— **Utilizing Social Networking Sites (SNSs) have become part of a daily life for many individuals. This means information previously kept offline, is now stored online. Since this information floats online, all it takes is for a tech savvy individual to gain access in order to cause chaos. Thus, due to the growth of users utilizing social networking sites, the threat of privacy and security invasion online becomes more prevalent. Users unaware of these threats are often subject to various forms of online invasion and can thus damage their own and their institution's reputation. This paper explores what an awareness model should constitute to caution and mitigate privacy and security invasion on SNSs. By employing an extensive literature review, logical reasoning and argumentations, social networking site threats/challenges, security mechanisms, and related systems are explored towards characterizing an awareness model to caution and mitigate privacy and security invasion on SNSs.**

Keywords— Challenges/Threats, Security Mechanisms, Awareness Model, Component Framework, Proof-of-concept prototype

I. INTRODUCTION

Social networking sites has taken previously private conversations from offline to online. Facebook and Twitter among others lead the charge onto an open and social web that can reveal information for all to see [1] Whenever an individual interacts with a computer, views websites, posts on Facebook and Twitter or "Likes" something, an online trail is left; also known as a digital footprint which makes it possible to easily form an image of the individual [2] that can subject them to certain harm or attack [3].

A large amount of information, including basic personal information and private conversations are all stored somewhere online and this warrants caution because any individual with tech savvy skills can gather this information without having to resort to illegal means, which is evident on sites such as Facebook and Twitter among others.

The problem is that too many individuals utilizing social networking sites are unaware of threats to their privacy and security due to vulnerability of information they share online and are thus subject to various forms of online invasion. It is the argument in this research that by equipping individuals with the requisite knowledge the rate of social networking invasion can be mitigated. Awareness principles suggests that having the right information can help to define activities and the expectation of users [4] especially if personalized context-ware content delivery is often prescribed [5]. Thus, a lack of insight or understanding of situational context can result in ignorance that can be very costly to an individual.

In order to determine how to address this issue of privacy invasion on SNSs, this research latches on the awareness principle. Thus, ensuring that individuals are well aware of the SNSs vulnerabilities, dangers and what they can do to mitigate thing will help instill a privacy and security culture. The next section elaborates on the threats and challenges associated with SNSs.

II. THREATS AND CHALLENGES OF USING SNSs

Sharing information with the masses online generally requires utilizing SNSs, which is the tool used to communicate online [6]. Because SNSs provide the ability to share information online, it has the ability to monitor all communication and information provided by the user. This unique position has sparked a lot of discourse as to the vulnerability of SNS users' privacy.

More and more SNSs develop the ability to observe and record everything that user's say and do while utilizing their site [7]. Thus, unaware users are increasingly relying on the ability of SNSs, not only to provide quality informational sharing services but also to store and manage private and confidential information.

However, as wonderful as the benefits of SNS's are, issues like privacy remain a challenge because the user's concerns are focused on different features of privacy and often don't live up to the required privacy preferences [8; 3]. Thus, are users remain vulnerable to some certain and immediate threats as will be highlighted in the following sub sections.

A. Cyber-crime

Cyber-crime is described as criminal activities carried out by the use of computers or the Internet. The growing list of SNS cyber-crimes include crimes against individuals, such as cyber-bullying, spamming, phishing and malware (viruses, spyware, etc.), and also crimes against companies, such as employees sharing information via SNSs, thus, exposing company computer networks to potential cybercrime threats [9; 3]. Nale [10] and Abdulahi et al., [8] highlight other crimes and extent to which cybercrimes are growing, such as blackmail/extortion and electronic harassment among others which can be used against both, individuals and companies.

It is important to understand that the same aspects which let us find friends on Facebook, LinkedIn and MySpace, is a great way of aggregating the identities of a given human and can lead to many of the aforementioned cybercrimes. Equally important is knowing that when a cyber-crime occurs, information such as when the user logged onto the internet and the specific IP address that was assigned to the user, can be used to locate the user who committed the cyber-crime. Herewith follows a brief description of various crimes/forms of invasion:

Fake Apps – Symantec's Haley, as supported by [11] explains that the apps appear to be legitimate but often they contain a malicious payload. The phony apps increasingly being designed for mobile devices, masquerade as released free versions of popular legitimate apps. Some of these apps use aggressive advertising tactics to sell the user's data and browsing habits to third-party advertising networks, as in the case of Facebook.

Like-jacking – Symantec's Haley, as supported by [11] explains that using fake "like" buttons, attackers trick users into clicking website buttons that install malware and may post updates on a user's newsfeed, spreading the attack. A common scam that attempts to get users to enable a phony Facebook "Dislike" button continues to get detected from time to time [12].

B. Complaints/issues associated with social networking sites

By mid-2010, Facebook recorded half a billion active users, making it not only the largest social networking site, but also one of the most popular destinations on the web [13]. The Sophos Press Release [13] continues by mentioning that unsurprisingly, the massive and committed user base is heavily targeted by scammers and cybercriminals, with the number and diversity of attacks growing steadily throughout 2010 – malware, phishing and spam on SNSs have all continued to rise during previous years. Hence, a logical assumption can be made that the same holds true for other successful SNSs such as MySpace and Twitter, to name but a few.

Thus, it is vital to increase awareness of unsuspecting users by means of exposing which SNSs these attacks frequently could occur on in a comparable and presentable manner as summarised in Table 1. This table has been formulated based on the foregoing discussion on complaints and forms of invasion associated with SNSs relating to the uncertainty with which unaware users utilizing SNSs might presume they could be attacked from.

Table 1: Complaints and forms of invasion relevancy to provided Social Networking Sites

Complaint/Form of Invasion	***Facebook***	***Twitter***	***MySpace***
Misleading, constantly changing privacy policies	√	x	x
Law and government, accessing your phone, email and website data to catch criminals	√	√	√
Accounts being hacked and used for illicit purposes	√	√	√
Use of the "like" button to gather information regarding your online habits and then building a social profile	√	x	x
Clickjacking	√	x	√

Besides the complaints and attacks/forms of invasion shown in Table 1, it can be viewed that various SNSs, share similar vulnerabilities, that when exploited, pose a threat to the privacy of unaware SNS users of whom are oblivious to which SNSs are frequently targeted.

C. Deep packet inspection (dpi)

A capability rather than a tool, Deep Packet Inspection (DPI) is built into firewalls and other network devices and along with packet capture technologies revolutionised network surveillance over the last decade by making it possible to grab information from network traffic in real time. A Belgian collection society, SABAM, had been trying to coerce various sites, including the locally-large social network Netlog, into installing a monitoring system that would automatically detect infringements of copyright [14].

Regardless of how DPI may be used, with the multitude of information being gathered, users' online behaviour and patterns are formed, enabling the construction of online user profiles based on certain criteria.

D. Online Profiling

Profiling internet users on the basis of their online behaviour and activities is perceived to be a valuable trading asset to many advertising companies and third parties. This information is rich with personal preferences and patterns users display while communicating online.

Data retained by SNSs holds various benefits and drawbacks for both aware and unaware users. The benefits and drawbacks of users' data retention and the SNS's necessity to increase awareness among users are summarised in Table 2. This table has been formulated based on the foregoing discussion on the unique abilities of SNSs relating to their popularity as well as their privacy concerns.

Table 2: The Benefits and Drawbacks of Data Retention by Social Networking Sites

Benefits	***Vulnerabilities/Drawbacks***
Opportunity to meet new people	Identity theft made easier
User friendliness	Overwhelming & addictive

Ability to join groups sharing interests & inviting people to events	Online bullying & cyberstalking
Free to use	Scammers trick people into downloading malware
Allows for professional social networks to establish their brand online	Social Profiling & 3rd party info disclosure without consent
Can act as an application platform	Can be a time waster
Assist in combating terrorism	No more freedom of expression

Besides the advantages shown in Table 3, it can be viewed that data retained by SNSs, including the potential for online profiling, poses a grave threat to the privacy of unaware SNS users. In order to identify what requirements are necessary to help mitigate or prevent the disadvantages as mentioned in table 2, existing models/frameworks are discussed in section IV. The methodology employed towards realizing the objective research objective is highlighted next.

III. METHODOLOGY

Through an in-depth literature review threats, countermeasures and best practices are analyzed with the lessons learnt serving as input to conceptualizing the awareness model. An extensive literature review is made to identify issues associated with SNS utilization as well as the strengths and limitations of existing counter measures to ensure privacy and security of SNSs. Through a comparative analysis and combinations of complementary strengths as projected in existing solutions, a holistic approach towards the mitigation of the security and privacy invasion on SNSs is contended. Based on logical reasoning the proposed awareness model is characterized and premised on lessons learnt from existing approaches that exist to protect users' online privacy and confidentiality among other things. The next section highlights on some of the existing approaches.

IV. EXISTING MODELS/FRAMEWORKS THAT SAFEGUARD PRIVACY

This section explores from literature existing models and frameworks that reflect privacy invasion management in some way. These models will aid in discerning what components are needed to help identify what the component model should consist of. The following subsection describe some of the models/frameworks, while explicating their lessons.

Information Technology Management (GS-2210) Competency Model - This model is designed to assist with the selection of the most applicable competencies to an individual's position. The model allows users to complete a self-assessment which can help identify areas in which to focus development as well as reviewing the competencies dictionary to identify additional non-technical competencies that are essential such as adaptability, advisory assistance, collaboration, communications, data gathering and analysis to name but a few.

Lessons learnt – GS-22110 the model suggests identifying the competency level of an individual is important in order to adequately, tailor the education training or awareness. Programmes. The Competency Model suggest that having policies and assessments in place can assist. Therefore, a conclusion can be made that having such capability will contribute to the awareness model as a whole.

Common Information Model (CIM) - Rice [15] states is an open industry standard that defines how the industry represents objects and the relationships between those objects in IT environments. Rice [15] further states that the CIM allows IT administrators to control hardware and software from different manufacturers in a universal way.

Lessons learnt – The CIM Model suggests the importance of procedures, information sharing and knowledge of working relationships can assist in providing renewed knowledge about available control mechanisms, and the possibility of increasing awareness.

ISO 27701 – international standard for data privacy – the first ISO 27000 series dedicated to privacy, explains how organisations can create a privacy information management system to meet best practices outlined in regulations such as the General Data Protection Regulation (GDPR). It ties into the ISO 27001, which deals information security, as a screw for privacy processing control. Essentially, it relays how organisations should collect personal data and prevent unauthorised use or disclosure.

Lessons Learnt – The ISO 27701 model suggest that When structuring an information security framework, organisations must take extra steps to ensure that privacy concerns are accounted for by for example not collecting excessive amounts of information on an individual, that can lead to privacy violation while ensuring that unauthorised employee or cyber-criminal don't get hold of the data. Furthermore, in addition to creating procedural and technical controls it suggests keeping apprised with Education training and awareness needs of stakeholders.

V. SECURITY MECHANISMS (COUNTER MEASURES)

According to Privacy Rights Clearinghouse [16], there are currently few laws maintained in the United States that can be interpreted as protecting information given to social networks. One such law, the Electronic Communications Privacy Act was passed in 1986 which holds that if information is stored on a server (such as on SNSs), then the law makes it easy for law enforcement or the government to access it via a court summons/subpoena [16].

South Africa (SA), recently began participating in the development of producing acts to prevent/mitigate cyber-crimes on SNSs. With the advent of acts such as: ECT; Popi; RICA; the Prevention of Organised Crime Act (POCA) and the Financial Intelligence Centre Act already available, it is believed that acts such as the ECT, should be tried and tested now rather than later for case laws to be followed [17].

Regarding the use of strategies, it is interesting to note that based on the results of a study done by Anabel Quan-Haase & Alyson Leigh Young [18] relevant to Facebook, many university students have taken the approach to protect their social privacy (The concern about controlling access to personal data) rather than their institutional privacy (The concern about how third parties will use personal data) more seriously. The results of the study concluded that strategies at protecting privacy included regulating access to tagged photos, restricting access to their walls, changing the visibility of their tagged photos to "only me", etc.

Thus, even though policies and controls may at a technical and corporate level prevent or mitigate online invasions, it does guarantee that without awareness, these online invasions will cease to occur and therefore, the importance of awareness models and/or programs need to be discussed. Highlighted in the next section are some types of crimes and how they apply to SNSs.

Table 3: Social networking site crimes/forms of invasion and their corresponding control mechanisms

Crime/ Form of Invasion	*Security Control Mechanism*
Like-Jacking	• Installation of like-jacking prevention extensions • Carefully review wall posts if tempted by potential scam • Review installed apps periodically
Evil Twin Attacks (Financial gain; Defamation; Stock churn; Cyber-bullying)	• Don't befriend the evil twin by ensuring person is who they say they are (Call person to ask whether they sent request; send email to verify the person themselves; etc. • Don't allow social networking usage at work • Educate employees on this type of attack and inform them on what they should do to help protect themselves • Assist employees of company in monitoring their social networking profiles for Evil Twin Accounts
Cyberbullying	• Don't acknowledge the message • Minors should report any abusive messages to adults immediately • Save and archive messages • Children; teens; and young adults should be educated Relevant to workplace • Introduce a policy that dictates acceptable usage with technology so as to safeguard companies from lawsuits
Malicious Software	• Do not click on unknown links • Never open e-mail attachments from people not known • Do not accept friends not known • Do not use apps of unfamiliarity • Ensure privacy settings are configured • Have antivirus installed and operating • Downloaded files should be scanned through antivirus • Have antispyware installed and operating • Utilize newest patches for software to remain up-to-date • Disable cookies • Do not connect to other sites while connected to your bank account **Relevant to workplace/corporations** • Implement security awareness program/model • Limit the use of peer-to-peer networks • Restrict administrative rights • Disable active content

Having examined the cyber-crime security control mechanisms shown in Table 3, and keeping in mind the information provided in Table 1, it should be noted that these crimes can still pose a threat to the privacy of unaware SNS users. Having discussed threats and countermeasures, and after investigating various existing models/frameworks, the requirements for the model that have been identified will be discussed in the next section.

VI. REQUIREMENTS FOR AWARENESS MODEL TO CAUTION/MITIGATE PRIVACY INVASION ON SNS

This section is based on the discussion of the previous section aided in identifying the requirements of an awareness model. Therefore, it must account for three (3) major types of control – Tactical, Mechanistic and Review/Support.

The tactical phase is the first instance of the strategic direction of mitigating the privacy and security risks of stakeholders which requires planning, information sharing/knowledge of working relationships, analysis, policies and procedures. Having a management component that will assist in assessing user awareness of current SNS threat knowledge which will strive to educate users on threats concerning their online reputation, advising on available countermeasures and available guidelines, hence, creating an awareness plan as far as this aspect is concerned, to adequately caution and mitigate privacy invasion. This must be accomplished by initiating the Security Education, Training and Awareness (SETA) aspect and is concerned with managing and manoeuvring how people should behave when dealing with certain online privacy threats when trying to safeguard their online reputation whilst utilizing social networks. Thus, the model proposed must account for them.

Secondly, the mechanistic/technical aspect include intrusion detection/prevention features, notification features, access control features and integrity check features. Having a mechanistic management/technical control component that will assist in handling implementation of protection technology and services to help execute the awareness plan. This aspect also involves SETA plan which involves educating the organization on the security best practises and the use of SNSs. This must be accomplished by requiring the implementation of controls to mitigate the threats as discussed in the cyber-crimes section and table 3. This aspect involves the implementation of controls from two perspectives namely, the administrative and the technical perspectives. Thus, the model proposed must account for them.

Thirdly, the review and support aspect include documentation, activity monitoring and logging, evaluation and audit controls. Its purpose is to assist in monitoring the maturity of the model, document any new changes and implement them to further develop the model's usefulness, to periodically review threats and update the model when necessary and to enforce audit checks to maintain the integrity of the model. This must be accomplished by incorporating both the tactical management and the mechanistic management/technical control aspect which is responsible for monitoring and reviewing both the previous aspects of the model to ensure effectiveness. Improving the previous two (2) aspects will ensure corrective and preventive actions are taken to meet due diligence. Thus, the model proposed must account for them.

VII. THE AWARENESS COMPONENT FRAMEWORK

Protecting the privacy confidentiality, and integrity of social media users is no longer just best practice when protection an entity's online reputation, but a legal requirement. As discussed in previous chapters, the need for a tactical management component is emphasized. And as indicated in the existing models/framework section it should consist of information sharing/knowledge of working relationships, analysis, policies and procedures. Also as indicated in the control mechanisms the necessity for a technical/mechanistic component needs to be in place too. Furthermore, these two components should be interconnected through a continual review process as shown in figure 1.

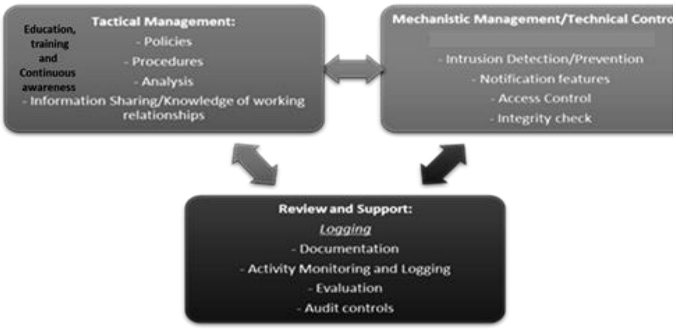

Tactical Management:
Education, training and Continuous awareness
- Policies
- Procedures
- Analysis
- Information Sharing/Knowledge of working relationships
Mechanistic Management/Technical Contro
- Intrusion Detection/Prevention
- Notification features
- Access Control
- Integrity check
Review and Support:
Logging
- Documentation
- Activity Monitoring and Logging
- Evaluation
- Audit controls

Figure 1: The component framework for Social Networking Security Awareness

The analysis in section 4 showed the need to combine different elements that complement each other in order to have a holistic and compliant social media privacy and security solution. The results of the investigation showed that although current models contain most components, there are gaps when dealing with raising users' awareness regarding privacy matters which need to be filled.

The framework as shown in figure 1 will help identify and analyse the core aspects and components that should be featured in social media awareness models in order to comply with security best practices and legal requirements. Furthermore, it provides requirements to aid in the design and building of a proof of concept prototype which could be implemented to help users gain more awareness regarding social media security issues and how to protect theirs and their institutions' online reputation. The aspects considered are discussed in the next sub sections.

Tactical Management phase (People Aspect) in figure 1 assesses user awareness of current SNS threat knowledge and strives to educate users on threats concerning their online reputation, advising on available countermeasures and available guidelines, hence, creating an awareness plan. More so behavioural approaches in form of procedures. This phase is important because of the threats discussed earlier and will be valuable in helping users identify more easily the dangers they face on SNSs and how to avoid it by becoming more aware of their current situation. It accomplishes this by integrating lessons from table 3 and section 6. Furthermore, subtle reminders in the form of pop ups on screen (tool from the mechanistic phase) should remind users to remain vigilant. Thus, setting conscious behaviour baseline for users especially if dealing with large users like staff and students of an institution.

Mechanistic Management/Technical Control phase (Artificial/Mechanical Aspect) in figure 1 handles implementation of protection technology such as intrusion detection, administrative and real time access control among others and services to help execute the awareness plan. Its purpose is to monitor the maturity of the model, document any new changes and implement them to further develop the model's usefulness, to periodically review threats and update the model when necessary and to enforce audit checks to maintain the integrity of the model. This phase is important because of the threats discussed earlier and will be valuable in helping users counteract these threats by using tools and tips to help mitigate these attacks. It accomplishes this by integrating lessons from table 3.

Review and Support phase (Review Aspect) in figure 1involves managing the security awareness model to serve the organisation or users' objectives. The aspect ensures that awareness tips and advice remain up-to-date and that SNS privacy threats don't go unnoticed. Thus, inform strategic policy changes that drives the tactical phase. This aspect is closely regulated while interchangeably changing between the tactical management and mechanistic management components to ensure they're properly enforced. This phase is important in aiding the review process to ensure new information is properly documented and new tips are integrated properly as well as determining maturity of the model over a course of time.

VIII. CONCLUSIONS AND THE FUTURE

In order to conceptualize model that will aid caution and mitigate privacy and security concerns among users an extensive literature review was conducted to determine and argue the necessary requirements and elements that make up the model. This objective was achieved by using argumentation and reasoning in order to conceptualize what the component model should consist of. It is believed that the model provides a stepping stone towards building a tool in the form of an add-on to a browser to help users manage privacy and security concerns by displaying tips and warnings as well as features to help caution and mitigate privacy invasion on SNSs. Thus, future work will focus on establishing a proof-of-concept prototype premised on the requirements to caution and mitigate privacy invasion on SNSs. More so, determine and evaluate the validity of the awareness model and whether users utilizing SNSs would subscribe to the proof-of-concept prototype as service which could be added on to browsers. Furthermore, whether the development of a game for youngsters/teens which can help develop their awareness early on. Other future research could focus on integrating a service which is custom made and tailored to specific institutions and/or businesses. As noted by [19]: “Forgetting is a truly pervasive problem while, As such, we must act with vigour and vigilance in order to ensure our privacy and security.

REFERENCES

[36] Bullas, J. (2014). Is Social Media a Serious Threat to Your Privacy? - Infographic. Retrieved Apr. 21, 2014, from Jeffbullas's Blog - Internet Marketing: http://www.jeffbullas.com/2012/02/23/is-social-media-a-serious-threat-to-your-privacy-infographic.

[37] Karena, C. (2013, May 30). Data collection a growing threat to our privacy. Retrieved Apr. 21, 2014, from Australian Breaking News Headlines & World News Online | SMH.com.au: http://www.smh.com.au/digital-life/digital-life-news/data-collection-a-growing-threat-to-our-privacy-20130529-2na5i.html.

[38] Keeley, B. Little, C., (2017). The state of the World Children 2017; Children in a Digital World. Retrieved Jan 12, 2020 from UNICEF: *https://www.unicef.org/publications/files/SOWC_2017_ENG_WEB.pdf*

[39] Röcker C. (2010) Information Privacy in Smart Office Environments: A Cross-Cultural Study Analyzing the Willingness of Users to Share Context Information. In: Taniar D., Gervasi O., Murgante B., Pardede E., Apduhan B.O. (eds) Computational Science and Its Applications – ICCSA 2010. ICCSA 2010. Lecture Notes in Computer Science, vol 6019. Springer, Berlin, Heidelberg

[40] Kirsch-Pinheiro, M., Gensel, J., & Martin, H. (2004). Representing Context for an Adaptative Awareness Mechanism. In G.-J. de Vreede, L. Guerrero, & G. Marín Raventós (Eds.), Groupware: Kofod-Petersen, A. (n.d.). Using Activity Theory to Model Context Awareness : a Qualitative Case Study.

[41] Dollarhide M.E. (2019, May 02). Social Media Definition. Retrieved Jan 15, 2020 from INVESTOPEDIA: https://www.investopedia.com/terms/s/social-media.asp.

[42] Abdulahi, A., Samadi, B., Gharleghi, B., (2014). A Study on the Negative Effects of Social Networking Sites Such as Facebook among Asia Pacific University Scholars in Malaysia. *International Journal of Business and Social Science, 5 (10).*

[43] Ntlatywa, P. (2012). Factors that Influence the Choice of Privacy Settings on Social Networking Sites. 18.

[44] Sophos Press Release. (2014). Security Threat Report 2011. Sophos Ltd. Retrieved May 07, 2014

[45] Nale, S. (2012). The 10 Most Common Internet Crimes Complex. http://www.complex.com/tech/2012/11/the-10-most-common-internet-crimes/blackmail-extortion..

[46] Westervelt, R. (2013). Top 5 social networking attacks you should dodge. Retrieved Jun. 9, 2014, from CRN: http://www.crn.com/slide-shows/security/240163136/top-5-social-networking-attacks-you-should-dodge.htm/pgno/0/5.

[47] Wisniewski, C. (2020, Jan 28). What is "Likejacking"?. Retrieved Jan 28, 2020, from Sophos:https://www.sophos.com/en-us/security-news-trends/security-trends/what-is-likejacking.aspx

[48] Sophos Press Release. (2017). Sophos Security threat Report Reveals Increase in Social Networking Security Threats. Retrieved May 09, 2014, from Sophos: http://www.sophos.com/en-us/press-office/press-releases/2011/01/threat-report-2011.aspx

[49] Geere, D. (2012, Feb. 16). Social Networks don't have to Police Copyright, Rules EU. Retrieved May 09, 2014, from Wired.co.uk: http://www.wired.co.uk/news/archive/2012-02/16/eu-social-networks-copyright.

[50] Rice, J. (2014). What is a common Information model. Retrieved May 30, 2014, from Ask: http://www.ask.com/question/what-is-a-common-information-model.

[51] Privacy Rights Clearingouse (PRC). (2019). *Social Networking Privacy: How to be Safe, Secure and Social.* Retrieved Jan 08, 2020, from Privacy Rights: https://privacyrights.org/consumer-guides/social-networking-privacy-how-be-safe-secure-and-social

[52] Reilly, K. O. (2013). South African Law Coming to Grips With Cyber Crime. De Rebus, 3.

[53] Quan-Haase, A., & Young, A. L. (2013). Privacy Protection Strategies on Facebook: The Internet Privacy Paradox Revisited. Information, Communication & Society, 16(4), 16.

[54] National Institutes of Health. (2014). Office of Human Resources. http://hr.od.nih.gov/workingatnih/competencies/occupation-specific/2210/

A Survey on Slow DDoS Attack Detection Techniques

Oluwatobi Shadrach Akanji
Department of Computer Science
Federal University of Technology
Minna, Nigeria
akanjioluwatobishadrach@yahoo.com

Opeyemi Aderiike Abisoye
Department of Computer Science
Federal University of Technology
Minna, Nigeria
o.abisoye@futminna.edu.ng

Sulaimon A. Bashir
Department of Computer Science
Federal University of Technology
Minna, Nigeria
bashirsulaimon@futminna.edu.ng

Oluwaseun Adeniyi Ojerinde
Department of Computer Science
Federal University of Technology
Minna, Nigeria
o.ojerinde@futminna.edu.ng

***Abstract*— The ease with which DDoS attack is being launched using publicly available tools has made DDoS to be a recurring security problem. However, given the immense work by researchers to stem the tide of volumetric DDoS, attackers have resorted to using a slow DDoS attack which is similar to benign traffic thus making detection and mitigation difficult. This paper seeks to provide the scholarly community with a survey on slow DDoS attack detection techniques worked upon by researchers over time. A low amount of work has been done when the work on slow DDoS detection is juxtaposed with that of volumetric DDoS. However, researchers who have worked on detecting slow attacks have achieved remarkable results. Machine learning detection technique has proven to be effective with random forest and K-Nearest Neighbour (KNN) being the major algorithms that have consistently achieved good results in terms of Area Under Curve (AUC), accuracy, and false positive rate. Other detection techniques of time series and performance model have also been effective against slow DDoS but need to be improved upon given the non-linearly separable nature of a slow attack and benign traffic. Most researchers resorted to using attack tools to generate attack data due to the absence of a standard data set. Recommendations for future studies include exploration of detecting slow table overflow attacks in SDN before a table overflow event occurs.**

Keywords—Slow DDoS, Slowloris, Slow POST, Slow Read, Slow attack detection, Slow HTTP

I INTRODUCTION

The threats to devices in a networked environment keeps on metamorphosing because of the variety in network devices, protocols, and configuration. Among these threats is Distributed Denial of Service (DDoS). DDoS attacks involve the use of a large number of Internet-enabled and connected devices to synchronously send illegitimate requests to a target thus overwhelming the target's capacity to respond to the requests [1]. The manipulation of data transfer rates which consequently consumes the target's resources is one of the strategies used to cause a DDoS. A DDoS situation is reached when the attacker maintains connections or sends data to the victim which results in the unavailability or improper functioning of the services offered by the victim to legitimate users. According to [2], volumetric and application layers are the major categories of DDoS attacks. Volumetric attacks are characterized by large data transfer rate launched against the targets which exhaust the bandwidth of the target's links or the memory storage and processing power of the target. Unlike volumetric attacks structured on the network and transport layers, the application layer attacks exploit the behaviour of application layer protocols thereby increasing detection difficulty and circumventing network and transport layers DDoS detection mechanisms. The application layer attacks could employ either fast or slow data transfer rate to achieve DDoS. The use of slow or low data transfer rate to achieve application layer DDoS requires establishing and maintaining connections with the victim for prolonged periods hence, hindering service availability to legitimate clients. Slow data transfer rate DDoS are also known as slow DDoS.

