

**13<sup>TH</sup>**

# **INTERNATIONAL** *Conference*



**Theme**

## **Information Technology Innovation for Sustainable Development**

**Conference Proceedings**  
**Volume 28**

Edited by:  
**Professor Adesola ADEROUNMU**  
**Professor Adesina SODIYA**

ISSN: 2141-9663

**ACKNOWLEDGEMENT**

The Nigeria Computer Society (**NCS**) acknowledges the immense and revered support received from the following organisations:

Chams Plc

Computer Professionals Registration Council of Nigeria (CPN)

Computer Warehouse Group

Data Sciences Nigeria Limited

Main One Limited

National Information Technology Development Agency (NITDA)

RLG Limited

Sidmach Technologies Limited

Systemspecs Limited

Zinox Technologies

We also recognize the laudable efforts of all others in cash or in kind towards the success of the 12<sup>th</sup> International Conference

**REVIEWERS:**

Prof. G. A. ADEROUNMU

Dr. A.S. SODIYA

Dr. A. O. Oluwatope

Dr. S. A. Akinboro

Dr. P. A. Idowu

Prof. 'Dele OLUWADE

Dr. (Mrs.) S. A. Bello

Prof. O. S. ADEWALE

Dr Mrs M. L. Sanni

Dr. F. T. IBRAHALU

Dr. I. Adeyanju

Dr. (Mrs.) S. A. ONASHOGA

Dr. A. P. ADEWOLE

Dr. I. K. Ogundoyin

Dr. S. E ADEWUMI

Dr. Wale Akinwunmi

Dr L. A. Akanbi

Dr. B. I. Akhigbe

Dr. F. O. Asahiah

Dr. Mrs I. O. Awoyelu

Dr. A. I. OLUWARANTI

Prof.. A. T. AKINWALE

Dr. E. O. OLAJUBU

Dr. O. A. OJESANMI

Dr. E. ESSIEN

Dr. Mrs I. O. Awoyelu

Dr. (Mrs.) O. R. VINCENT

Prof. Olumide B. LONGE

Dr. A. A. O. OBINIYI

Dr. F. T. IBRAHALU

Dr. (Mrs.) O. T. AROGUNDADE

Dr. (Engr.) A. Abayomi-Alli

Dr. O. R. VINCENT

Dr. R. G. Jimoh

Dr. O. J. OYELADE

Dr. A.A. Adeyelu

Dr. A. O. Ogunde

Dr. G. O. Ogunleye

**Publication Office:**

**Nigeria Computer Society (NCS):** Plot 10, Otunba Jobi Fele Way, Central Business District,  
Behind MKO Abiola Garden, Alausa, Ikeja – Lagos, Nigeria

P.M.B. 4800 Surulere, Lagos, Nigeria. Phone: +234 (0)8097744600, 09038353783

E-mail: ncs@ncs.org.ng Website: www.ncs.org.ng

## Forward

On behalf of members of the 19<sup>th</sup> National Executive Council (NEC) of the Nigeria Computer Society (NCS), it is with a great sense of responsibility that I welcome you all to this occasion, the 13<sup>th</sup> International Conference of our association. NCS is the national platform for advancing Information Technology Sciences and Practices in Nigeria.

We give thanks to God for making it possible for participants, presenters and resource persons to travel from all over the world to assemble in the beautiful city of Federal Capital of Nigeria for this most important event. We are here this year not just for an event but for a cause.

Sustainability issues have both local and global dimensions. A world in which poverty, inequalities, environmental threats and human insecurity thrive threatens the present and the future. Tech advances have obviously transformed the world but very significant divides – societal and digital – remain and are expanding. The challenges are not new but innovative IT oriented models, responses and culture are needed.

Our present economic challenges show clearly that oil is not our future. A strategic IT focus on sustainability through local content and other critical elements is the way forward. The conference will bring the issues of inclusion and sustainability to the front burner. The Global 2030 Agenda for Sustainable Development including the 17 Sustainable Development Goals (SDGs) is a huge agenda. Tackling these huge challenges using IT innovation is what this year's Conference entails. It is an agenda for current and future generations. The goals are ambitious but achievable. The IT community is equal to the task.

The theme of this year Conference is aptly titled: **“Information Technology Innovation for Sustainable Development”**. IT industry players, academics, researchers and thought leaders have been invited to speak, proffer solutions and challenge us all based on this theme.

As a multistakeholder event, the 13<sup>th</sup> International Conference brings stakeholders from government, industry, academia, the United Nations, multilateral agencies, International organisations, youth groups and civil society together to raise critical issues, debate, exchange information, conduct demos, share innovative ideas and produce clear and actionable recommendations that plot the way forward.

The organizers of the Conference owe special thanks to our national and international guests and lead paper presenters for accepting to be part of this year's Conference.

We also appreciate our partners – Computer Professionals Registration Council of Nigeria (CPN), Sidmach Technologies, Data Sciences, Systemspecs, Main One and National Information Technology Development Agency (NITDA). It is our prayer that together, we will move Nigeria to a greater height. I wish you all, very exciting and resourceful deliberations and Journey mercies back to your destinations at the close of this Conference.

We wish to appreciate our editors, Dr. A. S. Sodiya and Dr. A. O. Oluwatope; and all reviewers for their efforts at ensuring quality presentations at this Conference. The Chairman, Local Organizing Committee, Mr Rex Abitogun and all members of his committee who have taken the organization of this Conference as task that must be accomplished, thank you all. The National Executive Council of the Nigeria Computer Society is immensely indebted to the dynamic Conference Planning Committee, which worked assiduously, even against odds, to make this reality.

Thank you and God bless you all.

**Professor Adesola Aderounmu FNCS**  
President, Nigeria Computer Society

**TABLE OF CONTENTES**

- |                      |     |
|----------------------|-----|
| 1. Acknowledgement   | i   |
| 2. Forward           | ii  |
| 3. Table of Contents | iii |

**Session A: Traffic Technology**

- |   |    |
|---|----|
| 1. Multi-Agent Based Monitoring and Control of Power Distribution System - <b>M.K. Ahmed; A. Aliyuda; M. S. Bute</b>    | 2  |
| 2. Data Mining Driven Approach for Predicting Causes of Road Accident - <b>L. J. Muhammad; A.Yakubu; I. A. Mohammed</b> | 10 |

**Session B: Sustainable Healthcare for All**

- |   |    |
|---|----|
| 1. An Ontological Knowledge Framework for Diagnosing Breast Cancer Using on Reasoning Algorithm - <b>O. N. Oyelade; A. A. Obiniyi; S. B. Junaidu</b>  | 16 |
| 2. Design of a Framework for Healthcare Crime Investigation Using Big Data Analytics - <b>S. T. Yange; H. A. Soriyan</b>                              | 25 |
| 3. Framework for Development of Mobile Telenursing System for Developing Countries - <b>J. O. Adigun; J. O. Onihunwa; D. A. Joshua; O. O. Adesina</b> | 35 |
| 4. Reasoning Over Vague Ontologies Using Fuzzy Soft Set Theory In Medical Domain - <b>R. Salahudeen; A. F. Donfack Kana</b>                           | 49 |

**Session C: E-Government, Digital Development and E-Readiness**

- |   |    |
|---|----|
| 1. A Conceptual Framework for e-Census, e-Election and Good Governance - <b>A. A. Eludire</b> | 62 |
|---|----|

**Session D: Cybersecurity, Infrastructure Protection and Digital Privacy**

- |  |    |
|--|----|
| 1. A Chaos Based Image Encryption Algorithm Using Shimizu-Morioka System - <b>H. J. Yakubu; T. Aboiyar</b> | 75 |
| 2. An improved RSA Algorithm Based on Residue Number System - <b>Y.K. Saheed; K.A. Gbolagade</b>           | 84 |

3. Arithmetic Operations in Deterministic P Systems Based on the Weak Rule Priority - **C. M. Peter; D. Singh** 90

#### Section E: Sustainable Infrastructure for Innovation, Research and Development

1. An Auto Generated Approach for Generation Stop Words Using Aggregated Analysis - **Tijani O.D.; Akinwale A.T.; Onashoga S.A.; Adeleke E.O.** 99
2. Information Technology as a Tool for Attaining Food Security and Sustainable Development in Nigeria - **F. O. Okorodudu; G. O. Eloho; B. Ossai** 116
3. Using the Adjusted Weighting Function to Bridge the Networked Readiness Digital Divide - **P. K. Oriogun** 124

#### Session F: Digital Economies: Capacity Building, Start-ups and Youth Innovation

1. A Neuro-Fuzzy System for Characterising Saki Pattern in Woven Fabrics - **R. R. Madaki; Y. Baguda; L. Abdulwahab** 134
2. A Hybrid Dimensionality Reduction Model for Classification of Microarray Dataset - **M.O. Arowolo; S.O. Abdulsalam; R.M. Isiaka; K. Gbolagade** 139
3. Profit Maximization Product-Mix Problem of Small and Medium Enterprise - **Ademola O. Adesina; David O. Iyanda** 146

**13<sup>th</sup>**

**International Conference**



**Session A:**  
**Traffic Technology**

---

Full Paper

**MULTI-AGENT BASED MONITORING AND CONTROL OF  
POWER DISTRIBUTION SYSTEM**

---

**M.K. Ahmed**

Department of Mathematics,  
Gombe State University,  
Gombe  
kabirmka@yahoo.com

**A. Aliyuda**

Department of Mathematics,  
Gombe State University,  
Gombe  
aliyuda.ali6@gmail.com

**M. S. Bute**

Department of Mathematics,  
Gombe State University,  
Gombe  
msbute@gmail.com

**ABSTRACT**

In this Paper, a Multi-agent System platform that will allow agents to detect and report fault in Power Distribution System (PDS) with the minimum time delay is presented. Multi-agent System Engineering (MaSE) methodology was used to demonstrate how agents will be able to manage the complexity of PDS. The simulation of the framework was done using Java Agent Development Framework (JADE). The result indicates that Multi-Agent System technology can be applied to detect and report fault in Power Distribution System with the minimum time delay.

**Keywords:** Agent, Multi-Agent System, Power Distribution System, Multi-agent Software Engineering (MaSE)

## 1. INTRODUCTION

Power Distribution is the stage in which electric power is delivered to final consumers through a transmission system. The major concerns of Electricity Distribution Companies in Nigeria include detection of illegal connection, fault detection and so on. This research is motivated by the fact that problems requiring multi-agent solution include problems that are inherently distributed, for example geographically and problems whose solutions can be drawn from distributed units (Hyacinthet and Divine, 2001). As such, electricity distribution falls in this scope. In addition, Adler and Blue (2002), described the significant of agent technology in enhancing the design and analysis of problem domains under the following three conditions:

- i. The problem domain is geographically distributed;
- ii. The subsystems exist in a dynamic environment; and
- iii. The subsystems need to interact with each other more flexible.

The domain of electricity distribution is well suited to an agent-based approach because of its geographically distributed nature and the existence of its subsystems in dynamic environments where the subsystems interact with each other.

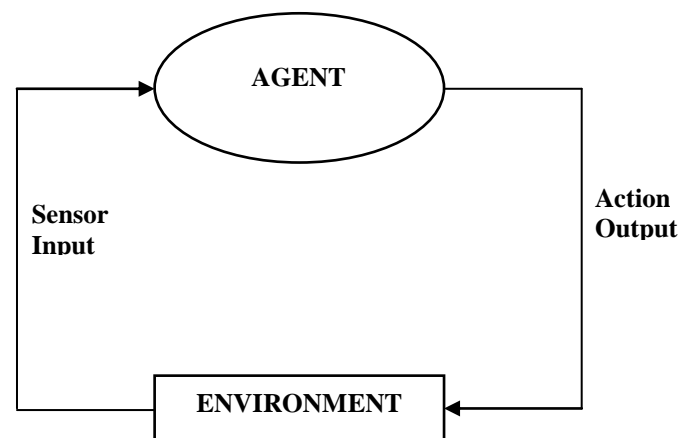
Agent-based computing is one of the powerful technologies for the development of distributed complex systems (Zambonelli and Parunak, 2003). In (Dimeas and Hatziargyriou, 2005) the authors presented agents as new paradigm for software development since object-oriented design, and the concept of intelligent agents has already found a diverse range of applications in manufacturing, real-time control systems, electronic commerce, network management, transportation systems, information management, scientific computing, health care, and entertainment, etc. The reason for the growing success of agent technology in these areas is that the inherent distribution allows for a natural decomposition of the system into multiple agents that interact with each other to achieve a desired global goal.

Agents can operate without the direct intervention of humans or others (Wang, 2005). This feature helps agents to monitor and detect fault in power distribution system in a particular area. In Multi-Agent System (MAS), agents communicate with other agents in a system to achieve a global goal.

Agents can also perceive their environment and respond in a timely fashion to environmental changes (Rosa et al., 2009).

### 1.1 An Agent

An agent is a computer system that is situated in some environment, and that is capable of autonomous action in this environment in order to meet its design objectives (Bo and Harry, 2010).



**Figure 1:** An Agent in Its Environment.

Figure 1 gives an abstract view of an agent. In this diagram, the agent takes sensory input from the environment, and produces as output actions that affect it. The interaction is usually an ongoing, non-terminating one. We can see the action output generated by the agent in order to affect its environment. In most domains of reasonable complexity, an agent will not have complete control over its environment. It will have at best partial control, in that it can influence it.

### 1.2 Multi-Agent System (MAS)

In (Dimeas and Hatziargyriou, 2005) the authors defined multi-agent system as a group of agents which sense the environment and act in order to achieve specific objectives. MAS can therefore be seen as a network of interacting software modules that bring together dispersed systems that collectively manage complex tasks that are beyond the capacity of any individual system on the network.

In (Wan-Yu Yu et al., 2015) the authors described that, the term multi-agent system implies more than one agent interacting with each other within an underlying communication infrastructure where



individual agents are often distributed and autonomous. Multi-agent systems are based on the idea that a cooperative working environment can cope with problems which are hard to solve using the traditional centralized approach to computation. Agents are used to interact in a flexible and dynamic way to solve problems more efficiently.

## 2. RELETED WORKS

For the design of our Agent-based framework we revised the state of the art of MAS platforms in relation to power distribution systems (Tao et al., 2004), (Ioannis and Dimitris, 2007), (Deshmukh et al., 2008), (Rumley et al., 2010), (Chinwuko et al., 2011) and (Wan-Yu Yu et al., 2012) We also revised the state of the art of Agent-Oriented Software Methodologies (Maalal and Addou, 2011) and (Silaghi, 2005). The purpose of this revision is to select the most appropriate and suitable agent platform and agent-oriented software methodology for this work.

In (Ioannis and Dimitris, 2007) a conceptual framework for a power system self-healing infrastructure is examined. In (Tao et al., 2004) and (Deshmukh et al., 2008) the authors presented a MAS designed for distribution systems restoration. These works abstract network buses as agents, along with a facilitation agent who is responsible for aiding negotiation processes among bus agents. In (Adejumobi and Pekum, 2009), agents have hierarchical levels and higher level agents coordinate a group of lower level agents. In (Rosa et al., 2010) each agent is considered to have its special functionalities e.g. acquiring data, analyzing data, managing situations, etc. In other words each step of decision making is done by a special agent. An intelligent agent-based environment to coordinate maintenance schedule discussions is introduced in (Rumley et al., 2010) and autonomous regional active network management system is introduced and discussed in (Chinwuko et al., 2011).

In general, the above literatures consist of a master agent which makes the final decision based on the data received from other agents. This cannot be considered as distributed control since the master agent is behaving like a control center.

In (Wan-Yu Yu et al., 2012) the authors proposed an agent committee approach to push the technology to some extent in solving the power restoration problems in a more distributed and efficient way.

The idea is to allow neighboring switch agents to be organized into a local power committee whose responsibility is to ensure that the local power demand can be satisfied as well as reconciled with an optimal solution to the conflicting issues proposed among committee members who are responsible for identifying a feasible power source to the local power demand problems. The advantage of this approach is to localize the global problem solving ability to a local committee of agents who could coordinate with neighboring committees to achieve an agreement that satisfies the global objective of the distribution system without relying on a pre-determined centralized optimization algorithm as traditional approaches. The authors illustrated a committee-based multi-agent system whose objective is to find a solution of power restoration problem that can maximize the service zones while minimizing the number of switch operations under the topological and operational constraints of power distribution systems.

While this approach contributed in pushing the technology to some extent, however, since a committee is an organization of agents, it requires a leader to actually conduct the decisions of the committee and communicate with its committee member and other committees. For this purpose, the system will encounter a problem whenever a break in communication occurs between the Local Power Committee leader and the central unit of the system.

## 3. MULTI-AGENT SYSTEM FRAMEWORK

In order to design a framework, we selected the MaSE methodology which is a general purpose methodology for developing multi-agent systems that is founded on the basis of software engineering principles (Maalal and Addou, 2011). MaSE divides the development process into two major phases: the analysis phase and the design phase (Silaghi, 2005). The analysis phase consists of the following stages: capturing goals, applying use cases, and refining roles and the design phase consists of the following stages: creating agent classes, constructing conversations, assembling agent classes, and system design. For each phase, MaSE provides a set of stages need to be performed. The main goal of MaSE is to guide a designer through the software lifecycle from a documented specification to an implemented agent

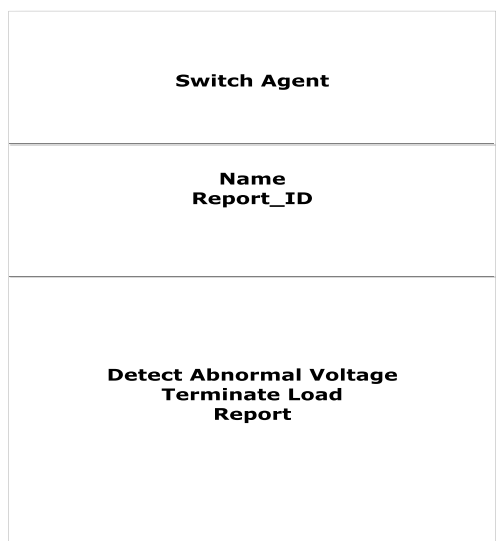
system, with no dependency of a particular MAS architecture, agent architecture, programming language, or message-passing system.

**3.1 The Switch Agent (Voltage Agent)**

Switch agents are passive agents who sense their environments and forward the information gathered to their Zone Agents along the topology. They are in charge of maintaining the flow of current in their areas in a normal state. Therefore, they must periodically update their knowledge of the area in order to adapt their power injection.

These agents put forth responsibilities that include monitoring system voltage and frequency to detect contingency situations or grid failures, and sending signals to the zone agent when fault is detected. In addition to that the switch agent is also responsible for load termination in case of high voltage to avoid damage of consumer appliances.

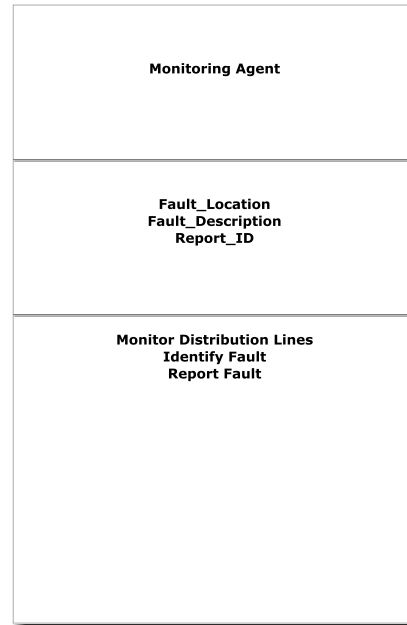
The goal of each switch agent is to update the topological and electrical information about the actual area whose initiating switch agent is the master. Figure 2 depicts switch agent architecture. In the architecture, the agent checks each active adjacent link and sends message information to its zone agent about the current states of attached links and waits for a response. The Zone Agent receiving this message will send an acknowledgement to the switch agent sending the message and will recursively forward it further to the control center (control agent). In case of no acknowledgement from the zone agent receiving the message, the switch agent will resend the same message to its zone agent for further action.