Slow DDoS attacks are generally application layer attacks that exploit application layer protocols of HTTP, FTP, IMAP, and SMTP. Unlike volumetric DDoS, it utilizes less bandwidth and small computational resource of the attacker [2]. The low bandwidth usage characteristic of slow DDoS enables it to evade detection because the data transfer rate bears semblance with that of either a legitimate user with a slow connection or one whose device has low data transmission capacity [3]. The attacker occupies most or all the service queues at the application level thereby causing incoming requests to be discarded [4]. Slow HTTP DDoS, an attack against web servers, is the most prominent in this category which can be attributed to the vast amount of web servers. A slow HTTP DDoS attacker establishes a connection with the webserver using the three-way handshake protocol after which the connection is maintained using a few amount of data [5]. Although it is true in some situations that slow DDoS attacks focus on slow data transfer rate, it also entails the use of few amount of data relative to the bulk of data requested for or being transferred to sustain a connection to the victim [4][6]. The advantages of the slow DDoS attacks which includes detection evasion, low attack resource requirement, and easy configuration endears it to DDoS attackers. Also, the ability to launch a slow HTTP DDoS attack from a mobile phone has compounded the problem of detection and mitigation given the wide use of mobile phones for Internet connectivity [7]. In general, these DDoS attacks are aimed at targets such as OpenFlow switches, web servers, file servers, and mail servers.

II. TYPES OF SLOW ATTACKS

Classification of slow DDoS attacks is based on either the application layer or the device an attack targets. The types of slow DDoS attacks with their targets are examined in this section.

a. *Slow HTTP DDoS*

A slow HTTP DDoS is a type of DDoS which exploits the way the HTTP protocol on web servers operates particularly the lack of time-bound active connection rules and the need to wait for the completion of requests [8][9]. HTTP is one of the most popular Internet protocols which executes on the TCP/IP protocol suite, the backbone of the World Wide Web (WWW). Whenever a HTTP request is sent to a web server, the request is accompanied with a header which contains information such as window size, protocol version, and window scaling necessary for the webserver to process the request and send the required response appropriately [9]. To launch a slow HTTP DDoS attack, a normal TCP or UDP connection is first established with the victim and then the attacker seeks to maintain every connection established by either sending or reading few amounts of bytes to or from the webserver. There are three types of slow HTTP DDoS: the slow read, the slow POST, and the slow GET. To hide the attack origin, the attacker may utilize HTTPS as the transport protocol to establish and maintain connections [9].

Slow Read

A slow read DDoS attack is aimed at causing the unavailability of web services to legitimate clients by requesting for data resource from the web server and then forcing the victim to send the reply at a low rate [5][10]. After establishing a connection with the webserver, the attacker requests for a resource while advertising a small TCP window. The TCP window defines the number of bytes readable by a client. The attacker ensures that the TCP window advertised is smaller than the web server's buffer size thus causing delays which fills up the webserver's buffer with responses waiting for dispatch [11]. In some cases, the attacker advertises a TCP window size of 0 bytes which makes the web server wait indefinitely for the client to be available for response receipt, however, timeout mechanisms and zero-byte window detection mechanisms implemented on the webserver makes the attack easy to detect. Hence, attackers may resort to using varying amount of bytes large enough to sustain the connection and evade detection mechanisms but small enough to cause a DDoS scenario [12]. The attacker continues to establish numerous connections to the webserver until it has occupied most or all of the available connections on the webserver. This ensures that there is an increase in the web server's response time or availability of the web service to legitimate clients is none existent. The method of operation for a slow read DDoS attacker is illustrated in fig. 1.

Slow POST

Unlike the slow read attack, the slow POST attack sends data to the webserver at a rate that maintains the connections established for a long period. The slow POST attack is also known as the *R-U-Dead-Yet* (RUDY) or slow body attack relies on sending a HTTP POST request which advertises a large content-header value. On receiving the request, the target server allocates resources necessary for the completion of the data transfer until the connection is completed or terminated by the client [11]. Since the webserver waits, as long as the connection is active until the specified length of data is received, the attacker resorts to sending small amounts of data to the server at intervals, regular or random, smaller than the timeout value of inactive connections. For the attack to be successful, the attacker launches several similar connections to the webserver and initiates the same data transfer method [13][9]. For instance, an attacker might have advertised a content-length of 5 megabytes (MB) for a POST request to a web server but sends about 20 to 30 kilobytes (KB) within the range of 15 to 25 seconds given that the timeout value for inactive connections on the webserver is 30 seconds. Accordingly, it will take approximately 2,500 seconds or 41 minutes per connection to complete such a request. Illustrated in fig. 2 is the slow POST attack process.

Slow GET

Similar to the slow POST attack, the slow GET attack also involves sending of data to the target web server. Slow GET attacks are also known as a slow header or slowloris attacks. A legitimate GET request is sent to the webserver after establishing a connection, however, on receiving a 200 OK message from the server which indicates that the server is ready to receive the headers, the attacker splits the header into several chunks which are sent at a low rate. In a normal scenario, the header consists of two Carriage-Return Line-Feed (CRLF) characters ("\r\n\r\n") which signify the end of the header and the beginning of the body to the webserver thus allowing the webserver to begin processing of the request. A single CRLF character signifies the end of a line and the beginning of another in the header request [14]. However, in an attack scenario, both CRLF characters are not transmitted thus causing the victim web server to keep the connections open as it waits indefinitely for the completion of the header requests [9][15]. The indefinite wait of the server causes the dropping of connection requests made by legitimate clients because the connection limit of the webserver has been reached. Slow GET attack description is shown in fig. 3.

Slow TCAM

The emergence of Software-Defined Networking (SDN) brought about the decoupling of the control and data planes into different devices thus allowing for a centralized view of the network. The controller of the network resides in the control plane as it maintains a unified view of the network while the switches reside in the data plane operating as packet forwarding devices. The switch maintains a Ternary Content Addressable Memory (TCAM) where it stores all the flow rules obtained from the controller whenever a new flow arrives at the switch. However, the TCAM has its limit as it can store rules from 1500 to 3000 flow rule entries [16]. A slow TCAM attack sends new flows to a switch thereby triggering flow rule requests from the switch and installation of the flow rules to the switch by the controller. The flow rules are then maintained by sending small amounts of data at intervals less than the TCAM inactive flow rule timeout value. An attacker seeks to establish numerous flow rules on the switch aimed at causing new flows from legitimate traffic to be dropped since the TCAM reaches its maximum amount of flow rules allowed and the flow rules in the switch are still active. This type of slow attack can be made effective through the recruitment of a large number of bots that send new flows to the switch at a low rate.

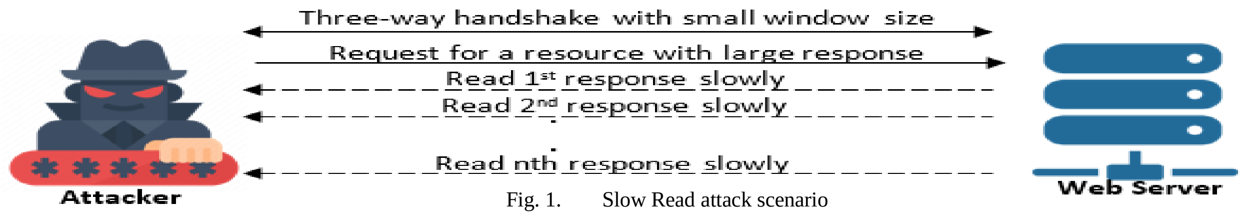


Fig. 1. Slow Read attack scenario

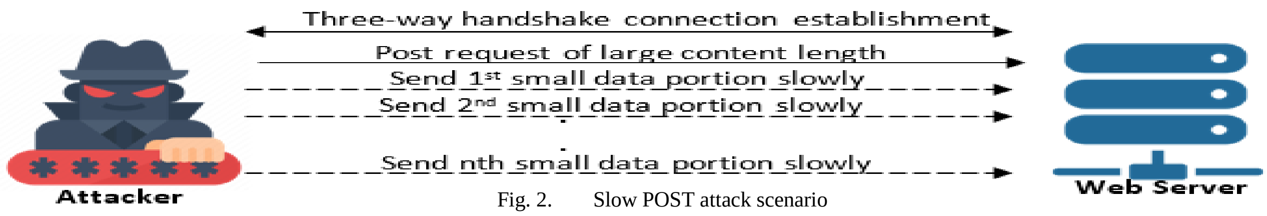

Three-way handshake connection establishment
Post request of large content length
Send 1st small data portion slowly
Send 2nd small data portion slowly
Send nth small data portion slowly
Attacker
Web Server

Fig. 2. Slow POST attack scenario

In fig. 4, the attacker makes an initial connection to a web server connected to the target switch in the SDN network. The initial network is then sustained by transferring data at a rate that evades any flow entry timeout mechanism set. Furthermore, the attacker increases the number of connections that passes through the switch until it exhausts the flow entry capacity of the switch. As illustrated in fig. 4, the limit of flow entries in the switch is m unique connections whereas the attacker attempts to make n unique connections where n is greater than m. This invariably leads to a table overflow on the target switch.

III DETECTION METHODS

The detection of slow DDoS attacks is difficult because the behaviour of the attack is similar to that of a slow client that sends legitimate traffic. Also, since the attacker establishes a connection to the webserver by adhering to legitimate connection rules in the case of slow HTTP DDoS or sends new flows to the SDN switch requesting for a resource in the SDN network in the case of slow TCAM, attack detection is challenging. Slow DDoS detection methods proposed by researchers can be classified into machine learning, time series, probability with distance metric, and performance models techniques. Detection techniques that employ machine learning methods seek to predict the class category of a new flow record or packet-based on previously identified records of benign and attack traffic or based on the similarity observed between previous traffic. The use of time series is aimed at harnessing the function of time progression to detect an attack. The possibility of traffic to be an attack traffic is considered using probability-based measurements. Similarly, distance-based measurements compute the possibility of a new traffic to be an attack traffic based on the closeness of the features of the new traffic to that of a previously established attack traffic. Since an attack changes the state and behaviour of a web server, performance model technique of attack detection calculates the behaviour of the webserver or data transfer rate under normal circumstances and seeks to identify any behaviour that deviates from the initially established behaviour. Table I presents a summary of the detection techniques with their strengths and weaknesses.

b. Machine Learning

Machine learning techniques of supervised and unsupervised learning were used in detecting slow DDoS attacks. Machine learning techniques under the supervised learning category which makes predictions based on previously observed features is the most prominent category used in the analysis.

The use of 5-NN, Naïve Bayes, multilayer perceptron, support vector machines, JRIP, Random forest, C4.5 decision trees, and logistic regression to detect DoS attacks of slow POST and slowloris was evaluated in [13]. The learners achieved high Area Under Curve (AUC) which was attributed in part to the use of Netflow feature set. The highest AUC value of 0.99905 with a class ratio of 50:50 was recorded in RF and the second highest AUC of 0.99904 with a class ratio of 65:35 was recorded in RF. Although their work showed good detection of slow POST and slowloris attacks, they employed the use of a DoS attack that originates from a source. Since DoS attacks are easier to detect compared to DDoS due to the variation in features such as source and destination IP address pair, the work charts a path for further research using DDoS. Furthermore, the similarity in the way slow POST and slowloris attacks are launched might have lent some degree of high detection rate to their experiment. However, their work buttresses the findings of other researchers about the random forest being a good machine learning technique to detect slow and volumetric DDoS. Also, 5-NN achieved high detection rate compared to other techniques used in their work however, it was surpassed by random forest.

Six classifiers of random forest, KNN, logistic regression, SVM, decision trees, and deep neural networks were used in [17] to detect slow HTTP attacks. KNN and Decision trees achieved high detection rates. KNN had an accuracy of 99.81%, false positive rate of 0.08%, and false negative of 1.09% while decision tree achieved an accuracy of 99.87%, false positive of 0%, and false negative rate of 0.03% when there was an equal composition of attack and legitimate traffic in the dataset. The achievement of KNN strengthens the view that KNN, an unsupervised learning algorithm, can be used to detect slow attacks. However, the detection time of KNN when an unbalanced dataset was used was 61.21 seconds which means that prompt detection of slow attacks when KNN is used is not always guaranteed.

Instead of using full packet captures, Netflow features were used in [8] with eight classifiers to detect slow read attacks in SDN networks. The use of Netflow was attributed to the low packet processing overhead associated with Full Packet Captures (FPC). The classifiers of random forest, C4.5 N, 5-Nearest Neighbour, C4.5D, MLP, JRip, SVM, and Naïve Bayes achieved an AUC of 96.76%, 96.72%, 96.69%, 96.62%, 95.06%, 94.71%, 89.22%, and 88.94% respectively.

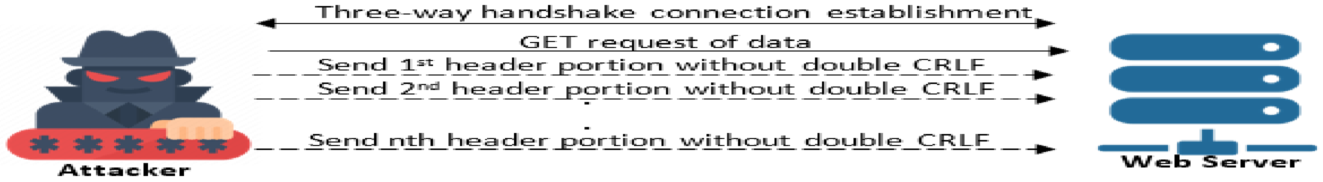
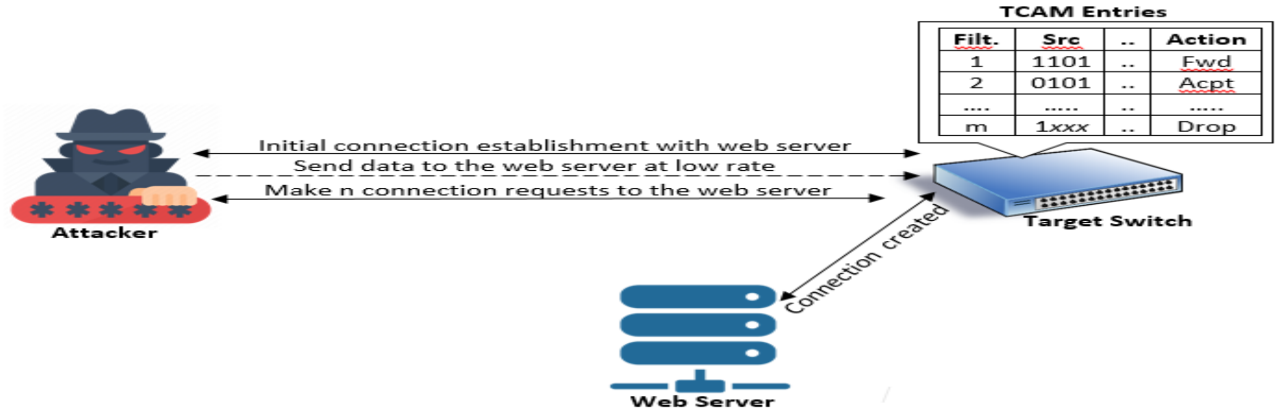

Three-way handshake connection establishment
GET request of data
Send 1st header portion without double CRLF
Send 2nd header portion without double CRLF
Send nth header portion without double CRLF
Attacker
Web Server

Fig. 3. Slow GET attack scenario


TCAM Entries
Initial connection establishment with web server
Send data to the web server at low rate
Make n connection requests to the web server
Attacker
Target Switch
Connection created
Web Server

Filt.	Src	..	Action
1	1101	..	Fwd
2	0101	..	Acpt
....	.....	..	.....
m	1xxx	..	Drop

Fig. 4 Slow TCAM attack scenario

Since high AUC reflects high TPR and low FPR, the random forest is seen as the best classifier for detecting slow read attacks. Here, random forest classifier proves to be the best classifier that detects slow read attacks.

Detection of slowloris and slow POST attacks in encrypted traffic by clustering extracted features and performing machine learning detection of anomalies was performed in [18]. Machine learning techniques used are single linkage clustering, k-means, fuzzy c-means, self-organizing maps, and DBSCAN. K-means, fuzzy c-means, and self-organizing maps achieved high detection rates of 99.9957% with detection rate.

In another work, machine learning techniques to detect RUDY attacks using features from bi-directional network instances false positive rate of 0.0043% for slowloris attacks. Also, K-means, fuzzy c-means, and self-organizing maps achieved high detection rates of 99.9931% with false positive rate of 0.0043% for slow post attacks. Kmeans, another unsupervised learning algorithm, achieves high

selected using an ensemble feature selection approach containing 10 different feature ranker methods aimed at extracting the most important features for the detection of RUDY attacks at the network level. It was observed that the usage of fewer features increases detection time and analysis accuracy. SANTA dataset that was obtained from the network of a commercial Internet Service Provider (ISP) together with the RUDY attack dataset obtained during pen-testing used in their work. Three classification methods of K-Nearest Neighbor (K-NN) where k is five and two forms of C4.5 decision trees (C4.5D and C4.5N) were used to build the predictive models. The selected features include features that represent three main characteristics of traffic size, packet similarity, and traffic velocity. When seven features were used, results obtained for the AUC metric shows that 99.83%, 99.96%, and 99.99% were achieved by C4.5N, C4.5D, and 5-NN respectively; for true positive rate, 99.07%, 98.90%, and 98.97% were achieved by C4.5N, C4.5D, and 5-NN respectively; and for

false positive rate, 0.029%, 0.041%, and 0.0265% were achieved by C4.5N, C4.5D, and 5-NN. When all the features were used, the AUC metric achieved results of, 99.88%, 99.40%, and 99.99% by C4.5N, C4.5D, and 5-NN respectively; for true positive rate, 98.73%, 98.66%, and 98.83% were achieved by C4.5N, C4.5D, and 5-NN respectively; and for false positive rate, 0.0282%, 0.0307%, and 0.0316% were achieved by C4.5N, C4.5D, and 5-NN respectively. The higher AUC value means higher TPR and lower FPR. 5-NN also achieves a good detection by having the highest AUC and the lowest FPR values when seven features were selected. The increase in AUC and FPR values in 5-NN when all the features were used points the effect of large feature usage on the detection rate of 5-NN. Although higher AUC was obtained, the corresponding increase in false positive rate cannot be substantiated given that the clustering algorithm flags a greater amount of legitimate traffic as malicious [19].

Usage of the random forest algorithm to detect slow read attacks in a cloud environment was performed in [12]. Raw TCP logs of a slow read attack were analysed and preprocessed before passing the data to the random forest classifier. The accuracy of the random forest classifier increases with an increase in the number of trees however, the computational complexity also increases. Pre-pruning of the trees has proven to increase the false negative rate to 50.10% compared to 1.90% when pre-pruning was not used. Accuracy of 83.34% was recorded when pre-pruning was used compared to 99.37% when pre-pruning was not used. However, 0% false positive rate was observed in either case. The use of pre-pruning of trees in random forest makes the solution not to be developed appropriately through the growth of the trees. The absence of pre-pruning sheds more light on the reason random forest classifiers have consistently shown its suitability in detecting slow attacks. It was also noted in their work that increasing the number of trees to improve performance gain may not be justified when the number of trees reaches a point where the computational cost of finding a solution affects the detection rate adversely.

HTTP count and delta time were used in [20] with other features to detect slow HTTP attacks using machine learning classifiers of naïve bayes, naïve bayes multinomial, multilayer perceptron, random forest, logistic regression, and radial basis function network. Results obtained indicate that naïve bayes multinomial has the best accuracy of 93.67%, true positive of 91.49%, and false positive of 3.10% compared to the results obtained for other machine learning techniques.

Detection of slow attacks using machine learning techniques has proven that although detection might be difficult, it is not impossible. The ability to detect slow attacks rely on the correct identification and tweaking of the classifier's parameters. As observed in KNN, using the value of K as five gives better result compared to other values of K. Also, the use of pre-pruning has been shown to affect random forest classification adversely.

c. Time Series

Detection of slow POST, header, and read DoS attacks based on a nonparametric CUSUM algorithm was introduced in [21]. It detects changes in the distribution of observed values. 13 different sampling techniques were used. Detection rate reduces as the threshold number increases. The threshold of 2500 achieved 100% detection rate with 0% false alerts. Selective flow sampling achieved the highest detection rate when the sampling rate is greater than 20%. The result obtained using selective flow sampling can be attributed to the selection of small flows for analysis rather than large flows. This ensures that the slow attacks that generate small flows are easily identified.

The use of spectral analysis to detect low rate DoS that affect Apache 2.2 servers was worked on in [22]. The spectral analysis is focused on the distribution of power over the frequency of a time series. In their work, a Discrete Fourier Transform was used to transform the signal to the frequency domain. It was observed that the beginning of an attack is more detectable than an ongoing attack using their method. Different detectability using different bot wait times was noticed as wait times also affect detectability. It was observed that detection using spectral analysis was possible when the attacker used fixed waiting times or floods the server with connection requests when starting the attack.

Time series decomposition that separates the time series into random and trend components on which the cumulative sum (CUSUM) technique and double autocorrelation technique were applied respectively in the work by [23]. Detection latency of 32 seconds was recorded with FPR and FNR of 4.3% and 9.8% respectively.

Time series method of detecting slow DDoS attacks have achieved good detection rate however, it is worthy of note that several factors affect the detection rate adversely compared to machine learning techniques.

d. Probabilistic with Distance-based Similarity Metric

Euclidean distance similarity metric was employed to detect slow attacks in [24]. The analysis of log files to calculate the similarity was used. Another distance similarity metric, Hellinger distance, was used in [25] to measure the distance between the probability distributions of the normal and attack traffic generated. Evasion of the detection system is inevitable if an attacker can generate packets whose probability distribution is similar to that of the normal traffic used as a benchmark.

Chi-square statistics was also used to detect slow rate DoS attacks. Selecting the appropriate threshold and interval time proved difficult as an increase in the interval time improves recall rate and causes a high false positive rate too but a reduction in interval time reduces recall rate and improves false positive rate [26].

The use of probability and distance-based similarity metric has not proven to be effective in detecting slow DDoS attacks yet. It can be attributed to the non-linearity of the attack type in contrast with volumetric attacks. Volumetric attacks are easily detected because the deviation of its features from benign traffic features is immense. The dilemma of using probability-based detection is evident in [26].

e. Performance Model

Packet inter-arrival time and window size analysis were used in [11] to detect slow HTTP DDoS attacks. It was identified that the average window size in client to server communication for normal traffic, slowloris, RUDY, and slow read attack are 34041, 14123, 14034, and 7241 respectively while in server to client communication the average window size recorded was 27022, 6854, 6856, and 0 respectively. It can be observed that the average values of slowloris and RUDY attacks are closer to each other which can be attributed to the similarity of their attack. It was also recorded that the average packet delta time in client to server for normal, slowloris, RUDY, and slow read attacks were 302.28, 75.16, 74.123, and 339.28 ms respectively while that of the server to the client was 151.12, 0.115, 0.561, and 28.759 ms.

A solution to slow HTTP DDoS attacks on OpenStack cloud platform was implemented in [15]. A packet pre-monitoring module identifies the behaviour of packets and the passes it to the classifier zone module. An allowed and blocked list is also maintained. All clients are placed in the allowed list until the client violates some conditions. The average network delay is calculated by sending 5 pings to the client and the average reply response time is calculated by taking into account the time the client responds to the ping messages. Once the delay between the HTTP requests exceeds five times the calculated network delay, the client is moved to the block list zone. The five times network delay is based on considering the processing time for applications. Frequent advertisement of TCP window of zero is monitored and placed in the block list. Furthermore, POST or GET requests sent to the webserver when 80% of the timeout value has elapsed are treated as an attacker and placed in the block list. In their work, slow body attacks were detected when connection requests reached 1700 while slow read attacks were detected when connection requests reached 1000.

Connection threshold that aids in detecting a slow attack in an SDN network was examined in [27]. A slow attack is detected when an incomplete HTTP request is made when the number of open connections on the web server exceeds the predetermined threshold number of concurrent connections being processed.

The TABLE FULL message generated in SDN when new rules cannot be installed due to a full TCAM was utilized in [16] to detect a TCAM attack which in turn activates a mitigation mechanism.

Reverse proxy was used in [28] to mitigate slowloris attack and detect the attack by measuring the stress at the server. The reverse proxy handles requests on behalf of the original server pending the completion of the request.

TABLE I. SUMMARY OF SLOW DDOS DETECTION TECHNIQUES

S/N	Author	Detection Technique	Strength and result	Weakness
1	Calvert and Khoshgoftaar [13]	Machine Learning	The high detection rate of 99.905% in random forest	Only DoS was examined Only slow header and slow POST were examined The high computational cost for generating trees
2	Siracusano *et al.* [17]	Machine Learning	Decision Tree accuracy of 99.87%	A small change in data can cause immense change in optimal solutions
3	Kemp *et al.*[8]	Machine Learning	Use Netflow for low packet processing overhead Random Forest had AUC of 96.76%	Only slow read was examined
4	Zolotukhin *et al.* [18]	Machine Learning	K-means achieved a detection rate of 99.9931%	Only slowloris and slow POST were examined
5	Najafabadi *et al.*[19]	Machine Learning	5-NN achieved AUC of 99.99% with false positive of 0.0265%	Only slow POST attack was examined
6	Shafieian *et al.*[12]	Machine Learning	Random forest without pre-pruning achieved 99.37% accuracy with 1.90% false negative, and 0% false positive	Only slow read was examined Tree creation computational cost
7	Singh and De [20]	Machine Learning	Naïve Bayes multinomial achieved an accuracy of 93.67%	A high false positive rate of 3.10% was recorded
8	Jazi *et al.*[21]	Time Series	Selective flow sampling achieved the highest detection rate of 100% when set to a sampling rate greater than 20%.	High resource consumption due to sampling rate
9	Brynielsson and Sharma [22]	Time Series	Detects the beginning of an attack	Attack wait times affected detectability The continuation of an attack may not be detected
10	Liu and Kim [23]	Time Series	Average attack detection time of 32 seconds	False positive rate of 4.3% and false negative rate of 9.8%
11	Cusack and Tim [24]	Probability with Distance-based Similarity	Low processing overhead	Detection of attack after havoc has been caused due to the use of log files for analysis
12	Tripathi *et al.* [25]	Probability with Distance-based Similarity	Simple probability distributions and Hellinger distances were utilized to detect attacks	Possibility of detection evasion by generating attack packets with probabilities close to the normal traffic probabilities
13	Tripathi and Hubballi [26]	Probability with Distance-based Similarity	0% false positive rate recorded for ΔT = 5 minutes 100% recall rate recorded for ΔT = 20 and 25 minutes	Large ΔT increases false positive rate but improves recall rate and low ΔT reduces recall rate but improves the false positive rate
14	Muraleedharan and Janet [11]	Performance model	Identification and recording of core features that signifies any of the slow HTTP attacks	Only DoS attacks were examined
15	Idhammad *et al.* [15]	Performance model	Effective in identifying slow connection masqueraders	Variable window size and data transfer interval small enough to cause DDoS can circumvent the detection technique
16	Hong *et al.* [27]	Performance model	Simple to implement	Difficulty establishing an appropriate threshold
17	Dantas *et al.* [16]	Performance model	Ease of implementation	The attack is detected only after the table overflow has occurred
18	Yeasir *et al.* [28]	Performance model	Ease of detection because attacks stress the server's resources	Only slowloris attack was considered
19	Shtern *et al.* [29]	Performance model	Ease of detection by using the performance of the webserver to identify attacks	The dilemma of when to establish the performance metric to be used for comparison

The establishment of a performance model using the central processing unit (CPU) utilization and time, workload, disk utilization and time, waiting time and throughput to form a baseline that signifies attack was explored by [29]. However, the dilemma of when to establish the baseline is an obstacle identified in their model. Perhaps, the baseline created might have been performed when an attack was taking place which makes it difficult to detect subsequent attacks easily using that established baseline.

Performance-based models of detecting slow DDoS attacks have proven to be good in detecting attacks however, they are not devoid of issues as evident in [29]. Selecting the appropriate threshold has been difficult to perform given the dynamic nature of an attack and benign traffic.