**Figure 2:** Switch Agent (Micro Level) Architecture

**3.2 Monitoring Agent**

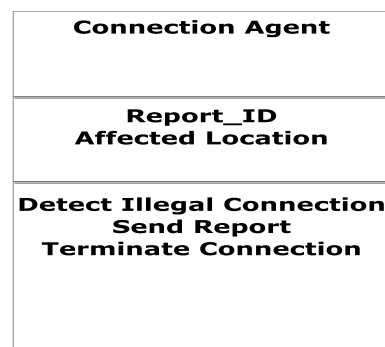
The Monitoring Agent is responsible for checking the distribution line to identify fault (line breakage) and report to its zone agent. Figure 3 below shows the agent micro level architecture.



**Figure 3:** Monitoring Agent (Micro Level) Architecture

**3.3 Connection Agent**

One of the major problems observed in Electricity Distribution Companies is illegal connection. This Connection Agent is responsible for detecting such connections, reporting it to zone agent and terminating the illegal connection. Figure 4 below shows the agent micro level architecture.



**Figure 4:** Connection Agent (Micro Level) Architecture

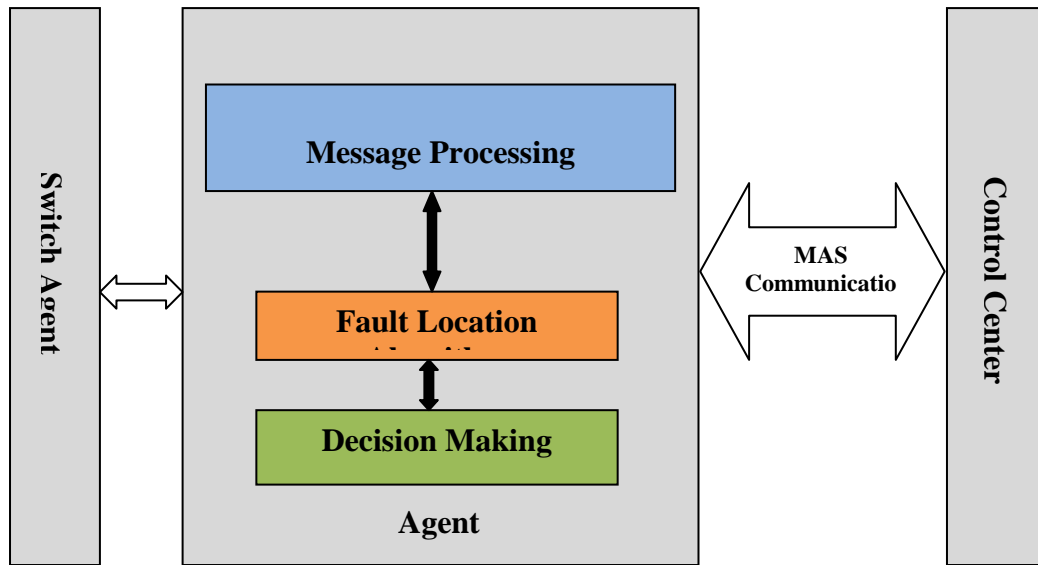


Figure 5: Single Zone Agent Architecture

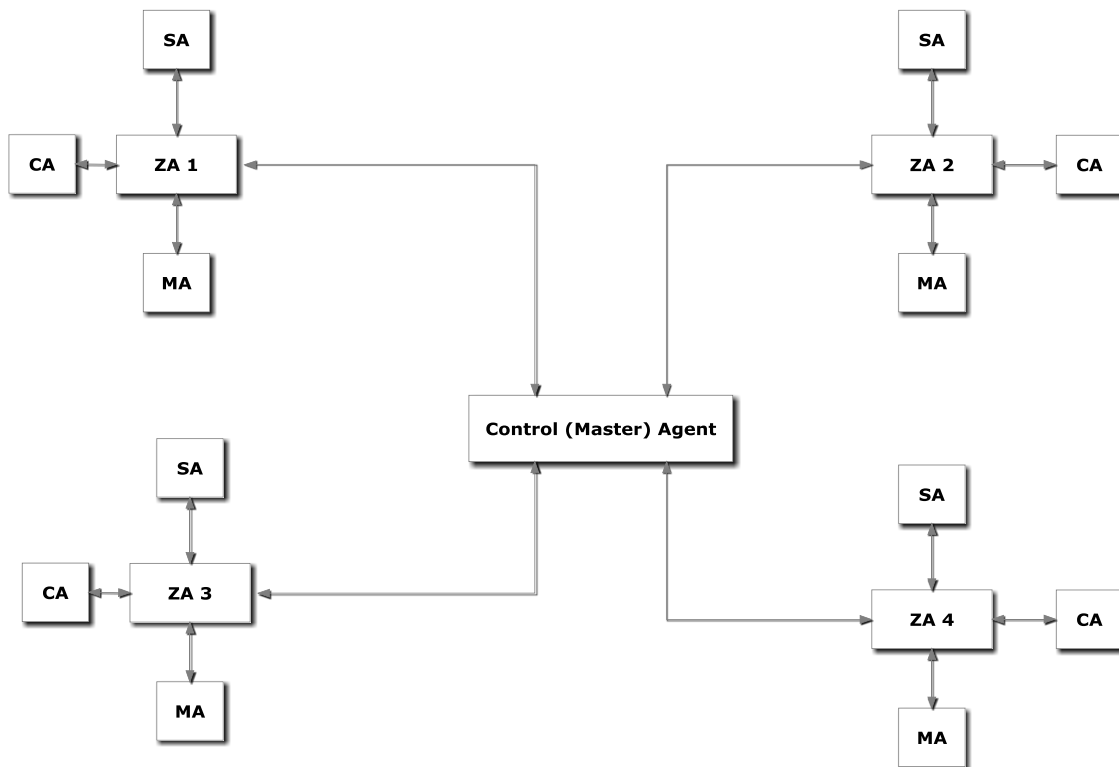


Figure 6: MAS Collaboration Diagram

### 3.4 The Zone Agent

At the beginning, each Zone Agent knows only the data of the Switch Agents located in the zone. As algorithm proceeds, Zone Agents gain additional data from the Switch Agents and use this information to fulfill the following goals:

- I. Determine where fault occurred in the zone using the message received from Switch Agents
- II. Sends acknowledgement to the switch agent sending the message
- III. Forwards the message received to the control center and waits for acknowledgement
- IV. Send the same message to the next Zone Agent in case of no acknowledgement from the control center. Figure 5 depicts single zone agent architecture.

### 3.4 Collaboration among Agents in MAS

Our framework consists of the Switch Agent (SA), the Monitoring Agent (MA) and the Connection Agent (CA). After the SA, MA and the CA specifications, the next step involves the formalization of agent roles and interactions applied to the problem at hand. To accomplish this task, a collaborative diagram which defines the interaction among agents and their interaction with the environment needs to be defined. The collaborative diagram of MAS is shown in Figure 6. The diagram illustrates three agents (SA, CA and MA) and their interactions with each other and the environment through their respective Zonal Agents (ZA). The SA interacts with the environment and communicates with the ZA which in turn forwards the state of the environment to the Control (Master) Agent. The Control Agent receives copies of all messages exchanged within the MAS and is responsible for interpreting and displaying these messages to users for appropriate action.

The initialization of the MAS is performed by the CA, the SA and the MA notifying the ZA of their presence. This includes notifying the ZA of their names and ID; the ZA then communicates with the Control Center of their presence. In

order to demonstrate the proposed MAS, a simulation test prototype is developed in JADE as a simplified distribution circuit.

## 4. RESULTS AND DISCUSSION

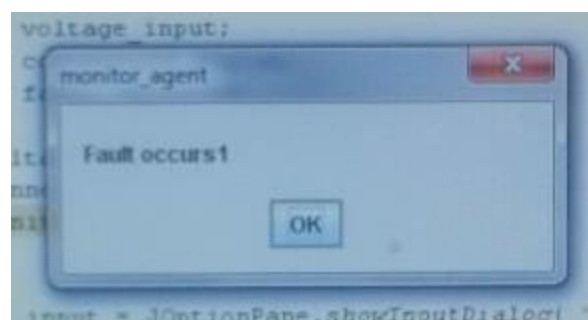
Figure 7 depicts a message generated by the Switch (Voltage) Agent when abnormal voltage is detected. This agent terminates the load and report to the Zonal Agent which in turn copies the message to the Control (Master) Agent.

The Monitoring Agent is responsible for checking the Distribution Line. Figure 8 depicts the Monitoring Agent detecting a fault in line. This message is reported to the Zonal Agent which in turn copies to the Control (Master) Agent.

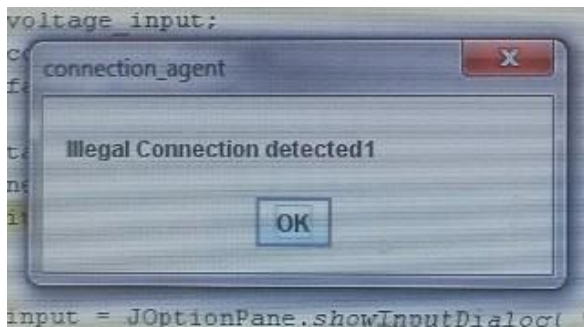
The Connection Agent is responsible for detecting an illegal connection by power consumers. Figure 9 depicts a scenario where an illegal connection is detected and reported accordingly.



**Figure 7:** Switch (Voltage) Agent detecting and reporting abnormal voltage



**Figure 8:** Monitor Agent detecting and reporting a fault



**Figure 9:** Connection Agent detecting and reporting an illegal connection

### 5.0 CONCLUSION

This research describes the design of distributed MAS architecture for detecting fault in Power Distribution System. The MAS consists of a SA, ZA and the CA. Agents exchange their messages based on the IEEE FIPA standard to ensure the system interoperability. The application implementation process is illustrated through simulated case study, which indicates that the proposed MAS can respond to its environment when fault is detected. In our future work, we intend to extend our MAS to have other additional Agents. The first one for ensuring customers are not charged amount higher than their power consumption. While the second agent will be responsible for finding a way of detecting a case of bypassing prepaid meter using an Agent as Sensor.

### 6 REFERENCES

- Adejumobi, I. A. and Pekum, A. J. (2009). "Software Application for Electrical Distribution System Reliability Studies: Box-Jenkins Methodology". *Pacific Journal of Science and Technology*. 10(2):377-387.
- Adler, J. L. and Blue, V. J. (2002). A cooperative multi-agent transportation management and route guidance system. *Transp. Res. Part C: Emerging Technol.*, vol. 10, no. 5/6, pp. 433-454
- Bo, C. R. and Harry, H. (2010). "A Review of the Applications of Agent Technology in Traffic and Transportation Systems". *IEEE Transactions on Intelligent Transportation Systems*, Vol. 11, NO. 2, pp 44 – 49.
- Chinwuko, E. C., Nwuba, U. and Mgbemena, C. O. (2011). "Optimum Reliability and Cost of Power Distribution System: a case of Power Holding Company of Nigeria". *International Journal of Engineering Science & Technology*, 12(5): 233 – 238.
- Deshmukh, R. K., Davidson, E. M. and McArthur, S. D. J. (2008). "Exploiting Multi-Agent System technology within an autonomous regional active network management system". *Proceedings of the 14<sup>th</sup> IEEE International Conference on Intelligent Systems Application to Power Systems*. Pp 57 – 61. New York.
- Dimeas, A. and Hatziaargyriou, N. D. (2005). "Operation of a Multi-agent System for Micro Grid Control". *IEEE transaction on power systems*, Vol. 20, No.3, 48 – 53.
- Hyacinthet, S. N. and Divine, T. N. (2001). *A Perspective on Software Agents Research*. Applied Research & Technology Department British Telecommunications Laboratories. Martlesham Heath, Ipswich, Suffolk, IP5 3RE, UK, pp 11 – 16.
- Ioannis, S. B. and Dimitris, P. L. (2007). "Implementing Multi-Agent Systems Technology for Power Distribution Network Control and Protection Management". *IEEE Transactions on Power Delivery*, Vol. 22, NO. 1, pp 14 – 19.
- Maalal, S. and Addou, M. (2011). "A New Approach of Designing Multi-Agent Systems". *International Journal of Advanced Computer Science and Applications*, Vol.2, No. 11, pp 67 – 72.
- Rosa, M. A., Miranda, V., Carvalho, L. and DaSiva, A. M. L. (2010). "Modern Computing Environment for Power System Reliability Assessment". *Proceedings of the 12<sup>th</sup> ACM Probabilistic Methods Applied to Power Systems*, Pp 29 – 35. Singapore.
- Rosa, M. A., DaSilva, A. M. L., Miranda, V., Matos, M. and Sheblé, G. (2009). "Intelligent Agent-Based Environment

to Coordinate Maintenance Schedule Discussions”. International Symposium on Intelligent Systems Applications to Power Systems, pp 22 – 27.

Rumley, S., Elvira, K., Hugh, R. and Alain, G. (2010). “Multi-Agent Approach to Electrical Distribution Networks Control”. 27<sup>th</sup> Annual IEEE International Computer Software and Applications Conference. Pp 62 – 68. New York

Silaghi, G. C. (2005). “Software Engineering Approaches for Design of Multi-agent Systems”. IEEE Transactions on Power Systems 17(2): 457–462.

Tao, Y., Nagata, T., Kimura, K., Sasaki, H. and Fujita, H. (2004). A Multi-Agent Approach to Distribution System Restoration”. Proceedings of the 47th IEEE Midwest Symposium on Circuits and Systems, Vol. 2, pp 51 – 58.

Wang, F. Y. (2005). “Agent-Based Control for Networked Traffic Management Systems”. IEEE Intell. Syst., vol. 20, no. 5, pp. 92–96.

Wan-Yu Yu, V., Men-Shen, T. and Yen-Bo, P. (2012). “Coordinating a Society of Switch Agents for Power Distribution Service Restoration in a Smart Grid”. IEEE Trans. on Power System, Vol. 25, No. 1, pp. 72 – 81.

Zambonelli, F. and Parunak, H. V. D. (2003). “Signs of a Revolution in Computer Science and Software Engineering”. In Proc. 3rd Int. Workshop Eng. Soc. Agents World, vol. 25, No. 9, pp. 13–21

## Full Paper

**DATAMINING DRIVEN APPROACH FOR PREDICTING CAUSES OF ACCIDENT OF KANO-WUDIL HIGHWAY****L. J. Muhammad**

Mathematics and Computer Science  
Department, Federal University,  
Kashere.  
mljtech@gmail.com

**A. Yakubu**

Mathematics and Computer Science  
Department, Federal University,  
Kashere.  
au.nlaro@gmail.com

**I. A. Mohammed**

Yobe State University, Damaturu,  
Yobe State, Nigeria  
ibrahimsallau@gmail.com

**ABSTRACT**

Road traffic accidents is the inadvertent crash involving at least one motor vehicle, occurring on a road open to public circulation, in which at least one person is injured or killed and it is indisputably one of the most frequent and most damaging calamities bedeviling human societies. It is therefore, of paramount importance to seek to identify the root causes of road traffic accidents in order to proffer mitigating solutions to address the menace. This research, aimed at predicting the likely causes of road accidents, its prone locations and time along Kano– Wudil highway in order to take all necessary counter measures is a step forward in this direction. In this study data mining decision tree algorithm was used to predict the causes of the accidents, its prone locations and time along Kano – Wudil Highway that links Kano State to Wudil Local Government Area, Kano State for effective decision making.

**Keywords:** Accident, Data mining, Decision tree, Id3 tree, Algorithm.

## 1. INTRODUCTION

Road Traffic Accidents killed more than 1.2 million people, and injured between 20 and 50 million others in 2004, thereby becoming the ninth most common cause of death in that year. Road traffic accidents remain among the most central public health problems in the world. A tragic fact is that among the young people aged between 15 and 29 years, road traffic accident is one of the most common causes of death worldwide (WHO, 2009). The incidence of fatal road accidents in Nigeria is phenomenal. Trend analysis of fatal road accidents between June 2006 and May 2014 using Nigeria Watch database shows that 15,090 lives were lost to fatal road accidents in 3,075 events. The highest fatality occurred in 2013 (2,061 deaths), a 2.8% increase from the 2012 record of 1,652 deaths. However, the figures were rising again in 2014, with fatality records of 964 deaths between January and May 2014 (FRSC, 2014).

Nigeria is ranked second-highest in the rate of road accidents among 193 countries of the world. Aside from the Boko Haram crisis, accidents are currently by far the main most violent cause of death in Nigeria. The World Health Organization (WHO) adjudged Nigeria the most dangerous country in Africa with 33.7 deaths per 100,000 population every year. According to their report, one in every four road accident deaths in Africa occurs in Nigeria (WHO, 2009).

The intensity of fatal road accidents varies from state to state. States are ranked by a low, medium, or high severity index. Ebonyi and Lagos states ranked low (0.00 – 2.99%) in the severity index of fatal road accidents that occurred between June 2006 and May 2014. Ebonyi had a severity index of 2.4 from 51 deaths in 21 crashes, while Lagos had a low severity index of 2.5 from 1,590 deaths in 620 crashes. States that are ranked medium (3.00 – 5.99%) include FCT (Abuja), Ekiti, Delta, Akwa Ibom, Plateau, Bauchi, Bayelsa, Kwara, Osun, Cross River, Taraba, Ogun, Abia, Nasarawa, Oyo, Anambra, and Sokoto. Those ranked with a high index (6.00 – 8.99%) are Imo, Jigawa, Benue, Niger, Edo, Gombe, Borno, Ondo, Enugu, Kano, Kaduna, Rivers, Zamfara, Kogi, Katsina, Kebbi, Adamawa, and Yobe. Highest in this category is Yobe State, with a severity index of 11.4. (FRSC, 2014).

In Nigeria today, the issue of road accidents has become a teething problem and there has been a major problem of the cardinal causes responsible for these accidents and their prone locations along high ways in Nigeria. This research harnessed data

mining algorithms to predict the causes of accident and accident prone locations along Kano- Wudil Highway for counter-accidents decision making by relevant stakeholders.

## 2. RELATED WORK

Many works have been carried out by different researchers on the prediction of the causes of accidents, their prone locations along various roads around the globe using data mining algorithms.

Chang and Chen, (2005) conducted data mining research focusing on building tree-based models to analyze freeway accident frequency. Using the 2001-2002 accident data of National Freeway 1 in Taiwan, the authors developed classification and regression tree (CART) and negative binomial regression models to establish the empirical relationship between traffic accidents and highway geometric variables, traffic characteristics, and environmental factors. The authors found that the average daily traffic volume and precipitation variables were the key determinants of freeway accident frequency.

Getnet (2009) investigated the potential application of data mining algorithms to develop models supporting the identification and prediction of major driver and vehicle risk factors that cause road traffic accidents. The research used the WEKA version 3-5-8 tool to build the decision tree (using the J48 algorithm) and rule induction (using PART algorithm) techniques. Performance of the J48 algorithm was slightly better than that of the PART algorithm. The license grade, vehicle service year, vehicle type, and experience were identified as the most important variables for predicting accident severity.

Dipo and Akinbola (2012) investigated the cause of accident and accident prone locations on Highways in Nigeria, Lagos- Ibadan highway as a case study using decision tree data mining algorithm. WEKA software was used to analyse accident data gathered along this road. The results showed that causes of accidents, specific time/condition that could trigger accident and accident prone areas could be effectively identified. There were 50 rules generated from this tree. Rule 1- 18 indicate the occurrence of accident in Location 3 and rule 19-50 also shows the occurrence of accident in location 2. This indicates that, location 2 (Above 10km – 20km, from Lagos to Ibadan) has the highest number of road accident occurrence with Heavy-vehicle in the after-noon and during the dry season. Rule 41 is the best one that can be used for prediction. The rule says that, Tyre bust is the cause of road accident



with heavy vehicle within location 2 in the day time and during the dry season.

This research harnessed the decision tree data mining algorithm to predict the causes of accident and accident prone locations along Kano- Wudil highway. The reason of using decision tree data mining algorithm for the research could be seen in the research of Dipo and Akinbola (2012), where a comparison of different Decision Tree algorithms and Artificial Neural Networks performance were analyzed using road accidents data set. The location is between the first 40 kilometers along the Ibadan-Lagos Express road. The work used Multilayer Perceptron as well as Radial Basis Function (RBF) Neural Networks, Id3 and Function Tree algorithms. Results shows that the Id3 tree algorithm performed better with higher accuracy rate, while Radial basis function performed better than multilayer perceptron in terms of time used in the building of the model and number of correctly classified instances. The result showed that, Decision Tree techniques outperformed Artificial Neural Networks with a lower error report and with a higher number of correctly classified instances and better accuracy rate generated. Tyre burst, broken shaft and loss of control variables were the three major causes of accidents where tyre burst represents the major cause of accidents.

### 3.0 Methodology

**3.1 Data Mining** is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data. It can also be defined as an interactive process of discovering valid and novel, useful and understandable patterns or models in large database. Data Mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. Data mining uses advances in the field of Artificial Intelligence (AI) and Statistics.

### 3.2 Decision tree

Decision tree is a “divide-and-conquer” approach to the problem of learning from a set of independent instances, which leads naturally to a tree-like style of representation called a decision tree. A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a

class label. The topmost node in the tree is the root node.

This research harnessed decision tree data mining algorithm to predict the causes of the accident, its prone locations and time along Wudil – Kano Highway.

### 3.3 Accidents along Wudil- Kano Highway

Kano to Wudil Highway is one of the busiest roads in Kano. It is a route that links Kano to Jigawa, Gombe, Bauchi, Adamawa, Taraba, Maiduguri and Yobe States and it is 44km away from Kano City, along Maiduguri Road. According to FRSC Traffic Digest, January 2014, total Road Traffic Crashes 164 cases increased by 24% in 2011 compared to 2010 figure of 132. Thereafter, a downward trend is observed in total Road Traffic Crashes Fatality (119) increased by 53% in 2013 over 2012 figure of 78 deaths. There was consistent increase in Traffic volume per hour from 2010 with 392 vehicles per hour to 542 vehicles per hour in 2013 representing 38% increase along Kano – Wudil Route. In January, 2014 Wudil - Kano Highway is among Top 20 Routes with highest no. of person killed and Top 20 Routes with Highest Number of Persons Injured. Chart 1.1 and Chart 1.2 depicts the top 20 routes with highest of person killed and that of highest number of persons injured in January, 2014.

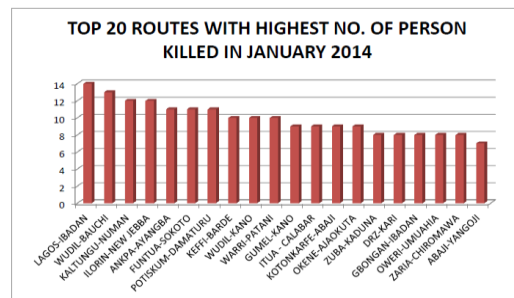


Chart 3.1 Source: Federal Road Safety Corp (2014)

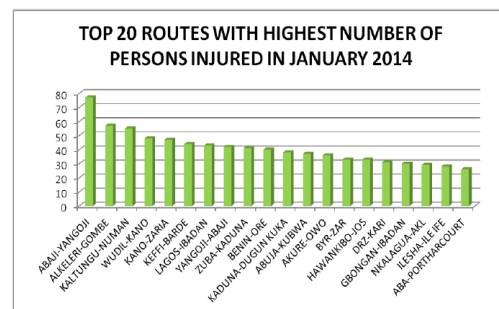


Chart 3.1 Source: Federal Road Safety Corp (2014)

The above statistics depicts that predicting the causes of accident, its prone locations and time along Kano – Wudil Highway is very important input for counter-accidents decision making of the high way. Table 3.1 depicts the statistics of the accident along the high way between January, 2014 to June, 2016.

**Table 3.1** Frequency of the Accident

SN	Month of the Accident	No. of Accidents
1	January, 2014	8
2	February, 2014	6
3	March, 2014	5
4	April, 2014	6
5	May, 2014	7
6	June, 2014	1
7	July, 2014	4
8	August, 2014	7
9	September, 2014	5
10	October, 2014	4
11	November, 2014	8
12	December, 2014	5
13	January, 2015	6
14	February, 2015	7
15	March, 2015	5
16	April, 2015	7
17	May, 2015	6
18	June, 2015	10
19	July, 2015	5
20	August, 2015	7
21	September, 2015	9
22	October, 2015	8
23	November, 2015	2
24	December, 2015	2
25	January, 2016	8
26	February, 2016	2
27	March, 2016	11
28	April, 2016	3
29	May, 2016	4
31	June, 2016	5
<b>Total Number of Accident</b>		<b>165</b>

#### 4. DATA MODELING

The research considered the data of accident record between 44 km Kano to Wudil L.G.A. The data were organized into a relation. The sample data used covered the period of 30 Months, from January 2014 to June, 2016 as indicated in table 3.1 The output

variable is the location and the locations can be divided into four distinct locations tagged location A, B, C and D. Location 1 – 11km is LocationA, above 11 km – 22km LocationB, above 22km – 33km LocationC and above 33 km – 44km LocationD.

**Table 4.1** Showing variables given both continuous and categorical value.

**Table 4.1** Variables

SN	Variables	Description	Value
1	Type of vehicle	Small Car	A
		Heavy Car	B
2	Time of accident	Morning	A
		Afternoon	B
		Evening	C
3	Causes of accident	OverSpeed	A
		LossOfControl	B
		WrongOvertaking	C
		TyreBlowouts	D
		PoorLights	E
		UncertainCause	F
		BrakeFailure	G
4	Location of Accident	LocationA	A
		LocationB	B
		LocationC	C
		LocationD	D

#### 5. RESULTS

Weka data mining software was used to mine the data using Id3 decision tree data mining algorithm. The algorithm is one of the most widely used and practical methods for inductive inference over supervised data. It represents a procedure for classifying and categorical data based on their attributes. It is also efficient for processing large amount of data, so is often used in data mining application.

##### 5.1 Causes of accident

Below are results obtained using Id3 decision tree for the causes of accident along Kano - Wudil Highway.

##### 5.1.1 Id3 tree

AccidentLocation = LocationA: UncertainCause

AccidentLocation = LocationB

| VehicleType = SmallCar: PoorLights

| VehicleType = HeavyCar: LossOfControl

AccidentLocation = LocationC

| AccidentTime = Morning: WrongOvertaking

| AccidentTime = Afternoon: null

| AccidentTime = Evening: WrongOvertaking

AccidentLocation = LocationD

| AccidentTime = Morning: OverSpeed

| AccidentTime = Afternoon: WrongOvertaking  
 | AccidentTime = Evening  
 | | VehicleType = SmallCar: WrongOvertaking  
 | | VehicleType = HeavyCar: BrakeFailure

### 5.2 Locations of accident

Below are results obtained using Id3 decision tree for the locations of accident along Kano - Wudil Highway.

#### 5.2.1 Id3 tree

AccidentCause = OverSpeed  
 | AccidentTime = Morning: LocationD  
 | AccidentTime = Afternoon: null  
 | AccidentTime = Evening: LocationC  
 AccidentCause = LossOfControl  
 | VehicleType = SmallCar: LocationC  
 | VehicleType = HeavyCar: LocationB  
 AccidentCause = WrongOvertaking  
 | AccidentTime = Morning: LocationD  
 | AccidentTime = Afternoon: LocationD  
 | AccidentTime = Evening: LocationC  
 AccidentCause = TyreBlowouts: LocationD  
 AccidentCause = PoorLights: LocationB  
 AccidentCause = UncertainCause: LocationA  
 AccidentCause = BrakeFailure: LocationD

### 5.3 Time of accidents

Below are results obtained using Id3 decision tree for the time of accident along Kano - Wudil Highway.

#### 5.3.1 Id3 tree

AccidentCause = OverSpeed  
 | AccidentLocation = LocationA: null  
 | AccidentLocation = LocationB: null  
 | AccidentLocation = LocationC: Evening  
 | AccidentLocation = LocationD: Morning  
 AccidentCause = LossOfControl: Evening  
 AccidentCause = WrongOvertaking  
 | AccidentLocation = LocationA: null  
 | AccidentLocation = LocationB: null  
 | AccidentLocation = LocationC: Evening  
 | AccidentLocation = LocationD: Afternoon  
 AccidentCause = TyreBlowouts: Evening  
 AccidentCause = PoorLights: Evening  
 AccidentCause = UncertainCause: Morning  
 AccidentCause = BrakeFailure: Evening

## 6. DISCUSSION

There are 7 identified causes of accidents along the Kano-Wudil Highway which include; over speed, loss of control, wrong overtaking, tyre blowouts, poor lights, uncertain causes and brake failure. The result

showed that out of the 165 instances of the accident, between January, 2014 to June, 2016, 81 instances of the accident occurred as a result of wrong overtaking, 33 as a result of over speed, 16 instances as a result of loss of control, 10 instances as a result of tyre blowout, 10 instance also as a result of poor light, 9 instances as a result of brake failure and 6 instances of the accident was uncertain. The best decision tree result was obtained with Id3 with 165 instances, 120 instances were correctly classified and 45 instances were incorrectly classified, which represent 72.7273%, 27.2727% respectively. The mean absolute error is 0.0989, root mean squared error is 0.2235, relative absolute error is 49.0673 % and root relative squared error is 70.7409 %.

For the prone locations of the accident, the result indicates that, out of 165 instances of accident 84 occurred at location D, 59 accidents occurred at Location C, 16 accidents occurred at Location B and 6 accidents occurred at Location A. . The best decision tree result was obtained with Id3 with 165 instances, 133 instances were correctly classified and 32 instances were incorrectly classified, which represent 80.60661%, 19.3939% respectively. The mean absolute error was 0.0951, root mean squared error was 0.2231, relative absolute error was 31.3524 % and root relative squared error was 57.4578 %. For the prone times of the accident, the result also indicates that, out of 165 instances of accident 97 occurred in the evening, 51 accidents occurred in morning and 17 accidents occurred in afternoon. However, the best decision tree result was obtained with Id3 with 163 instances, where 127 instances were correctly classified and 38 instances were incorrectly classified, which represent 76.9697% and 23.0303 respectively. The mean absolute error is 0.1821, root mean squared error is 0.3036, relative absolute error is 49.5963 % and root relative squared error is 70.9891 %.

## 7. CONCLUSION

The historical data collected for the accidents occurred between January, 2014 to June, 2016 long Kano-Wudil Highway was analyzed using WEKA data mining software using Id3 decision tree and predicted the causes of the accident, its prone location and time. The result showed that mostly the cause of the accident is wrong overtaking, followed by loss of control, then tyre blowout, poor lights, uncertain causes and brake failure. The result indicated that, accident mostly occurred in location D, followed by Location C, then B and the mostly at evening time.

## 7. REFERENCES

- Chang, L. and Chen , W. (2005). Data mining of tree based models to analyze freeway accident frequency. *Journal of Safety Research* 36: 365- 375
- Data Mining - Decision Tree Induction (2016) accessed date: 8<sup>th</sup> July, 2016 [http://www.tutorialspoint.com/data\\_mining/dm\\_dti.htm](http://www.tutorialspoint.com/data_mining/dm_dti.htm)
- Dipo, T. A. and Akinbola, O. (2012). Using Data Mining Technique to Predict Cause of Accident and Accident Prone Locations on Highways. *American Journal of Database Theory and Application*, 1(3): 26-38
- Federal Road Safety Corp (2014). Traffic Digest Report on Road Traffic Crashes (RTC) Involving Buses on Nigerian Roads (2014)\
- Getnet, M. (2009). Applying data mining with decision tree and rule induction techniques to identify determinant factors of drivers and vehicles in support of reducing and controlling road traffic accidents: the case of Addis Ababa city. Addis Ababa Addis Ababa University
- Global status report on road safety: time for action, WHO, 2009.
- Han, J. and Kambe, M. (2001) Data mining Concepts and Techniques Morgan Kaufmam, *Academic Press*
- Vitus, N. U. (2014) Trends and patterns of fatal road accidents in Nigeria (2006- 2014) *IFRA-Nigeria working papers series*
- Olutayo, V. A. and Eludire, A. A. (2014). Traffic Accident Analysis Using Decision Trees and Neural Networks. *I.J. Information Technology and Computer Science*, 2014, 02, 22-28

**13<sup>th</sup>**

**International Conference**



**Session B:**

**Sustainable Healthcare for All**

---

## Full Paper

# AN ONTOLOGICAL-BASED KNOWLEDGE FRAMEWORK FOR DIAGNOSING BREAST CANCER

---

**O. N. Oyelade**

Department of Computer Science,  
Faculty of Physical Science,  
Ahmadu Bello University,  
Zaria.  
onoyelade@abu.edu.ng

**A. A. Obiniyi**

Department of Computer Science,  
Faculty of Physical Science,  
Ahmadu Bello University,  
Zaria.  
aaobiniyi@gmail.com

**S. B. Junaidu**

Department of Computer Science,  
Faculty of Physical Science,  
Ahmadu Bello University,  
Zaria.  
sahalu@abu.edu.ng

**ABSTRACT**

Clinical decision support systems are successfully being used to improve healthcare service delivery. However, these decision systems depend essentially on large clinical knowledgebase which contribute to the overall precision of diagnosis process. Hence, the need for efficient clinical knowledge engineering and management processes. Clinical knowledge engineering, though carried out through different approaches, involves knowledge creation, update and making it available for decision making. However, most of the approaches used in knowledge engineering process limit the resulting clinical knowledge developed. This dysfunctional knowledgebase are characterized by their incompleteness, not structurally designed, incorrect, difficult to update as knowledge evolves and inconsistent. Therefore, this paper takes on the approach of ontology as a pattern for modeling clinical knowledgebase particular for diagnosing breast cancer. This approach first categorizes breast cancer knowledgebase into two: taxonomy and causal knowledgebase. These are then further broken down into four levels of reasoning namely abduction, deduction, induction and abstraction. The resulting knowledge framework provides clinical decision support systems with a well-structured and tacit-enabled ontological knowledgebase. Finally, a demonstration of the use of our clinical knowledgebase, in reasoning with it at the four levels stated here are shown. Result shows that this ontological approach enhances decision making process by 20%, 5%, 16%, and 11% at the abstraction, abduction, deduction, and induction respectively as against the simplified clinical knowledgebase of diagnosis-symptom approach.

**Keywords:** Ontology, Inference making, Ontology, Clinical decision support systems, Clinical knowledgebase

## 1. INTRODUCTION

Knowledge representation is very important to support reasoning process, and it is a means that helps modelers represent the (medical) knowledge that enables the clinical decision support systems (CDSS) to deliver appropriate decision-support services during the care process (Peleg, 2006). Davis *et al.* (1993) suggest that knowledge representation plays five distinct roles it plays: fundamentally a surrogate, a set of ontological commitments, it is a fragmentary theory of intelligent reasoning, it is a medium for pragmatically efficient computation, and it is a medium of human expression. These roles enable the knowledge engineer to effectively model their knowledge framework.

Now, knowledge representation of medical knowledge is usually dynamic due to the evolving nature of medical knowledge. Moreover, another side to knowledge representation medicine is the representations of medical reasoning models Lucas (1993). These reasoning models enable CDSS to exploit the evolving medical knowledge in its framework, in providing appropriate recommendations.

On the other hand, ontology is a specification of a conceptualization Gruber (1993) which enables its designers to represent concepts in a given domain, and further shows the relationships among these concepts. A domain could be taken in life and modeled using the related data or concepts in that domain so that one could either keep for use. Ontology is represented using a specific language such as Web Ontology language (OWL), OWL2, Resource description framework (RDF) and RDF schema (RDFS). Some of these languages, particularly RDFS and OWL provide features that enable one to make inferences from the data retrieved. Domain of choice is usually selected when using ontology languages to model their concepts. These domain ranges from medicine, engineering, humanity, and other domains. However, modeling for medicine alone can result in a very large

knowledgebase and this could yield an inconsistent knowledge framework. Hence the position of this paper in modeling the knowledge framework for breast cancer diagnoses.

In this paper, an ontological knowledgebase framework is designed to take advantage of the inference making capability of ontologies, thereby, generating more findings or results during search over the knowledgebase.

## 2. RELATED WORKS

Fernando and Henskens, (2013) used the Select and Test (ST) medical reasoning model in diagnosing multiple ailments, but model the knowledgebase using the simplified clinical diagnosis-symptom approach – more like a tabular knowledge layout pattern. The author Chang *et al.* (2015) developed an ontology model based on the terminology used by medical experts to describe depression. This model was then used with a Bayesian network to infer the probability of a user/patient becoming depressed. Also, Wang and Tanzel (2013) designed a framework that utilizes ontological knowledge model for case base reasoning and a simplified ontology medical knowledge model. These models were used for pre-diagnosis when a patient's symptoms and signs are entered, and in pattern matching to gather candidate diseases in diagnosis, with case-based reasoning used to refine diagnostic decision. Reyes-Ortiz (2013), the author presented a computational model of representation of medical knowledge to support decision-making task during a medical consultation to reduce wrong diagnosis. This ontological model is able to infer a list of clinical diagnoses from the data of signs, symptoms, risk factors and medical background. Romero-Tris (2010) created three ontology-based tasks used in prospective and retrospective diagnoses. The health care ontology was built for the care of chronically ill patients that was created and validated in the k4care project.

### 3. THE PROPOSED ONTOLOGICAL KNOWLEDGE REPRESENTATION FRAMEWORK

Use In this section, the ontological knowledge framework is presented and a general explanation is given. Figure 1 illustrates the ontological knowledge framework that is being proposed in this paper. This framework, as earlier stated, is categorized into two: the taxonomy ontology and the causal ontologies. Within the first category (taxonomy), we have the abstraction reasoning knowledgebase, while the second category (causal) consists of the abduction, induction and deduction reasoning knowledgebase. The taxonomy ontology simply lists out acceptable and relevant medical terms used by oncologist specifically in breast cancer treatment. The ontology provides the reasoning and diagnosing process a database or thesaurus from which terms are chosen and verified as against patients/user's inputs.