III. DISCUSSION

As shown in Table I, machine learning detection techniques have proven to be effective and efficient in detecting slow DDoS attacks in computer networks. Eight works of literature on slow DDoS detection using machine learning were examined. Prominent among the supervised and unsupervised learning categories are the random forest and KNN techniques respectively. However, the computational overhead of random forest and the slow detection time of KNN in the presence of unbalanced datasets are their shortcomings.

Performance models and time series techniques of detecting slow attacks trail behind machine learning techniques as evident in Table I in terms of results achieved and an improvement in their approach of detecting DDoS attacks is needed. Six performance model technique and three time series detection technique were evaluated. The advantage these aforementioned techniques have over machine learning techniques are their ability to extract features easily, perform detection faster, and conserve the detection system's resources because they do not rely on external modules to detect attacks.

The use of probability with distance-based similarity technique has not yielded any remarkable result yet. Although three research works were identified, the studies either encountered hitches of either possible attack detection circumvention or inability to detect the attack appropriately. This is due, in part, to the inability to represent slow attack traffic using concrete values.

All the techniques used in detecting slow DDoS attacks as examined in this study have shortcomings however, the use of machine learning detection technique offers more prospect in detecting attacks than any of the other techniques studied.

IV. CONCLUSION

As observed, research into the field of slow DDoS attack detection is low compared to that of volumetric attacks. This could be attributed to the ease with which researchers can perform experiments that detect volumetric attacks without resorting to other techniques of feature extraction and technologies that may be beyond their scope. This lack of adequate research on detecting slow attacks is also evident through the absence of standard datasets of slow attacks compared to that of volumetric DDoS which has CAIDA, NSL-KDD, and DARPA datasets. In the absence of the dataset, slow DDoS attack researchers have resorted to creating their dataset by simulating the attack using tools such as slowHTTPTest, slowloris.py, and OWASP switchblade amongst others.

For further studies, researchers can develop a standard data set for slow attacks and also improve upon performance model and time series detection techniques. Also, the adjustment of parameters in machine learning techniques together with feature selection should be explored. Furthermore, studies on slow table overflow attack detection and mitigation are needed given that detecting and mitigating a table overflow attack after it has wreaked havoc is not efficient and proactive.

In summary, although some researchers were able to demonstrate how slow DDoS attacks can be detected, more needs to be done in the field of slow DDoS attack detection and mitigation considering its detection difficulty, low attack resource usage, and the ability to launch one from a mobile phone.

REFERENCES

[1] M. Dabbagh, B. Hamdaoui, M. Guizani, and A. Rayes, "SOFTWARE-DEFINED NETWORKING SECURITY : PROS AND CONS," no. June, pp. 73–79, 2015.

[2] R. Swami, M. Dave, and V. Ranga, "Software-defined Networking-based DDoS Defense Mechanisms," *ACM Comput. Surv.*, vol. 52, no. 2, p. 36, 2019.

[3] J. Boite, P. A. Nardin, F. Rebecchi, M. Bouet, and V. Conan, "Statesec: Stateful monitoring for DDoS protection in software defined networks," in *2017 IEEE Conference on Network Softwarization: Softwarization Sustaining a Hyper-Connected World: en Route to 5G, NetSoft 2017*, 2017.

[4] E. Cambiaso, G. Papaleo, G. Chiola, and M. Aiello, "Slow DoS attacks: definition and categorisation," *Int. J. Trust Manag. Comput. Commun.*, vol. 1, no. 3/4, p. 300, 2013.

[5] J. Park, "Analysis of Slow Read DoS Attack and Countermeasures on Web servers," *Int. J. Cyber-Security Digit. Forensics*, vol. 4, no. 2, pp. 339–353, 2015.

[6] E. Cambiaso, G. Papaleo, and M. Aiello, "Slowcomm: Design, development and performance evaluation of a new slow DoS attack," *J. Inf. Secur. Appl.*, vol. 35, pp. 23–31, 2017.

[7] P. Farina, E. Cambiaso, G. Papaleo, and M. Aiello, "Understanding DDoS Attacks From Mobile Devices," 2015.

[8] C. Kemp, C. Calvert, and T. M. Khoshgoftaar, "Utilizing netflow data to detect slow read attacks," in *Proceedings - 2018 IEEE 19th International Conference on Information Reuse and Integration for Data Science, IRI 2018*, 2018, pp. 108–116.

[9] S. Suroto, "A Review of Defense Against Slow HTTP Attack," *JOIV Int. J. Informatics Vis.*, vol. 1, no. 4, p.

127, 2017.

[10] D. Ameyed, F. Jaafar, and J. Fattahi, "A slow read attack using cloud," in *Proceedings of the 2015 7th International Conference on Electronics, Computers and Artificial Intelligence, ECAI 2015*, 2015, pp. SSS33–SSS38.

[11] N. Muraleedharan and B. Janet, "Behaviour analysis of HTTP based slow denial of service attack," in *Proceedings of the 2017 International Conference on Wireless Communications, Signal Processing and Networking, WiSPNET 2017*, 2018, vol. 2018-Janua, pp. 1851–1856.

[12] S. Shafieian, M. Zulkernine, and A. Haque, "CloudZombie: Launching and detecting slow-read distributed denial of service attacks from the Cloud," in *Proceedings - 15th IEEE International Conference on Computer and Information Technology, CIT 2015, 14th IEEE International Conference on Ubiquitous Computing and Communications, IUCC 2015, 13th IEEE International Conference on Dependable, Autonomic and Se*, 2015, pp. 1733–1740.

[13] C. L. Calvert and T. M. Khoshgoftaar, "Impact of class distribution on the detection of slow HTTP DoS attacks using Big Data," *J. Big Data*, 2019.

[14] O. Yevsieieva and S. M. Helalat, "Analysis of the Impact of the Slow HTTP DoS and DDoS Attacks on the Cloud Environment," p. 5, 2017.

[15] M. Idhammad, K. Afdel, and M. Belouch, "Detection System of HTTP DDoS Attacks in a Cloud Environment Based on Information Theoretic Entropy and Random Forest," *Secur. Commun. Networks*, vol. 2018, 2018.

[16] Y. G. Dantas, I. E. Fonseca, and V. Nigam, "Slow TCAM Exhaustion DDoS Attack," vol. 1, pp. 17–31, 2017.

[17] M. Siracusano, S. Shiaeles, and B. Ghita, "Detection of LDDoS Attacks Based on TCP Connection Parameters," in *Global Information Infrastructure and Networking Symposium*, 2018.

[18] M. Zolotukhin, T. Hamalainen, T. Kokkonen, and J. Siltanen, "Increasing web service availability by detecting application-layer DDoS attacks in encrypted traffic," in *2016 23rd International Conference on Telecommunications, ICT 2016*, 2016.

[19] M. M. Najafabadi, T. M. Khoshgoftaar, A. Napolitano, and C. Wheelus, "RUDY attack: Detection at the network level and its important features," in *Proceedings of the 29th International Florida Artificial Intelligence Research Society Conference, FLAIRS 2016*, 2016, pp. 282–287.

[20] K. J. Singh and T. De, *Emerging Research in Computing, Information, Communication and Applications*. 2015.

[21] H. H. Jazi, H. Gonzalez, N. Stakhanova, and A. A. Ghorbani, "Detecting HTTP-based application layer DoS attacks on web servers in the presence of sampling," *Comput. Networks*, vol. 121, pp. 25–36, 2017.

[22] J. Brynielsson and R. Sharma, "Detectability of low-rate HTTP server DoS attacks using spectral analysis," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015*, 2015, pp. 954–961.

[23] H. Liu and M. S. Kim, "Real-time detection of stealthy DDoS attacks using time-series decomposition," in *IEEE International Conference on Communications*, 2010.

[24] B. Cusack and Z. Tian, "Detecting and tracing slow attacks on mobile phone user service," in *Proceedings of the 14th Australian Digital Forensics Conference, ADF 2016*, 2016, no. December, pp. 4–10.

[25] N. Tripathi, N. Hubballi, and Y. Singh, "How Secure are Web Servers? An empirical study of Slow HTTP DoS attacks and detection," in *Proceedings - 2016 11th International Conference on Availability, Reliability and Security, ARES 2016*, 2016, pp. 454–463.

[26] N. Tripathi and N. Hubballi, "Slow rate denial of service attacks against HTTP/2 and detection," *Comput. Secur.*, vol. 72, pp. 255–272, 2018.

[27] K. Hong, Y. Kim, H. Choi, and J. Park, "SDN-Assisted Slow HTTP DDoS Attack Defense Method," *IEEE Commun. Lett.*, vol. 22, no. 4, pp. 688–691, 2018.

[28] M. Yeasir, M. Morshed, and M. Fakrul, "A Practical Approach and Mitigation Techniques on Application Layer DDoS Attack in Web Server," *Int. J. Comput. Appl.*, vol. 131, no. 1, pp. 13–20, 2015.

[29] M. Shtern, R. Sandel, M. Litoiu, C. Bachalo, and V. Theodorou, "Towards mitigation of low and slow application DDoS attacks," in *Proceedings - 2014 IEEE International Conference on Cloud Engineering, IC2E 2014*, 2014, no. Vm, pp. 604–609.

Probing Attack Detection Using JRIP Classifier

Olomi Isaiah Aladesote
Department of Computer Science
Federal Polytechnic, Ile Oluji,
Ondo State, Nigeria
isaaladesote@fedpolel.edu.ng

Ebenezer Akinyemi Ajayi
Faculty of Information Science and Technology, Melaka,
Melaka State, Malaysia
ORCID iD: 0000-0001-5745-1687
ebeseun@gmail.com

ABSTRACT— The services provided by the internet has made it an inevitable part of our life. The increased usage of mobile devices has further increased the number of internet users. However, this advancement in network technologies has paved the way for intruders to gain illegitimate access to a network. This paper, therefore, extracts records of both probing and normal attacks from an improved version of Knowledge Discovery and Data Mining 1999 (KDD '99) tagged NSL-KDD dataset, based on thirteen attributes, a result of an existing work adopted in this research. A copy of the dataset was run on data deduplication software written in C# to remove the duplicate records; the result formed a second set of the dataset. The two sets of the dataset were run on the Waikato Environment for Knowledge Analysis (WEKA) 3.7.13 using JRIP algorithm. The experimental result shows that the classification rate, sensitivity, specificity and false alarm rate for the first dataset are 97.75%, 89.96%, 99.62% and 0.38% respectively. In contrast, classification rate, sensitivity, specificity and false alarm rate (FAR) second dataset are 99.03%, 84.67%, 99.87% and 0.13% respectively. It can be concluded that the result of the second dataset (dataset run on data deduplication) outperformed the first dataset based on classification rate, specificity and FAR.

Keywords— Probing, Data deduplication, WEKA, Propositioner Rule Learner, NSL-KDD

I INTRODUCTION

The growth and improvement in Information Technology (IT) has brought about security challenges in network [1]. This has paved a way for attackers to seek ways of gaining access to the network through the back door. Thus, there is need for a system or approach to secure the network to prevent the activities of attackers [2]. Intrusion detection system is a security means of guarding entire networks from being accessed illegally. In this perspective, intrusion can be seen as any act that breaches the policies of information security or Confidentiality, Integrity and Availability (CIA) triad in a network [3].

Intrusion detection plays an active part in network defense by helping network administrator or personnel in alerting them of malicious behaviours. An effective Network Intrusion Detection System (NIDS) would yield low false-alarm rates and high intrusion-detection rates [4]. A false alarm occurs when normal traffic is misidentified as malicious. A good and standard benchmark evaluation dataset is a vital constituent for developing an effective Intrusion Detection System (IDS). Selection of a suitable dataset brings about efficient and effective IDS, since the performance of any IDS relies solely on the efficiency and accuracy of the dataset [5].

Data deduplication can be described as a method of removing duplicate records. Thus, allowing only a copy to be kept [6]. There are many intrusion datasets; however NSL- KDD dataset is selected for the research.

We summarized the contributions of this paper as follows:

- Removal of the redundant or duplicate records from the dataset
- Classify both normal and attacks records using JRIP technique
- carry out a performance evaluation on the two sets of dataset

The rest of the paper is structured as follows: section 2 presents related works. Method used in the research is presented in Section 3. The result is presented and discussed in section 4. System Performance and Evaluation is presented in section 5 while conclusion and future work are presented in section 6.

II. RELATED WORKS

Performance evaluation between Gain Ratio and Correlation Based Feature Selection was carried out for classifying Pima Indian Diabetic Database (PIDD) [7]. The authors used C4.5 and genetic algorithm for Gain Ratio and Correlation Based Feature (CBF) respectively. The result showed that feature selection by CBF performed better than Gain Ratio.

In the work of [8], Gain Ratio and Principal Component Analysis (PCA) were employed to extract relevant attributes of KDD ’99 dataset that are discrete and continuous in nature respectively. The two techniques were implemented using C# Programming language. The result revealed that the forty-one (41) attributes of the dataset were reduced to thirteen attributes.

[9] worked on Feature Reduction using PCA for Effective Anomaly–Based Intrusion Detection on NSL-KDD. The authors proposed a hybrid Principal Component Analysis Neural Network Algorithm (PCANNA) which was implemented in JAVA programming language. The result showed that the forty-one features of KDD ’99 was reduced to eight (8) features.

Many research works had been carried out in the field of network security with one weakness or the other. The main and common weakness is that the duplicate or redundant records were not removed from the dataset after feature extraction. This research work is to apply data deduplication in addition to feature or attribute selection reduction to KDD ‘99 datasets.

III. METHODOLOGY

The work of [8] on Feature or Attribute Extraction for Intrusion System using Gain Ratio and PCA was studied, which resulted to thirteen relevant and highly significant attributes were adopted in the paper. The records of Probing attacks and normal were extracted from NSL-KDD dataset. A data deduplication program was developed in C#. The probing dataset with normal traffic contains twelve thousand one hundred and thirty two (12132) records. A copy of the dataset was run on WEKA 3.7.13 using JRIP rule based classifier while another copy was first run on data deduplication program before it was run on WEKA using the same classifier.

IV RESULT AND DISCUSSION

This section presents the experimental result of the research

A. **Dataset without Data Deduplicaton**

This copy of the dataset was only run on WEKA 3.7.3, a machine learning tool. JRIP rule based classification was selected. 8492 (70%) records were used for a training dataset while 3640 (30%) as test data.

TABLE 1. CONFUSION MATRIX OBTAINED FROM JRIP CLASSIFIER ON TEST DATA

	Ipsweep	***Mscan***	***Nmap***	***Normal***	***Portsweep***	***Saint***	***Satan***
Ipsweep	42	0	0	0	0	0	0
Mscan	0	290	0	0	0	0	0
Nmap	0	0	22	0	0	0	
Normal	0	0	0	2922	0	0	11
Portsweep	0	0	0	5	29	0	2
Saint	0	1	0	7	0	55	16
Satan	0	0	0	12	4	24	198

All records of ipsweep, Mscan and Nmap were correctly classified. The result has shown that ipsweep is 100% accurate and 100% reliable while Mscan is 100% accurate and 99.66% reliable. The system has revealed that Nmap is 100% accurate and 100% reliable. Out of 2933 records of Normal traffic that were classified, 2922 were correctly classified while 11 were wrongly classified as Satan. The classification result showed that Normal is 99.63% accurate and 99.17% reliable. Out of 36 Portsweep, 29 records were correctly classified while 5 and 2 were wrongly classified as Normal and Satan respectively. The result showed that Portsweep is 80.56% accurate and 87.88% reliable. Out of 79 Saint classified, 55 were correctly classified while 7, 16 and 1 were wrongly classified as Normal, Satan and Mscan respectively; this showed that Saint is 69.62% accurate and reliable. Out of 238 Satan, 198 were correctly classified, 12, 4 and 24 were wrongly classified as Normal, Portsweep and Saint respectively; Satan is 83.19% accurate and 87.22% reliable.

B. Dataset with Data Deduplicaton

This copy of the dataset was first run on Data Deduplication software before the result was classified using JRIP rule based classifier on WEKA 3.7.3 as shown in Table 2

TABLE 2: RESULT OF DATA DEDUPLICATION SOFTWARE ON PROBING ATTACK AND NORMAL TRAFFIC

Attack Type	Before Deduplication	After Deduplication	% Data Deduplication
Ipsweep	141	13	90.8
Mscan	996	291	70.8
Nmap	73	5	93.2
Normal	9711	7757	20.1
Portsweep	157	24	84.7
Saint	319	70	78.1
Satan	735	78	89.3
Total	12132	8238	

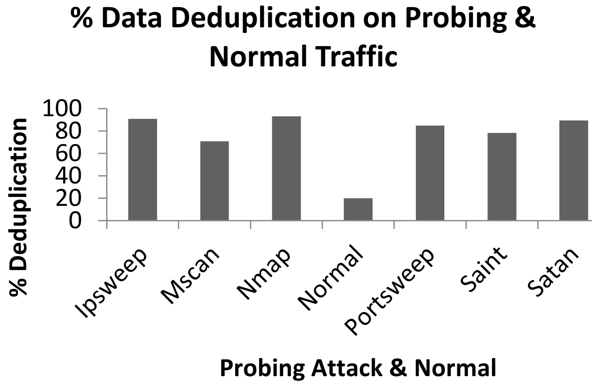


Figure 1: % Data Deduplication on Probing Attack & Normal

Table 2 and Figure 1 above showed that 141 records of Ipsweep were reduced to 13, which is 90.8% reduction. 996 records of Mscan were reduced to 291, resulting to 70.8% reduction. 73 records of Nmap were reduced to 5, leading to 93.2% reduction. 9711 records of Normal traffic were reduced to 7757, resulting to 20.1% reduction. 157 records of Portsweep were reduced to 24, which led to 84.7% reduction. 319 records of Saint were reduced to 70, leading to 78.1% reduction while 735 records of Satan were reduced to 78, which amounted to 89.3% reduction.

C Classification of Data Deduplicated Records

The remaining records of Probing & Normal traffic were only on WEKA 3.7.3 using JRIP rule based classifier. 5767 (70%) records were used a training dataset while 2471 (30%) as test data

Table 3: Confusion Matrix obtained from JRIP classifier on test data

	Ipsweep	*Mscan*	*Nmap*	*Normal*	*Portsweep*	*Saint*	*Satan*
Ipsweep	4	0	0	0	0	0	0
Mscan	0	90	0	1	0	0	0
Nmap	0	0	1	0	0	0	0
Normal	0	1	0	2331	1	0	1
Port sweep	0	0	0	2	6	0	0
Saint	0	1	0	5	0	3	4
Satan	0	0	0	2	0	6	12

All records of Ipsweep and Nmap were correctly classified. The result has shown that both Ipsweep and Nmap are 100% accurate and 100% reliable. Also, Mscan is 98.9% accurate and 97.83 reliable. Out of 2933 records of normal traffic that were classified, 2922 were correctly classified while 11 were wrongly classified as Satan. The system has shown that normal is 99.87% accurate and 99.57% reliable. 29 records of Processtable were correctly classified while 5 and 2 were wrongly classified as Normal and Satan respectively. Thus, Processtable is 75% and 85.71% accurate and reliable respectively. 55 records of Saint were correctly classified while 1, 7 and 16 were classified as Mscan, Normal and Satan respectively. Also the result has shown that Saint is 23.08% accurate and 33.33% reliable respectively. 198 records of Satan were correctly classified while 12, 4 and 24 were wrongly classified as Normal, Portsweep and Saint respectively. Hence Satan is 60% and 70.59% accurate and reliable respectively.

5.0 SYSTEM PERFORMANCE AND EVALUATION

In order to evaluate the effectiveness and viability of the system, comparison analysis between the two sets used was carried- out using Classification Rate, False Alarm Rate (FAR), Sensitivity and Specificity. The True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) of dataset without data deduplication are 636, 2922, 11 and 71 respectively while the TP, TN, FP and FN of dataset with data deduplication are 116, 2331, 3 and 21 respectively.

Table 4: System Performance and evaluation result

	Classification Rate (%)	Sensitivity (%)	Specificity (%)	False Alarm Rate (%)
Dataset without Data Deduplication	97.75	89.96	99.62	0.38
Dataset with Data Deduplication	99.03	84.67	99.87	0.13

Figure 2: Graphical representation of system performance and evaluation

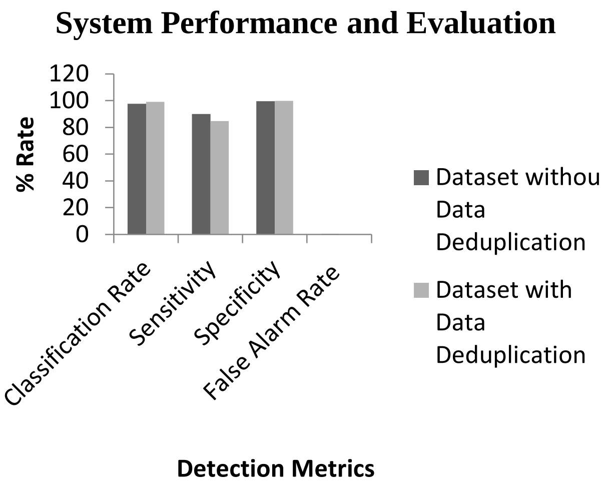


Figure 2: Graphical representation of system performance and evaluation

6.0. CONCLUSION AND RECOMMENDATION

The experimental result showed that dataset with Data Deduplication outperform the other dataset (without deduplication) using certain detection metrics. It is hereby recommended that Data deduplication should be combined with feature or attribute extraction technique for effective intrusion detection system**.**

REFERENCES

[1] P. Amudha, S. Karthik, and S. Sivakumari, “A hybrid swarm intelligence algorithm for intrusion detection using significant features,” *Sci. World J.*, vol. 2015, no. 1, 2015, doi: 10.1155/2015/574589.

[2] V. Jaiganesh, P. Sumathi, and S. Mangayarkarasi, “An analysis of intrusion detection system using back propagation neural network,” *2013 Int. Conf. Inf. Commun. Embed. Syst. ICICES 2013*, pp. 232–236, 2013, doi: 10.1109/ICICES.2013.6508202.

[3] O. I. Aladesote, B. K. Alese, and F. Dahunsi, “Intrusion detection technique using hypothesis testing,” *Lect. Notes Eng. Comput. Sci.*, vol. 1, pp. 393–397, 2014.

[4] R. Zuech, T. M. Khoshgoftaar, N. Seliya, M. M. Najafabadi, and C. Kemp, “A new intrusion detection benchmarking system,” *Proc. 28th Int. Florida Artif. Intell. Res. Soc. Conf. FLAIRS 2015*, no. McHugh, pp. 252–255, 2015.

[5] H. M. Imran, A. Bin Abdullah, and S. Palaniappan, “Towards the Low False Alarms and High Detection Rate in Intrusions Detection System,” *Int. J. Mach. Learn. Comput.*, vol. 3, no. 4, pp. 332–336, 2013, doi: 10.7763/ijmlc.2013.v3.332.

[6] “a D Va N C E D D Ata D E D U P L I C At I O N T E C H N I Q U E S a N D T H E I R,” 2013.

[7] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, “Comparative Study of Attribute Selection Using Gain Ratio and Correlation Based Feature Selection,” *Int. J. Inf. Technol. Knowl. Manag.*, vol. 2, no. 2, pp. 271–277, 2010.

[8] O. Isaiah, A. Olutola, and O. Olayemi, “Feature or Attribute Extraction for Intrusion Detection System using Gain Ratio and Principal Component Analysis (PCA),” *Commun. Appl. Electron.*, vol. 4, no. 3, pp. 1–4, 2016, doi: 10.5120/cae2016652032.

[9] S. Lakhina, S. Joseph, and B. Verma, “Feature Reduction using Principal Component Analysis for Effective Anomaly–Based Intrusion Detection on NSL-KDD,” *Int. J. Eng. Sci. Technol.*, vol. 2, no. 6, pp. 1790–1799, 2010.

Intelligent Road Traffic Control System: A Case Study of Sango Intersection Ibadan

[1]Abdulhameed, I.A. [2]Azeez, S.A. [3]Mamudu, A.O. [4]Oluwaseun, S.A.
Department of Computer Engineering Technology,
The Polytechnic, Ibadan.
Idrab4all@yahoo.com, abdulhameed.adedamola@polibadan.edu.ng

ABSTRACT - Unarguably, transportation is a non-separable part of any society as it exhibits a very close relation to the style of life, the range and location of activities and the goods and services which will be available for consumption. Traffic congestion is a severe problem in many major cities across the world and it has become a nightmare for the commuters. Simply put, traffic congestion means there are more vehicles trying to use a given road facility than it can handle without exceeding acceptable levels of delay or inconvenience. This research is to alleviate traffic congestions, and improve the levels of service and efficiency transportation system by developing an intelligent technology control system which controls congestion in real time. The result shows that the traffic signals glows red signals on two sides of the intersections and green signal for the remaining intersection thus allowing vehicle to pass at the intersection. The red signal glows for 5 seconds, the yellow signal glows for 5 seconds then in between transition from the yellow signal to the green signal, the green signal glows for 15 seconds. It does this for each intersection.

Keywords: Arduino, technology, congestions, traffic, delay

I. INTRODUCTION

Advances in transportation has made possible changes in the way of living and the way in which societies are organized and therefore have a great influence in the development of civilization (Biplav Srivastava, 2010).

An intelligent transportation system (ITS) is an advanced application which aims to provide innovative services relating to different mode of transport and traffic management and enable users to be better informed and make safer, more coordinated, and 'smarter' use of transport networks.

They vary from basic management systems such as car navigation; traffic signal control systems; container management systems; variable message signs; automatic number plate recognition or speed cameras to monitor applications, such as security CCTV systems; and to more advanced applications that integrate live data and feedback from a number of other sources, such as parking guidance and information systems and weather information (Wikipedia, 2018). They work by transmitting real-time information to improve traffic safety and relieve traffic congestion (Sumit Mallik, 2014). On the other hand, Arduino is the most popular programmable board that helps to complete any communications and electronics based projects. Basically, arduino is an open source platform which has “one click compile or upload” feature.

Traffic congestion is a severe problem in many major cities across the world and it has become a nightmare for the commuters. Congestion is particularly associated with motorization and the diffusion of the automobile, which has increased the demand for transportation infrastructure. Sango road traffic control has been deployed for years, but not intelligent to control vehicles.

Therefore, this paper focuses on building an intelligent technology in road traffic control, considering Sango Intersection as a case study. However, the specific objectives are to design a smart traffic control system; to test for its functionality in order to reduce traffic problems.

II. LITERATURE REVIEW

A. Overview of Intelligent Transportation System

Intelligent Transportation System (ITS) applies advanced technologies of electronics, communications, computers, control and sensing and detecting in all kinds of transportation system in order to improve safety, efficiency and service, and traffic situation through transmitting real-time information to improve traffic safety and relieve traffic congestion (Sumit Mallik, 2014).

The Highway Safety Act of 1970 established the National Highway Traffic Safety Administration (NHTSA). While coming at these various issues and technologies from different places, the two sectors have often converged in their approaches over time, resulting in joint projects and investments that have provided a variety of benefits (Shaheen and Finson, 2013).

Traffic signal controls can be set to optimize one or more desired goals (e.g. time savings and energy reduction). More recently, traffic signal control has become a predominant component of intelligent transportation system, and its impacts have been evaluated in simulation and real-world settings. Intelligent transportation system applications for traffic signals include communication systems, adaptive control systems, real-time data collection and analysis (Shaheen and Finson, 2013).

Another related literature is **IMAGE processing based intelligent traffic controlling authored by** Sathuluri et al *(2019).*

B. An overview of Arduino

Arduino is the most popular programmable board that helps to complete any communications and electronics based projects. Basically, arduino is an open source platform which has "one click compile or upload" feature. It is an open source electronics prototyping platform that easy to use and flexible both for the software and hardware. Arduino is able to observe its surroundings through the light, controlling motor and other activators. It is very easy to write and upload codes to the I/O board to run any programs due to its open source platform features. The use of arduino board as a control panel of the system is a very effective device to use due to its unique nature operation.

III. METHODOLOGY

This research includes the use of Arduino Uno board, Resistor, LED. The researchers designed a traffic signal with the green led, red led and yellow led used as the signal indicator which is then connected to an Arduino board.

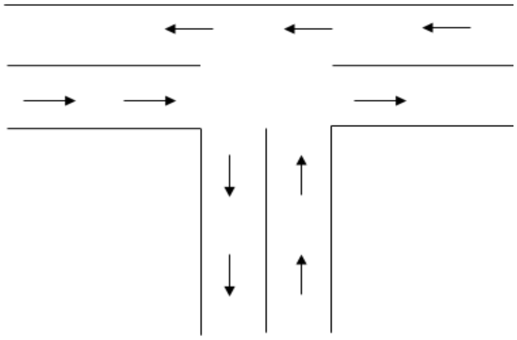


Fig. 1: Sango Intersection

The case study has a Three-way road intersection: three phases mathematically. Each Phase is representing transition from green to red signal between the intersections. There are various types of Arduino boards that can be used for different purposes.

Fig. 2: Arduino UNO R3 Board (Tutorialspoint, 2018)

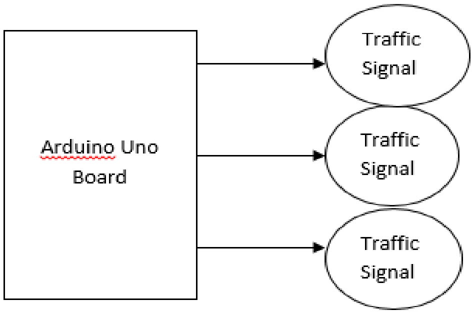


Fig. 3: Full circuit design

Components Used in the System

To design the circuit of the project we have used the following component:

1. An Arduino Uno Board
2. Resistors
3. LED

C. Principle of Operation

As reported earlier, the intelligent traffic control system will be able to control the movement of vehicles in fixed time basis using time interval. Such a control is done by using red, yellow and green indicators.

The traffic control system operates thus:

1. The traffic signals glows red signals on two sides of the intersections and green signal for the remaining intersection thus allowing vehicle to pass at the intersection. The red signal will glow for 5 seconds, the yellow signal glows for 5 seconds then in between transition from the yellow signal to the green signal, the green signal glows for 15 seconds. It does this for each intersection.
2. It makes use of the red signal to stop vehicles, yellow signal to prepare waiting vehicles for ready and green signal to pass vehicles. All the operation are controlled by the programmed Arduino circuit board.

IV. RESULT AND DISCUSSION

After connecting all the parts as shown in the circuit design diagram we uploaded the programming code to the Arduino UNO Board. After uploading it was observed to have a positive result. The circuit works properly as it was designed. After testing all the components, we combined all the components together as shown in the circuit design diagram (fig. 4) and found that the application works correctly with the programmed code uploaded. The traffic signals glows red signals on two sides of the intersections and green signal for the remaining intersection thus allowing vehicle to pass at the intersection. The red signal will glow for 5 seconds, the yellow signal glows for 5 seconds then in between transition from the yellow signal to the green signal, the green signal glows for 15 seconds. It does this for each intersection.

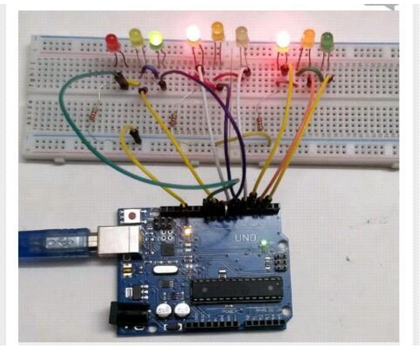


Fig. 4: Final Development of Circuit

With the development of the circuit it must be operated with some guarding precautions which are:

1. A stable power supply must be provided for the device
2. The operation must be in fixed time system

VII Result Presentation

After completing the circuit and simulating the intelligent traffic control system using the case study of Sango intersection by connecting it into a demo Three way intersection it worked correctly. As a result we can say that the traffic control system we have developed is ready to use commercially.

VCONCLUSIONAND RECOMMENDATIONS

The Intelligent Traffic Control System was designed and developed to decrease traffic congestions. We have used 5volts and GND from Arduino UNO Board. We have used Red, Yellow and Green LED as a traffic signal indicator. At the end we have designed and developed a Microcontroller based Intelligent Traffic Control System, and fixed the problem that we had before. In this project we succeeded in minimizing the traffic congestions created by manual traffic control system with the help of microcontroller and improved algorithm. This works every time of the day and night unlike manual traffic control.

Recommendation

Stakeholders especially the Federal Government should take this project more seriously which is seen as a virtual improvement in future of transportation in our nation. With the help of this project there is an opportunity of doing a big project in future by the inclusion of a traffic management system which will be able to consider impact on citizens such as pedestrians. Also, it is recommended that further study should be carried out on effective use of wireless technology and high speed micro controller to provide smooth and clear flow of traffic for ambulance clearance and stolen vehicle detection.

REFERENCES

Ashley R.K., Soon J.L., Yoo J.K. (2013) Traffic Signal Systems: A Review of Current Technology in the United States. Volume 3 Page 33-41, 2013

Biplav, S. (2010), A New Look At the Traffic Management Problem and where to Start. November, 2010

http://www.tutorialspoint.com (2018) Arduino uno description, Article retrieved on 13th May, 2018

http://www.wikipedia.org/wiki/Intelligent transportation_system, Article retrieved on 13th May, 2018

http://www.google.com.ng/search?q=schematic+diagram+of+arduino+uno+r3,Article retrieved on 14th November, 2018

Joachim W. (2012) Information in Intelligent Transportation Systems, June 2012

Salim B.I., and Mahmudur R. (2012) Design and Developing of a Microcontroller Based Intelligent Traffic Control System. January, 2012.

Shaheen S.A., and Finson R. (2013) Intelligent Transportation Systems, Reference Module in Earth Systems and Environmental Sciences, September, 2013.

SSumit M. (2014) Intelligent Transportation System, International Journal of Civil Engineering Research. Volume 5, Page 367-372, April 2014

An Alternative Data Mining Management Framework using Artificial Neural Network for Crime Prediction and Preventions

Usman Dahiru Haruna 1
Modibbo Adama Univesity of Technology, Yola.

Kabiru Ibrahim Musa 2
Abubakar Tafawa Balewa University Bauchi

Abstracts

Data mining is powerful technology with great potential to help both developing and developed countries focus on the most important ideas hidden in their data repositories. It's automates the detection of relevant patterns in a data stores, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends for management decision making. Crime prediction and prevention is one of the application areas of data mining, Crimes are social nuisance and cost our society dearly in several ways. Its observed that, the major challenge facing all law-enforcement and intelligence-gathering organizations is how to accurately and efficiently analyze the growing volume of crime data to make informed decision. Though developed countries have addressed it by subjecting their stable and structured dataset in to data mining framework. This paper looked at the developing countries like Nigeria designed a robust data mining management framework for crime prediction and prevention by reviewing the data mining frameworks available. Data from crime related data warehouse, telecommunication companies and social site was identified as the input dataset for the framework to produce the prediction variables for crime management. The paper recommends subjecting the framework in to used and enhancement toward improving the framework to handle the some of the limitation identified.

1. INTRODUCTION

The extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies in both developing and developed countries focus on the most important information in their data warehouses. It uses machine learning, statistical and visualization techniques to discovery and present knowledge in a form which is easily comprehensible to humans. Various popular data mining tools are available today. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge -driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve [1]. Data mining automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends. Because data mining tools predict future trends and behaviors by reading through databases for hidden patterns, they allow organizations to make proactive, knowledge-driven decisions and answer questions that were previously too time-consuming to resolve. The large volumes of crime-related data existed in police departments and also the complexity of relationships between these kinds of data has forced the traditional crime analysis methods to become obsolete. These methods require a considerable amount of time and human resources on one hand and on the other hand, they are not able to get all effective parameters/relationships involved due to their high amount of human interference.

1.2 Statement of the Problem

Crime has become the major challenge in both developed and developing countries, but developed countries have put in place ways in curbing the intensity of crime and went further to predict the future occurrence of crime by subjecting structured dataset into some data mining algorithms, the study paper revealed that both Traditional data mining techniques and advanced data mining techniques works on structured dataset for pattern recognition and also identification of crime relationship [3]. In an attempt to build crime prediction and prevention framework, Shruti et al. [4] designed a framework for crime prediction and prevention and the framework analyze data, looks for patterns and correlations using past crime data. however, the framework fails to identify regions for crime hotspots which is needed for prediction, to identify which areas are more prone to crimes and the type of crime. Alongside the increasing use of the computerized systems to track crimes, Since existing framework are successful on clean and structured data, then It is against these backdrop that this paper seeks to improve on an existing data mining algorithm to manage crime and crime related issues in developing country like Nigeria that will provide a mechanism for identifying crime hotspots and also which will work on data that is unstructured and incorporated with human interference.

Aim and Objectives

The main aim of this study is to design a robust data mining management framework for crime prediction and prevention/ decision making. By Perform a systematic evaluation of existing data mining algorithms and to propose a new data mining management framework for crime prediction and prevention that will work on unstructured dataset which will be limited only to accept dataset from any crime related organization data warehouse, telecommunication companies and social site.

LITERATURE REVIEW

A. *Crime as related to data mining*

Bhargava et al. [5], defined crime as an act or the commission of an act that is forbidden, or the omission of a duty that is commanded by a public law and that makes the offender liable to punishment by that law, they also said crime can be property crime, violent crime, cyber-crime and Others. Prabakaran et al. [7] A crime is an unlawful activity for which a man can be penalized by law. Crime against a person is called personal crime like murder, robbery, etc. Crimes are one of the major factors that affect various important decisions of an individual's life like moving to a new place, roaming at right time, avoiding risky areas, etc. Crimes affect and defame the image of a community. Crimes also affect the economy of a nation by placing the financial burden on government due to the need for additional police forces, courts etc. [8].

B. *Data mining*

Data mining is one of the most powerful ways of knowledge extraction or we can say it is one of the best approaches to detect underlying relationships among data with the help of machine learning and artificial intelligence techniques. Crime Detection is one of the hot topics in data mining where different patterns of criminology are identified. It includes variety of steps, starting from identification of crime characterization till detection of crime pattern [3]. Savita et al. [11] looked at data mining as the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use [12]. Data mining has two ways in dealing with data namely Classification and Prediction [13]. Various techniques exist which include: Entity extraction, Clustering techniques, Association rule mining, Sequential pattern mining, Deviation detection, Classification technique and Artificial neural networks Shamaila et al. [3]

Table 1 Showing Literature Mapping for Crime Prediction and Prevention Framework

N	Author	Title	Problem Address	Methodology used	Dataset source	Outcome	Weakness
1	Devakunchari et al. [14]	Analysis of Crimes Against Women in India Using Regression	Crime against women was given less concern, in the past, the measures are still not effective, therefore there is need to identify the effectiveness of the measure and age group that need awareness	Regression algorithms	National Crime Records Bureau (NCRB) crime data	predict the crime rate across the country and predict age group	Data used as the data set is unreliable because the data was not pre-processed
2	Vishnupriya et al. [15]	An Effective Data Mining Techniques for Analyzing Crime Patterns	Managing Hight rate increase in crime data stored from various sources by employ many learning algorithms to extort hidden knowledge from huge volume of data.	Clustering, Classification and watermark	Traffic Violation, Border Control, the Narcotics, and Cyber Crime dataset	crime changes over time increase identifies crime behavior manage large amount of crime data with certain degree of data security	detects only hyper spherical clusters and it can't resolve the number of clusters in the crime data.
3	Ginger et al [16]	An exploration of crime prediction using data Mining on open data	there are fewer approaches focusing on predictive frameworks of crime and the frequency of anti-social behavior crimes.	instance-based learning algorithm, Linear Regression and The M5P algorithm	UK police and contain over 600,000 records	decision trees (M5P algorithm) can be used to reliably predict crime frequency in general, as well as anti-	time frame for prediction, reliable prediction framework for particular types of crime

						social behavior frequency	
4	Swadi et al. [2]	Crime Data Analysis Using Data Mining Techniques To Improve Crimes Prevention Procedures	Discovering patterns and trends, making forecasts, finding relationships and possible explanations, mapping criminal networks and identify possible suspects	Classification, Association, Prediction, outliers and link analysis	dataset police departments from the Internet	associations and relationships between attributes were discovered, and linkage between attributes including 5crime type and criminal age, job, history and others was found	Can't work on huge dataset, and need a serious pre-processing approach to make dataset reliable
5	Mona et al [17]	Building Unstructured Crime Data Prediction Model (Practical pproach)	Crime prediction on unstructured data using scikit-learn Python	classification technique, and TF-IDF as a vector space model	police incident report of the city of Madison with details for all crime types	predict the crime type	Can't provide crime hotspot and can't predict the next expected crime region
6	Vijayalakshmi et al. [10]	Crime Pattern Recognition and Prediction Using Optimized K-means and SVM	There is general believe that, crime increase at high rate and affects many dimensions like education which affect their level of competitiveness	k-means algorithm and SVM	Chicago Police Department	Relationship established and Possible next crime was predicted,	Can't handle a large part of data elements
7	Kang et al. [18]	Prediction of crime occurrence from multi-modal data using deep learning	Previous studies have used data from multiple domains but their	feature-level data fusion method deep neural	Chicago because it has both a large	facilitated accurate and effective	unable to provide information

			prediction models treat data from different domains equally. These methods have problems in crime occurrence prediction, such as difficulty in discovering highly nonlinear relationships, redundancies, and dependencies between multiple datasets	network (DNN).	population and a high crime level	prediction of crime occurrences and works well on high dimensional and multi-model data compared to the previous once	regarding a specific crime type at a given time slot.
8	Sivaranjani et al. [9]	Crime prediction and forecasting in Tamilnadu using clustering approaches	Thought, other frameworks exist, the need arose for timely and more effective prediction and forecasting framework	K-Means clustering, Agglomerative clustering and Density Based Spatial Clustering with Noise (DBSCAN) algorithms	National Crime Records Bureau (NCRB) of India	outline of large crime data and simplify in handling, searching and retrieving of the preferred crime information.	The agglomerative clustering will never redo any step i.e. once attached object can never be separated
9	Deepika et al. [6]	An Approach to Crime Data Analysis: A Systematic Review	need for the advancements in the data storage collection, analysis	CRISP-DM Cross industry standard process for data mining (NNDT)	location based social networks like Facebook, twitter, blogs, surveillance	prediction and then evaluation are performed and visualization of results is done by graphs	tackling the variety of data formats from multiple data sources and transforming the data into a desirable form

E. THE PROPOSED FRAMEWORK

The proposed framework is designed base on Alkesh et al. [19] crime analysis methodology that stipulate Data Collection, Data pre-processing, Classification, Pattern Identification, Prediction and Visualization as phases. It uses different visualization techniques to show the trend of crimes and various ways that can predict the crimes using machine learning algorithms.

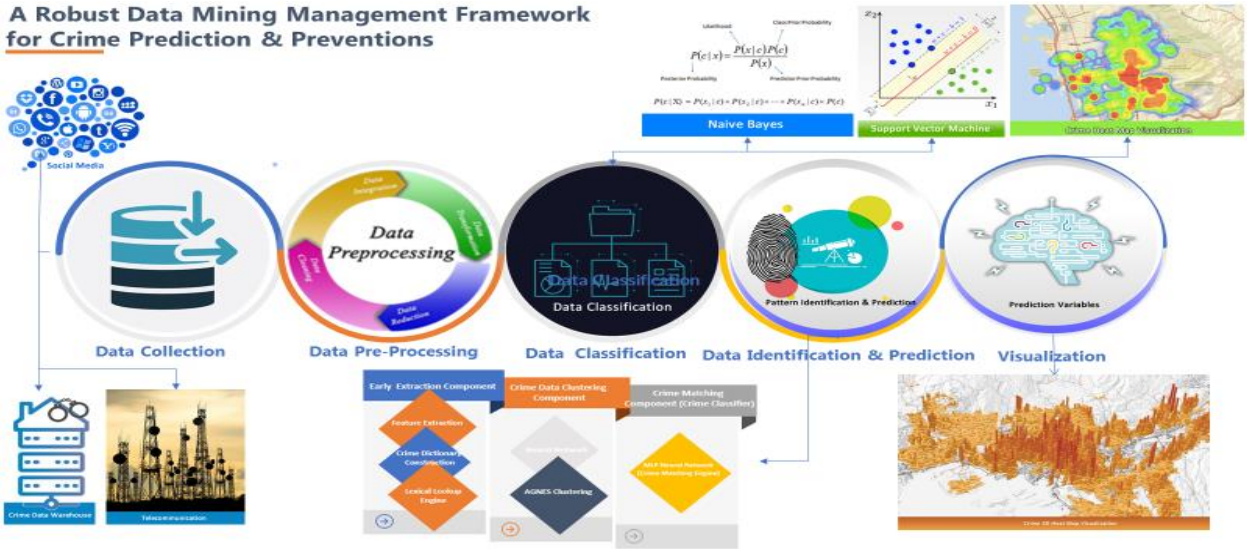

A Robust Data Mining Management Framework for Crime Prediction & Preventions
Social Media
Data Collection
Crime Data Warehouse
Telecommunication
Data Preprocessing
Data Integration
Data Transformation
Data Reduction
Data Cleaning
Data Pre-Processing
Data Classification
Data Classification
Naive Bayes
Likelihood
Class Prior Probability
Posterior Probability
Predictor Prior Probability
$P(c|x)=\frac{P(x|c)P(c)}{P(x)}$
$P(c|X)=P(x_1|c)\times P(x_2|c)\times\cdots\times P(x_n|c)\times P(c)$
Support Vector Machine
x_2
x_1
Crime Heat Map Visualization
Early Extraction Component
Feature Extraction
Crime Dictionary Construction
Lexical Lookup Engine
Crime Data Clustering Component
AGNES Clustering
Crime Matching Component (Crime Classifier)
MLP Neural Network (Crime Matching Engine)
Pattern Identification & Prediction
Data Identification & Prediction
Prediction Variables
Visualization
Crime 3D Heat Map Visualization

Fig 1: A proposed Robust Data Mining Management Framework for Crime Prediction and Preventions (Base on Deepika et al. [6] framework)

The dataset for the framework will be collected from crime related organization data warehouse (like Police), telecommunication companies and social media data. In the pre-processing phase, removal of the inconsistent data (such as missing values, redundant information, etc.) through data, integration, transformation, reduction and clustering so that required data required for the predicting will be classified, then algorithm called Naïve Bayes which is a supervised learning method as well as a statistical method for classification and support vector machine will be used for classification, then Apriori algorithm. Apriori was be used for pattern identification and for prediction, the framework used the decision tree concept. which helps the algorithms to make better decisions about variables? So, for getting the crime prone areas we pass current date and current attributes into the prediction technique, the result is shown using some visualization mechanisms. Based on this information the framework provides the use of the following algorithms. Entity Extraction as a branch of text mining, Crime data clustering and Neural Network Techniques to make the prediction

F. Conclusions and Recommendations

Crime trends and patterns identification to predict what kind of offenses might occur next in a particular district within a specific period of time and season. So, combined techniques are required to build a better crime prediction by integrating multiple models to solve single problem for improving prediction performance of single classifier that helps to predict what kind of crime might occur next in a particular district within a certain period of time and identifies the Season and time factor at which crimes are occur more frequently happening of crime. Hence crime prediction helps people stay away from the districts at a certain time of the day, month and season along with saving living style. In addition, having this kind of knowledge would help people to improve their living and travelling place choices. And the paper recommends full adoption and implementation of this framework as an integrated enterprise software. for evaluation purpose and success quantifiable matrix visibility and Further improvement should be done on the framework to accommodate Misspelling and grammatical mistakes.

G. References

[1] Pallavi, W. and Shelke, R., R. (2016). A study of data mining tools in knowledge discovery process. International Journal for Research in Applied Science and Engineering Technology, 4(III), 74-79.

[2] Swadi B. Al-Janabi, Kadhim and Fatlawi, Hayder (2016). Crime data analysis using data mining techniques to improve crimes prevention procedures. Iraqi conference for Information technology, 1(3), 34-39.

[3] Shamaila Q, and Hafsa S. D. (2018). A survey of data mining techniques for crime detection university of sindh. Journal of Information and Communication Technology, 2(1), 1-6.

[4] Shruti S.Gosavi and Shraddha S. Kavathekar (2018). A survey on crime occurrence detection and prediction techniques. International Journal of Management, Technology and Engineering, 8(XII), 56-62.

[5] Bhargava R, Rathore P. and Sangwa R. (2018). A contemplated approach for criminality data using mining algorithm. International Journal on Future Revolution in Computer Science and Communication Engineering, 4(2), 236-240.

[6] Deepika T. and Sanjiv S. (2018). An approach to crime data analysis: a systematic review. International Journal of Engineering Technologies and Management Research, 5(2), 67-74.

[7] Prabakaran and Shilpa, M. (2018). Survey of analysis of crime detection techniques using data mining and machine learning. Journal of Physics, 1(10), doi: 10.1088/1742-6596/1000/1/012046.

[8] Hitesh, R., T., Bhavna S. and Ginika M. (2018). Crime prediction and monitoring framework based on spatial analysis. Aje. J and Baek. S (Eds.), International Conference on Computational Intelligence and Data Science in Procedia Computer Science (pp. 696–705). Elsevier Ltd.

[9] Sivaranjani S, Dr. Sivakumari S. and Aasha M. (2016). Crime prediction and forecasting in tamilnadu using clustering approaches. IEEE International Conference on Emerging Technological Trends, 978(1), 13-19.

[10] Vijayalakshmi M, Shivansh Bhatt, and Himanshu Goyal (2018). Crime pattern recognition and prediction using optimsed k-means and svm. International Journal of Pure and Applied Mathematics, 118(22), 581-586.

[11] Savita M. and Gurjit K. (2017). Introduction to data mining and data warehouse. International Journal of Advanced Research in Computer Science, 8(4), 398-400.

[12] Georgina N., O. and Arthur, U. (2014). Leveraging data mining and data warehouse to improve prison services and operations in Nigeria. Information and Knowledge Management, 4(5), 27-34.

[13] Sankar, J (2014). A literature review on data mining. International Journal of Research in Computer Applications and Robotics, 2(7), 95-101.

[14] Devakunchari R, Bhowmick S, Bhutada S. P, and Shishodia Y. (2019). Analysis of crimes against women in india using regression. International Journal of Engineering and Advanced Technology, 8(2S2), 124-127.

[15] Vishnupriya V and Valarmathi M. (2019). An effective data mining technique for analyzing

crime patterns. Journal of Computer Engineering, 1(7), 26-30.

[16] Ginger Saltos and Mihaela Cocea (2017). An exploration of crime prediction using data mining on open data. International Journal of Information Technology and Decision Making, 16(05), 1-27.

[17] Mona Mowafy, Rezk A, and El-bakry H. M. (2018). Building unstructured crime data prediction model: practical approach. International Journal of Computer Application, 8(4), 1-7.

[18] Kang H.W, and Kang H.B (2017). Prediction of crime occurrence from multi-modal data using deep learning. Journal of Pone, 12(2), doi: 10.1371/journal.pone.0176244.

[19] Alkesh, B., and Sarvanaguru, R., A. (2018). Crime prediction and analysis using machine learning. International Research Journal of Engineering and Technology, 5(9), 1037-1042.

M-Agricultural Networks (M-AN): An Information System Management Approach Towards Food Security and Sustainability in Nigeria

Osang[1], Francis Bukie and Umoren, I.[2]
[1]Department of Computer Science, National Open University of Nigeria, Abuja
[2]Department of Computer Science, Akwa Ibom State University, Ikot Akpaden, Nigeria
fosang@noun.edu.ng *+234 7057129566*, imehumoren@aksu.edu.ng *+234 8036813637*

*****Corresponding author: fosang@noun.edu.ng +234 7057129566

Abstract

Information Systems (IS) have become critical for restructuring organizational processes, and as a result, the nature of the roles and duties of IS professionals have changed. It has become increasingly important for IS professionals to understand how the objectives of an IS relate to the organizational goals. The paper offers an effective approach for managing agricultural information; discusses latest development on applications and developments in the use of Information Technologies. The system is developed and implemented using Visual Studio.Net, SQL Compact 3.5, Corel graphics and Word processor based on Data Driven approach to help farmers achieve high productivity efficiency and timely solution to problems with cost effectiveness. The paper adopted structured interviews to collect data from a randomly selected farming members and some non-members of farming association as well as the Unified Process Methodology (UPM) towards the realization of a sustainable agricultural development. Hence, the relevance of the studies on agricultural information system is undoubtedly increasing. The work also provides an appropriate forum for agricultural information specialists for dissemination of information, exchange and knowledge sharing. Basically, this paper presents a platform to effectively manage our data, and how to transform the data into information, useful for decision making, and how that information eventually becomes significant knowledge in impacting the Niger Delta region of Nigeria towards food security and sustainability. Firstly, the information technology such as the Internet that is now dynamically changing our life style and social consciousness provide the best tool for information sharing and mutual communication in agricultural services. Secondly, information science enables an effective and stable agricultural production through several models such as crop growth prediction and decision support system.

Keyword: M-Agricultural Networks, food security, information system, automated system etc.

1. Introduction

The mobile revolution is finally here. Certainly, the role of ICT to enhance food security, sustainability and support rural livelihoods is gradually recognised and was formally endorsed at the World Summit on the Information Society (WSIS) 2003-2005. Wherever one looks, the evidence of mobile penetration and adoption is irrefutable. Though, the Nigerian population has not yet experience explosion, clues of food crisis can be seen; that is, the lack of food in some rural communities is chronic and even major States are facing the difficulties to balance the agricultural productivity and the environmental requirement. As the balance of food supply and demand are now inevitably under the strategy of the world trading mechanism and control of the developed countries, it is almost meaningless to solve the crisis within the local communities and the state. Absolutely, the national and international involvement and collaboration is required for sustainable food productivity on the basis of provision of relevant Agricultural information and mutual understanding which could reduce the level of food insecurity. The Nigerian population is growing geometrically in the world and the requirements on food is abruptly changing from quantity to quality as may be required. In Nigeria, there are several individual agricultural features that are not common among the developed countries. Yams, Beans, cassava or rice-dependency and farming scale are the typical examples. In this work, sharing such common features is of necessity. The researchers, biotechnologist, administrators and farmers who are interested in the Nigerian-Agro informatization and world agriculture, may join together, exchange information and discuss about agricultural bioscience information and agro-informatics which will help many nations find some key areas to solving nation -wide food crises.

Stimulation of agricultural production in Nigeria requires more effective agricultural information and management system. Rolls *et al.* (1994) conducted studies to analyse the information system for smallholder farmers in Malaysia. They put farmers central to the information system and found their roles as producer, inventor and communicator. There was a considerable information exchange among the actors in the system and the farmers in particular were active in disseminating innovative information and technology. Availability, accessibility and utilization are very crucial in the development efforts of any nation’s agricultural sector. According to (Demiryurek, 2006), agricultural information can be transferred to large numbers of farmers through mass media simultaneously and at a lower cost per farmer than other extension methods (group or individual). However, the availability of these sources limits the farmers' access as well as their usefulness. Again, the mass media has weaker feedback potential than other conventional extension methods and the capacity of these sources at the stage of adoption of agricultural innovations is limited (Demiryurek 2006). Therefore, the success of farm business actually requires production record as an instrument that is required for performance measurements. According to Ackerman (2016), "The advances that have taken place in calculating equipment and methods make it possible to determine the relationship between ultimate yields, time of harvest and climatic conditions during the growing season. Relationship between the perspective and actual yields and changing prices can be established. With such information at hand the farmer should be in a position to make a decision on his prediction with a high degree of certainty at mid-season regarding his yield and income at harvest time." (Mishra *et al.,* 1999; Muhammad *et al.,* 2004, allege that, modern agricultural production has been attributed partly to the rapid spread of information and the ability of the developing countries to utilize the research result. Ciborra (2005), proposes that information systems "deal with the deployment of information technology in organizations, institutions, and society at large". The skilful and conceived management is one of the most important success factors for today's farms. Only when a farm is well managed, it can generate the funds to finance its sustainable development and thereby its survival in today's fast changing environment. Developed countries are using Management Information System to assist deferent task for their end users. The use of agricultural website helps in dissemination of vital agriculture information such as online detailed content, crops, crop management techniques, fertilizers and pesticides and many other agricultural related materials. Fundamentally, the challenge in southern Nigeria shows that, despite having agricultural information systems to help farmers in production, most people do not have this information at their fingertips because it requires farmers the use of Internet and other advanced technologies to access such information. This paper aimed at developing an automated system that will give timely solutions to problems based on sets of data presented by the farmers using some related set of rules to generate knowledge for efficient decision making.