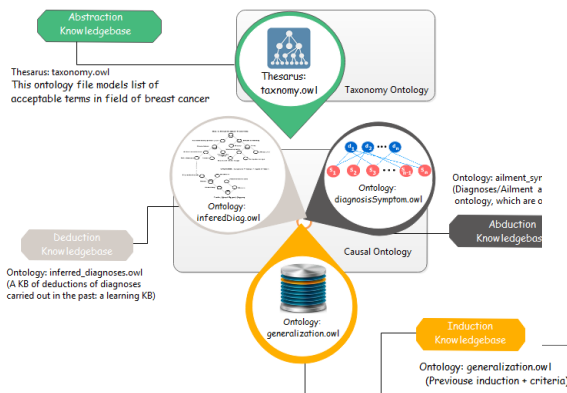


Figure 1: The Ontological Knowledge Framework for Breast Cancer Diagnosis

#### 3.1 Abstraction Reasoning Knowledge base

The national cancer institute (NCI) in America, has over the years maintained thesaurus for cancer. While carrying out this research, this thesaurus was downloaded and studied. This document, modeled with OWL DL sized up to about 320 MB. Findings on this thesaurus revealed that it combines all forms of cancer,

and all possible concepts in cancer. However, this research finds it too generic and large for consideration during implementation. Hence, this research embarks on following the clinical protocol of breast cancer in modeling a sizable thesaurus for this it. The acceptable medical terms for the domain of breast cancer were collected and formulated as a thesaurus. This thesaurus captured by Figures 2a, 2b, and 2c forms the taxonomy of medical terms used for accepting input into the other reasoning levels. Observe that the words or terms are arranged in a hierarchical pattern with some sub-classing some other terms. Note also that these super-classing and sub-classing relationship captured in the figures are model to illustrate the relation among such terms as they are used by oncologist.

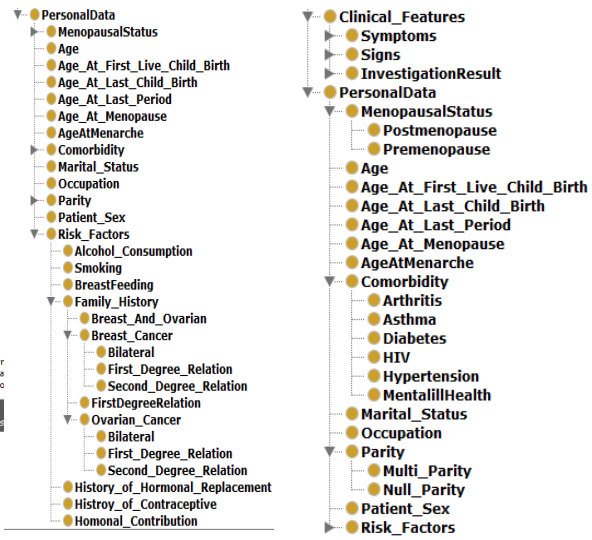


Figure 2a: Breast cancer thesaurus





from the GOAL to FACT end of vice versa. While the knowledgebase supplies the unconnected model, the deduction reasoning process provides the inference making process. While the FACT end models the known facts in diagnosing breast cancer, the GOAL end of the knowledgebase models basic inputs expected from user.

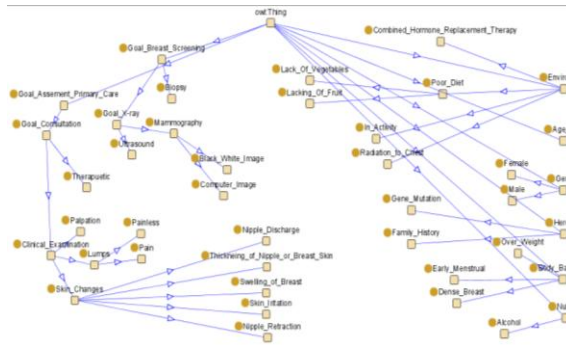


Figure 4: Deduction layer knowledge base

3.4 Induction Reasoning Knowledge base

The knowledge representation for reasoning at the induction level consists of knowledge of facts that models certain criteria that must be met to conclude on an ailment being the result of a diagnoses process. Here, the aim of inductive reasoning is to check if likely diagnoses (breast cancer) meets it diagnostic criteria as specified by patient/user input.

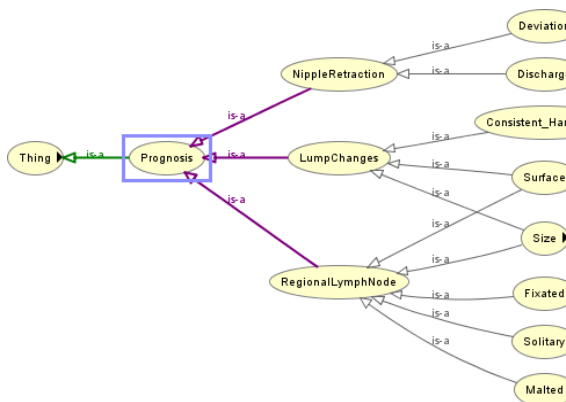


Figure 5a: The induction module knowledgebase

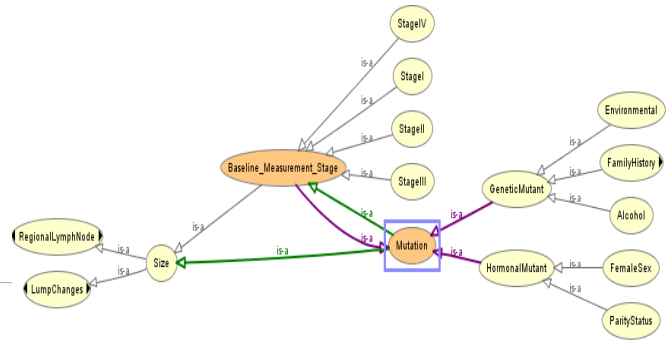


Figure 5b: The induction module knowledgebase

The knowledge base modeled for the induction reasoning level is broken down into two so as to detail the representation. This is why the representation is captured by Figures 5a and 5b. While that of Figure 5a shows the likely issues to be considered in obtaining the prognosis breast cancer, the knowledge representation of Figure 5b drills down into the hierarchy of the knowledge base. Figure 5a simply implies that to satisfy the criteria for breast cancer, it is necessary to consider some changes around the nipple, lump presence, and changes around the regional lymph nodes. Figure 5b for example models how changes in the sizes of both breast lump and regional lymph nodes can help in finding the diagnostic criteria. When the knowledge model in Figure 6b is further traced down the hierarchy, it will be observed that more facts links up parameters that sum into necessary criteria for making breast cancer diagnoses.

4. DISCUSSIONS

In this section, we shall be demonstrating how the levels of reasoning utilize our knowledge framework in carrying out diagnosis process. First we shall give sample patient’s input and then feed it into the four reasoning levels.

In Table 1, this paper presents three parameters for which the input into the four reasoning levels is collected. These parameters are: breast cancer risk factors, symptoms presented by patient, and investigations carried out by clinicians.

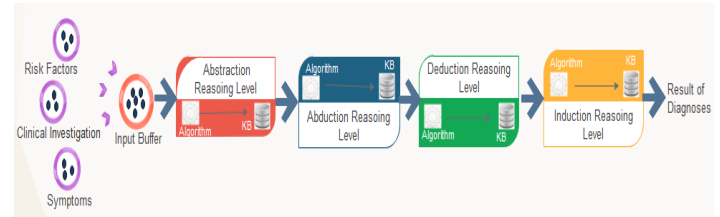
After all these parameters were presented/investigated from the patient, the assumed findings are tabularized correspondingly in Table 1. Hence, these findings are then used as sample input into the reasoning pattern for the four reasoning levels. The reasoning levels into logically mines out information from their corresponding knowledgebase model in the proposed framework of this paper.

**Table 1: Sample Patient Input**

S/N	Parameter	Findings by both Patient & Medical Personnel
1	<b>Risk Factors</b>	Female, Family history of ovarian cancer, Age at first live child birth was 17yrs old.
2	<b>Symptoms</b>	One Breast Lump felt, Nipple Discharge, Nipple Retraction/Deviation, Breast swelling, Breast Pain, Generalized body weakness, and Cough
3	<b>Clinical Investigations</b>	Biopsy: malignant growth,

Figure 6 demonstrates the interface that exists between each reasoning level and its data source. First, the values from the three parameters are combined into one general input buffer and then sent into the abstraction reasoning level. The aim of the reasoning task carried out at this level is to check if the expected symptoms in likely diagnoses are found in patient as presented in the input buffer. However, the reasoning process at this level must check the input from the buffer against the related list of acceptable terms modeled in Figure 2. Once these terms are reconciled and inferred to appear in the list of acceptable medical terms for this domain, then the list of this symptoms are sent as input into the next

level of reasoning – Abduction reasoning level.



**Figure 6:** An illustration of interfacing the reasoning levels and their corresponding knowledge base

The purpose of the abduction reason level is to get all the diagnoses related to symptoms which were the output of the abstraction reasoning level. As the abduction reasoning level produces an explanation that best accounts for the patient’s symptoms, it relies solely on its ontology knowledgebase modeled in Figure 3. The goal of this level is to arrive at the correct diagnosis for a given patient. Next now are the internal workings of the Deduction reasoning level with relation to its knowledgebase modeled in Figure 4. Here, the reasoning patterns aims to get all the symptoms related to diagnoses selected at the abduction level. Though a list of diagnosis is generated from the last level, but for the sake of brevity, the knowledgebase is only structured to project the symptoms of breast cancer only. Lastly, we have the induction reasoning level which checks if the likely diagnoses meet their diagnostic criteria. The diagnostic criteria of breast cancer have being modeled in Figure 5a and 5b. Hence, the knowledgebase is searched to authenticate that the criteria of diagnosing breast cancer are satisfied by the input from the previous reasoning level – Deduction reasoning level.

**5. RESULT PRESENTATION**

In this section, we compare the impact of the knowledge framework proposed in this paper as against the knowledge framework of [5]. They represented all the relationship between diagnoses and symptoms in their

knowledgebase as sets  $D = \{d_1, d_2, \dots, d_n\}$  and  $S = \{S_1, S_2, \dots, S_m\}$  respectively. This relationship was represented in a two-layer graph, in which each arc connecting every  $d_i$  and  $S_j$  is associated with a value  $L_{ij}$  representing the likelihood ( $L$ ) that  $S_i$  implies  $d_j$ . Though our approach established such a relationship, however our knowledge engineering models the knowledgebase in a way to support inference making. By this we mean that given  $S_j$  and  $d_i$  to be connected not just by a likelihood value only, but also by an association or relation. For example, in Figure 2, it was model that **ClinicalFeatures** is a term which in turn has subclasses **Signs** and **Symptoms**. And these subclasses also have their own subclasses. Hence, the relation or association that exist between **ClinicalFeatures** and **Signs** is called *isa* (inheritance relation), and the same relation or association exists between **ClinicalFeatures** and **Symptoms**. So, by inference finding a **Sign** could infer finding and **Symptom**. The ontological inference

capability is exploited in generating reasonable and sufficient result during search of any information in our knowledge framework.

Therefore, the mathematical representation of our knowledge framework compare to that of [5] will be given as  $D = \{D_1, D_2, \dots, D_n\}$ ,  $S = \{S_1, S_2, \dots, S_m\}$ , and  $R = \{R_1, R_2, \dots, R_n\}$  such that an instance  $I = D \otimes S$ . where given any  $d_i$  to have a relation  $R_k$  with  $S_j$ , then because of the relation  $R_k$ ,  $D_i$  by inference, may now have more symptoms ( $s$ ) to point to apart from  $S_j$  alone. And by this inference capability, enough information is gathered with one query or search on the knowledgebase.

When the knowledgebase at each levels reasoning are represented using the likelihood value alone as in [5], we plot the hit value of our work against their knowledge representation pattern with this result of this research and presented them with Table 2 and the graphs in Figures 7a-d.

Table 2: A Comparison of hit value over two knowledge frameworks

Level-Based Knowledgebase	Abstraction		Abduction		Deduction		Induction	
	This Paper	Fernando (2013)	This Paper	Fernando (2013)	This Paper	Fernando (2013)	This Paper	Fernando (2013)
Comparison								
Hit Values	38	7	18	13	41	20	11	3
Sample Queries	Types of clinical features		List of breast cancer symptoms		Facts about breast cancer		Breast cancer diagnosing criteria	

In the graphs below, the results of the comparison with another representation of knowledge at these four different reasoning levels are shown.

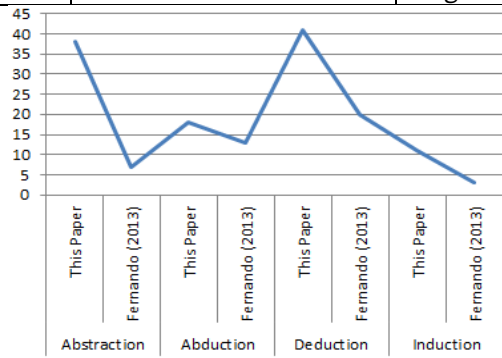


Figure 7a: Reasoning levels knowledgebase representation of the comparison table

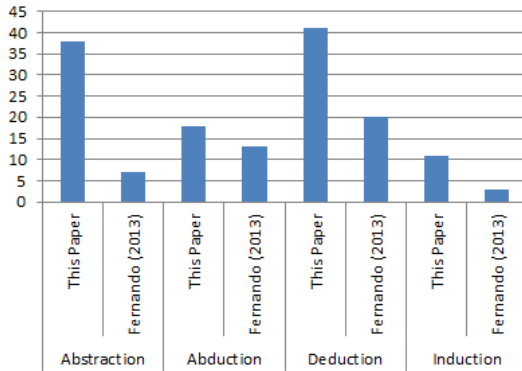


Figure 7b: Bar chart representation of the comparison table

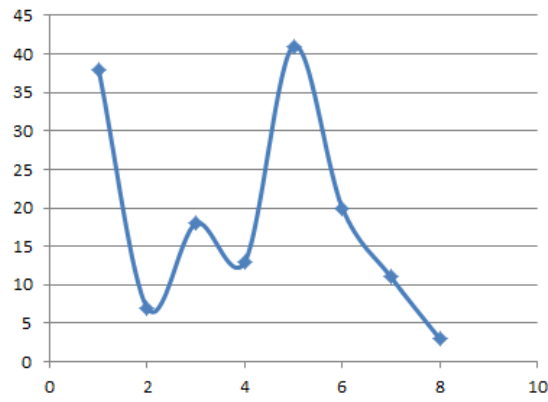


Figure 7c: Representation of the comparison table

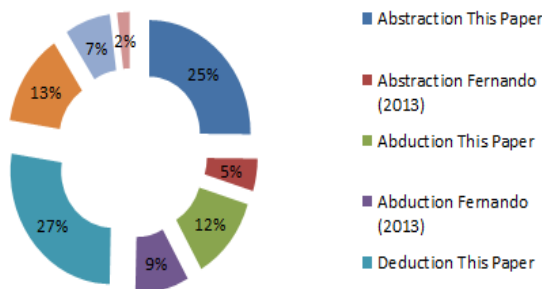


Figure 7d: Percentile representation of the comparison table

6. CONCLUSION

In conclusion, in this paper, an ontological knowledge framework for supporting reasoning process of diagnosing breast cancer at four different levels of reasoning

was presented. These four levels of reasoning are abstraction, abduction, deduction, and induction. Furthermore, the communication pattern between the knowledgebase of each of these reasoning levels are was also illustrated and discussed. And finally, a comparison was made against another knowledge representation at these four levels by other authors and result shows that an ontological representation gave an enhancement of 20%, 5%, 16%, and 11% at the abstraction, abduction, deduction, and induction respectively.

7. REFERENCES

Fernando, I. and Henskens, F. A. (2013). ST Algorithm for Medical Diagnostic Reasoning. R. 23. *Polibits* (48) 2013. ISSN 1870-9044; pp. 23-29.

Chang Y. S., Fanb C. T., Loc W. T., Hunga W. C., Yuanb S. M. (2015). Mobile cloud-based depression diagnosis using an ontology and a Bayesian network. *Future Generation Computer Systems* 43-44(2015)87-98

Wang H. T. and Tanzel A. U. (2013). Composite Ontology-Based Medical Diagnosis Decision Support System Framework. *Communications of the IIMA*, Volume 13 Issue 2.

Reyes-Ortiz A. J., Jiménez A. L., Cater J., Meléndez C. A., Márquez B. P., García M. (2013). Ontology-based Knowledge Representation for Supporting Medical Decisions. *Research in Computing Science* 68 (2013). pp. 127-136

Romero-Tris C., Riaño D., Real F. (2010). Ontology-based retrospective and prospective diagnosis and medical knowledge personalization. *Proceeding KR4HC'10 Proceedings*

- of the ECAI 2010 conference on Knowledge representation for health-care, pages 1-15
- Peleg M. Decision support, knowledge representation and management in medicine. *Yearb Med Inform* 2006;45:72-80. PMID:17051298.
- Davis R., Shrobe H., and Szolovits P. (1993). What is a Knowledge Representation? *AI Magazine*, 14(1):17-33, 1993.
- Gruber, T., (1993). A Translation Approach to Portable Ontologies, *Knowledge Acquisition*, Vol. 5, No. 2, pp. 199-220.
- Lucas P. (1993). The Representation of Medical Reasoning Models in Resolution-based Theorem Provers. *Artificial Intelligence in Medicine*, 5(5), 395{414,1993}

---

## Full Paper

# DESIGN OF A FRAMEWORK FOR HEALTHCARE CRIME INVESTIGATION USING BIG DATA ANALYTICS

---

### S. T. Yange

Department of  
Mathematics/Statistics/Computer Science,  
University of Agriculture, Makurdi, Nigeria  
lordesty2k7@gmail.com

### H. A. Soriyan

Department of Computer Science and  
Engineering,  
Obafemi Awolowo University,  
Ile-Ife, Nigeria.

### ABSTRACT

One major challenge encountered during crime investigation via automated systems is the inability of conventional data analysis techniques to adequately handle the enormous data that are made available during the investigation. Existing crime investigation frameworks are built on orthodox data analysis techniques which cannot sufficiently manage the unprecedented size and variety of data available today, not to mention the significantly more anticipated data in the near future. This has affected the healthcare industry where data is predominantly multi-structured and is growing at a considerably faster rate. To address this, a big data analytics model based on deep learning was designed using enterprise application diagrams. This model is intended to be implemented using Apache Hadoop, a big data implementation framework. This model creates a platform that handles a phenomenon affecting millions of people world-wide. It provides security intelligence by shortening the time of correlating and deriving evidence from large volume of data during healthcare crime investigation. Finally, this research also enabled the healthcare systems to systematically use big data analytics to identify inefficiencies and best practices that improved care delivery and reduce costs.

**Keywords:** Crime, Hadoop, Deep Learning, Investigation Data Analytics, Health Insurance

## 1 INTRODUCTION

Big data usually includes datasets whose sizes are beyond using conventional data analysis tools to manage. The analysis of big data commonly known as big data analytics is the process of collecting, organizing and analysing large, diverse dataset that involves different types such as structured and unstructured, and streaming and batch, with sizes from terabytes to zettabytes to discover patterns and other useful information (Ularu *et al.*, 2012). Big data analytics can be applied in information security which involves the ability to gather massive amounts of digital information to analyse, visualize and draw insights that can make it possible to detect crime. This can transform security analytics by improving the maintenance, storage and analysis of security information. Big data analytics correlate the data drawn from multiple sources such as network traffic, log files, financial transactions, healthcare claims etc. into a coherent view so as to identify the anomalies and suspicious activities of the criminals. Big data is ideal for investigating information security issues; and detecting a crime is largely about uncovering data patterns that are not ordinary from log files. Applying big data techniques will ease such analysis to reveal anomalies that point to a data breach. With this powerful strength, big data analytics could lead to the discovery of big crime which invariably could culminate into 'big arrest'.