2. Related Literature

According to Aina (2004), IT is an omnibus term that combines computer and telecommunication technology; hence, it is sometimes called Information and Communication Technology (ICT). It is concerned with the technology used in handling, acquiring, processing, storing, and disseminating information (Afolabi, 2012). The computer is useful for processing information while the telecommunication facilities provide means for information communication or transfer using networks. However, for computers to be able to communicate with one another there has to be a network which provides a link, and when this link is across the globe, an international network called Internet results. The Internet is a connection of millions of computers all over the world by networks (Ogbomo, 2004; Ibegwam, 2002).

A variety of innovations that integrate ICTs into the dissemination of agricultural information to farmers (Farmers Information Services –FIS) have been developed at local, national and regional levels. They have currently demonstrated a promising field of new research and application in agriculture whilst bringing new sources of information and new tools for local knowledge dissemination. They are increasingly enabling farmers to focus, search and extract useful and up-to-date market information. Because of its potential to ameliorate this old rural farming problem an evaluation of its usage among farming communities becomes necessary (Muriithi *et al.* 2009).

Jain *et al.,* (2010) have found out that many farmers are not completely utilizing the full potential of the ICT. The integration of information and communication technology (ICT) in agriculture can be utilized for providing accurate, timely, relevant information and services to the farmers, thereby facilitating environment for more remunerative agriculture production. According to (Afolabi, 2012), the availability of the Internet as a major component of ICT has improved access to information tremendously. Hence, information users and Information Technology is the buzz technology now-a-days. It is the technology that is helping to exchange the information in fast and easier way. Due to this technology the distance between or the difference between the nations is reduced and now world is becoming a global village. This technology provides an opportunity to the developing nations and under developed nations so that can build up their strategies and compete with the developed nations. Agricultural decisions on: timely land preparation, planting, weeding, irrigation, harvesting, storage and marketing have always been central concerns to agricultural stakeholders. ICT especially mobile telephones can speed the way farmers in rural areas of Nigeria get, exchange and manipulate information. They rework the way farmers interact with markets and cities. According to Krishna (2012), the agricultural sector is faced with major challenge of increasing production to feed a growing and increasingly prosperous population in a situation of decreasing availability of natural resources.

According to Ozowa (2010), over the years, deliberate, though ineffective efforts have been made by donors and African countries to bring about agricultural development without much to show for it. Much of the failure is attributed to non-integration of agricultural information with other development programmes to address the numerous related problems that face farmers**.**

Jain *et al,* (2010), have found out that many farmers are not completely utilizing the full potential of the ICT. The integration of Information and Communication Technology (ICT) in agriculture can be utilized for providing accurate, timely, relevant information and services to the farmers, thereby facilitating environment for more remunerative agriculture production.

Ugwuishiwu et al (2012), carried out a research on the application of ICT in Crop Production and developed an Agro- Information System. The system architecture defines the key components of the system together with the interactions between these components. The system was modelled in a way that when a user want to view information in the system, the user has to enter a name of a particular crop, from there he selects the variety of the crop, and view the requirements of that variety.

According to Mtega and Msungu (2013), ICTs facilitate the accessibility of agricultural information services and thus is a channel necessary for building local capabilities, integrating new and traditional knowledge and increases profit from agriculture. The role of ICTs in supporting agricultural production system has been identified to play vital role in the transfer of technology and in sharing modern agricultural practices with the farmers. Nwagwu and Soremi (2015), posit that in increasing access and

exchange of information, ICTs offer the potential to increase efficiency, productivity, competitiveness and growth in various aspects of agricultural sector.
Cornelia (2016), Carried out a research and implemented a system that carries out automated management of crop production which has been proven effective for proper management of information as regards crop production and supports adequate decision-making.

1) 3. Sustainable Food Systems

A healthy, sustainable food system is one that focuses on Environmental Health, Economic Vitality, and Human Health & Social Equity as illustrated in figure 1.

Figure 1: Mobile Agriculture Information Network

- **Environmental Health** – ensures that food production and procurement do not compromise the land, air, or water now or for future generations.
- **Economic Vitality** – ensures that the people who are producing our food are able to earn a decent living wage doing so. This ensures that producers can continue to produce our food.
- **Human Health & Social Equity** – ensures that particular importance is placed on community development .and the health of the community, making sure that healthy foods are available economically and physically to the community and that people are able to access these foods in a dignified manner.

4. Innovation IT Project

4.1 Mobile Agriculture (m-agriculture) Nigeria: A National Program to Realize Nigeria Mobile Agricultural outline

In M-Nigeria, we proposed an extensive mobile Internet application and services infrastructure development project, designed to ensure that high-speed, wireless broadband access is available almost everywhere on Nigeria. In an attempt to promote the development of Nigeria's Agriculture information and communications networks, M-Nigeria focuses on the three main areas: M-Service, M-Life, and M-Learning with the goal of establishing wireless broadband M-metropolitan (M-Cities) and providing wireless access services to several subscribers Osang (2016). Through the widespread adoption of wireless applications, M-Nigeria aims to improve Nigeria's international ranking for agric-information mobile Internet access and the international competitiveness of its cities Osang and Ngole (2013). It also intends to go one step further to reach its goal of helping Nigeria export not only agriculture products, but also total system solutions to the world. Figure 2 shows the framework of Mobile Agriculture Information (MAI) as may be applicable.

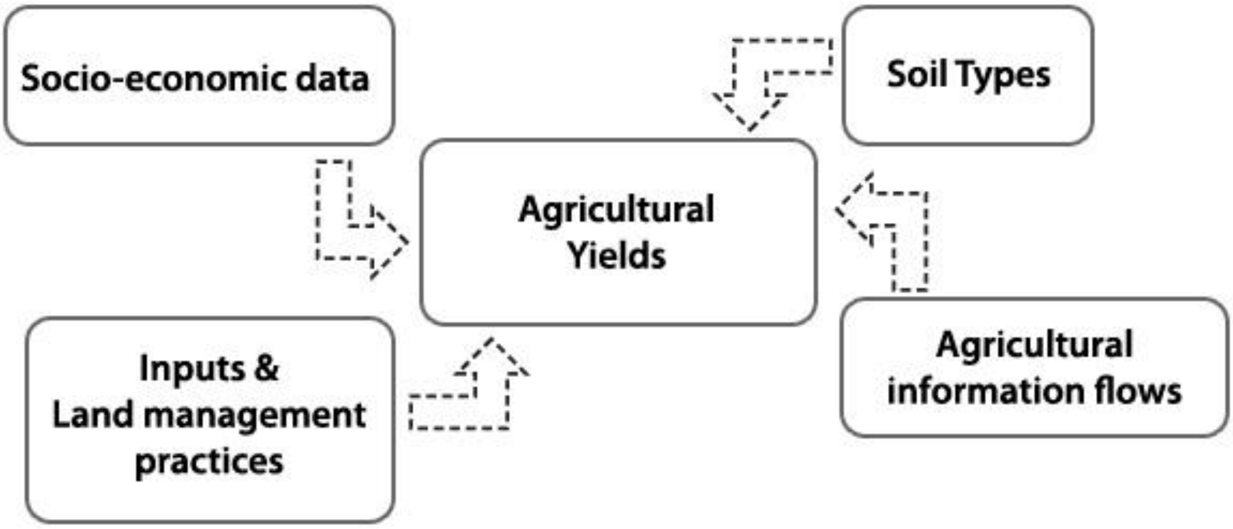

Socio-economic data
Soil Types
Agricultural Yields
Inputs & Land management practices
Agricultural information flows

Figure 2: The framework for Mobile Agriculture Information

Figure 3 demonstrate the Architectural Diagram of the Proposed System as may be required.

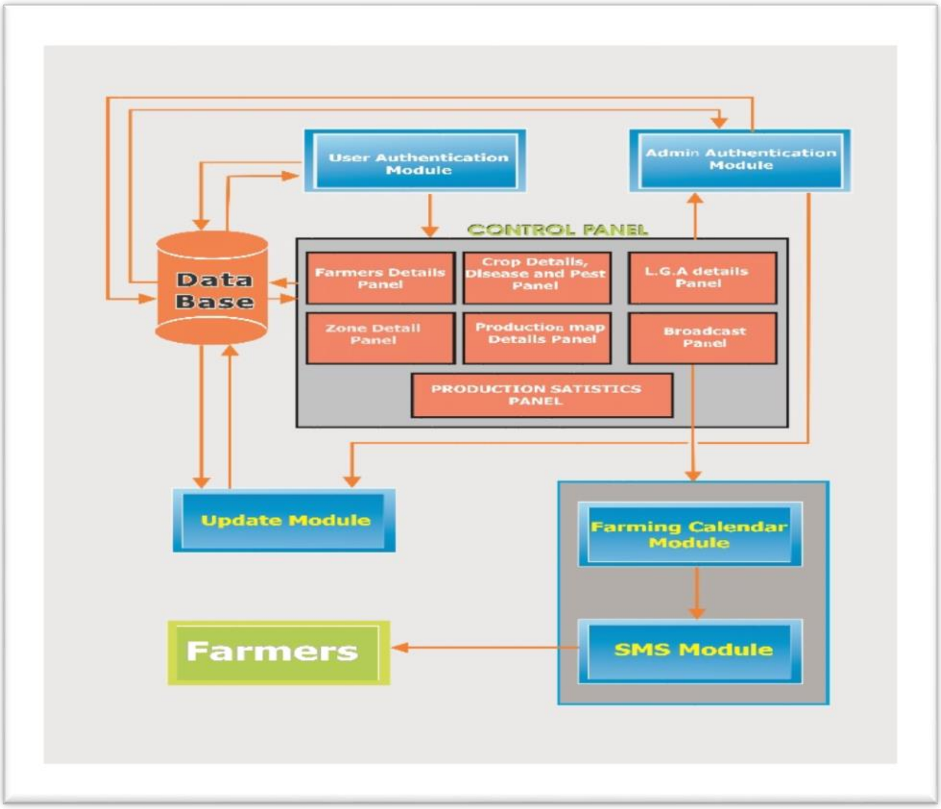


Figure 3: The Architectural Diagram of the Proposed System

4.2 System Model

An innovative and sustainable national development through Information Tchnology (IT) centred largely on the economy which rests on three pillars: ‘knowledge’, ‘information’ and ‘technology’. The structural framework is model as shown in the figure 4.

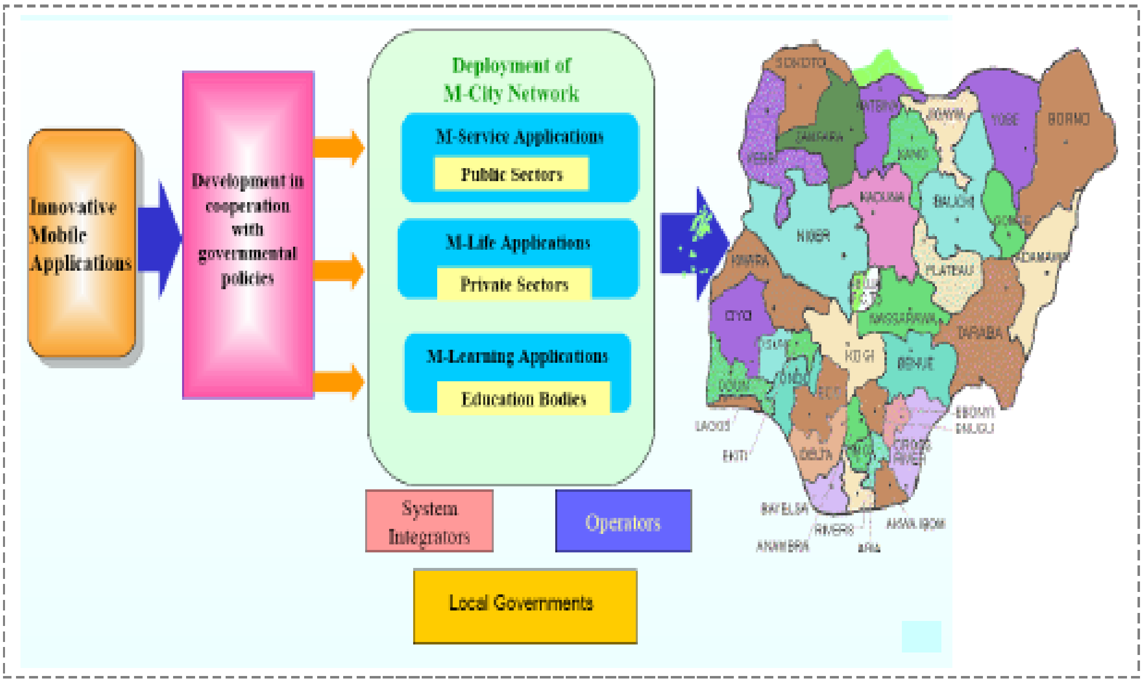


Figure 4: Mobile Agriculture Architecture

5. Opportunities and Challenges for InnovativeInformation Technology (IT) Development in the Agricultural Sector

The agricultural sector is confronted with the major challenge of increasing production to feed a growing and increasingly prosperous population in a situation of decreasing availability of natural resources. Factors of particular concern are water shortages, declining soil fertility, effects of climate change and rapid decrease of fertile agricultural lands due to urbanisation. However, the growing demand, including for higher quality products, also offers opportunities for improving the livelihoods of rural communities. Realising these opportunities requires compliance with more stringent quality standards and regulations for the production and handling of agricultural produce. New approaches and technical innovations are required to cope with these challenges and to enhance the livelihoods of the rural population. The role of ICT to enhance food security and support rural livelihoods is increasingly recognised and was officially endorsed at the World Summit on the Information Society (WSIS) in 2003-2005. This includes the use of computers, internet, geographical information systems, mobile phones, as well as traditional media such as radio or TV. Although it is a relatively new phenomenon, evidence of the contribution of ICT to agricultural development and poverty alleviation is becoming increasingly available.

Low data usage: In Nigeria, mobile data usage is low. There are moderate users of mobile data services, but these are largely based in the developed regions. In some rural regions schooling may be limited and therefore a large majority of people - of all age groups - are unfamiliar with using technology, which severely reduces the demand for mobile data services. If operators are to grow mobile data revenues in these regions, they must first educate these customers on the benefits of mobile data services.

Enhancing agricultural production: Increasing the efficiency, productivity and sustainability of small scale farms is an area where ICT can make a significant contribution. Farming involves risks and uncertainties, with farmers facing many threats from poor soils, drought, erosion and pests. Key improvements stem from information about pest and disease control, especially early warning systems, new varieties, new ways to optimise production and regulations for quality control. Improving market access: Awareness of up-to-date market information on prices for commodities, inputs and consumer trends can improve farmers' livelihoods substantially and have a dramatic impact on their negotiating position. Such information is instrumental in making decisions about future crops and commodities and about the best time and place to sell and buy goods. In many countries, initiatives have appeared that seek to address this issue . Simple websites to match offer and demand of agricultural produce are a start of more complex agricultural trade systems. These sites tend to evolve from local selling/ buying websites and price-information systems, to systems offering marketing and trading functions. Typically, price information is collected at the main regional markets and stored in a central database. The information is published on a website, accessible to farmers via information centres. To reach a wider audience, information is broadcast via rural radio, TV or mobile phone, thereby creating a 'level playing field' between producers and traders in a region.

Provision of standard project that displays prices on light boards at major markets. The sustainability of these systems requires attention, with an important role for the private sector and organised producer groups. Web-based trading platforms offering one-stop shop facilities are emerging, especially for main commodities. In recent years, short message and text services have taken up and effectively deliver prices and trading information via mobile phone to farmers. Partner organisations are supported in adding ICT to core processes. For instance, in Ghana, IICD supports the Social Enterprise Foundation of West Africa (SEND) in linking rural soybean producers to mills, through the use of satellite, databases and mobile phones, thereby ensuring a fair income for producers and a steady supply of raw materials for the mills. Capacity-building and empowerment: Communities and farmer organisations can be helped through the use of ICTs to strengthen their own capacities and better represent their constituencies when negotiating input and output prices, land claims, resource rights and infrastructure projects. ICT enables rural communities to interact with other stakeholders, thus reducing social isolation. It widens the perspective of local communities in terms of national or global developments, opens up new business opportunities and allows easier contact with friends and relatives. A role is also played by ICT in making processes more efficient and transparent. It helps in making laws and land titles more accessible. Global Positioning Systems (GPS) linked to Geographical Information Systems (GIS), digital cameras and internet, help rural communities to document and communicate their situation. The Rural communities benefit from better access to credit and rural banking facilities. Recent mobile banking initiatives offer further scope to reduce costs and stimulate local trade.

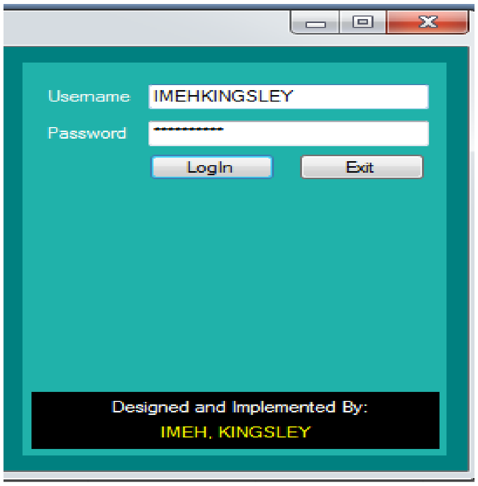
6. Evaluation and Discussion of Results.

The result of this research is very encouraging and rewarding. Although this system is for agricultural production enhancement, the design is security conscious and this is seen in three aspects.

a) Creating a User
b) Updating farmers and other system information
c) Information dissemination

Most of these security areas can only be performed by an administrator.

The log-in page is a proof of the first level User authentication as seen in figure 5.

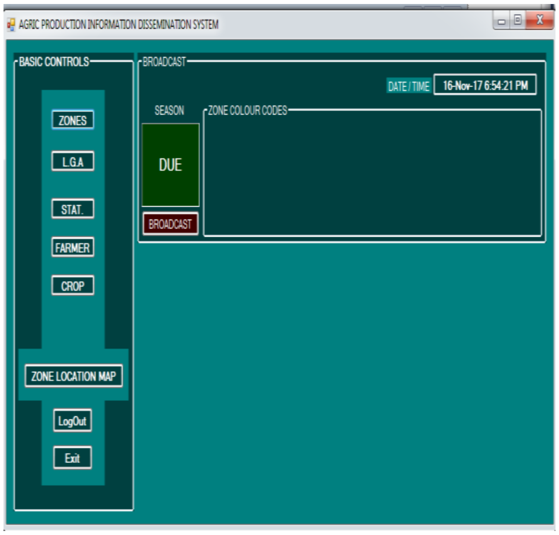


Username: IMEHKINGSLEY

Designed and Implemented By:
IMEH, KINGSLEY

Figure 5: Application Login page

Every intended user must be an enrolled user. Access can only be granted to a user if only it is an existing user, otherwise access is denied. The System's Control Panel is a bus route to every activities of the system. This interface is as shown in figure 6.

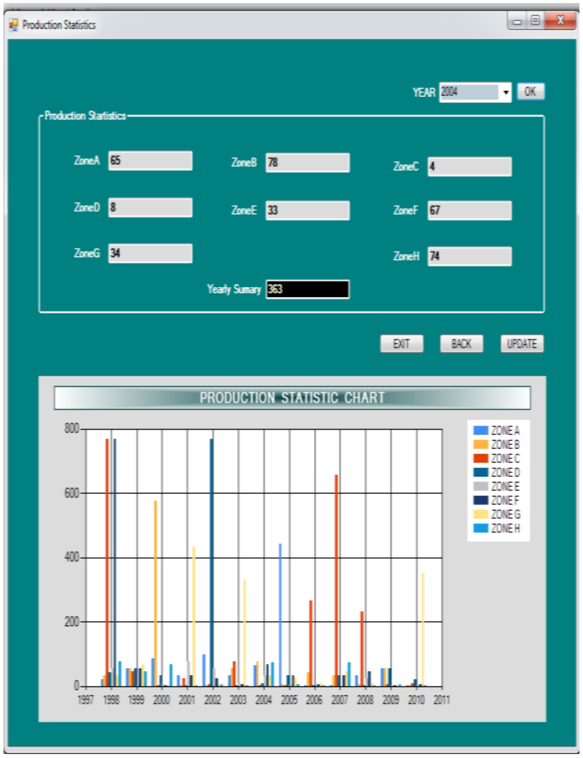


DATE / TIME: 16-Nov-17 6:54:21 PM

SEASON: DUE

Figure 6: The System Control Panel

This system is able to give some valuable and critical information. This information can be seen as shown in the production statistic panel, as shown in figure 7.



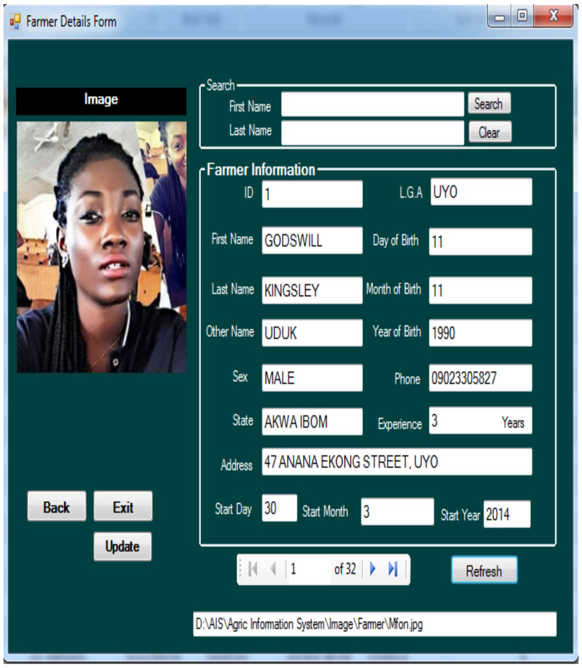
YEAR: 2004

ZoneA	65	ZoneB	78	ZoneC	4
ZoneD	8	ZoneE	33	ZoneF	67
ZoneG	34			ZoneH	74

Yearly Sumary: 363

Figure 7: Production Statistic Panel

One of the most important modules of this system is the Farmer-Information module. This module holds important information of every enrolled farmer in the state of study and can be queried for some desired information at a particular time. With this provision, a farmer can be searched based on the First Name and the Last Name. The screen shot of this system module is as shown in figure 8.



Farmer Information

ID	1	L.G.A	UYO
First Name	GODSWILL	Day of Birth	11
Last Name	KINGSLEY	Month of Birth	11
Other Name	UDUK	Year of Birth	1990
Sex	MALE	Phone	09023305827
State	AKWA IBOM	Experience	3 Years

Address: 47 ANANA EKONG STREET, UYO

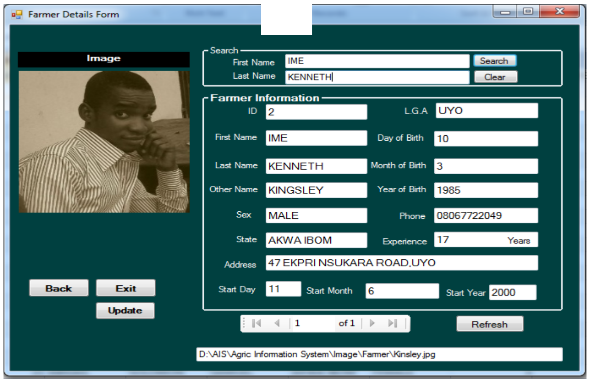
Start Day: 30 Start Month: 3 Start Year: 2014

1 of 32

D:\AIS\Agric Information System\Image\Farmer\Mfon.jpg

Figure 8: Farmers details panel

Suppose I have a farmer with the first name 'FABIAN' and Last Name 'EDEM', This means that, having the information of this farmer within some nanosecond is very achievable. This possibility is demonstrated in figure 9.



Farmer Details Form

Search
First Name: IME
Last Name: KENNETH

Farmer Information
ID: 2
L.G.A: UYO
First Name: IME
Day of Birth: 10
Last Name: KENNETH
Month of Birth: 3
Other Name: KINGSLEY
Year of Birth: 1985
Sex: MALE
Phone: 08067722049
State: AKWA IBOM
Experience: 17 Years
Address: 47 EKPRI NSUKARA ROAD,UYO
Start Day: 11
Start Month: 6
Start Year: 2000

D:\AIS\Agric Information System\Image\Farmer\Kinsley.jpg

Figure 9: Search result on farmer screen shot

6.1 New User Creation

The User Creation Control makes it possible for a New User to be instantiated right from the Logon Page, through ***File-Menu*** and ***Create*** as a sub-menu. This functionality of the system requires the Administrator's acknowledgement in form of ***Username*** and ***Password***. In the event of successful credentials supplied to the criteria interface, the New User Creation Dialog with **Menu-Click from Homepage is** presented for the creation process as may be required. See figure 10, figure 11 and figure 12 respectively.

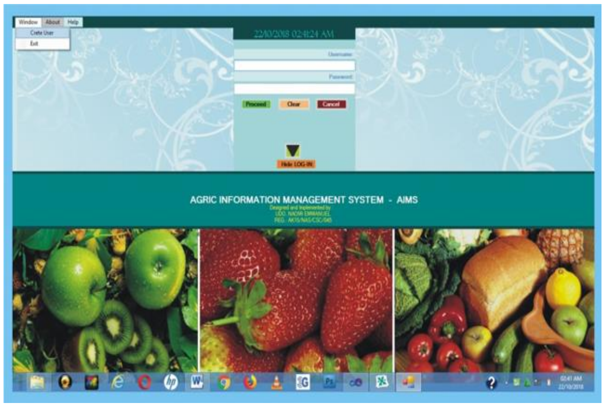


Figure 10: Create New User Menu-Click from Homepage

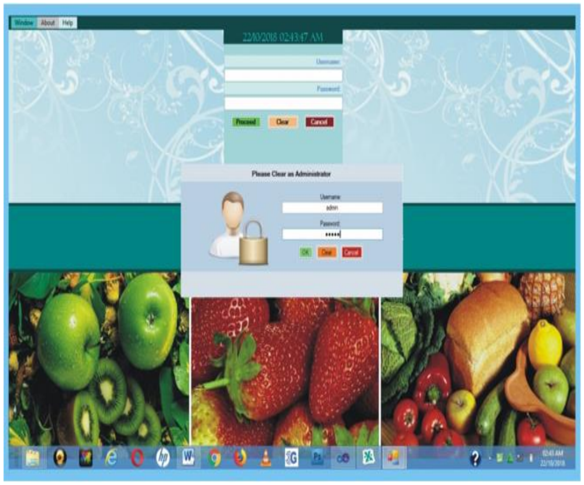
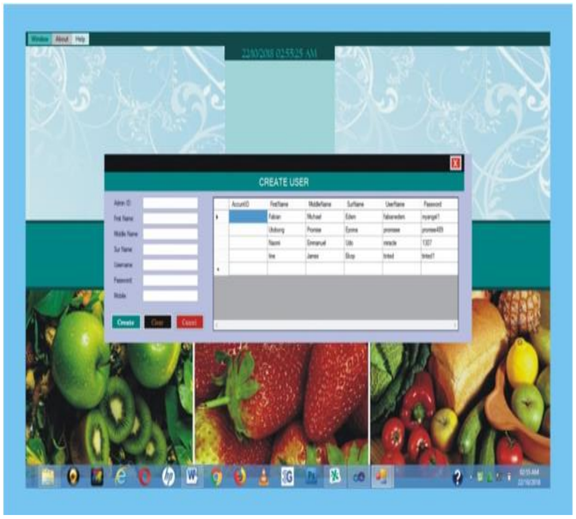


Figure 11: Administrator's Credential Screen



CREATE USER

Admin ID
First Name
Middle Name
Sur Name
Username
Password
Mobile

AccountID	FirstName	MiddleName	SurName	UserName	Password
	Fabian	Michael	Edem	fabianedem	myangel1
	Utibong	Promise	Ekoma	promisee	promise405
	Naomi	Emmanuel	Udo	miracle	1307
	Ime	James	Etop	tinted	tinted1

Figure 12: New User Creation Screen

6.2 User Login

High-level security provision checks an unauthorized usage in the developed system. This means that a user, if only must interact using the System, must be either a previously enrolled user with valid credentials or newly created by an Administrator. Hence, a login trial with the wrong credentials bounces off such unauthorized user. Figure 13 and Figure 14 shows the above discussed results.