Crime encompasses a wide range of illicit practices and illegal acts involving intentional deception or misrepresentation. Crime is any illegal act characterized by deceit, concealment, or violation of trust (IACA, 2014). These acts are not dependent upon the threat of violence or physical force. Crimes are perpetrated by parties and organizations to obtain money, property, or services; to avoid payment or loss of services; or to secure personal or business advantages." In other words, crime is a harmful act or omission against the public which the society wishes to prevent and which, upon conviction, is punishable by fine, imprisonment, and/or death. No conduct constitutes a crime unless it is declared criminal in the laws of the country (Yunusa *et al.*, 2014; Dutta and Hongoro, 2013). Some crimes (such as theft or criminal damage) may also be civil wrongs (torts) for which the victim(s) may claim damages in compensation. Crime and fraud are synonymous, therefore in this research, the two words will be used interchangeably.

Crime impacts organizations negatively in several areas including financial, operational, and psychological. While the financial loss owing to crime is significant, the full impact of crime on an

organization can be overwhelming. The losses to reputation, goodwill, and customer relations can be also devastating. The society is strongly affected by crime, both due to the cost of crime, as well as the decline in the quality of life that citizens suffer as a consequence of crime. As crime can be perpetrated by any employee within an organization or by those from the outside, it is important to have an effective crime management programme in place to safeguard the organization's assets and reputation. Crime and society are closely linked-for better and for worse and is as old as humanity, and occurs in different degrees of severity. However, society can also play a role in reducing and deterring crime. Many agencies and programmes in crime management are based on societal and community efforts. The magnitude of criminal activities can be perceived in all spheres of life (IACA, 2014).

For instance, the healthcare sector is among the most information intensive industries. Its information, knowledge and data keep growing on a daily basis and the ability to extract useful information that will improve the quality of healthcare services rendered is very crucial. Crime in this sector involve the intentional deception or misrepresentation for gaining some shabby benefits in the form of health expenditures (Dutta and Hongoro, 2013). This can be anything like providing false and intentionally misleading statements to patients, submitting false bills or claims for services, falsifying medical records or reports, lying about credentials or qualifications, unnecessary medical treatment or drug prescription; which seriously drain the finances in the healthcare system. This severely deters the healthcare industry from providing quality and safe care to legitimate patients; and it has called for an effective crime management system so as to reduce this illegal behaviour with the intention of improving the quality and reducing the cost of healthcare services. Owing to the large number of cases reported, investigated and prosecuted, it has been identified as a "high-risk" area in many regions such as the UK, the US, Romania, Nigeria etc. (Dutta and Hongoro, 2013).

Healthcare crime exist in many forms: dishonest providers, organized criminals, collusion with patients, and patients who misrepresent their eligibility for health insurance coverage. It can be categorized into: health insurance crime, drug crime and medical crime. Due to the confidentiality of the medical records, data for healthcare crime comes mostly from health insurance crime; and it occurs when a company or an individual defrauds an insurer or government healthcare programme. In



this paper, a survey of the existing healthcare investigation approaches is carried out and a new approach is designed.

## 2 RELATED WORKS

Crime in the healthcare insurance involve three parties (Dora and Sekharan, 2015; Dutta and Hongoro, 2013): healthcare service provider (i.e., the physician, pharmacist, laboratory scientist, healthcare centre, pharmacy, laboratory, and even ambulance companies) which render healthcare services; healthcare service consumer or beneficiary or insurance subscriber (i.e., patient) which receive healthcare service from the provider; and the healthcare insurance carrier which receive regular premiums from subscribers and make the commitment to pay healthcare cost on behalf of beneficiaries. These parties exchange information amongst them in the course of care delivery. This is basically in the form of service requested by the subscriber (patient visit) to the provider, explanation of benefits which contain the detail services rendered by the provider to the subscriber, claim/bill which is sent to the carrier for the services rendered to the subscriber by the provider, and the payment to the provider based on the claim submitted to the carrier.

Among these three types of fraud, the one committed by health service provider's accounts for the greatest proportion of the total healthcare fraud. Although a vast majority of service providers are honest and ethical, but a few dishonest ones may have various possible ways to commit fraud on a very broad scale, thus posing great damage to the health care system. Some service providers' fraud, such as that involving medical transportation, surgeries, invasive testing, and certain drug therapies, even places patients at a high physical risk.

As the number of beneficiaries (patients) of this scheme increases, high volume of data is generated by both the providers and the carriers. Consequently, some fraudulent activities (such as billing services that were never rendered, performing medically unnecessary services, misrepresenting non-covered treatments as medically necessary covered treatments, and misrepresenting applications for obtaining lower premium rate) are carried by these actors (beneficiary, provider and insurer). This give rise to the need to investigate such acts in an attempt to identify perpetrators, and this requires a proper analytics tool (Dutta and Hongoro, 2013; Li *et al.*, 2008).

Recent development of new technologies eased production, collection and storage of high dimensional and complex data. Healthcare has been no exception. Modern medicine generates a great deal of data which is stored in medical databases. Medical databases are increasing in size in three ways (Li *et al.*, 2008): the number of records in the database, the number of fields or attributes associated with a record, and the complexity of the data itself. Extracting pertinent information from such complex databases for inferring potential fraudulent activities has become increasingly important for fraud detection. Dutta and Hongoro (2013) gives an account of the amount of information involved in the reimbursement process for healthcare insurance scheme, which supports the cost of prescription medications to seniors and the disabled in the US. In such a complex process, involving many actors, the possibility of fraud cannot be overlooked. At the same time, quality of medical records should be ensured to avoid, for instance, fraudulent claims.

With these enormous amount of data that is generated in the healthcare industry, traditional methods of detecting healthcare crime are time-consuming and inefficient due to the complicated nature of medical processes and the complexity of the data have made crime to have a favourable niche in the healthcare systems as most crimes go undetected. Conventional analysis methods are not suitable due to the limitations to manage the volume, velocity, variety, veracity, value and complexity of the data in the healthcare (Dora and Sekharan, 2015; Ekin *et al.*, 2013; Bagul *et al.*, 2016; Bagde and Chaudhari, 2016; Fashoto *et al.*, 2013; Jacquelin, and Shrijina, 2013).

### 2.1 Healthcare Crime Investigation Approaches

Jacquelin and Shrijina (2013), proposed the use of cluster analysis for geographical investigation of potential fraud. The emphasis of the work was on types of fraud committed by a single party. As pointed out by Dora and Sekharan (2015) some frauds in the health insurance involves more than one party: conspiracy or conspiratorial frauds. A typical conspiratorial fraud scenario is that patients collude with physicians to fabricate medical service and transition records to deceive the insurance company to whom they subscribed. This can be very rewarding owing to its complexity, increasing popularity, and severe consequences.

Fashoto *et al.* (2013) develop a data mining technique for fraud detection in health insurance scheme using knee-point k-means algorithm. The research considered the National Health Insurance Scheme (NHIS) as the case study. The work, focuses

on the application of some computer-based techniques that could help to properly target investment in the healthcare sector and also reduce health insurance fraud by healthcare providers. To this effect, the knee-point k-means clustering method was employed, which was capable of detecting fraudulent claims by health service providers. Cluster-based outliers were examined. Health providers' claims submitted to a health maintenance organisation (HMO) were grouped into clusters. Claims with similar characteristics were grouped together. The claims were grouped into two clusters: fraudulent and non-fraudulent. This research did not classify the fraud detected, whether it was provider, consumer or insurer frauds; it used only the unsupervised technique (K-Means algorithm) for clustering; and the data was collected from only one HMO which cannot yield a perfect result.

In a survey on hybrid approaches for fraud detection in health insurance by Ekin *et al.* (2013), the act committed with the intent to obtain a fraudulent outcome from an insurance process was carefully examined. According to the researchers, when a claimant attempts to obtain some benefits or advantages to which they are not entitled to. A hybrid framework that applied some data mining techniques to detect frauds was proposed. This framework considered the analysis of the characteristics of healthcare insurance data, some preliminary knowledge of healthcare system and the fraudulent behaviours. The framework harnessed the advantages of both the supervised and unsupervised learning techniques to detect fraudulent claims. This framework did not consider the high volume, velocity, variety, veracity etc. of data and it was not implemented. In the same vein, Travaille (2011), investigated the benefits of big data technology and the main methods of analysis that can be applied to the cases of fraud detection in public health insurance system in Romania. The research outlined the benefits of using big data technology in combating crime in the healthcare industry.

Consequently, methods for identifying and preventing fraud must always be adjusted and ready to rediscover the fraudulent actions (Yunusa *et al.*, 2014; Dutta and Hongoro, 2013). To add to the lapses in legislation of the country, each country has unique economic, political, social, and institutional opportunities for and barriers which makes fraud examination different amongst countries. A crucial and peculiar issue in the Nigerian National Health Insurance Scheme is the high level of corruption in

the sector, lack of accountability and clear sense of irresponsibility (Etoebe and Etoebe, 2010).

Bologa *et al.* (2010), proposed a model using big data in investigating real time crime in the health insurance in the cloud. This approach utilizes fraud management solution to detect potential frauds in the cloud. The solution was based on a high volume of historical data, predictive statistical models and social media analytics. It renders its services through client components like apps and web-services. Just like the Ekin *et al.* (2013), Travaille *et al.* (2011) and Bologa *et al.* (2010) did not implement any working system for the research. Also, as opined by Ekin *et al.* (2013), healthcare crimes are country specific and Nigeria has not adopted the cloud services, and there are no healthcare laws relating to the data on the cloud, this model cannot be used to investigate crime in Nigerian healthcare system.

Dutta and Hongoro (2013) in a bid to address this issues provides an approach to detect and predict potential frauds by applying big data, Hadoop environment and analytics methods which led to rapid detection of claim anomalies. The solution was based on a high volume of historical data from various insurance company data and hospital data of a specific geographical area. With the voluminous, diverse, and varying nature of the data used, the distributed and parallel computing tools collectively termed big data were employed. The work demonstrated the effectiveness and efficiency of the open-source predictive modelling framework used to describe the results from predictive model. The research was able to detect erroneous or suspicious records in submitted healthcare datasets and proved how the hospital and other healthcare data are helpful for the detecting healthcare insurance fraud. The research also used the decision tree algorithm.

In Konasani *et al.* (2012), a fraud detection approach in the health insurance using data mining techniques was developed. This approach used SVM (Support Vector Machine) and Evolving Clustering Method (ECM) in health insurance field for fraud detection. In the research, SVM algorithm was used for classification and ECM algorithm was used for clustering. The SVM was used to train the system to determine decision boundary between legitimate and fraudulent claims classes while the ECM identifies new data point that comes in, it clusters them by modifying the position and size of the cluster (i.e., used to cluster dynamic data hence it find out newly incoming fraudulent claims).

Rawte and Anuradha (2015) developed a model for detecting healthcare fraud and abuse using the

supervised and unsupervised data mining techniques. According to them, the supervised methods applied to healthcare fraud and abuse detection are decision tree, neural networks, genetic algorithms and Support Vector Machine (SVM); while the unsupervised methods that have been applied to health care fraud and abuse are clustering, outlier detection and association rules. This paper concluded that outlier detection is an unsupervised method and routine online processing task as supervised learning method.

In our research, we considered the further classification of fraud so as to report the actual fraud. We also considered deep learning which combined both supervised and unsupervised techniques, since hybrid methods are proven to yield better results (Ekin et al., 2013; Konasani et al. 2012; Rawte and Anuradha, 2015; Joudaki et al., 2015). We collected data from different stakeholders in the healthcare insurance relating to fraud. This research also used big data analytics techniques to carry out the investigation since the anticipated data is very large in size.

### 2.2 Nigerian National Health Insurance Scheme

A beneficiary got enrolled in the NHIS programme after an eligibility screening is performed by the NHIS. The person's circumstances and income is verified and if the criteria are met, the person can enrol in the NHIS programme. When the beneficiary is ill, he/she will go to the hospital where he/she will get the necessary treatment. The provider of the services provided to the beneficiary would submit the bill to the HMO in the form of claims. It is generally assumed that all of the providers have an agreement with NHIS and that they participate in the programme. The providers send the claim to the HMO which is reviewed and processed for the payment of the services of participating providers to the beneficiaries (Etobe and Etobe, 2010).

### 3 METHODOLOGY

This paper employed the hierarchical structure of the deep learning (deep belief network) architecture. The deep belief network learns high level representations, and complicated structure automatically from complex health insurance data, collected from the various health facilities. This data is stored in the Hadoop Distributed File System. The MapReduce framework provide algorithms that will be implemented for the distributed processing of the data. It provides a framework that allows distributed processing of complex datasets using simple programming models. The multiple layers of this architecture continuously abstract features of the data undergoing processing from one layer to

another, making the search for fraudulent claims simpler. The pictorial view of this model was designed with enterprise application diagram. It uses the Nigerian Health Insurance Scheme (NHIS) as a test bed.

### 4 RESULTS

This paper studies how the big data framework can be leveraged to extract, preprocess and analyse data from the NHIS with the aim of identifying fraud. Apache Hadoop is the Big Data framework considered. The Hadoop Distributed File System (HDFS) implementation of the Hadoop is used as an alternative to store data and the MapReduce to process extremely large data sets on commodity hardware. In addition, the research used Hive as an open-source data warehousing solution which is built on top of Hadoop. The design of the system is shown in Figures 1 and 2.

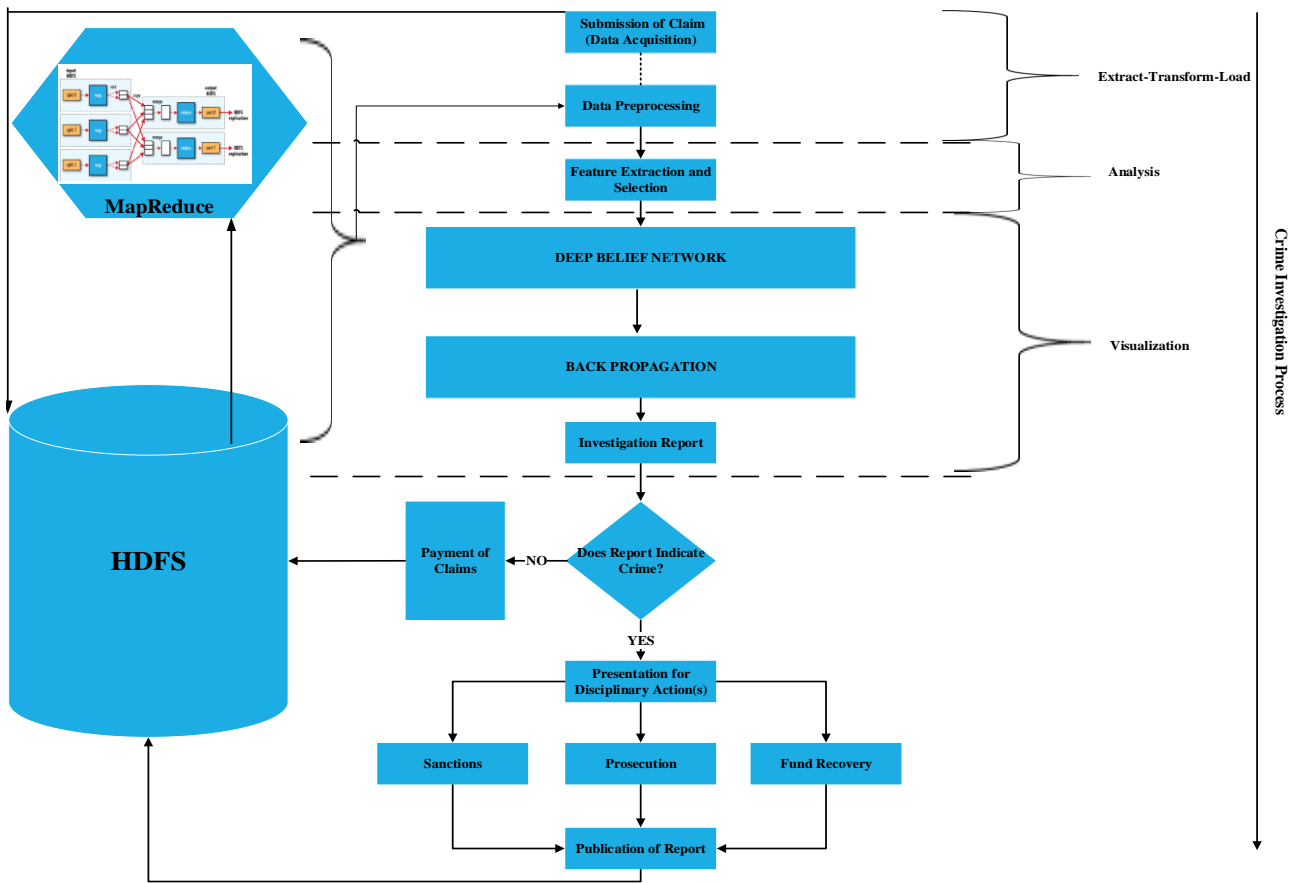


Figure 1: Schematic View of the Model

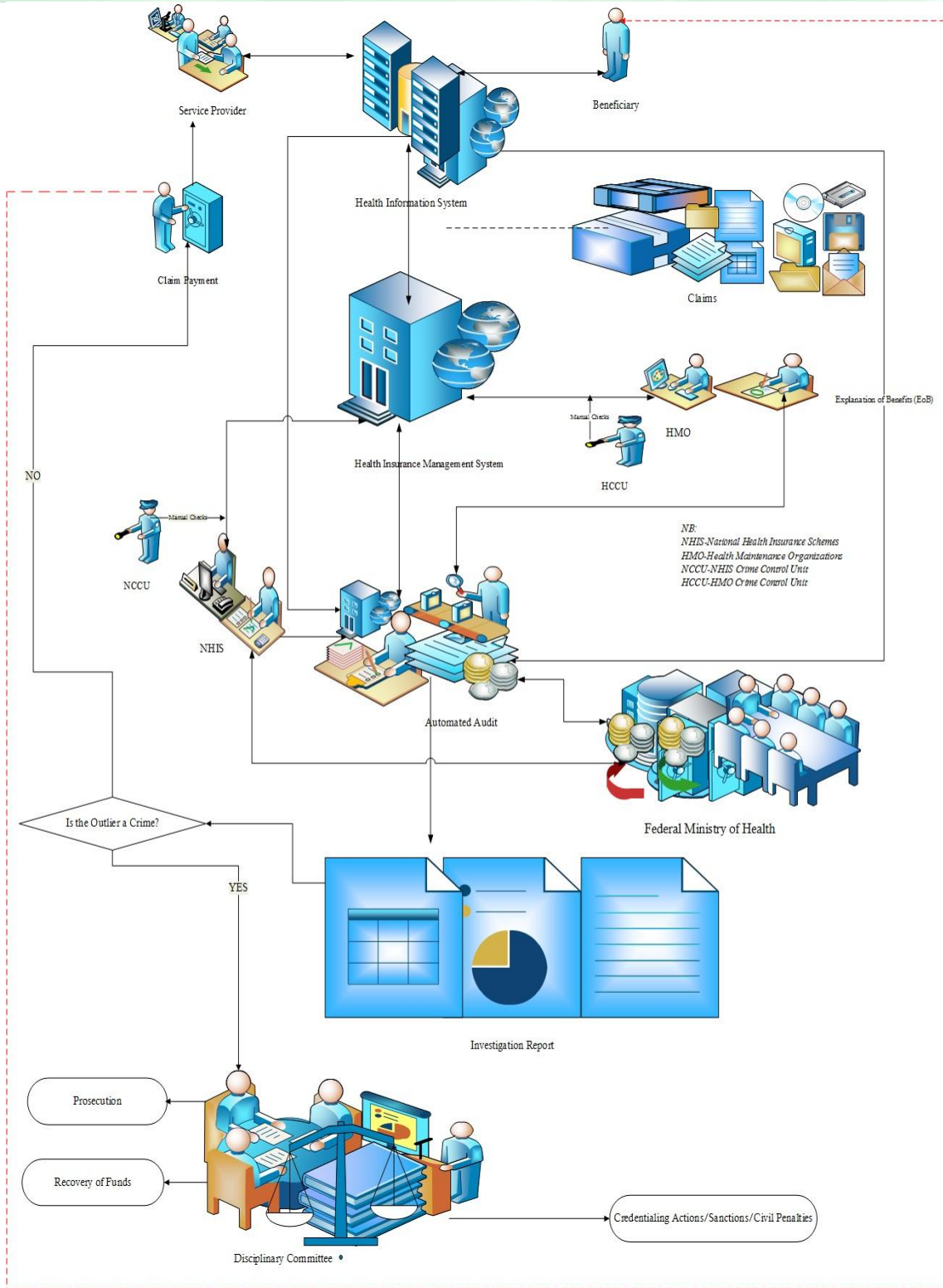


Figure 2: Model Flow

Hive supports queries expressed in a SQL like declarative language-HiveQL. RHadoop is a bridge between R, a language and environment to statistically explore data sets, and Hadoop, a

framework that allows for the distributed processing of large data sets across clusters of computers. The conceptual view of the framework is shown in Figure 1 and Figure 2. The model took

into cognizance the high volume of data from this sector.