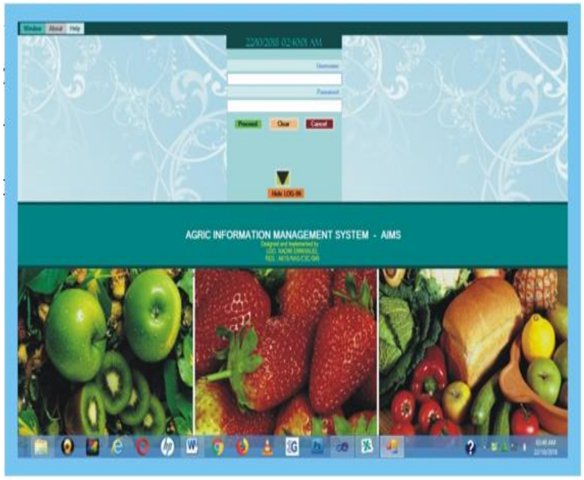
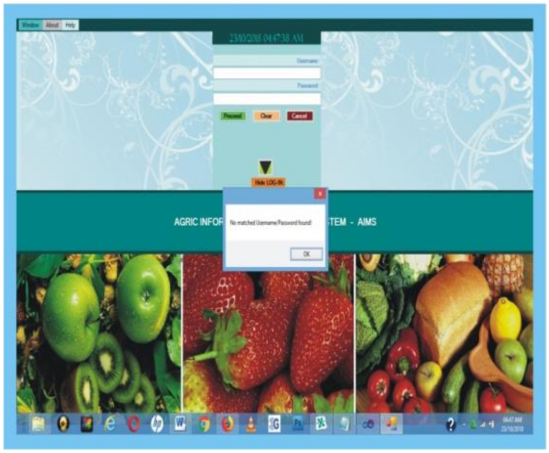


Figure 13: *User Login Screen*

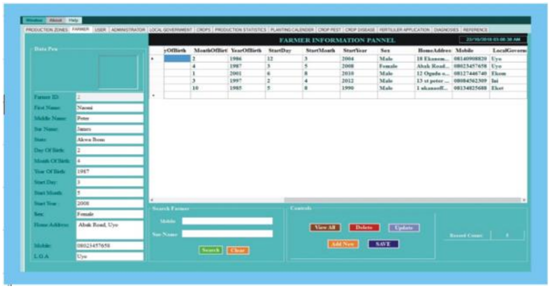


No matched Username/Password found!

Figure 14.: User Login Screen with Bounced Feedback Due to Wrong Parameter input(s)

6.3 Database Entity Manipulations

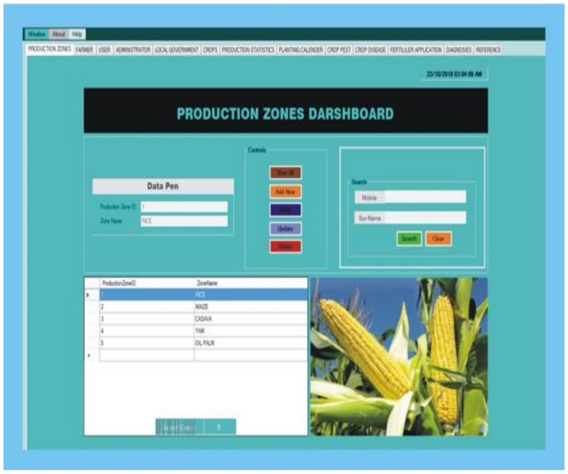
The system uses database to store data and processed information for effective retrieval and manipulations. Data manipulation in the developed system is very optimal and efficient. It has also been evaluated true of Creation, Reading, Update and Deletion (CRUD) Processes in the system. Results is as shown in Figure 15, Figure 16 and Figure 17.


Window About Help
PRODUCTION ZONES FARMER USER ADMINISTRATOR LOCAL GOVERNMENT CROPS PRODUCTION STATISTICS PLANTING CALENDER CROP PEST CROP DISEASE FERTILILER APPLICATION DIAGNOSIES REFERENCE
FARMER INFORMATION PANNEL
Data Pen
Search Farmer
Mobile
Sur Name
Search Clear
Controls
View All Delete Update
Add New SAVE
Record Count

Farmer ID: 2
First Name: Naomi
Middle Name: Peter
Sur Name: James
State: Akwa Ibom
Day Of Birth: 2
Month Of Birth: 4
Year Of Birth: 1987
Start Day: 3
Start Month: 5
Start Year: 2008
Sex: Female
Home Address: Abak Road, Uyo
Mobile: 08023457658
L.G.A: Uyo

yOfBirth	MonthOfBirt	YearOfBirth	StartDay	StartMonth	StartYear	Sex	HomeAddres	Mobile	LocalGoverm
	2	1986	12	3	2004	Male	18 Ekanem...	08140908820	Uyo
	4	1987	3	5	2008	Female	Abak Road...	08023457658	Uyo
	1	2001	6	8	2010	Male	12 Ogudu o...	08127446740	Ekom
	3	1997	2	4	2012	Male	13 st peter ...	08084502309	Ini
	10	1985	5	8	1990	Male	1 ukanaofE...	08134825688	Eket

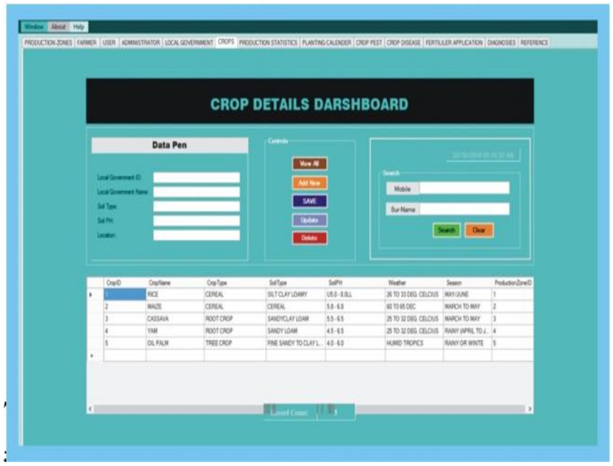
Figure 15: Farmer Information Panel


Window About Help
PRODUCTION ZONES FARMER USER ADMINISTRATOR LOCAL GOVERNMENT CROPS PRODUCTION STATISTICS PLANTING CALENDER CROP PEST CROP DISEASE FERTILILER APPLICATION DIAGNOSIES REFERENCE
PRODUCTION ZONES DARSHBOARD
Data Pen
Controls
View All
Add New
Update
Delete
Search
Mobile
Sur-Name
Search Clear
Record Count

Production Zone ID: 1
Zone Name: RICE

ProductionZoneID	ZoneName
1	RICE
2	MAIZE
3	CASAVA
4	YAM
5	OIL PALM

Figure16: Production Zones Dashboard

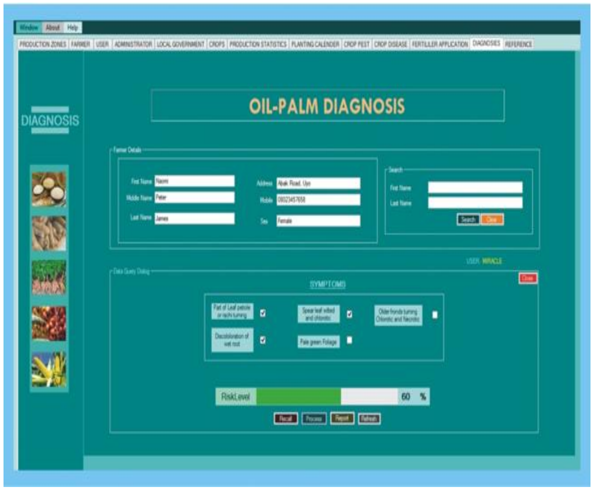


CropID	CropName	CropType	SoilType	SoilPH	Weather	Season	ProductionZoneID
1	RICE	CEREAL	SILT CLAY LOAMY	US 0 - 8.0LL	26 TO 33 DEG. CELCIUS	MAY/JUNE	1
2	MAIZE	CEREAL	CEREAL	5.8 - 6.8	60 TO 65 DEC	MARCH TO MAY	2
3	CASSAVA	ROOT CROP	SANDYCLAY LOAM	5.5 - 6.5	25 TO 32 DEG. CELCIUS	MARCH TO MAY	3
4	YAM	ROOT CROP	SANDY LOAM	4.5 - 6.5	25 TO 32 DEG. CELCIUS	RAINY (APRIL TO J...	4
5	OIL PALM	TREE CROP	FINE SANDY TO CLAY L...	4.0 - 6.0	HUMID TROPICS	RAINY OR WINTE	5

Figure 17: Crop Details Dashboard

6.4 Problem Diagnosis Screen

Diagnosis is entirely a very active unit of the System and deriving knowledge from the database by some intelligent modules to actualize Correct Solution Prediction to target problem as seen in Figure 18 for the results.


Window About Help
PRODUCTION ZONES FARMER USER ADMINISTRATOR LOCAL GOVERNMENT CROPS PRODUCTION STATISTICS PLANTING CALENDER CROP PEST CROP DISEASE FERTILILER APPLICATION DIAGNOSIES REFERENCE
DIAGNOSIS
OIL-PALM DIAGNOSIS
Farmer Details
Search
First Name
Last Name
Search Clear
USER: MIRACLE
Data Query Dialog
Close
SYMPTOMS
Recall Process Report Refresh

First Name: Naomi
Middle Name: Peter
Last Name: James
Address: Abak Road, Uyo
Mobile: 08023457658
Sex: Female

Part of Leaf petiole or rachi turning ☑
Spear leaf wilted and chlorotic ☑
Older fronds turning Chlorotic and Necrotic □
Discoloration of wet root ☑
Pale green Foliage □

RiskLevel 60 %

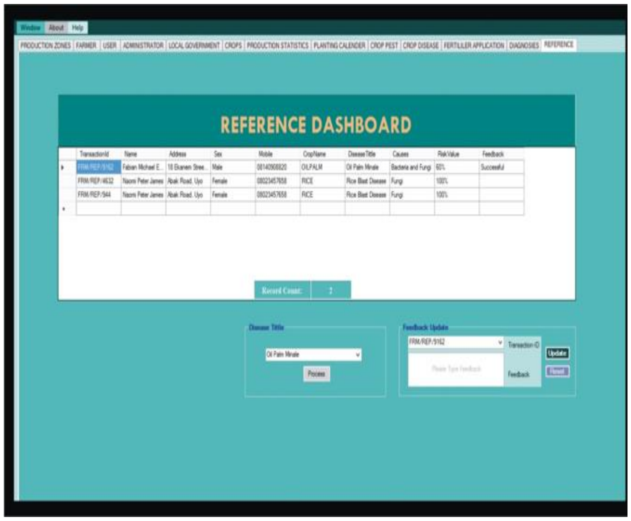
Figure 18: Oil-Palm Diagnosis Screen

6.5 Diagnosis Solution Report and Printing Screen

Diagnosis is one of the key components of the system, which helps in administering relevant solution to problems registered by the Farmer. It also generates a corresponding report sheet to the Farmer, carries very sensitive information concerning every unique transaction for onward evaluation and implementation. It is on this note that the Farmer must give a corresponding feedback for accurate and timely update as demonstrated in the results shown in Figure 20.

6.6 Diagnose Transaction Reference Tracking Screen

This result is a Proof of Diagnose transaction tracking for Prior Reverencing. IT provisions reference issuance to any affected Farmer based on Disease Tittle, the results is as shown in Figure 19.



TransactionId	Name	Address	Sex	Mobile	CropName	DiseaseTitle	Causes	RiskValue	Feedback
FRM/REP/9162	Fabian Michael E...	18 Ekanem Stree...	Male	08140908820	OILPALM	Oil Palm Mirale	Bacteria and Fungi	60%	Successful
FRM/REP/4632	Naomi Peter James	Abak Road, Uyo	Female	08023457658	RICE	Rice Blast Disease	Fungi	100%	
FRM/REP/944	Naomi Peter James	Abak Road, Uyo	Female	08023457658	RICE	Rice Blast Disease	Fungi	100%	

Record Count: 2

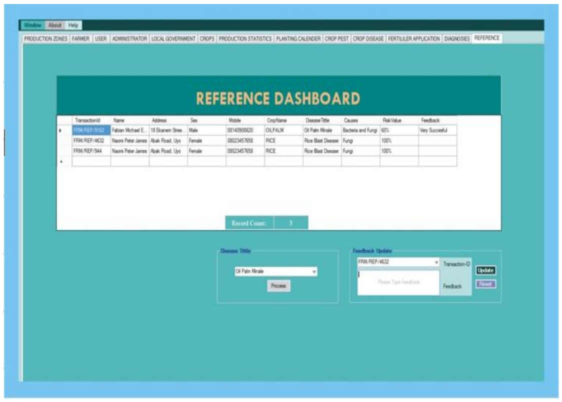
Disease Tittle: Oil Palm Mirale

Feedback Update: Transaction ID FRM/REP/9162; Feedback: Please Type Feedback

Figure 19: Transaction Reference Screen

6.7 Feedback Update

Feedback of every implemented solution generated by the system to the Farmer has to be properly captured to reflect Success of any Solution generated by the system to the Farmer base on Disease Tittle. See Figure 4.17 for Result.



TransactionId	Name	Address	Sex	Mobile	CropName	DiseaseTitle	Causes	RiskValue	Feedback
FRM/REP/9162	Fabian Michael E...	18 Ekanem Stree...	Male	08140908820	OILPALM	Oil Palm Mirale	Bacteria and Fungi	60%	Very Succeful
FRM/REP/4632	Naomi Peter James	Abak Road, Uyo	Female	08023457658	RICE	Rice Blast Disease	Fungi	100%	
FRM/REP/944	Naomi Peter James	Abak Road, Uyo	Female	08023457658	RICE	Rice Blast Disease	Fungi	100%	

Record Count: 3

Disease Tittle: Oil Palm Mirale

Feedback Update: Transaction ID FRM/REP/4632; Feedback: Please Type Feedback

Figure 20: Feedback Update Screen

6.8 Production Statistics Screen

An analytical graph of the yearly yields of each of the production zones is presented on this dashboard. Every individual Zones is updated based on the Field aggregate track record of the production made yearly by the zone. The Bar Chart reflects the characterized figure for Zone-A and their related yield representation in Bars Coloured with Blue.

7.0 Conclusion

Generally, this study adopted Agricultural information system theory approach to understand the situation holistically and to identify the strengths and weaknesses of the system. Many possibilities in the area of innovative and sustainable M-Agriculture for farmers were observed in this work. Diverse issues covering technology advances such as PDA's, laptops, mobile phones and smart phones were considered. The reference also covered m-learning framework for food security, information system and automated system accordingly. M-Agricultural Networks, Information System and Automated System are expected to be seen to act as enablers in the future for successful platform for food security and sustainability. But for M-Agriculture to be commercially successful, educational institutions may have to promote and continue effective research into this idea. Currently, distance is no longer a barrier as people can research and study across large distances. However, it also depends on the continuing development of mobile devices, hardware and software across mobile networks. Nevertheless, a major challenge yet to be overcome is the cost of mobile hardware, software, connection and usage charges. The lack of sustainability of many mobile learning projects indicates that this may well be the major difficulty to implementing M-Agriculture on a national scale. Hence, there is great need for the investigation of low-cost solutions to implementing M-Agriculture so that it can be sustainable. On the whole, M-Agricultural Networks (M-AN) is an approach towards efficient provisioning of food security and sustainability in Nigeria.

8.0 Recommendations

Basically, continued improvement in the performance, usability and connectivity of mobile devices and gradual understanding of the affordances of mobile learning in the wider context of technology supported learning is to be encouraged. Subsequently, this paper considered strategies that can take wireless and mobile learning to sustainable and substantial position in Nigeria. The way forward for an institution will obviously depend on expertise, enthusiasm and the level of institution's resources. Nevertheless, there are some tactics that will enhance the success of a wireless and mobile learning strategy:

a) The work will reward mobile agricultural learning positive local visibility.

b) High-level 'buy-in', managers seen using wireless and mobile devices, will increase credibility and status of learning.

c) Identification and exploration of potential revenue streams will enhance some aspect of wireless and mobile learning.

d) Recognition that mobile and mobile devices are ‘personal’ and encourage ‘ownership’ amongst farmers and mobile devices will develop familiarity, expertise and confidence.

e) Reliable and robust technical support, infrastructure, network access and hardware will mean that agric scientist can innovate – especially in front of ‘live’ classes - without risk.

References

Afolabi, M. (2012). The Use of Information and Communication Technology in Agricultural Research in Nigerian Universities. PNLA Quarterly, the official publication of the Pacific Northwest Library Association, Pp:1-12. Volume 76, number 3. PNLA Quarterly 76:3. (http//: www.pnla.org).

Aina, L.O. (2004). Library and Information Science Text for Africa, Third World Information Services Ltd, Ibadan.

Ciborra, C. (2002:5). *Labyrinths of information,* Oxford: Oxford University Press.

Cornelia, (2016). Managing Crop Production and support services using Management Information System approach, Department of Computer Science Students research project, AKSU, Nigeria.

Demiryurek, K. (2006). Distance education for rural people in developing countries: Turkish experience. *Journal of Extension Systems,* **22**(2), 83-94.

Food Security Network (2012). http://www.foodsecuritynews.com/about-food-security Life Sciences Research Office (LSRO) (1990): Food security

Jac Stienenwith, Wietse Bruinsma and Frans Neuman (2003-2005), International Institute for Communication and Development (IICD), World Summit on the Information Society (WSIS).

Krishna K. P.S. (2012). Mapping & Preliminary Evaluation of ICT Applications Supporting Agricultural Development retrieved from www.acdivoca.org/

Leeuwis, C. (2004). *Communication for rural innovation: rethinking agricultural extension* (3rd. ed.). Oxford: Blackwell Science.

Mtega, W. P. and Msungu, A. C. (2013). Using information and communication technologies for enhancing the accessibility of agricultural information for improved agricultural production in Tanzania. The Electronic Journal of Information Systems in Developing Countries, 56(1): 1-14.

Muriithi, G, Bett E, & Ogaleh SA. (2009). Information Technology for Agriculture and Rural Development in Africa: Experiences from Kenya. Conference on International Research on Food Security, Natural Resource Management and Rural development, Tropetag: University of Hamburg

Ogbomo, M. (2004). Web page design.In Madu, E.C. (Ed.). Technology for information management and service: Modern libraries and information centres in developing countries. Ibadan: Evi-Coleman

Olaniyi, O. A. (2013). Assessment of utilization of information and communication technologies (ICTs) among poultry farmers in Nigeria: An emerging challenge. Journal of Animal Science Advances, 3(7): 361-369. Available online @ www.grjournals.com

O’Farrell, C. (2005). Information and Communication (ICT’s) for sustainable livelihoods http://www.rdg.ac.uk/AcaDepts/ea/AERDD/ICTBriefDoc.pdf .Accessed

Osang, F.B. (2016). An Integrative Model for Determining Smartphone Services Adoption Decisions in Sub-Saharan Africa: A Structural Equation Modelling Approach. International Journal of Natural and Allied Sciences 11(1), 66-78.

Osang, F. B. & Ngole, J. (2013). Prospects and Challenges of Mobile Learning Implementation in Nigeria: Case Study National Open University of Nigeria (NOUN). 5th International on ICT4Africa, Harare (Zimbabwe).

Ozowa, V.N. (2010). Information needs of small-scale farmers in Africa: (http://www.worldbank.org/html/cgiar/newsletter/june97/9nigeria.html).

Rolls, M.J., Hassan, S.H.J., Garforth, C.J. & Kamsah, M.F. (1994). *The agricultural information system for smallholder farmers in Peninsular Malaysia.* Reading: AERDD, University of Reading. (Rural Extension and Education Research Report No.1).

Ugwuishiwu C.H., Udanor C., Ugwuishiwu B.O (2012). Application of ICT in Crop Production, International Journal of Soft Computing and Engineering (IJSCE)ISSN: 2231-2307, Volume-2 Issue-4.

Williams E. Nwagwu and Opeyemi Soremi (2015). ICT Use in Livestock Innovation Chain in Ibadan City in Nigeria, International knowledge Sharing Platform, ISSN (Paper) 2224-7181 ISSN (Online) 2225-062X. (http://www.iiste.org).

World Summit on the Information Society (WSIS, 2003-2005). Towards knowledge societies: UNESCO world report , Geneva Phase: 10-12 and Tunis Phase: 16-18.

Design of a System that uses Information Communication Technology (ICT) to Manage Solar Energy, Reduce Climate Change and Increase Poultry Production

Okeke Godswill C., Ajah Ifeyinwa A., and Eke Vincent O.C.

ABSTRACT

The deleterious effect of burning fossil fuel for electricity is an issue of concern in the present century. Also of global concern is the need for food security in an ever increasing world population. Although efforts have been made to address these issues, existing researches have not considered a microprocessor-based electricity controller with embedded units for managing, monitoring and controlling the generated solar-based electricity aimed at reducing climate change and increasing poultry production. A bottom-up approach was used to design the proposed multi-sensor-based and multi-user interface-based prototype that monitors and controls the lighting and temperature of the poultry house, while wirelessly alerting the farmers of malfunctions that may lead to a system failure. We recommend that ICT be used in this manner, in conjunction with solar electricity to curb climate change and increase poultry production.

Keywords: climate change, ICT, microcontroller, poultry house, solar energy

I. INTRODUCTION

Climate change is the result of emitting gases into the atmosphere that result in new weather patterns that last for at least a few decades, and maybe for millions of years. Climate change results in global warming which is an increase in the average air and ocean temperatures caused mainly by emissions of greenhouse gasses. Man's activities that result in the emission of greenhouse gases include: burning of fossil fuel when cooking, driving, generating electricity, and waste products elimination by incineration. Information and Communication Technology (ICT)-based efforts made to curb the above problems might include but not limited to: reducing travel through online videoconferencing; using e-ticketing, e-billing and online reading media like Kindle, instead of printed paper which might eventually be disposed of by burning; and reducing the burning of fossil fuel for electricity generation through the use of home automation – a system for monitoring and controlling home appliances to reduce their energy consumption.

Many Nigerian entrepreneurs these days, especially those in rural areas, go into poultry farming as a business. Two important requirements for the birds to stay healthy are light at night and a well-regulated temperature. With light, the birds are able to eat and move around at night. When the temperature is too cold, the birds get sick and die off easily. Not all poultry farmers are able to provide adequate light and a precise temperature control for the birds, especially those in rural areas. Due to the erratic nature of mains electricity supply in Nigeria, electricity used in some of these poultry houses are usually sourced from fuel powered electricity generators. These generators emit gases which affect the health of animals – including birds and man – and contribute to climate change.

Considering the geographical location of Nigeria on the globe, solar energy is a better alternative to the energy derived from the fuel powered generators. A major challenge in the use of solar energy is that when the sun sets, the battery stored energy has to be efficiently managed until sunrise. The focus of this research is to provide a means of harnessing ICT in managing the use of solar energy in order to improve poultry production and curb the problem of climate change. When solar electricity is generated, stored and managed, the constant power it supplies will allow the poultry house appliances, including light and temperature controls, to work as and when required, thereby increasing poultry production and reducing emissions due to fuel powered generators. By helping poultry farmers switch from polluting fuels, such as kerosene, petrol and diesel, to clean solar energy, the proposed design will help to improve poultry production and tackle climate change, thereby improving lives.

I. LITERATURE REVIEW

Poultry production is one of the fastest growing sectors of livestock industry in developing countries and environmental variation is one of the major factors that affect the sustainability of livestock production systems in tropical climate (Sinha *et al.*, 2017). The climatic factors that affect poultry production include light, temperature, humidity, and wind. Information and communication technology (ICT) systems can be used to provide adequate control of these factors to increase exercise, eating and drinking, and thereby reduce skeletal and metabolic disorders among birds.

Lighting patterns for broilers are aimed mainly at stimulating and controlling feed intake. This is mostly important for 1-5 days old chickens which need 24 hours constant light with 5-60 minutes blackout training and 6-10 days old chickens which need 23 hours constant light (Ajay, 2016).

An experimental study was conducted by a joint research program between Aviagen Broiler Breeders and the University of Saskatchewan. The research described the impact of 14, 17, 20 and 23 hours of light per day with all darkness provided in one period, on broiler production and meat yield parameters, welfare and bird health. It was discovered that providing broilers with 20 hours of light a day gave the highest growth rate at all ages. Broilers given 20 hours of light a day also had the highest feed intake. On broiler welfare, data on the effects of lighting on broiler production shows that the best production occurs in broilers given between 17 and 20 hours of light (Karen *et al*, 2010). These outcomes buttress the relevance of an automatic system for managing the lighting of the poultry house.

The changes in the biomass as result of the continuous global warming has a deleterious effect on domestic animals in general, and birds in particular. Excess heat or cold can cause the death of chicken and adversely affect the return from the enterprise. Birds, like mammals are homoeothermic, they produce heat to maintain a relatively constant body temperature and may permit certain variations within their temperature range without significant perturbation (St-Pierre *et al*, 2003).

But a bird can only give off heat to its environment if the temperature of that environment is cooler than the bird. If heat produced by the birds is not moved away from them and out of the poultry house quickly, it will be more difficult for them to avoid heat stress (Sottnik, 2002). Ambient temperature (AT) above 25 ^oC is stressful for birds, but more stressful is the fluctuations caused by this environmental thermal changes, especially when it is accompanied by high relative humidity (RH), as these unleash various pathophysiological response in birds (Sritharet *et al*, 2002; Simon, 2003).

The following temperatures are recommended for broilers (Jan, 2019).

Table 1. Recommended temperatures for broilers

Period	Temperature
1st day	32-34 °C
1st week	30 °C
2nd week	26 °C
3rd week	22 °C
4th week and above	20 °C

Under high environmental temperature, birds change their behavioural and physiological responses to maintain their body temperature through seeking thermoregulation. Birds subjected to heat stress conditions spend less time in feeding, more in drinking, panting, and wings elevation, move towards cooler surfaces (Mack *et al.*, 2013).

Modification of surrounding environment, ventilation system, bird density and nutritional management can be used to reduce the heat stress in poultry (Dayyani and Bakhtiyari, 2013). The surrounding environment can be controlled by using various things such as fans, fogger with fan, cooling pads, curtain, static pressure controllers and thermostats. Air movement inside the house is important for efficient ventilation (Sinha *et al.*, 2018).

Mechanical ventilation systems and air movement is produced by fans and exhaust fan in the building. Good ventilation system is essential for heat stress management. It removes the moisture loaded air from the poultry house and enters equal amount of fresh air from outside. Ventilation system should be maximized as the air movement assist removal of ammonia, moisture and carbon dioxide from the poultry house and enter fresh oxygen from outside (Butcher and Miles, 2012).

Widyaningrum and Pramudita (2017) did a smart home using an automatic lamp and fan control based on microcontroller. An Arduino Mega 2560 microcontroller was used to sense the physical conditions through connected sensors. The light turns on when the room is dark and the fan turns on when the temperature is above 25 ^oC. Likewise, the light turns off when the room is bright and the fan turns off when the temperature is below 25 ^oC.

Poonam and Gupta (2015) did a paper that presents the automatic control of home appliances including room light and fan controller using microcontroller AT89C51. The system is a reliable circuit that takes over the task of controlling the room fan and room lights as well as counting number of persons/visitors in the room very accurately. The counter increments by one whenever somebody enters the room and the light and fan are switched ON. The counter decrements by one whenever somebody leaves the room. Whenever the value of the counter is zero, then the light and fan is switched off. The automatic system is aimed at energy conservation.

The above systems do not give the user the ability to change the temperature and lighting thresholds of the systems. This is particularly important since different ages of birds may require different duration of light and degrees of temperature.

II. PROPOSED SYSTEMS ARCHITECTURE

The proposed system will involve the construction of a 3-way extension socket. The first socket will control the poultry house lights and will be user-programmable to turn on or off at predefined periods of time with the help of a real time clock (RTC). A light sensor will inform the user of the brightness of the poultry house. The second and third sockets will control a fan and a heater respectively and will operate based on the temperature of the poultry house as sensed by a temperature sensor. The heater will turn on when the ambient temperature is too cold for the birds, and turn off when the ambient temperature is not too cold. The exhaust fans will turn on to move hot air out when the temperature of the poultry house is too high and turn off when the temperature is not too high.