## 5 DISCUSSION

The model designed in this paper, harnessed the big data's capability for crime investigation. It is of great value especially in the healthcare insurance scheme. It can investigate breaches in security, determine compliance with established policies and operational procedures, and enable the reconstruction of sequences of events affecting the healthcare insurance domain to enable auditors do their work efficiently. It considered the volume and complexity of this data which made it impossible for humans and other traditional means to be sufficient enough to identify the crime perpetrated by the hoodlums in the healthcare insurance industry. The deep belief network algorithm of deep learning was used to "learn" normal activities so as to fish out any unwholesome activity in the healthcare insurance. The model created room to capture and process the data, help to visualize its flow and apply automatic learning techniques capable of discovering patterns and detecting anomalies for proper investigation of the activities of fraudsters. With this, common repetitive errors that are "hidden" inside huge repositories of data which would go undetected in the absence of big data technologies because the orthodox techniques not being capable to correlate the huge quantities of data available in the medical sector could easily be identified and corrected.

The deep learning architecture combined two machine learning theories: unsupervised and supervised theories, one is used in the pre-training while the other is used in fine-tuning the network. In the pre-training of a deep belief network, the unsupervised learning theory is used. This is aimed at finding clusters of similar inputs in data without being explicitly told that these data points belong to a different class. With the aid of this theory, unlabelled NHIS data was used to initialize the network in the pretraining phase. The supervised theory is aimed at classifying inputs data with the aid of the target output. This theory implements the Back Propagation (BP) Algorithm which expressed the logic behind it. The idea behind BP algorithm is quite simple, output of the network is evaluated against desired output. If results are not satisfactory, connection (weights) between layers are modified and the process is repeated again and again until the error is small enough to be ignored. This theory aided the fine-tuning of all the weights and biases in the network pretrained by the unsupervised theory.

As opined by (Rawte and Anuradha, 2015; Dutta and Hongoro, 2013; Ekin et al., 2013) and other researchers, this combined hybrid approach in the deep learning makes it easier to detect erroneous or suspicious records in submitted healthcare datasets and proved how the hospital and other healthcare data are helpful for the detecting healthcare insurance fraud.

Components of the system are discussed below.

**Hadoop Framework:** This is the combination of the MapReduce and the Hadoop Distributed File System (HDFS). This is the backbone of this model. It enables distributed parallel processing of huge amounts of data across inexpensive, industry-standard servers that both store and process the data, and can scale without limits. It can handle all types of data from disparate systems: structured, unstructured, log files, pictures, audio files, communications records, email - regardless of its native format. Even when different types of data have been stored in unrelated systems, it is possible to store it all into Hadoop cluster with no prior need for a schema.

**Data Acquisition and Preprocessing:** Real world data that is collected from different sources is noisy and heterogeneous (different format) in nature. The heterogeneity in the healthcare data is responsible for the prevalence of missing values and inconsistencies which poses a great challenge leading to an inaccurate result if not addressed at the beginning. Raw data must be processed (this task is associated with segmentation, normalization and noise removal algorithms) into a form that is acceptable. This helps in constructing the homogeneous data set.

**Analysis:** This step aims to define new features out of the original attributes, to maximize the discrimination power of the machine learning method in separating fraudulent and legitimate cases. The feature extraction and selection procedures aim at finding the minimum number of discriminative features that are considered. The term feature selection refers to algorithms that select the best subset of the input feature set, whereas methods that create new features based on transformations or combinations of the original feature set are called feature extraction algorithms. Feature extraction takes in a pattern and produces features values. Feature extraction may provide a better discriminative ability than the best subset of given features, but these new features (a linear or a nonlinear combination of given features) may not have a clear physical meaning. Health insurance is a complex phenomenon governed by multiple

variables because there is no universal factor that can be used to predict the fraud.

**Visualization:** Visualization present large volumes of data, provide interactivity to explore the data, make visual patterns easy to see and make multivariate analysis simple and easy to comprehend.

**Other Components:** The investigation report is produced after this stage. If the report does not indicate any fraudulent act, payment is made, but if it is otherwise, the report is presented for disciplinary action (sanction, prosecution in the court of law and fund recovery) from the disciplinary committee. This stage marks the end of the investigation and everything is recorded in the database. Depending on the success of an investigation, further action may result in the form of civil cases, criminal prosecutions or both. Such actions can result in monetary penalties (fines), recoveries of funds (belonging to the public trust or to private payers) and industry sanctions (such as revocation of privileges or prohibition against administering services to NHIS patients) against offenders.

## 6 CONCLUSION

This research designed a model that automatically identify the general patterns of suspicious behaviour of criminals in the healthcare insurance claims. It employed advance data analysis techniques of big data which can aid auditors in investigating breaches in security, determine compliance with established policies and operational procedures, and enable the reconstruction of sequences of events affecting the healthcare insurance domain. This brought humans and computers to collaborate and work closely together in crime investigation. With this, it will reduce the time it takes to uncover fraudulent activity, and also shrink the negative impact of significant losses owing to fraud. It has created a platform that will handle a phenomenon that is affecting millions of people all over the world. This is the first of its kind to use big data analytics techniques in healthcare crime investigation in Nigeria. The work provided security intelligence, by shortening the time of correlating and deriving evidence from large volume of data, for healthcare crime investigation purposes. Finally, this research also enabled the healthcare systems to systematically use big data analytics to identify inefficiencies and best practices that improve care delivery and reduce costs.

## 7 REFERENCES

Jacquin, M.J. and Shrijina, S. 2013.

Implementation of Data Mining in

- Agba, A M.O. et al. 2010. National Health Insurance Scheme (NHIS) and Employees' Access to Healthcare Services in Cross River State, Nigeria. *Global Journal of Human Social Science*, Vol. 10, No. 7, pp. 9-16.
- Bagde, P.R. and Chaudhari, M.S. 2016. Analysis of Fraud Detection Mechanism in Health Insurance Using Statistical Data Mining Techniques, *International Journal of Computer Science and Information Technologies*, Vol. 7, No. 2, pp. 925-927.
- Bagul, P.D. et al. 2016. Survey on Hybrid Approach for Fraud Detection in Health Insurance. *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 4, No. 4, pp. 6918-6922.
- Bologa, A. et al. 2010. Big Data and Specific Analysis Methods for Insurance Fraud Detection. *Database Systems Journal*, Vol. 1, No. 1, pp. 30-39.
- Dora, P. and Sekharan, G.H. 2015. Healthcare Insurance Fraud Detection Leveraging Big Data Analytics. *International Journal of Science and Research*, Vol. 4, No. 4, pp. 2073-2076.
- Dutta, A. and Hongoro. C. 2013. Scaling Up National Health Insurance in Nigeria: Learning from Case Studies of India, Colombia, and Thailand. Washington, DC: Futures Group, Health Policy Project.
- Ekin, T. et al. 2013. Applications of bayesian methods in detection of healthcare frauds. *Chemical Engineering Transactions*, Vol. 33, pp. 151-156.
- Etobe, E.I. and Etobe, U.E. (2015). The National Health Insurance Scheme and Its Implication for Elderly Care in Nigeria. *International Journal of Science and Research (IJSR)*, Vol. 4, No. 2, pp. 128-132.
- Fashoto, S.G. et al. 2013. Application of Data Mining Technique for Fraud Detection in Health Insurance Scheme Using Knee-Point K-Means Algorithm. *Australian Journal of Basic and Applied Sciences*, Vol. 7, No. 8, pp. 140-144.
- IACA (International Association of Crime Analysts) 2014. Definition and Types of Crime Analysis. White Paper released by Standards, Methods, & Technology (SMT) Committee.

- Medical Fraud Detection. *International Journal of Computer Applications*, Vol. 69, No. 5, pp. 1-4.
- Joudaki, H. et al. 2015. Using Data Mining to Detect Health Care Fraud and Abuse, *Global Journal of Health Science*, Vol. 7, No. 1, pp. 1-10.
- Konasani, V. et al. 2012. *Healthcare Fraud Management using Big Data Analytics*. An Unpublished Report by Trendwise Analytics, Bangalore, India.
- Li, J. et al. 2008. A survey on statistical methods for health care fraud detection. *Health Care Management Science*, Vol. 11, pp. 275-287
- Musal, R. 2010. Two models to investigate Medicare fraud within unsupervised databases. *Expert Systems with Applications*, Vol. 37, No. 12, pp. 8628-8633
- Rawte, V. and Anuradha, G. 2015. Fraud Detection in Health Insurance using Data Mining Techniques. *Proceedings International Conference on Communication, Information & Computing Technology (ICCICT)*, Jan. 16-17, 2015.
- Travaille, P. et al. 2011. Electronic Fraud Detection in the U.S. Medicaid Healthcare Program: Lessons Learned from other Industries. *Proceedings of the Seventeenth Americas Conference on Information Systems*. Detroit, Michigan August 4th-7th 2011.
- Ularu, E.U. et al. 2012. Perspectives on Big Data and Big Data Analytics. *Database Systems Journal*, Vol. 3, No. 4, pp. 3-14.
- Wozniak, M. et al. 2014. A Survey of Multiple Classifier Systems as Hybrid Systems. *Information Fusion Journal*, Vol. 16, pp. 3-17.
- Yunusa, U. et al. 2014. Trends and Challenges of Public Health Care Financing System in Nigeria: The Way Forward. *IOSR Journal of Economics and Finance (IOSR-JEF)*, Vol. 4, No. 3, pp. 28-34.



---

## Full Paper

# FRAMEWORK FOR DEVELOPMENT OF MOBILE TELENURSING SYSTEM FOR DEVELOPING COUNTRIES

---

**J. O. Adigun**

Department of Computer Science,  
Federal College of Wildlife Management,  
New Bussa  
Niger State, Nigeria  
sunkanmisegun@gmail.com

**J. O. Onihunwa**

Department of Computer Science,  
Federal College of Wildlife Management,  
New Bussa  
Niger State, Nigeria  
johnonihunwa@yahoo.com

**D. A. Joshua**

Department of Computer Science,  
Federal College of Wildlife Management,  
New Bussa  
Niger State, Nigeria  
dejjoshuade@yahoo.com

**O. O. Adesina**

Department of Computer Science,  
Federal College of Wildlife Management,  
New Bussa  
Niger State, Nigeria  
bbumbum96@yahoo.com

**ABSTRACT**

Proliferation and increased performance of mobile computing systems has brought about development of myriad of mobile applications including mobile surveillance, mobile news, mobile games, mobile learning, mobile health etc. Mobile health has many sub-fields one of which is mobile nursing. In this paper, framework for provision of effective mobile nursing system was developed, the framework aid sustainable healthcare provision in developing countries by enabling telenurses to make clinical decisions based on expert advice and carry out some medication administration functions like medication usage monitoring etc. This is with the view to improve health care quality, thereby expanding access to affordable care at reduced health care cost of patients.

Development of this framework involved the establishment of stakeholders required by the system: these included the care centre, telenurse, telepatients and teleconsultants. The detailed attributes and functions of these stakeholders as well as relationship and interaction between the stakeholders were specified. The requirement statement gathered was transformed into use case diagram of the mobile nursing system wherein the design of the framework of the mobile nursing system was designed on. Furthermore, the flow chart of the mobile application which implements the framework designed was detailed.

The mobile nursing system was shown to require low start-up cost as it only requires a central server and mobile phones running Android operating system already possessed by the telenurse, telepatients and teleconsultants. It is believed that the system is cost effective complements to traditional nursing that will reduce the problem of nurse understaffing and reduce rural marginalization in terms of nursing staffs by enabling nurses take ubiquitous clinical decisions about their patients.

**Keywords:** MOBILE HEALTH, MOBILE NURSING, CLINICAL DECISION.

## 1. INTRODUCTION

Computers have found applications in virtually all fields including e-surveillance, e-news, e-games, e-learning, e-commerce, e-library, e-business, entertainment, e-photography and e-health etc. This is particularly true now that some computing devices are being developed to be so minute that they can be embedded in clothing and even humans, thus, computing is drifting away from just being concentrated on computers stationed at some location and relates more and more towards society, its infrastructures and its people to whom computing is made available anytime and everywhere using any device in any location and in any format (Abowd, 2005; Reza, 2005; Ballagas et al, 2006; Lazakidou and Iliopoulou, 2012; Adewale et al, 2014; Jordanova, 2010).

Computers have found applications in virtually all fields including e-surveillance, e-news, e-games, e-learning, e-commerce, e-library, e-business, entertainment, e-photography and e-health etc. This is particularly true now that some computing devices are being developed to be so minute that they can be embedded in clothing and even humans, thus, computing is drifting away from just being concentrated on computers stationed at some location and relates more and more towards society, its infrastructures and its people to whom computing is made available anytime and everywhere using any device in any location and in any format (Abowd, 2005; Reza, 2005; Ballagas et al, 2006; Lazakidou and Iliopoulou, 2012; Adewale et al, 2014; Jordanova, 2010).

Mobile computing is a ubiquitous computing paradigm associated with the mobility of hardware, data and software applications (Reza, 2005). Mobile computing systems are computing systems that can be used while the user is on the move, they possess easily moveable hardware (in terms of size, portability etc.) and possess software constrained in their usage while the user is on the move; they include laptops, personal digital assistants (PDAs), mobile phones, etc.

Earlier, mobile systems are constrained in their usage due to limited power supply, storage capabilities, connectivity to network, etc. however; mobile systems are now overcoming these constraints and are almost levelling up with the

capabilities of stationary systems. More so, mobile systems possess capabilities that enable their usage in myriad of applications that could not be sensibly implemented on stationary systems, for example, mobility capability of mobile systems enables them to be used in location tracking.

Mobile software applications have therefore been applied to many fields that the traditional computing systems have found applications including the mobile surveillance, mobile news, mobile games, mobile learning, mobile commerce, mobile library, and mobile business and for the purpose of this study mobile health etc.

The increase in population without proportionate increase in medical resources/facilities availability in the many developing countries is putting a lot of stress on the available medical resources. Information communication technology (ICT) has however recorded breakthroughs in breaching this gap and one of such breakthroughs include its provision for mobile health (mHealth) for various medical fields. The original concept behind mHealth is to support healthcare delivery and clinical decision via wide application of all available mobile technologies – mobile phones, personal digital assistants (PDAs), monitoring devices, etc. (Jordanova, 2010). It is already proven that distant care management service can enhance self-care, change health-related behaviours and improve outcomes in patients with a number of long-term conditions (McNeil et al., 2008). Mobile Health will soon become a necessity and a fantastic challenge for the future.

Nursing practice is one of many paramedical fields whereby mobile health has found applicability in developed countries. However, most developing countries including Nigeria have not been able to annex mNursing on a large scale beyond simple telephone triage and electronic messaging (Iluyemi and Briggs, 2010). Thus, developing countries have not been able to leverage on the capability of mobile nursing to complement traditional nursing care provision.

Mobile Telenursing or Mobile nursing (or simply mNursing) in developing countries could be described as the interaction between patient and nurse that takes place exclusively through mobile phones calls or Short Message Service (SMS).

mNursing however exceed mere interactions to include considerable traditional functions (explained in Carruthers, 2007) of nurses including telecare, tele-assessment, patient education, drug administration, drug usage reminding/monitoring, and crisis intervention through referral to appropriate health centre.

Mobile telenursing or simply mobile nursing (mNursing) aids in getting the patient to the right level of care with the right provider in the right place at the right time (AAACN, 2007). It is the use of mobile technology to provide or support nursing care practice or manage and coordinate nursing care to mobile patients from a distance using a broad range of telecommunication modalities (Hebda and Czar, 2013). These, mobile technology and telecommunication modalities include Short Message Service (SMS), MMS, multimedia conferencing, e-alerts, customised internet applications and telemonitoring equipment etc.

The mobile nursing system will enable telenurses to carry out some traditional nursing functions (including telediagnosis, medication usage monitoring etc.) to patients in a ubiquitous and real time environment. It will reduce self-medication yet keep more patients out of hospital thus reducing the stress on the scarce hospital facility at minimal cost to the patients.

### 1.1 Statement of the problem

In Nigerian hospitals currently, the nurses carry out most of their functions manually moving from one in-patients' bed to another and also attend to the out-patients sorting their appointments with doctors, educating the maternity patients etc. While the WHO recommends 1 nurse to a population ratio of 700, Nigeria currently has 148,129 nurses to cater for its population of above 150 million (Nursing and Midwifery Council of Nigeria (NMCN, 2012), with nurse population ratio of 1 to 1012 people thus most Nigerian hospitals are understaffed as far as nursing staffs is concerned (Lamothe *et al*, 2006; Oyeleye *et al*, 2013; WHO, 2013), this results in over utilization of available nurses (Oyetunde and Ayeni, 2014). In fact, these challenges are even more pronounced in rural areas as nurses in the rural areas even found themselves performing functions of medical doctors as some

rural areas do not have any qualified medical doctor in the primary health centres.

Also, the topography of Nigeria shows a country with high proportion outskirts, under served, under privileged rural population, with poor access road network, disparate distribution of health facilities and health personnel. About 80% of the population lives in the outskirts, whereas about 20% of health facilities and skilled personnel practices in the urban areas (Ogini and Nwoke, 2010). These have led to rural marginalization in terms of medical specialists including doctors and nurses etc. in the remote areas. Actors in the government and medical field have been working to bridge this gap for many years without uniform, replicable success. Ogini and Nwoke (2010) believed that the rapid advancement in information and communication technology and its universal reach present opportunity to bridge this gap, mobile telenursing system will be a great potential to extend medical care efficiently to over 40% of the population where nursing care is almost non-existent.

Mobile telenursing system presents the possibility of promoting patients' self-care hence allows nurses to divert substantial attention from this responsibility thus increasing nurses' effectiveness in handling the nurse – patients' ratio situation in the nation and reduce rural – urban health care marginalization. To make necessary provisions for the inherent benefit of telenursing in Nigeria is the need to develop mobile telenursing and drug administration system for use on mobile systems running Android OS.