Wireless (WiFi) and Global System for Mobile (GSM) modules will be included in the system to allow users to conveniently monitor and control the poultry house appliances, the solar panels and battery bank voltages. The users will have access to the system through their phones or PCs. Information about the states of the system components will be visible on a web browser while system alerts will be sent to the user's and technician's phones via SMS. Moreover, the user will be able to monitor and control the appliances via phone call and SMS sent to the system. In order to reduce the cost of operating the system wirelessly, access to the system's web-based interface will be through WiFi when the user is around the farm. The control will be through GSM and GPRS when the user is far from the farm. The web server will allow the microcontroller to periodically log information about the system components and appliances on the internet using a web hosting service. This information will be available on the users' and technicians' web browsers over the internet.

The user will be able to set the thresholds for the duration of light, degree of temperature and level of voltage monitoring and controlling sub-units using the web-based interface. The microcontroller will monitor the solar panels and battery bank voltages and then alert the user and technician when an under-voltage or over-voltage is detected for troubleshooting.

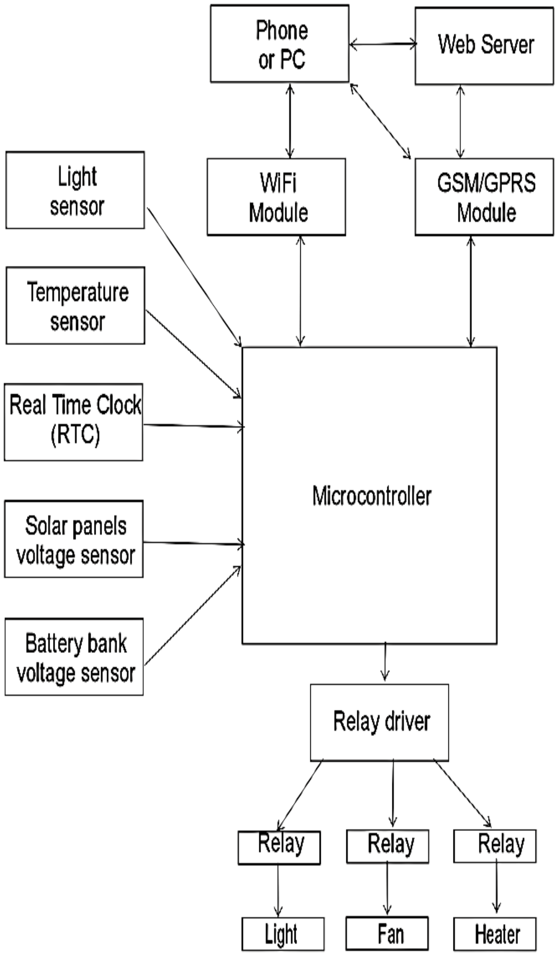


Figure 1: Block Diagram of the Proposed System

USING THE LIGHT SENSOR AND REAL TIME CLOCK (RTC) TO MONITOR AND CONTROL THE LIGHT

In the proposed automatic light subsystem, the microcontroller will use the data from the light sensor to inform the user about the brightness of the poultry house, and use the real time clock to turn on or turn off the light. The brightness of the poultry house will be measured using lux. Lux is the SI derived unit of luminance and luminous emittance. A family living room light is about 50 lux (Alan, 1998). The light turns on when the label of any of the checked checkboxes on the web page is the same as the current hour on the real time clock. The light turns off when the label of any of the unchecked checkboxes on the web page is the same as the current hour on the real time clock. This feature will be used to set the periods of time in the day when the lights of the poultry house must be on or off.

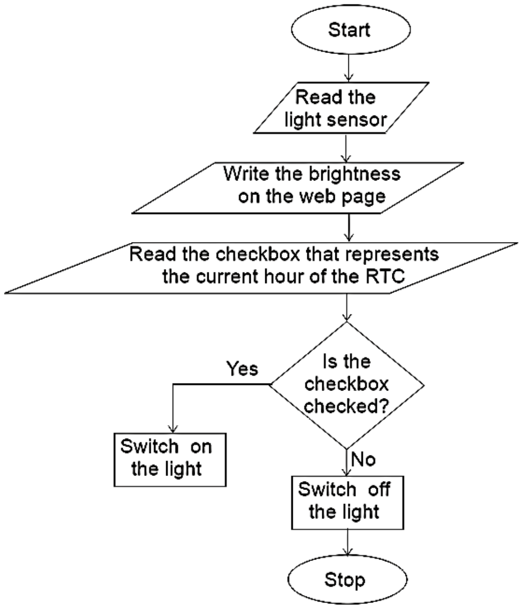


Figure 2: Programme Flow Chart for Monitoring and Controlling the Light

USING THE TEMPERATURE SENSOR TO MONITOR AND CONTROL THE FAN AND HEATER

In the proposed automatic temperature subsystem, the input will be obtained from the temperature sensor. The microcontroller will use the data from the sensor to turn on or turn off the fan or the heater. The heating element of the poultry house will be turned on when the temperature of the poultry house is below the default value of 25 ^{o}C and it will be turned off when the temperature is 25 ^{o}C and above. When the temperature is above 25 ^{o}C, the exhaust fans will be turned on to blow out the warm and moist air in the poultry house.

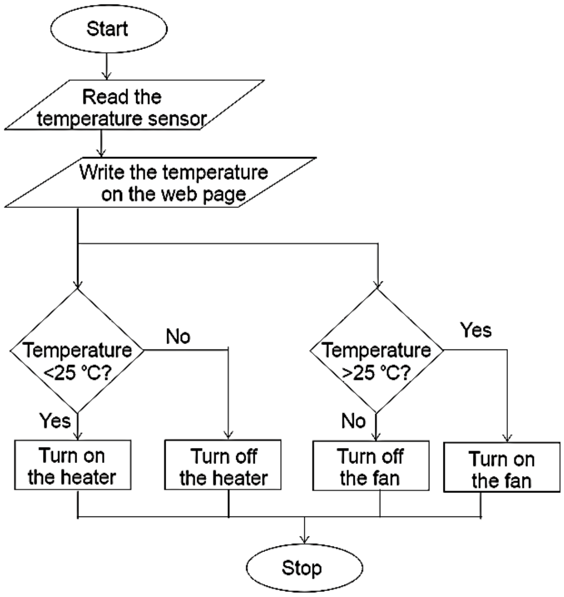


Figure 3: Programme Flow Chart for Monitoring and Controlling Temperature

a. MAINTENANCE ALERT

The microcontroller will alert the user and technician through SMS when an abnormally low or high voltage is sensed from the solar panels or battery bank. The normal solar panels' output voltage for a 12 volts solar system is within the range of 17 to 22 volts (Amy, 2016). Thus 15 and 24 volts will be considered as being too low and too high respectively. "SOLAR PANELS' VOLTAGE IS TOO LOW" will be sent to the user/technician via SMS when the solar panels' voltage falls below 15 volts and "SOLAR PANELS' VOLTAGE IS TOO HIGH" message will be sent as SMS when the voltage of the panels rises above 24 volts. The charging voltage of the 12 volts battery is 14 volts and for longevity, the battery should not be discharged below 12 volts (PowerSafe, 2014). Thus the system will alert the user and the technician when the battery bank's voltage falls below 12 volts or rises above 15 volts. The SMS to be sent to the user/technician when the battery bank's voltage is too low or too high is "BATTERY BANK VOLTAGE IS TOO LOW" or "BATTERY BANK VOLTAGE IS TOO HIGH" respectively.

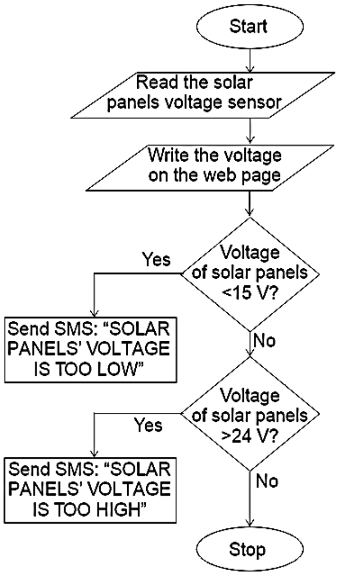

Start
Read the solar panels voltage sensor
Write the voltage on the web page
Voltage of solar panels <15 V?
Yes
Send SMS: "SOLAR PANELS' VOLTAGE IS TOO LOW"
No
Voltage of solar panels >24 V?
Yes
Send SMS: "SOLAR PANELS' VOLTAGE IS TOO HIGH"
No
Stop

Figure 4: Programme Flow Chart for Sending Maintenance Messages

USER INTERFACE DESIGN

The system will be operated via a web-based graphical user interface that allows the user and technicians to wirelessly gain access to the monitoring and control of the poultry house appliances. The web-based application is accessed over a network connection using hypertext transfer protocol (HTTP), rather than existing within the user's device's memory. The application runs inside a web browser. The choice of a web-based user interface is because every WiFi enabled phone or PC comes with at least one web browser. This allows the user to connect to the system without the need to install any additional software on his WiFi enabled phone or PC. Apart from the web-based interface, the hardware control unit consisting of the 3 sockets and the real time clock is another interface that allows the user to see and set the time of the real time clock. The third interface is the call and SMS interface which the user can use to monitor and control the poultry house appliances when he is outside the WiFi range. The system will also use this interface to send an over-voltage or under-voltage alert to the user. Therefore, the user interface design is done in three parts: the physical user interface, the web-based user interface and the call and SMS interface.

(1) The physical user interface

In the proposed system, the user interacts with the hardware interface by pressing the 2 buttons on the control panel to set the time of the real time clock. He receives feedback from the system through the 7-segment displays that shows the current time. The user of the system is expected to use the control panel only once in a while but is expected to see the current time from a distance, thus the use of large 7-segment displays.

The input devices used to accomplish the physical user interface are 2 push buttons named Hour Setting Button and Minute Setting Button. They are used to set the real time clock.

The output devices are:

i. 3 sockets named Light Socket, Fan Socket and Heater Socket. The black dot on each socket is a LED that shines when the socket is on.

ii. 4 7-segment displays: 2 are named Hour Digits and the other 2 are named Minute Digits.

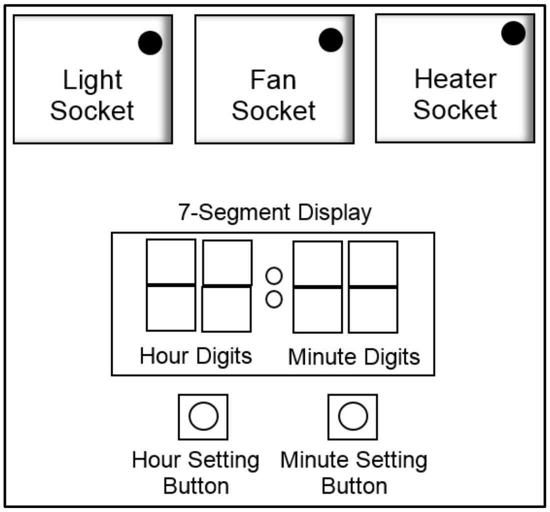


Figure 6: Physical User Interface of the Proposed System

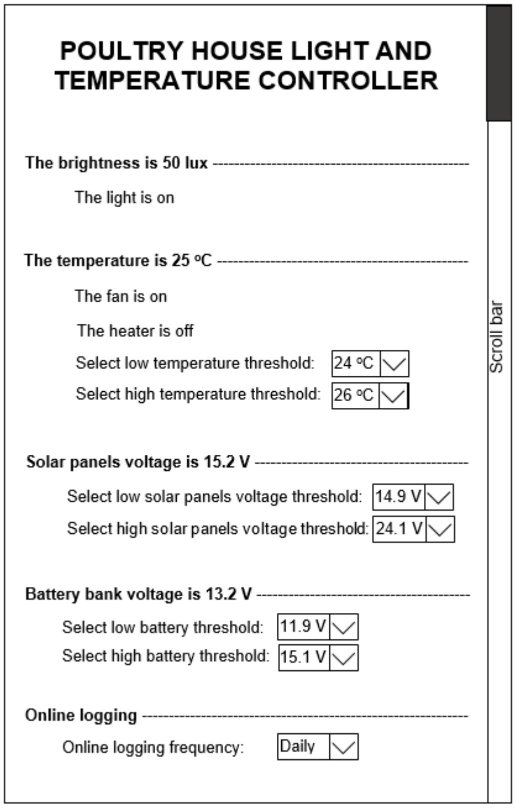
(2) The web-based user interface

This is another user interface that allows the user to easily and comfortably access the control of the poultry house appliances wirelessly. The components of the web-based interface will be textboxes, checkboxes, list boxes and the submit button.

Table 2: Components of the web-based user interface

Component	Data Type	Function
Textboxes	String	Used to display: i. heading of the web-page ii. settings confirmation information iii. system components and appliances information iv. voltage protection information v. WiFi SSID/Network Name vi. WiFi password vii. Users' phone numbers
Checkboxes	Boolean	Used to select/deselect hours of the day
List boxes	String	Used to select threshold values for the sensors: temperature, solar panels voltage and battery bank voltage Used to select online logging frequency
Button	GET request	Used to submit the user's settings to the microcontroller

A WiFi hotspot is an 802.11 b/g/n WiFi protocol (Espressif Systems, 2018), which can be received by WiFi enabled phones and PCs within the range of 70 meters. The programme displayed on the user's web browser is a html page stored in the microcontroller. As soon as the user connects to the system using its Service Set Identifier (SSID), password and IP address, the html page is transmitted to the user's browser. The design of the html pages to be displayed on the user's browser is shown below.



POULTRY HOUSE LIGHT AND TEMPERATURE CONTROLLER

The brightness is 50 lux

The light is on

The temperature is 25 ºC

The fan is on

The heater is off

Select low temperature threshold: 24 ºC

Select high temperature threshold: 26 ºC

Solar panels voltage is 15.2 V

Select low solar panels voltage threshold: 14.9 V

Select high solar panels voltage threshold: 24.1 V

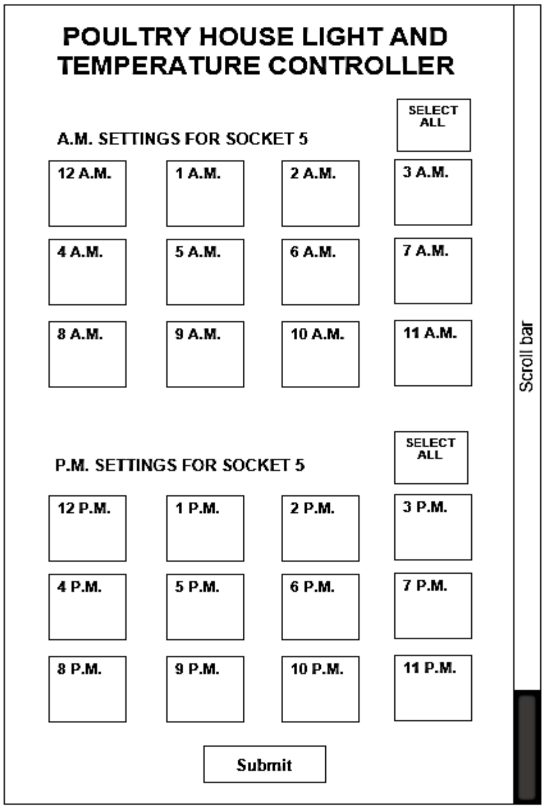
Battery bank voltage is 13.2 V

Select low battery threshold: 11.9 V

Select high battery threshold: 15.1 V

Online logging

Online logging frequency: Daily



POULTRY HOUSE LIGHT AND TEMPERATURE CONTROLLER

A.M. SETTINGS FOR SOCKET 5

SELECT ALL

12 A.M. 1 A.M. 2 A.M. 3 A.M.
4 A.M. 5 A.M. 6 A.M. 7 A.M.
8 A.M. 9 A.M. 10 A.M. 11 A.M.

P.M. SETTINGS FOR SOCKET 5

SELECT ALL

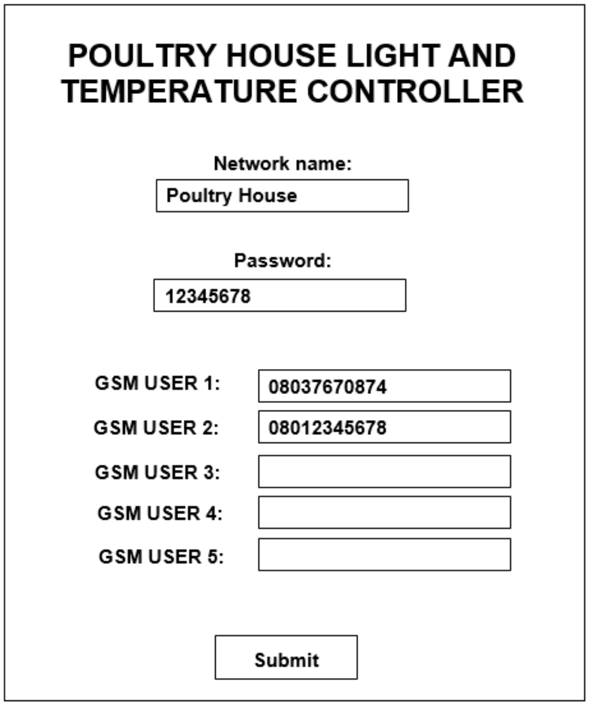
12 P.M. 1 P.M. 2 P.M. 3 P.M.
4 P.M. 5 P.M. 6 P.M. 7 P.M.
8 P.M. 9 P.M. 10 P.M. 11 P.M.

Figure 7: Web-Based Graphical User Interface

In the left-hand image above, the user is provided with the brightness of the poultry house. The user is shown the temperature of the poultry house and allowed to select the minimum temperature at which the heater turns on and the maximum temperature at which the exhaust fan turns on. The user is shown the voltages of the solar panels and battery bank and allowed to select the minimum and maximum voltages at which the user is alerted of an under-voltage or over-voltage. The web page also shows the online logging frequency and allows the user to change the time interval at which the microcontroller sends information about the system components and appliances to the internet.

The right-hand image shows the settings for the RTC-controlled socket meant for the lights of the poultry house. The boxes labelled 12 A.M. through 11 P.M. represent checkboxes that represent the 24 hours of the day. The 2 boxes labelled SELECT ALL represent checkboxes which will be used to check/uncheck all the 12 adjacent A.M. or P.M. checkboxes at the same time. A checkbox that is checked indicates that the appliance is expected to be on at the hour represented by the checkbox, while an unchecked checkbox indicates that the appliance will be off at the hour represented by the checkbox. The user can click the checkboxes to check or uncheck them and then click the submit button at the end of the page to submit the page. On receiving the user's settings, the microcontroller will store them and then transmit a confirmation page back to the user with a message that reads SETTINGS UPDATED SUCCESSFULLY. The confirmation page sent to the user will still contain the checkboxes whose checked/unchecked states represent the states of the appliances, thus allowing the user to make further changes to his settings if he wishes.

The ADVANCED SETTINGS page below will be used to display and change the WiFi security settings and GSM users of the system. The page will be displayed by long pressing the Minute Setting Button on the physical user interface while launching, submitting or reloading the web page. This is a security measure that ensures that the advanced settings can only be changed by someone with access to both the physical user interface and web-based user interface. When the page is loaded, the user will see the network name/Service Set Identifier (SSID) and password and be able to modify them. The page will also display the list of registered GSM users and allow the user to add or remove them.



POULTRY HOUSE LIGHT AND TEMPERATURE CONTROLLER

Network name: Poultry House

Password: 12345678

GSM USER 1: 08037670874

GSM USER 2: 08012345678

GSM USER 3:

GSM USER 4:

GSM USER 5:

Figure 8: Web-Based Graphical User Interface continued

(3) The call and SMS interface

The system will use this interface to alert the user whenever some maintenance is required. When the solar panels or battery bank voltage is too high or too low, the microcontroller alerts the user. This interface will also be used by the user or technician to monitor and control the poultry house appliances. When the phone number of the system is called by a registered user, the system replies with an SMS containing the states of the 3 appliances. Likewise, when the user sends an SMS to turn on or turn off any of the appliances, the system turns the appliance(s) on or off and sends back a confirmation SMS showing the states of all the appliances. Below is the list of messages that can be sent by the system and the events that trigger the messages.

Table 3: Call and sms control messages

	Message sent by the system	**Cause of the message**
1.	SOLAR PANELS' VOLTAGE IS TOO LOW	When the voltage of the solar panels is too low
2.	SOLAR PANELS' VOLTAGE IS TOO HIGH	When the voltage of the solar panels is too high
3.	BATTERY BANK VOLTAGE IS TOO LOW	When the voltage of the battery bank is too low
4.	BATTERY BANK VOLTAGE IS TOO HIGH	When the voltage of the battery bank is too high
5.	LIGHT IS ON/OFF FAN IS ON/OFF HEATER IS ON/OFF	When a registered user dials the phone number of the system
6.	LIGHT IS ON/OFF FAN IS ON/OFF HEATER IS ON/OFF	When a registered user sends an SMS to turn on/off any of the 3 appliances. Example of SMS: LIGHT=ON, FAN=OFF, HEATER=OFF
7.	UNRECOGNIZED KEYWORD	When a registered user sent an unrecognized message

b. CONCLUSION

The reviewed literature has determined that adequate lighting and temperature control are very important in poultry production. Thus the present design will adopt the use of a microcontroller whose function is to monitor and control the lighting and temperature of a poultry house in order to increase poultry production. The use of solar energy is hereby recommended and the design for monitoring the solar panels, batteries and appliances is hereby provided to protect the controller system and the poultry house appliances from damages that could be caused by under-voltage or over-voltage.

The design includes a WiFi and a GSM module which will afford the user the convenience and flexibility of monitoring and operating the system components and appliances. The modules will allow technical data about the solar panels and batteries to be tracked and monitored by the users and technicians from remote locations. The system will alert the technicians whenever maintenance is required. Moreover, the users and technicians will be able to log into the web-based control interface to obtain information concerning the components of the system, system faults and other vital data. Users will be able to register their phone numbers on the system. When the system is called by dialling its phone number, the states of the appliances will be sent as SMS to the caller provided he or she is a registered user. Registered users can also send SMS to the system to remotely turn on or off any of the connected appliances. This solar-based remote monitoring and controlling technology will allow potential maintenance problems to be addressed quickly and could increase the system's durability while increasing poultry production and reducing climate change.

c. REFERENCES

Ajay, P. (2016). Significance of Light in Poultry Production: A Review. *Advances in Life Sciences 5(4), 1154-1160.*

Alan, P. (1998). Strategic Study of Household Energy and Greenhouse Issues. Brighton, England: Sustainable Solutions Pty Ltd.

Amy, B. (2016). How to read the solar panel specifications. Boxborough, US: Alternative Energy Store Inc. Retrieved 14/07/2019 from https://www.altestore.com/blog/2016/04/how-do-i-read-specifications-of-my-solar-panel/#.XUhBw9h7mcw

Butcher, G.D. and Miles, R. (2012). Heat stress management in broilers. VM65 series of the Veterinary Medicine Large Animal Clinical Sciences Department, Florida Cooperative Extension Service, Institute of Food and Agricultural Sciences, University of Florida

Dayyani, N. and Bakhtiyari, H. (2013). Heat stress in poultry: background and affective factors. *International journal of Advanced Biological and Biomedical Research.* 1(11): 1409-1413

Espressif Systems (2018). *ESP8266EX Datasheet: Version 6.0.* Espressif IOT Team, Espressif Inc.

Jan, H. (2019). *Climate in poultry houses.* Netherlands: Poultry Hub. Retrieved 05/07/2019 from http://www.poultryhub.org/production/husba ndry-management/housing-environment/climate-in-poultry-houses/

Karen, S., and Hank, C. (2010). *Lighting for Broilers*. USA: ROSS TECH.

Mack, L.A., Felver-Grant J.N., Dennis, R.L. and Cheng, H.W. (2013). Genetic variation aiter production and behavioral responses following heat stress in 2 strains of laying hens. *Poult. Sci.*, 92: 285-294.

Poonam, L. and Gupta, R.P. *(2015*). Microcontroller Based Automatic Control Home Appliance*s. International Journal of Innovative Research in Science, Engineering and Technology Volume 4.*

PowerSafe (2003). *Safety, Storage, Installation, Operation & Maintenance Manual: Flooded Lead-Acid Batteries*. USA: EnerSys Inc.

Simon, M.S. (2003). Reducing heat stress problems. *World Poult 19*(3):16-17.

Sinha R., Kamboj, M.L., Lathwal, S.S., and Ranjan, A. (2018). Effect of housing management on production performance of crossbred cows during hot-humid season. *Indian J. Anim. Res.*, 52 (7): 1091-1094.

Sinha, R., Lone, S.A., Ranjan, A., Rahim, A., Devi, I and Tiwari, S. (2017). The impact of climate change on livestock production and reproduction: ameliorative management. *International Journal of Livestock Research.*, 7(6): 1-8.

Sottnik, J., (2002). *Climatical factors and their effect on production in animal housing*. ASAE Annual International Meeting/ CIGR XVth World Congress; Chicago, Illinois, USA: ASAE editors.

Sritharet. N., Hara, H., Yoshida, Y., Hanzawa, K., and Watanabe S. (2002). Effect of heat stress on histological features on pituicytes and hepatocytes, and enzyme activities of liver and blood plasma in Japanese quail (Coturnix japonica). *J Poult Sci; 39*:167-178.

St-Pierre, N.R., Cobanov, B., and Schnitkey, G.X. (2003). Economic Losses from Heat Stress by US Livestock Industries. *J Dairy Sci; 86*(E):E52-E77.

Widyaningrum, V.T., Pramudita, Y.D (2017).. Automatic Lamp and Fan Control Based on Microcontroller. Presented at the *2nd International Joint Conference on Science and Technology (IJCST).*

Predictive Analysis of Human Activities Using Supervised Learning

Cecilia Ajowho Adenusi, Olufunke Rebecca Vincent and Abiodun Folurera Ajayi
Department of Computer Science
Federal University of Agriculture
Abeokuta, Ogun State.
ceceresearch@gmail.com, vincent.rebecca@gmail.com, folureraajayi@gmail.com

ABSTRACT - Human activities are what every individual engaged in daily, weekly, monthly and annually. The art of typing in this paper contents through the use of a laptop is an activity, so every one of us engaged in a minimum of four to five activities per day according to the result of this research work. The activities dataset that was used for this paper are seventy-five in numbers, some are; jumping, sitting, lecturing, working on the computer, ascending and descending of stairs, working on the computer, use of a bathroom, use of social media and the likes and they were analysed using exploratory data analysis, but not all were used at prediction stage. Those with high responses were used to predict using Random Forest, the activities that such may likely engage in in the coming week and the result was shown in probabilities and percentage mode. Therefore from the result gotten through the use of Weka software, which was used to carry out a prediction on the analysed data, average human engaged in more than four activities daily. The results were further compared on five diverse models namely; Decision Table, Linear regression, Random Tree, Random Forest and ZeroR classifier models.

Keywords- Activities, Human, Questionnaire, Social Media, Recognition, Supervised.

I. INTRODUCTION

The human activity analysis is the key to technology for human activity recognition (HAR) and the essence is to select an appropriate sensor that captures user’s activity and deploy it through a monitor. The two categories involve sensor-based HAR and video-based HAR. The one with the use of video cameras for monitoring human body activities is video-based HAR while the one based on time series data collected with a sensor like the wearable sensor is sensor-based HAR. This system can be used in health or behavioural monitoring, smart home and so on and are usually time series, such as, normalization, smoothing and separate gravity component from acceleration data, extraction of feature data, classify with the use of classification algorithm which requires manual means of recognizing the human activity by using the tradition method of machine learning. The extracted features include; mean, variance, minimum, maximum, media and the likes while domain features include discrete cosine and fast Fourier transform coefficient. Once the feature extracted is too high Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA) are often used in order to reduce the dimension to make it more robust. The method used for classification of the artificial feature extracted vector; naive Bayes (Long X et al., 2009), Hidden Markov model (Lester J et al., 2005), decision tree (Fan L et al., 2013), K Nearest Neighbour, KNN) (Preece S J et al., 2009) are being implemented. This method learns, omits and extract complicated steps in tradition HAR research, which makes it superior to manual ways of feature extraction. For the analysis of complex activities from time series, a convolutional neural network is often used. (Hen Y et al., 2015) often study the feature engineering workload through automatic learning and continuous extracting features, due to the varieties in human activities, the input sensor sequence data is becoming a challenging task. At the moment, both automatic HAR and manual feature extraction method use technology for breaking down sensor signal to time periods that is much smaller, the techniques involved here are event and behaviour definition and fixed sliding window which is what most of the researcher adopted for dividing signal to fixed window length, which is labelled to reduced accuracy of HAR. The problem from the type of method is known as multi-classes window problem (Yao R et al., 2017) and is a common problem on time series in HAR and has a great impact on human activity sequences. The way to improve recognition accuracy is to predict human activities in the real-time, which is the need for this paper. And the methods that will be used are classification methods for human activities analysis and Random Forest for the feature extraction of variables needed and for prediction, the datasets are to be tested using exploratory data analysis method and the rest of the paper provide related work, analysis and algorithm description and assessment of result and conclusion is being given.