### 1.2 Aim and Objectives

The research aimed at developing a framework for mobile-nursing system that will be adequate for developing countries. The specific objectives include:

- i. To analyse the requirement and specification of the mobile nursing system
- ii. To design the framework of the system.
- iii. To implement the designed system modules on mobile phones running Android OS.

### 1.3 Scope of the Study

The concept of telenursing involves many players, many nursing functions and forms a large field to explore. However, the focus of this research was the drug administration duties of telenurses via mobile platform. Only nursing practice that imparts on drug administration (directly or indirectly) was taken into account during this research.

## 2. LITERATURE REVIEW

Several works (Moore and Robinson, 2006; Naditz, 2009; Schlachta-Fairchild, Elfrink, and Deickman, 2009; Jordanova, 2010 and Purc-Stephenson, 2013) have been done in the field of developing telenursing application on mobile devices; some of the works are discussed below:

Some literatures that presented issues on telenursing often equate mobile telenursing with telephone triage. Purc-Stephenson (2013) reviewing recent trends, emerging issues and evolving practices says: “telephone triage and advice services or telenursing, is an evolving model of care delivery facilitated by technology”. Rolland, Moore and Robinson (2006) presented telenursing or telephone triage as involving collecting and screening a caller’s health-related symptoms via telephone to determine the urgency of the problem and to advise on the best course of action, such as directing patient to an emergency department, making an appointment with a general practitioner, or using self-care that may involve direction on drug administration after the telenurses has related the case to the appropriate channel. However, Rolland *et al* (2006) whereas many health issues are better described through visuals rather than through telephone calls. This implies that symptoms may be misrepresented to the nurses which in effect may misdiagnose such patients.

Schlachta-Fairchild, Elfrink, and Deickman (2009) presented issues on Patient Safety, Telenursing, and Telehealth. They noted that compliance and adherence problems are among the many issues that necessitate adoption of telenursing. This is because, after a patient leaves a provider’s office or a hospital, the patient is responsible for his or her own health care at home. However, many patients often do not follow a treatment plan as directed by a physician or provider due to several factors, including: miscommunication or faulty

understanding of the treatment plan, lack of access to facilities needed for the treatment plan, and a complex treatment regimen that the patient cannot comprehend without additional guidance. They therefore concluded that telemonitoring of drug usage adherence could reduce compliance and adherence problems. However, Schlachta-Fairchild *et al* (2009) only made a description, they didn’t design neither did they implement the system.

Naditz (2009) reporting on telenursing: frontline applications of telehealth delivery related mobile telenursing to nurse – experienced nurse/specialists interaction in the telehealth clinic opened by department of veterans affairs in the US in which the closest doctor was 150 miles away. However, the nurse – patients’ interaction is not fully operational in this telehealth clinic as the nurses only seldom perform telephone triage. The nurses mostly serve as the eye of the doctor at the 150 miles distance hospital which the nurses relay cases to via teleconferencing. The nurses use the telehealth equipment to make assessment which is being telemonitored by the doctor who instruct the nurses to make the necessary examination and instruct basic treatment e.g. drug administration etc. need.

A winning strategy was developed in the Centre for Psychotherapy Research Stuttgart, Germany, for after-treatment of patient with Bulimia Nervosa based on SMS. The intervention consists of weekly messages from the patients on their bulimic symptomatology and a corresponding weekly feedback that is a mixture of pre-programmed parts and individually tailored information. Results indicate that the program is technically feasible, well-accepted by patients and helpful for patients with bulimia nervosa to readjust to everyday life after finishing inpatient treatment (Bauer *et al.*, 2009). However, the research did not go beyond simple SMS alert and feedback.

Another success story reported by Jordanova (2010) is the “On Cue” project in South Africa sending SMS reminders to patients with tuberculosis for drug regimen compliance. SMS were sent out every half hour within a chosen time-frame to remind patients to take medicine. As of January 2003, the city of Cape Town has paid only \$16/patient/year for SMS reminders. In this pilot, only 1 patient out of 138 was non-compliant (99.3% compliance rate). However, the research too like

(Bauer *et al.*, 2009) did not go beyond simple SMS alert and feedback.

Villaini *et al.* (2013) reported on clinical and psychological telemonitoring and telecare of high risk patients with chronic heart failure through wireless technologies, dedicated software was developed to collect patients' data on a smartphone, PDA, and allow their subsequent wireless transmission to the remote server for storage and analysis. The system was designed to communicate with each patient, asking simple questions about his/her symptoms and giving information and counselling through visual and acoustic reminders. The remote control included: (1) a patient front-end, (2) a medical front-end, (3) a software web-based system for assistance in the clinical decision and a reminder system. The patient front-end operated through a PDA that each patient received at discharge from the hospital. At a defined prescheduled time, patients are required to transmit their blood pressure twice a day, the cardiologist analysed the information received by each *patient's front-end*, evaluated their clinical priority and could call the patient or send an SMS to modify the therapy if a great variability was observed, to tailor the timing of drugs' administration, ask more information or recommend a clinical control. The system created a graphic trend of the variables preliminarily chosen by the cardiologist, so that a more frequent control could be asked for those that were found more difficult to control (i.e. a patient could be asked to transmit his/her blood pressure twice a day if a great variability was observed, to tailor the timing of drugs' administration). The telemonitoring device was seen to help in better treatment adherent. However, the system did not include the role of nurses thereby breaking the chain of healthcare delivery from diagnosis to treatment.

This paper presents framework that take telenursing beyond Rolland *et al.* (2006) that limited their work on mobile nursing system to use of telephone calls, it is also a step beyond the work of Bauer *et al.* (2009) and Jordanova (2010) that limited theirs to simple SMS because the framework presented in this paper take telenursing beyond calls and SMS to use of multimedia (text and visuals). The framework presented in this paper follows the specification described by Schlachta-Fairchild *et al.* (2009). The framework also included nurses in the chain of telehealth care delivery as

opposed to the work of Villaini *et al.* (2013). The framework specifies implementing the mobile nursing system on mobile devices running Android operating system for both nurse – patients (reported by reported by Naditz (2009)) and nurse – specialist/experienced nurse interaction. The patients can be reached anywhere, anytime and vital signs on drug administered (adverse effects of drugs, improvement noticed) can be relayed by the patients to the telenurses who then decide on next line of action as pertaining to the nursing care needs of patients after necessary consultation which prevents the chain of healthcare delivery from being broken from diagnosis to treatment. The system will also include an automated reminder system for reminding patients of drug usage as at when due.

### 3. METHODOLOGY

The research employed the experimental methodology in the design of the framework. The framework prevented the chain of healthcare delivery from being broken from diagnosis to treatment by allowing home health nurses, nurses in trains and aeroplane during a journey, nurses in rural areas, etc. collect and screen a patient's health-related symptoms via mobile devices and communicate the same to the more experienced specialist/physician and allows the physician/specialist to determine the kind of the problem and to advise on the best course of action.

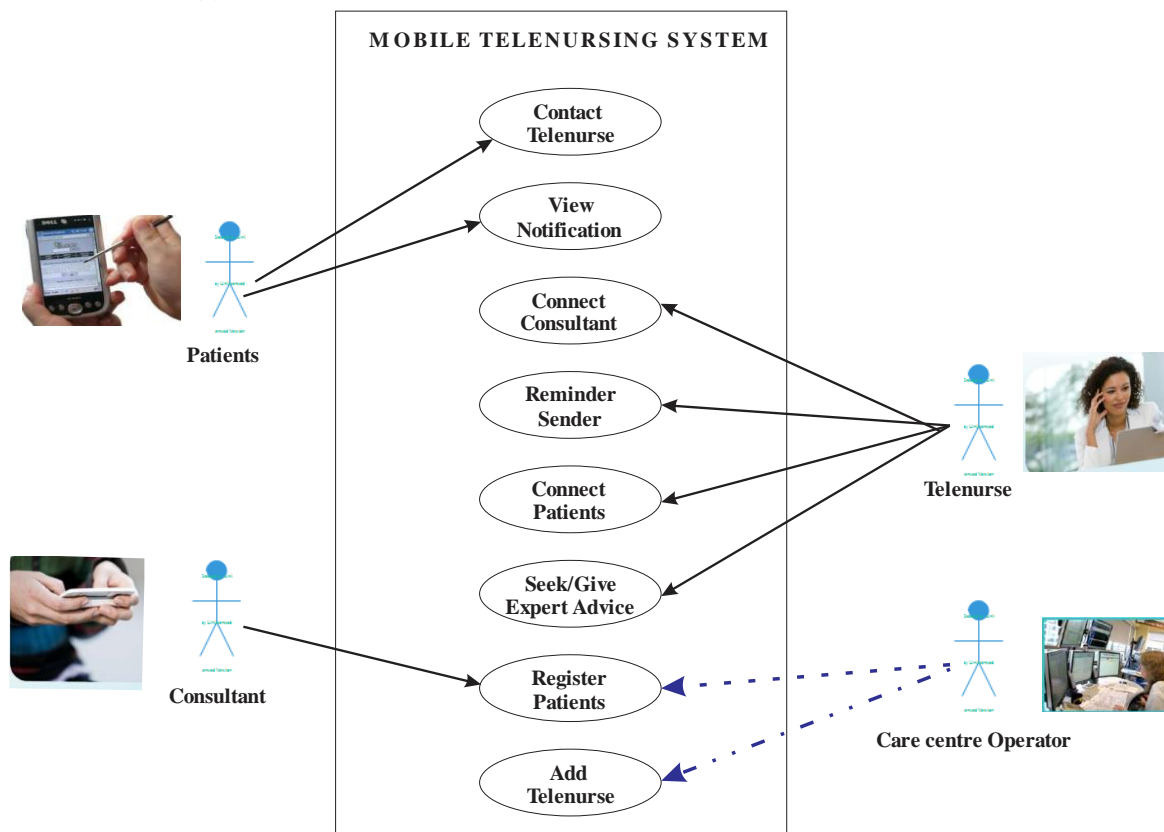
#### 3.1 System Analysis

This deals with stakeholder establishment and their relationship and/or interaction. The most important stakeholders in this project included the project customers (that is the care centre), the telenurses, and the telepatients and/or their care givers and for telediagnosis purpose consultants/specialists or tele-doctors. These stake holders' interaction is depicted in the use case presented in figure 1.

In figure 1, the care centre is the dedicated hospital facility to offering telenursing services, it is a department in the hospital where specialists are brought into or where communication can be established with consultants/specialists as the need may arise. The care centre telenurse operator perform additional function of assigning telenurses to registered patients at the point of the patients' registration and provides backup services to

telenurses when there is need to make consultation with nurses presents at the care centre or with consultants that may be needed to be brought in to the care centre, or make connections to the consultants via their own mobile devices as the case may be. In simple terms, the care centre function as both the server room and/or call centre, as such, telenurse(s) that served as the care centre

operator is expected to be present at the care centre to provide backup services for the mobile nurses. At the points of patient's registration, relevant data about the patients are collected and saved on the care centre database, and the telepatients is assigned a telenurse whom relevant patients' data will be forwarded to.



**Figure 1:** Use case depicting stakeholders' interaction in the proposed system

The telenurses are nurses that have been trained to nurse patients at a distant, among the responsibilities of the telenurses include accessing patient's current clinical condition and relaying it, patients' known allergies, and medical history etc. to available consultant for treatment. Drug prescription given are relayed to the patients and they monitor the patients for usage adherence through drug usage reminder module, they also ensure maintenance of adequate medication record, documents and reports medication incidents and adverse medication reactions through treatment feedback received from patients. The telenurses are to be armed with

mobile devices running Android OS on which the developed Android application was installed via which communication is made with the telepatients and the care centre as the case may be. The mobile devices requires touch pad or touch screen for entering data and has internet connectivity used by the system to enable the phones to keep in touch with the server. The telenurses were mobile and sometimes make consultations with the nurses at the care-centre as at when required.

The telepatients are patients receiving tele-treatments from the care centre. The telepatients had to register with the care centre. Patients' health records are collected from the patients at the point of registration and these forms the patients' initial medical records. The telepatients

were also armed with mobile devices running Android OS on which was installed the Android applications via which communication was made with the telenurses. The telepatients don't have direct contact with the care centre via the android application but can only contact the care centre via the telenurse assigned to them or via phone call. Sometimes, some telepatients required a care giver to serve as intermediary between them and their telenurses; their roles were sometimes necessitated when a patient (due to reason of inadequate ability to use mobile devices, incapacitation etc.) could not adequately use the mobile devices. In such cases, the care giver acts in proxy for the patient by relating directly with the telenurses about the patients' situation.

Communication within the system is mainly through the internet, that is, the consultant, care centre operator, the mobile patient and telenurse mobile devices used the mobile operators' internet services to communicate and internet connectivity through the various mobile operators was required for telecommunication between the stakeholders. The application provided the patients with a series of graphical interfaces which allowed them to enter manually, store and forward vital signs via the mobile operator communication links to the telenurse assigned to the patients. The nurse makes an initial assessment of the patients; send the assessment to the care centre server and either directs the patients to continue with medication or in circumstances, where changes to the dose are considered necessary after consultation with the consultants, the nurse send feedback with advice on needed treatment modifications to the patients. The telenurse ensures treatment adherence by programming automatic drug usage reminder periodically for the patients.

### 3.2 System Requirements Specifications

The functional and non-functional requirements of the system as gathered from the analysis in the previous section are depicted below:

#### Functional Requirements

- Registration of the telenurses and patients is done by care centre operator.
- Both telenurses and patients are given unique username and password.
- Patients are assigned to telenurses during patients' registration.

- After logging in, patients can read reminder/notification
- Patients can also contact telenurse to relay symptoms and any issues/cases to telenurses.
- After logging in, telenurses can contact consultant to relay issues that require expert advice to consultant.
- Telenurses can also connect patient to monitor patient's to ascertain if there are any/no issues that needs attention so as to give them adequate advice.
- Telenurses can also send drug usage reminder and other form of reminder.

#### Non-functional requirements:

- 24 X 7 available text-based chat-like platform that shows chat in real time
- Reliable store and forward multimedia upload platform to communicate visualizations
- Both web and database server that possess 24 X 7 availability
- Widely used mobile operating system in the region of deployment
- Efficient assignment of telenurses to patients
- Better component design to get better performance at peak time
- Flexible service based architecture will be highly desirable for future extension
- Efficient 2G/3G and GSM/GPRS internet connectivity.

### 3.3 System Design

The mobile telenursing system (MTS) framework based on the system requirements specification defined in previous section is presented in figure 2.

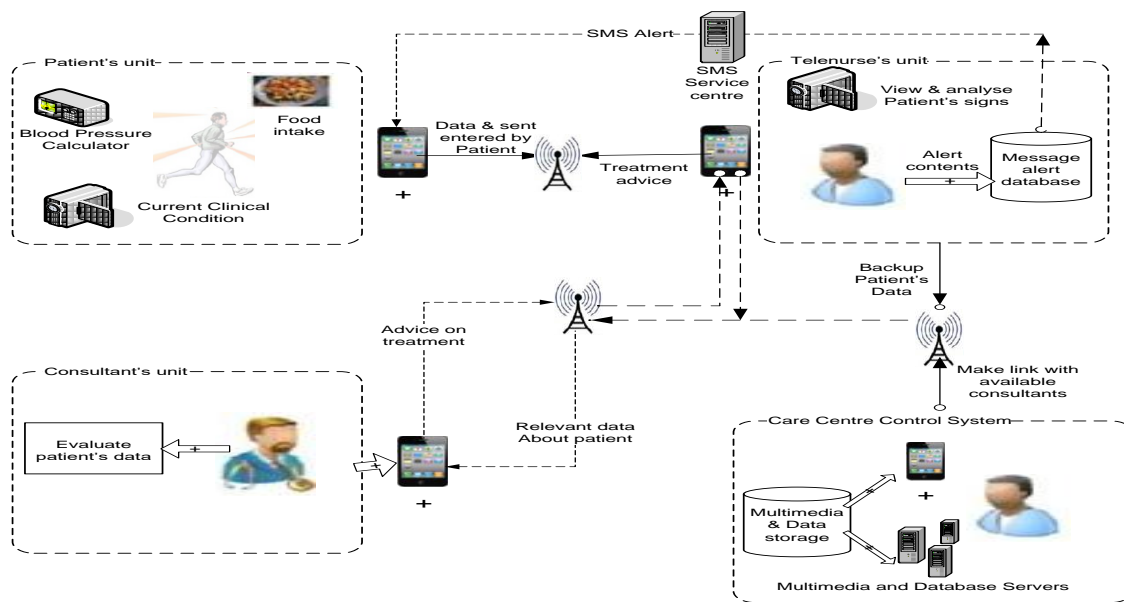
Figure 2 shows that the telenurses, patients and consultants all possess mobile devices, the system include client side and server side applications; more so, communications is mainly through the mobile carrier operator 2G/3G internet service. The patients' mobile devices possess components of the MTS through which they can present vital signs to the telenurse; the chat-like components for presenting vital signs (e.g. blood pressure measure, etc.) can be initiated by the patients in case of emergency vital signs. The patients also receive drug usage reminder and other form of reminder in

form of SMS alert through SMS service centre of his/her phone's mobile carrier operator.

It is the telenurses that is anticipated to often initiate patients' monitoring through the chat-like platform, the nurses then view and analyse patients' clinical condition and decide on treatment or link up with consultants in case of complications to receive expert advice through similar chat-like platform. The chat like platform allows for multimedia upload to telenurses or consultants.

The nurses have the option of setting up SMS reminder and send such as SMS alert to their assigned patient(s) at the specified time.

The consultants analyse data sent about a patient and make necessary recommendation through the same chat like platform to the telenurses. The care centre control system monitors the servers and provides back-up activities to telenurses as at when needed.



**Figure 2:** Framework of the mobile telenursing system (MTS) system

### 3.3.1 System Flow chart

From the framework presented in previous section, the patients' mobile devices possess components of the MTS through which they can present vital signs to the telenurse, also, the patients' unit can be initiated by the patients in case of emergency vital signs. However, it is the telenurses that is

anticipated to often initiate the application which present him/her with options to either chat with patients. Also, it is the telenurse that also established chat with the consultants and send SMS reminder to the patients. Therefore, the telenurse unit is the most important part of the framework; its flow chart is presented in figure 3.