II. RELATED WORKS

The manual feature extraction from time series and classification of it through the use of classification algorithm are the early work done on HAR. But the discriminative and generative model (Ng A Y et al., 2002) is the current algorithms for HAR. The discriminative model includes KNN, ANN, DT and SVM (He Z Y et al., 2008) extract coefficient of data accelerometer as the activity recognition characteristics for classifying human activity. The average result of SVM result in jumping, running, walking, and standing is 92.25%. (Fan L, 2013) uses five activities with a built-in accelerometer on the DT algorithm for constructing an independent activity model recognition. (Khan A M et al, 2010) structure a hierarchical scheme for the autoregressive model to generate a feature vector through LDA analysis and ANNs, the accuracy average of 15 diverse activities is 97.9%. (Preece S J et al., 2009) for the classification and analysis of daily activities based on the acceleration time series adopted individual cross-validation method with the use of the nearest neighbor classifier, the accuracy average is 95%.

Human activity recognition (HAR) review aimed at recognizing human activities from a sequence of observations on environmental conditions and actions of subjects. Discriminative and generative models are classification categorization template-based methods, the existing works of literature were classified in detailed taxonomy which include representation and classification and as well as datasets being used. The problems encountered are (i) complex and various backgrounds in applications like video surveillance, (ii) multi-subject interactions and group activities like waving hands, jumping and running, a good example here is in a football game. (iii) Finally is in long-distance and low-quality videos, present dataset in which the target person is obvious and clear cannot be compared to surveillance cameras installed in high places and the subject is rather small due to long-distance of cameras which then make it difficult to analyse activities of the torso (Efros et al., 2003).

There are common characteristics in the review between the sensor or wearable computing-based system (Chen et al., 2012) and the human activity based on a vision that does extract details from video cameras with the aid of computer vision techniques. The focus here is considering visual data in visualizing human activity (Zhang et al., 2013), and the notion is being observed by limiting it to a camera field and view that is based on sensor system which often depends on the availability of sensors such as radio frequency identification (RFID), smartphones or wearable sensors, this is being added to observation in order to get person’s movement represented (Turaga et al, 2008). Techniques performed in this study are a supervised method which is not a solution for the long term for an autonomous robot that should have minimal supervision after implementing it in the real world. To solve the problem, objects and humans needed to be observed and estimate poses in a qualitative spatial representation.

The distinction in human activity analysis using vision-based, which was extracted from video cameras through wearable or sensor computer vision techniques. This technique relays on the availability of wearable sensors, smartphones or small sensors that can be attached to human activities for observation on the human movements but with the use of a single camera’s field of view for visual data. These techniques perform supervised learning, that is, the training data requires ground-truth label with manual segmentation, of which, is not a feasible solution for mobile robots deployed in the real-world with minimal supervision. For such a task, it is best to used unsupervised learning techniques because it does not require manual annotation or time-consuming. The existing study used probabilistic LSA (Hofmann, 2001) and LDA (Blei et al., 2003), Latent Semantic Analysis (Deerwester et al., 1990) on an unsupervised setting for human activity categories.

A. RELATED MODELS

The related activity model learn from the existing approaches on human activities using machine learning and data mining techniques are tabulated below:

Table 1.1: Summary of reviewed related models

SN	Model name	Strength	Weaknesses
1	probabilistic models (Martinez et al., 2009)	Quantifies the integration of data in order to capture distribution over state transmission in samples of a batch run.	analytics for likelihood occurrences instead of the actual data in the events search and data points
2	Statistical Analysis. Dodge, 2006; Romijn, 2014)	Very fast to compute, doesn’t require building models	Independent of the model
3	K-nearest neighbor. Campilho, 2009	classifiers are memory-based	do not require a model to fit
4	Decision Tree. (Bao and Intille, 2004)	perform classification without requiring much computation	prone to errors in classification problems with many class
5	Support vector machines (SVM) Nantasement et al, 2013b	Model non-linear decision boundaries	Are memory intensive?
6	Deep	Good classifying for	It requires more

	Learning. Nasrabadi, 2007	audio, image data and text	expertise to train.
7	Random Forest	It handles well high-dimensional spaces	Lack of interpretation
8	Linear Regression	Easy to explain and understand	Poorly performs when there is non-linear relationship
9	K-Means	Its fast, surprisingly and simple clustering algorithm	Numbers of clustering must be specify by the user
10	DBScan	The performance is scalable	The hyper parameters must be tune by the user
11	Naïve-Bayes-Nearest-Neighbor (NBNN) Depth. Yang and Tian [2012]	Model performs surprisingly well in practice	Model are often birth by model, due to sheer simplicity.
12	Hierarchical Clustering	Clustering not to be spherical in shape	The level of hierarchy needed to be chosen by user
13	Logistic Regression	Regularization of the algorithm to avoid over fitting	With multiple or non-linear decision boundaries it tends to underperform
14	Affinity propagation	There is no need to specify number of clusters	Difficult to scale
15	Fuzzy Logic	Fuzzy rough set hybrid model, it can deal hybrid decision systems (datasets).The weakness is required high computation time.	Its required high computation time.

The above table entails the summary of the reviewed models carried out in the paper, including, the strengths of each model with their respective weaknesses.

Human activities, multimodal data, and gesture are recognition method that deals with modalities that process 2D or depth data as 3D. Stacking frames into volumes at the first stage with convolutional layers (D. Wu et al., 2016) but the 3D positional joints are fed into the diverse networks when there is availability or presence of additional pose data (D. Luvizon, 2018). Raw video and fusing pose modality is done as early or late fusion through fusion layers while pre-process pose is often reported in some situations to have improved performances. This new method, leverage on regularity during training and does not require pose when testing. Making the recurrent model local is the recent trend in contemporary work on activity recognition through the means of employing a recurrent neural network. The separation of a memory cell of LSTM network into sub-cells and individual representations of each part of the network long term was done by Partaware LSTMS (A. Shahroudy et al., 2016) anatomical hierarchy (Du et al., 2015) by making use of bi-directional LSTM to fit it in, the sub-networks focused on the anatomic parts such as; torso, legs, and arms that the skeletons are divided into. In this new model, the creation of a manual graph is not required neither does it depend on the external hacker, for the tasks between trackers nor objects are automatically learned.

Wearable sensor data of human activities recognition has been an attractive topic in research, due to the fact that it's applicable in a smart environment and health care sector. The recommendations and limitations of human activity recognition using a mobile phone to online activity recognition are well described by (Shoaib et al., 2015). The word online means an evaluation of classification pipeline (classifying and describing signal) on a mobile phone, but it does not consider convolution neural network methods which are the most suitable employed methods (Sojeong Ha and Seungjin Choi, 2016). Meanwhile, a comprehensive study on these methods in the light of wearable sensors was done by (Wang et al., 2017). Numbers of stacked autoencoders, deep learning methods and recurrent neural networks (RNN) were surveyed by the authors. The influence of devices that are heterogeneous on the last performance of classifiers on activity recognition. For this reason, the activities were represented by using common classifiers like support vector machine, random forest, and nearest neighbor and handcrafted features.

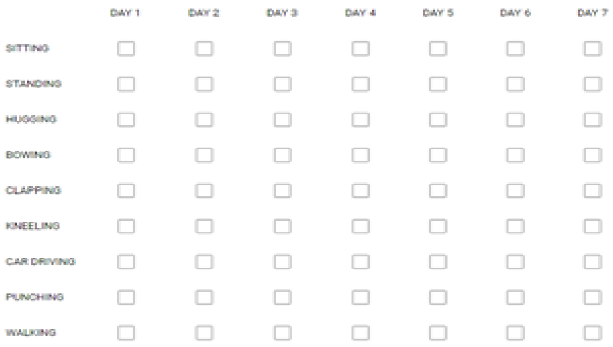
The applications of human activities recognition have covered many fields, such as security and surveillance, entertainment, health and intelligent. Environments in recent years. Diverse approaches such as object tagged, wearable device free have been used by researchers to identify human activities. This review present the research work over the period of 2010 to 2018 and the main focus is on device-free approach. This approach tagged the environment with devices to capture the human activities and the new taxonomy for categorizing and divide the review literature are motion-based, interaction and action-based. The focus here is only on one type of activity with the comparison of research works in diverse sub-areas. A survey of research work using device-free radioactivity recognition presented (Scholz et al.,2011), this survey categorizes the previous work on device-free radio activity recognition (DFAR) such that, the review is further divided as machine learning, threshold, and statistical modelling based DFAR and device-free localization(DFL) such that researchers provide details of diverse topics as spatial coverage, radio tomographic, presence detection and statistical modelling.

III. RESEARCH METHODOLOGY

This encapsulates the method used for collection of data, analysis, data manipulation, dataset used, model to be implemented and algorithm with the scientific method to be used.

A. Data Collection

Data is being collected from an individual via the use of a questionnaire, which was the method designed for this paper for data collection. The questionnaire was designed through the use of a google survey form and the sample of it is in figure 1.1. URL link was generated on google page and it was further shortened, so as to have a short URL. This URL link to the questionnaire webpage was then distributed to an individual via the use of social media, precisely, WhatsApp. Each individual clicks on the link to fill the questionnaire by just clicking the needed responses and their results were collation online and were downloaded in .csv format for further analysis.



	DAY 1	DAY 2	DAY 3	DAY 4	DAY 5	DAY 6	DAY 7
SITTING	□	□	□	□	□	□	□
STANDING	□	□	□	□	□	□	□
HUGGING	□	□	□	□	□	□	□
BOWING	□	□	□	□	□	□	□
CLAPPING	□	□	□	□	□	□	□
KNEELING	□	□	□	□	□	□	□
CAR DRIVING	□	□	□	□	□	□	□
PUNCHING	□	□	□	□	□	□	□
WALKING	□	□	□	□	□	□	□

Fig 1.1. Day-To-Day Human Activities Analysed

B. DATA MANIPULATION

The downloaded data collated result was furthered processed and needed features were extracted from it to be able to perform the analysis. Features like bio data were totally removed and there were 71 datasets in all but only 20 were extracted to be used for the prediction analysis. The twenty datasets that were selected were the likely needed ones that will give more accurate results than the other and also in order to save time and go ahead with the prediction, which is the goal of this paper.

C. DATASET TO BE USED

The data was manipulated and only 20 datasets will be used for the predictive method. This dataset features were further extracted and redesigned on an excel workshop and converted to .csv format for further processes. Therefore 20 day-to-day human activities were used as a sample for all the activities that humans may engage in, and seven instances will also be considered to run through the datasets.

D. TECHNIQUES ADOPTED

The techniques to be adopted are supervised learning techniques that can infer function from a labeled training set of data. This technique is a machine learning that learns a function that can perform input-output pairs and the steps to be taken are:

STEP I: The type of data to be used as a training set

STEP2: The representation of the training set in real-world use of a function.

STEP3: Representative of the learning function for a feature input

STEP4: Structure of a learning algorithm in determining the structure of the learned function.

STEP5: Executing and running the completed design by specifying some control parameters

STEP6: Measuring the performing of the function on a test set and evaluate the performance of a learned function

Under this techniques is classification method which will be used to carry out the analysis and the methods that will be used to classify out a dataset is exploratory methods (EDA), this methods analyze the datasets, summarize the set characteristics, explore the data and the formulated hypothesis that may lead to another data collection and experiments. (John Tukey, 1961) promoted EDA to encourage statisticians by designing procedures for analyzing data, means of planning the gathering of data to make its analysis easier, techniques for interpreting results, and all machinery and results which apply to analyze data. Turkey promotes the use of a summary of five number which is; maximum and minimum, quartiles, median, mean and standard deviation. All these were designed to complement the analytic theory of testing hypotheses (Laplacian) emphasis on experimental families (Morgenthaler, 2000). The fig1.3 further explains this method and a model was formulated at the end

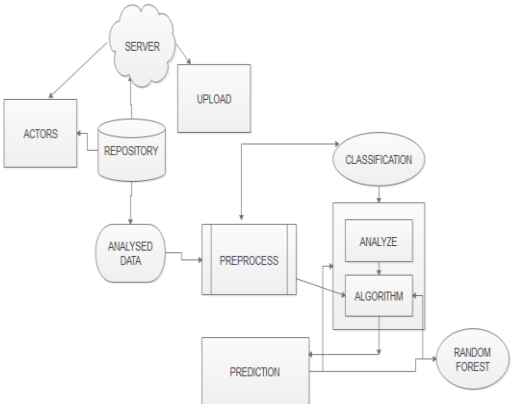


Fig 1.2 HUMAN ACTIVITIES MODEL

The above fig 1.2 entails the overall methods used in this paper and the steps are:
STEP I: Extraction and designing activities datasets
STEP II: Uploading it on Google server
STEP III: Each participant (actors) responding by checking through the dataset
STEP IV: Responses uploaded back to google server
STEP V: Download responses from google repository
STEP VI: Datasets analyses and process further
STEP VII: Using classification methods for analysis
STEP VIII: Using Random Forest for the prediction

The technique that will be used for prediction is Random Forest, which is a method used for classification, regression and outputting classification and also for prediction. Each has a decision trees' habit of training set over fitting. The training algorithm applies the techniques of bagging or bootstrap aggregating to tree learners.

Using a training set Z= z1, …,zn and responses P = p1, …, pn with Q times repeatedly bagging. The choose a sample randomly with replacement and fits trees to the sample, having:

For c = 1… Q

1. N training sets from Z, P; call Z_c, P_c, Sample, Replacement.
2. Train classification tree k_c on Z_c, P_c.

After this, predictions for samples unseen c', by averaging the predictions on all individual classification there on c'.

$$\mathrm{k} = \frac{1}{Q}\sum_{q=1}^{Q} kq(c')$$

This procedure decreases the model variance because it leads to better performance without increasing the bias, which means, while the prediction of a tree is sensitive to training set noise, the average of these trees is not in as much that the trees are not correlated. Therefore the standard deviation can be an estimate of the uncertainty of prediction from all individual classification tree on c'.

$$\sigma = \sqrt{\frac{\sum_{c=1}^{Q}(kc(z') - k))^2}{Q-1}}$$

The optimum number of trees Q can be found with cross-validation, the mean prediction error on ci, using with the trees that did not have ci in their tree.

Random Forest then used the modified learning algorithm tree that selects features at each spilled in the learning process by correlating the trees in a bootstrap sample. Once one or a few features are strong predictors for target output, such features will be selected in many of Q trees which make them be correlated.

```
{\displaystyle {\hat {f}}={\frac {1}{B}}\sum _{b=1}^{B}f_{b}(x')}
{\displaystyle \sigma ={\sqrt {\frac {\sum _{b=1}^{B}(f_{b}(x')-{\hat {f}})^{2}}{B-1}}}.}
```

III. RESULTS AND DISCUSSION

The implementation of the analysis was done using a java software tool called Weka with machine learning algorithm to solve the real life mining challenges on both classification algorithms and feature selection methods. The results are:

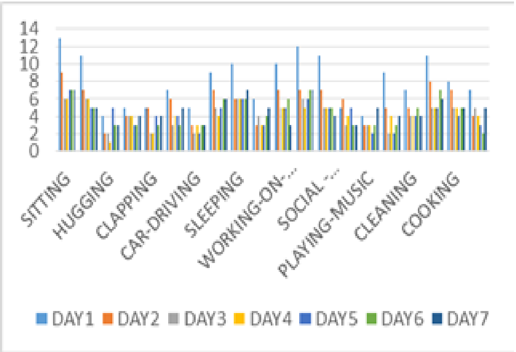


Fig 1.3: Seven Days Human Activities Analyzed

There were 71 activities dataset that was used for this paper and 20 out of the activities were used to plot the above graph, showing the analysis results of the used datasets. The color light-blue denotes day1, orange for day2, ash for day3, yellow for day4, blue for day5, light-green for day6 and naïve-blue for day7. From the result, the highest activities engaged in Day1 are Sitting, Praying, Social-media-usage, Standing and use-the-bath activities followed by day2 via day7.

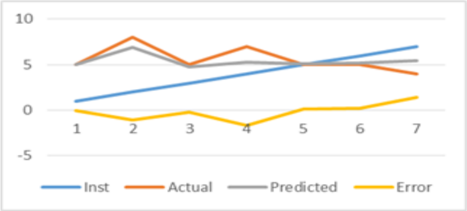
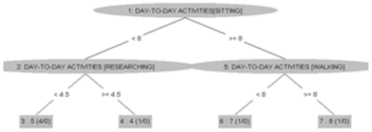


Fig. 1.4: Predicted Result Evaluated

This figure shows the results of predicted human activities value, using the previous 7days as a test data to predict the future and the error gotten from the predicted value was also included in the chart. The seven instances that was used are in blue colour followed by the actual values in orange and the predicted values with errors are in ash and yellow colours respectively. From the result, the prediction of human activities using supervised method of Random forest was successful with the given above result.


1: DAY-TO-DAY ACTIVITIES[SITTING]
< 8
>= 8
2: DAY-TO-DAY ACTIVITIES [RESEARCHING]
5: DAY-TO-DAY ACTIVITIES [WALKING]
< 4.5
>= 4.5
< 8
>= 8
3 : 5 (4/0)
4 : 4 (1/0)
6 : 7 (1/0)
7 : 8 (1/0)

This tree denotes how decision are being made, right from the first day-to-day activities down to the last one. The feature extraction used are sitting, researching and walking datasets.

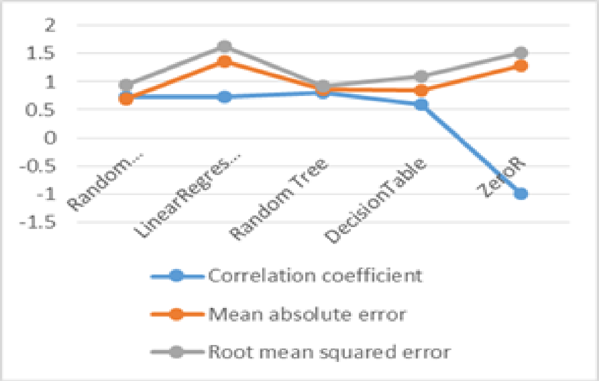


FIGURE 1.6 Predictive Classifier Models Of The Algorithms

The cross-validation measures were plot against the five classifiers models. The horizontal axis entails the other methods that were used to compare with the random forest initial method and the result is in the graph above. And the two best methods that can be used for this paper are the ones with lower root mean square and absolute error with are random forest and random tree. Linear regression is the one with the highest percentage absolute error while the random forest is the one with the lowest.

Table 1.2 RANDOM FOREST PREDICTIVE TRAINING SET RESULT

Inst	Actual	PREDICTED	ERROR
1	8	7.18	-0.82
2	7	6.35	-0.65
3	5	4.94	-0.06
4	5	4.94	-0.06
5	4	4.43	0.43
6	5	5	0
7	5	5.06	0.06

This table makes use of Random Forest method to predictive the next seven days human activities and their possibilities of occurrences were tabulated above on a training set category. The likely errors generated that may occur if the prediction exists were also included.

Categories of Selection	**Factors**	**Classifier Models**				
		Random Forest	Linear Regression	Random Tree	Decision Table	Zero R
Cross-Validation	Correlation coefficient	0.7268	0.7223	0.8076	0.5998	-1
	Mean absolute error	0.6829	1.3634	0.8571	0.8452	1.2857
	Root mean squared error	0.9475	1.624	0.9258	1.1024	1.5092
	Relative absolute error	53.1111%	106.0408 %	66.6667 %	65.7407 %	100 %
	Root relative squared error	62.7809%	107.6014%	61.3438%	73.0436%	100%
Test Split	Correlation coefficient	1	1	1	0	0
	Mean absolute error	1.055	1.2167	1.5	1.5	1.5
	Root mean squared error	1.1131	1.2828	1.5811	1.5811	1.5811
	Relative absolute error	70.3333%	81.113 %	100 %	100 %	100%
	Root relative squared error	70.4003%	81.1289%	100%	100%	100%
Training Set	Correlation coefficient	0.9969	1	1	0.9564	0
	Mean absolute error	0.2971	0	0	0.2857	1.102
	Root mean squared error	0.4294	0	0	0.378	1.2926
	Relative absolute error	26.963%	0 %	0 %	25.9259 %	100%
	Root relative squared error	33.1923%	0%	0%	29.2174%	100%

This table entails the categories of methods selected for the predictive analysis in order to be able to pick the best method or the one that is most appropriate for the problem to be solved and comparison was carried out on the classifier models with their tabulated results above.

V CONCLUSION

The day-to-day human activities are a vital aspect of every living creature on the surface of the earth, therefore, the need to look into such is very crucial. The results or responses gotten from here can be further processed in every area of human activities, be it in industry, or school, or banking and the likes. Seventy-One (71) dataset was the activities that was extracted out from day-t-day human activities and that was what was used for this paper. The datasets were the ones used to carry out the predictive analysis. Some of the activities are; running, jumping, researching, lecturing/teaching, waving, personal grooming, sitting, ascending and descending stairs, lecturing, use of a bathroom, use of public transport, typing, use of social media and the likes and the data collated for all these activities were done through the use of google form. The Featuring section was carried out on the analyzed human activities and the prediction was done using the Linear Regression Model.

But not all the activities collated were actually used using the analysis and predicted. The ones with high responses weekly were more considered than those with low responses.

VI REFERENCES

[1] Long, X., Yin, B., &Aarts, R. M. (2009, September). Single-accelerometer-based daily physical activity classification. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 6107-6110). IEEE.

[2] Lester, J., Choudhury, T., Kern, N., Borriello, G., & Hannaford, B.(2005). A hybrid discriminative/generative approach for modelling human activities.

[3] Fan, L., Wang, Z., & Wang, H. (2013, December). Human activity recognition model based on decision tree. In *2013 International Conference on Advanced Cloud and Big Data* (pp. 64-68). IEEE.

[4] Preece, S. J., Goulermas, J. Y., Kenney, L. P., Howard, D., Meijer, K., & Crompton, R. (2009). Activity identification using body-mounted sensors—a review of classification techniques. *Physiological measurement*, *30*(4), R1.

[5] Chen, C. H., Azari, D. P., Hu, Y. H., Lindstrom, M. J., Thelen, D., Yen, T. Y., &Radwin, R. G. (2015). The accuracy of conventional 2D video for quantifying upper limb kinematics in repetitive occupational tasks. *Ergonomics*, *58*(12), 2057-2066.

[6] Shahi, A., Deng, J. D., & Woodford, B. J. (2017, May). A streaming ensemble classifier with multi-class imbalance learning for activity recognition. In *2017 International Joint Conference on Neural Networks (IJCNN)* (pp. 3983-3990). IEEE.

[6] Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems* (pp. 841-848).

[7] Shiratori, T., &Hodgins, J. K. (2008, December). Accelerometer-based user interfaces for the control of a physically simulated character. In *ACM Transactions on Graphics (TOG)* (Vol. 27, No. 5, p. 123). ACM.

[8] He, Y., & Li, Y. (2013). Physical activity recognition utilizing the built-in kinematic sensors of a smartphone. *International Journal of Distributed Sensor Networks*, *9*(4), 481580.

[9] Lee, M. W., Khan, A. M., Kim, J. H., Cho, Y. S., & Kim, T. S. (2010, August). A single tri-axial accelerometer-based real-time personal life log system capable of activity classification and exercise information generation. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology* (pp. 1390-1393). IEEE.

[10] Efros, A. A., Berg, A. C., Mori, G., & Malik, J. (2003, October). Recognizing action at a distance. In *null* (p. 726). IEEE.

[11] Chen, L., Hoey, J., Nugent, C. D., Cook, D. J., & Yu, Z. (2012). Sensor-based activity recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(6), 790-808.

[12] Borges, P. V. K., Conci, N., &Cavallaro, A. (2013). Video-based human behavior understanding: A survey. *IEEE transactions on circuits and systems for video technology*, *23*(11), 1993-2008.

[13] Turaga, P., Chellappa, R., Subrahmanian, V. S., &Udrea, O. (2008). Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video technology*, *18*(11), 1473.

[14] Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, *42*(1-2), 177-196.

[15] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., &Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391-407.

[16] Ha, S., & Choi, S. (2016, July). Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors. In neural networks for multimodal gesture segmentation and recognition. *IEEE transactions on pattern analysis and machine intelligence*, *38*(8), 1583-1597.

[20] Luvizon, D. C., Picard, D., &Tabia, H. (2018). 2d/3d pose estimation and action recognition using multitask deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5137-5146).

[21] Shahroudy, A., Ng, T. T., Gong, Y., & Wang, G. (2017). Deep multimodal feature analysis for action recognition in rgb+ d videos. *IEEE transactions on pattern analysis and machine intelligence*, *40*(5), 1045-1058.

[22] Du, Y., Wang, W., & Wang, L. (2015). Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1110-1118).

[23] Scholz, M., Sigg, S., Schmidtke, H. R., &Beigl, M. (2011, December). Challenges for device-free radio-based activity recognition. In *Proceedings of the 3rd workshop on Context Systems, Design, Evaluation and Optimisation (CoSDEO 2011), in Conjunction with MobiQuitous* (Vol. 2011).

[24] Lara, O. D., & Labrador, M. A. (2012). A survey on human activity recognition using wearable sensors. *IEEE communications surveys & tutorials*, *15*(3), 1192-1209.

[25] Edwards, M., Deng, J., &Xie, X. (2016). From pose to activity: Surveying datasets and introducing CONVERSE. *Computer Vision and Image Understanding*, *144*, 73-105.

2016 International Joint Conference on Neural Networks (IJCNN) (pp. 381-388). IEEE.

[17] Shoaib, M., Bosch, S., Incel, O., Scholten, H., &Havinga, P. (2015). A survey of online activity recognition using mobile phones. *Sensors*, *15*(1), 2059-2085.

[18] Jordao, A., Nazare Jr, A. C., Sena, J., & Schwartz, W. R. (2018). Human activity recognition based on wearable sensor data: A standardization of the state-of-the-art. *arXiv preprint arXiv:1806.05226*.

[19] Wu, D., Pigou, L., Kindermans, P. J., Le, N. D. H., Shao, L., Dambre, J., &Odobez, J. M. (2016). Deep dynamic
[26] Su, X., Tong, H., &Ji, P. (2014). Activity recognition with smartphone sensors. *Tsinghua science and technology*, *19*(3), 235-249.

[27] Attal, F., Mohammed, S., Dedabrishvili, M., Chamroukhi, F., Oukhellou, L., &Amirat, Y. (2015). Physical human activity recognition using wearable sensors. *Sensors*, *15*(12), 31314-31338.

[28] Twomey, N., Diethe, T., Fafoutis, X., Elsts, A., McConville, R., Flach, P., & Craddock, I. (2018, June). A comprehensive study of activity recognition using accelerometers. In *Informatics* (Vol. 5, No. 2, p. 27). Multidisciplinary Digital Publishing Institute.

[29] O'Reilly, M., Caulfield, B., Ward, T., Johnston, W., & Doherty, C. (2018). Wearable inertial sensor systems for lower limb exercise detection and evaluation: a systematic review. *Sports Medicine*, *48*(5), 1221-1246.

ACADEMIA IN INFORMATION TECHNOLOGY PROFESSION (AITP)

(An Interest group of Nigeria Computer Society)

Website: *www.aitp.org.ng*

Email: *academia.infotech@gmail.com*

Information Technology in Education and Development

ITED 2020

4TH INDUSTRIAL REVOLUTION

ISBN: 978-012-679-1