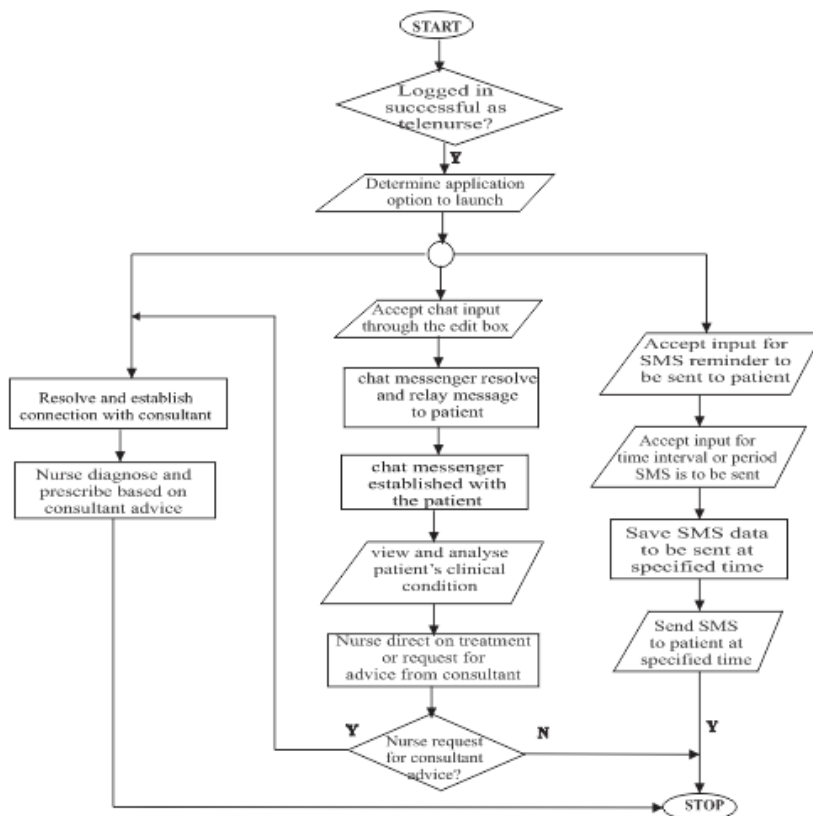


Figure 3: Flow chart of the mobile nursing system

The figure shows that the telenurse after authenticating himself or herself determines the application to launch. The nurse selects whether to establish chat with patient or send SMS reminder. Through the chat with patient, the nurse examines and analyse patients' clinical condition and directs on treatment. Patient's clinical condition that requires the intervention of a consultant however requires the telenurse to connect with consultant and directs on treatment based on consultant's recommendation. The process of sending of the SMS reminder is a simple database active transaction that checks for time difference between a time a reminder is to be sent to patient and the current time. Once the difference is zero, the database fields are sent to the mobile numbers of the patients.

#### 4. SYSTEM IMPLEMENTATION

It is noteworthy to state that the prototype of the system has not yet been completed as at the time of compilation of this paper, hence the screen shot of the prototype of the system is not presented in this paper, however, the system's implementation is well spelt out in this section.

The mobile telenursing system (MTS) based on the requirements and design defined in previous section include both client and server side applications.

The android application is to be built using the native android SDK with support for android version 3 to version 6, hence, compatible devices included any mobile phone running Android OS which has a touch screen with adequate size and supports 2G/3G and GSM/GPRS transmission network protocols. Other requirements of the phone included SIM cards from any of the available mobile network operator for receiving drug usage reminder, internet access and/or cellular communication, wireless connectivity via Wi-Fi for internet connectivity through hotspot. 2G or 3G network can be used and the mobile devices have base stations which connect to the MSC of the mobile operators. The communications is mainly through the

gateway provided through the any cloud platform as a service (PAAS) server such as heroku server. From there the communication is managed and controlled by the control server managed by the PaaS apps and necessary backups are done on the PaaS server.

Pusher real-time messaging library can be used for implementing real-time chats between the various users of the application such as between a telenurse and patients/consultants. The Backend Architecture (i.e. the database) used for storing the various information on the application including chats, telenurse details, patient details etc. is a MongoDB server hosted on mlab.com. MongoDB is a document object store that allows easy storage of unstructured data and scaling of the database. The backend server is based on the nodejs server which allows the javascript programming language be used on the server. ExpressJS serves as the framework of choice for routing and connecting to mongoDB. The asynchronous nature of nodejs is a perfect fit for the telenurse project considering the backend server has to handle multiple API calls from the android application. The project is hosted and deployable in the PAAS container such as heroku container to facilitate continuous integration and deployment whenever new features are added or bugs are fixed in the codebaseTelenurse Project.

The SMS reminder module is to be built on the SMS framework library of the android stack, therefore, the module relied on the users' (e.g. telenurse etc.) mobile operator as the SMS service centre. The telenurses initiates the reminder system by entering SMS to be sent out to the patients, along with other information such as the patient's phone number, the time the reminder is to be sent and the SMS detail, these are stored in the database that was built upon the SQLite library of the android stack. The SMS is pulled from the database at the specified time and sent to the SMSC (Short Message Service Centre) which directs it to the appropriate mobile device. Before sending the message, the SMSC finds the roaming customer by consulting the

HLR (Home Location Register) of the specified patients' mobile operator.

The android application included four main subsystems: patient's module, telenurse module and Doctor's/Specialist's module and the care centre control system.

#### 4.1 Patient's module

The user after entering appropriate login with username and password starts the navigation through the MTS patient's module. The patient's personal information as inputted during registration is displayed to him/her after which the next activity allows the patient to select an activity to perform among the following:

- i. Contact Telenurse: This module was used to relay symptoms and any issues/cases to telenurses.
- ii. Notifications: This module on the other hand was used to view notifications sent by telenurses.

#### 4.2 Telenurse module

During log in, the care centre operator has to check that he/she is the care centre operator after which the next interface allows the care centre operator to select an activity to perform among any of the following:

- i. Connect Consultant: This module on the other hand was used to relay issues that require expert advice to consultant.
- ii. Register Patient: This module was used to register patients and assign each patient to a telenurse at the point of registration.
- iii. Connect patient: This module was required by telenurses to monitor patient's health by contacting them to ascertain if there are any/no issues that needs attention so as to give them adequate advice.
- iv. Add Telenurse: This module on the other hand was used to register telenurses.
- v. Reminder Sender: This module enables telenurses to send drug usage reminder or any other form of reminder.

The difference between activities that can be performed by an assigned telenurse and the telenurse that acts as the care centre operator is that the assigned telenurses can neither add patients nor add another telenurse. It is the duty of the telenurse assigned to the care centre to register patients and assign a patient to an available telenurse during patient's registration. He/she can however attach a patient to himself or herself also. The assigned telenurse was made responsible for telemonitoring of his/her assigned patients, and the one whom the patients contact as at when due.

#### 4.3 Specialist's module

The module available to the specialists or doctors enable the doctor to access the patient's medical history on the care centre control server and allows them to have visualization of the patients' clinical condition sent to them by the telenurse and send recommendation to the nurse which the nurse follows to advise the patients accordingly.

#### 4.4 Implication of framework to health informatics in developing country

The authors had users (patients, nurses etc.) in developing countries in mind by ensuring the system introduces minimal cost implication to users in terms of nursing care. This is because higher population in developing countries are poor and may not accept such system if it presents much disparity in terms of nursing care cost compared with what they have been used to. The minimal cost implication is the reason why the researchers did not use real-time monitoring devices (e.g. real-time blood pressure and temperature monitoring device), hence, the framework developed incorporates the system with mobile phones which many people in the developing countries already possess, hence the system requires little or no extra healthcare cost to the users (especially patients) aside data charges and minimal installation and maintenance charges.

The framework afford health care providers in developing countries have better access and management of patients' health care

information because the care centre server hosted on the PaaS server tracks and store every communication within the system. The information retrieved from the care centre server affords better access to patients' medical history thus fostering better collaborations among various health care providers in the developing country.

## 5. Conclusion and Recommendation

### 5.1 Conclusion

In this paper we have developed a framework for mobile telenursing system which we called MTS, which is an effective complement to traditional nursing care made available to nursing personnel, patients and consultants via mobile phones using android technology. The framework specifies SMS reminder and the clinical condition monitoring systems. It is an attractive system that has the potentials of providing drug usage adherence, remote nursing care, triage, patient's evaluation, education, monitoring and after treatment nursing care etc. to patients, fills a service gap among those that have limited access to nursing care and provides guided self-care to patients thereby reducing patients' visit to hospital facility. It is believed that the system is cost effective complements to traditional nursing that will reduce the problem of nurse understaffing and reduce rural marginalization in terms of nursing staffs by enabling nurses take ubiquitous clinical decisions about their patients.

The MTS framework took care of the setback of the framework offered by Rolland et al (2006) which poses the challenge of patients misrepresenting symptoms to the nurses. The MTS framework reduces the chance of misrepresenting symptoms to the nurses as it enables patients to present health issues through text and visualizations rather than through telephone calls. The MTS framework also overcame the shortcoming of the framework developed by Villaini et al (2013) as the framework explicitly included the functions of telenurses. Also the framework extends the description of the framework by

Schlachta-Fairchild et al (2009) because the description made was converted to formal specification which was used to design the framework of the MTS.

However, in this paper, we do not anticipate use of real-time monitoring devices (e.g. real-time blood pressure and temperature monitoring device) which is as a result of anticipated provision of cost effective system that will be readily acceptable to the patients at initial stage of implementing the system.

### 5.2 Recommendation and Future Work

Due to the cost effectiveness of this system, its affordability to both nurses and patients and its promising prospects, there is the need to massively embrace the system in the developing countries.

In this paper, we do not anticipate use of real-time monitoring devices, thus subsequent research may need to include this probably after widespread acceptance of telenursing system in the developing countries.

Once the prototype is completed, the authors plan to pilot test the software in some strategically selected hospitals in Nigeria, and evaluate the performance of the developed system based on suitability, effectiveness and acceptability through user experience.

## 6. REFERENCES

- American Academy of Ambulatory Care Nursing (AAACN), 2007. *Telehealth Nursing Practice Administration and Practice Standards*. Pitman, NJ: Anthony J. Janetti.
- Abowd G.D., Iftode L., Mitchell H. 2005. "The Smart Phone: A First Platform for Pervasive Computing", *Pervasive Computing*. IEEE, vol. 4, pp. 18-19, March 2005.
- Adewale A. A., Abdulkareem A., Adelakun A. A., 2014. Development of An SMS Based Alert System using Object Oriented Design Concept. *International Journal of Scientific & Technology Research*. Vol. 3, Issue 5, May 2014.

- Ballagas R., Borchers J., Rohs M., and Sheridan J. G., 2006. "The Smart Phone: A Ubiquitous Input Device", *Pervasive Computing*. IEEE, vol. 5, pp. 70-77, March 2006.
- Bauer, S., Percevic, R., and Kordy, H., 2009. The use of short message service (SMS) in the aftercare treatment for patients with Bulimia Nervosa", *Presented at Med-e-Tel* 2009, [www.medetel.lu/download/2009/parallel\\_sessions/abstract/0422/the\\_use\\_of\\_short\\_message\\_service.doc](http://www.medetel.lu/download/2009/parallel_sessions/abstract/0422/the_use_of_short_message_service.doc)
- Carruthers, E. P., 2007. "Nursing." Microsoft® Student 2008 [DVD]. Redmond, WA: Microsoft Corporation, 2007.
- Hebda, T. & Czar, P., 2013. *Handbook of informatics for nurses & healthcare professionals*. (5th ed). Boston: MA. Pearson.
- Iloyemi A. and Briggs J. S., 2010. Wireless Access and Connectivity for Community Based Health Workers in Developing Countries: Models. *Question 14-2/2 Final Report: Mobile eHealth solutions for Developing Countries*. pp. 22–25
- Jordanova M., 2010. Mobile Health: m-Health, mHealth, or Mobile Health – which one is correct? *Question 14-2/2 Final Report: Mobile eHealth solutions for Developing Countries*. pp. 1–5
- Lamothe L, Fortin J. P., Labbe F., 2006. *Impacts of telehomecare on patients, providers, and organizations*. *Telemed J E Health* 2006; vol. 12, pp. 363-369.
- Lazakidou A. and Iliopoulou D., 2012. Useful Applications of Computers and Smart Mobile Technologies in the Health Sector. *Journal of Applied Medical Sciences*. vol. 1, no.1, 2012, 27-60
- McNeil, I., Wales J. and Azarmina, P., 2008. Satisfaction: the effect of a telephone based care management service on patient outcomes in the UK, in Jordanova M. & Lievens F. (Eds.) *Med-e-Tel: The International Educational and Networking Forum for eHealth, Telemedicine and Health ICT, Electronic Proceedings*, Publ. Luxexpo, Luxembourg, 2008, pp. 415-420.
- Naditz A., 2009. *Telenursing: frontline applications of telehealth delivery*. Mary Ann Liebert, telemedicine and e-health publications. Vol. 15, No 9
- Ogini A. N. and Nwoke E., 2010. Tele-Nursing and Nursing Informatics in Developing Countries: A Case Study of Nigeria.
- Oyeleye, O., Hanson, P., O'Connor, N. and Dunn. D., 2013. Relationship of Work Incivility, Stress, and Burnout on Nurses' Turnover Intentions and Psychological Empowerment. *Journal of Nursing Administration*, 43, 536-542. <http://dx.doi.org/10.1097/NNA.ob013e3182a3e8c9>
- Oyetunde M. O. and Ayeni O. O., 2014. Exploring Factors Influencing Recruitment and Retention of Nurses in Lagos State, Nigeria within Year 2008 and 2012. *Open Journal of Nursing*, 2014, 4, 590-601. Available at: <http://dx.doi.org/10.4236/ojn.2014.48062>
- Purc-Stevenson R. J., 2013. Telenursing: A Review of Recent Trends, Emerging Issues and Evolving Practices. *Ann Emerg Disp Resp* 2013; 1(2): pp. 6-11
- Reza B'Far, 2005. *Mobile Computing Principles. Designing and Developing Mobile Applications with UML and XML. Published in the United States of America by Cambridge University Press, New York*. pp. 1–5
- Schlachta-Fairchild L., Elfrink V. and Delchman A., 2009. *Patient safety, Telenursing and telehealth – Patient safety and quality – NCBI bookshelf* retrieved on 31<sup>st</sup> March, 2015 from [www.ncbi.nlm.nih.gov/books/NBK2687](http://www.ncbi.nlm.nih.gov/books/NBK2687)
- Schwaab B, Katalinic A, Riedel J and Sheikhzadeh A., 2005. Pre-hospital diagnosis of myocardial ischemia by telecardiology: Safety and efficacy of a 12-lead electrocardiogram, recorded and transmitted by the patient. *J Telemed Telecare* 2005; vol. 11, no. 1: pp. 41-44.
- Villani A, Malfatto G, Compare A, Della R. F., and Rella V. (2013): Clinical and Psychological Telemonitoring and Telecare of High Risk Patients with Chronic Heart Failure through Wireless Technologies:

**13<sup>th</sup>**

# **International Conference**



The Icaros Project. *J Clin Exp Cardiol* 4:

260. doi:10.4172/2155-9880.1000260

---

Full Paper

**REASONING OVER VAGUE ONTOLOGIES USING FUZZY  
SOFT SET THEORY IN MEDICAL DOMAIN**

---

**R. Salahudeen**

Department of Computer Science,  
Ahmadu Bello University,  
Zaria  
sridwan@abu.edu.ng

**A. F. Donfack Kana**

Department of Computer Science,  
Ahmadu Bello University,  
Zaria.  
donfackkana@gmail.com

**ABSTRACT**

Description Logic (DL) being a knowledge representation language that is widely used in building ontologies is a crisp language and lacks representation and reasoning for vague concepts in an ontology for some real-world applications like medical records which has been confronted with the challenge of analysing, interpreting or processing medical data involving vague concepts to obtain inferred diagnosis efficiently. On the other hand, fuzzy ontology can effectively help to handle and process uncertain data and knowledge. In this paper, we propose a reasoning algorithm based on fuzzy soft set theory in order to reason about the uncertain aspect of a medical ontology of vague domain. The proposed algorithm was evaluated by applying it on some vague ontologies in a medical domain and the result was compared with the tableaux based and the soft set ontology reasoning techniques. The obtained result shows that, the proposed algorithm is satisfiable when fuzzy concepts and assertions are involved in an ontology representation while such fuzzy conceptions are not handled by both tableaux based and soft set ontology procedures.

**Keywords:** Vague Ontologies, Medical Domain, Uncertainty, Description Logic, Fuzzy Soft Set

## 1. INTRODUCTION

The use of information from different sources is an intelligent task that requires human being who has a background knowledge of the information. However, due to existence of several information sources, human being processing speed cannot be relied upon for speedy information processing. In contrast, computers are known in dealing with voluminous information as long as their processing do not require human intelligence. Therefore, for information to be processed efficiently, its processing must be automated. For the semantic processing of information to be possible, systems must be able to understand the meaning of data they are processing and then, perform the processing semantically (Kana & Akinkunmi, 2014).

On the other hand, information technology today is widely adopted in solving many complicated problems in various disciplines. Modern medical practice employs usage of digitized equipment in capturing patient's data, analysing patient's record and using the record stored about a patient to infer possible diagnosis and even prescribe necessary drugs, laboratory investigations and further medical actions to be taken.

Most of the medical records are stored independently with individual formats and are distributed across various heterogeneous databases. As such, there is an emerging demand for the integration and exploitation of heterogeneous biomedical information for improved clinical practice, medical research and personalized healthcare (Anjum, et al., 2007).

Also, most of the traditional tools for formal modeling, reasoning, and computing are crisp, deterministic, and precise in nature and medical record is not an exception of this fact. However, there are many complicated problems in medical sciences that involve data which are not always all deterministic (Maji, et al., 2003).

As works with semantics grows more motivating, there is an increasing appreciation of the need for principled approaches in representing and reasoning under uncertainty. Uncertainty is the situation which involves imperfect and/or unknown information. The term "uncertainty" encompasses a variety of aspects of imperfect knowledge, including incompleteness, vagueness, ambiguity, and others (Laskey, et al., 2008).

Over the past, lots of researchers have made efforts to achieve the goal of solving uncertainty problem in ontologies, among the techniques used are theories of rough sets, soft sets, fuzzy sets and

some probabilistic approaches. (Straccia, 2001) focused on SHOIN (D), and added fuzziness to SHOIN (D) which was shown to have more representation and reasoning capabilities beyond classical SHOIN (D). (Sanchez & Tettamanzi, 2006) introduced fuzzy description logic with extended qualified quantification called  $ALCQ^{\#}$ , which allows for the definition of fuzzy quantifiers of the absolute and relative kind by means of piecewise linear functions on  $\mathbb{N}$  and  $Q \cap [0,1]$  respectively. (Stoilos, et al., 2007) made a fuzzy extension of SHIN DL to ALC DL, they focus on improving the expressiveness of ALC by allowing SHIN constructors such as: transitive role axioms (S), inverse roles (I), role hierarchies (H) and number restrictions (N). (Kana & Akinkunmi, 2014) defined a way of instantiating ontologies of vague domains using the concept of soft sets initiated by Molodtsov and the concept of rough sets introduced by Pawlack. They defined ontological algebraic operations and their properties while taking into consideration the uncertain nature of domains. They showed that, by doing so, intra ontological operations and their properties are preserved and formalized as operations in a vague set of objects and can be proved algebraically. (Jiang, et al., 2010) used the soft theory to deal with uncertainty in ontologies by parameterizing concepts in user queries and use soft set operations to achieve an optimal decision. It was however pointed out in (Roy & Maji, 2007) that classical soft set is not appropriate to deal with imprecise and fuzzy parameters, due to this claim, (Maji, et al., 2001) introduced the concept of the fuzzy soft set, a more generalized concept, which is a combination of fuzzy set and soft set.

At the same time, there has been some progress concerning practical applications of fuzzy soft set theory, especially the use of fuzzy soft sets in decision making. (Roy & Maji, 2007) used an application of fuzzy soft set theory in object recognition problem. The recognition strategy is based on multi-observer input parameter data set. The algorithm involves construction of comparison table from the resultant fuzzy soft set and the final decision is taken based on the maximum score computed from the comparison table. (Kong, et al., 2009) noted that the algorithm for the selection of optimal object in (Roy & Maji, 2007) is incorrect. They gave a counter-example to illustrate that using the algorithm the right choice cannot be obtained in general. (Feng, et al., 2010) upheld (Roy & Maji, 2007) method of using maximum scores for selecting optimal decision as a general useful



method and concluded that the counterexample given by (Kong, et al., 2009) is based on an improper understanding of maximum choice value as the optimal decision, they however shows the limitations of (Roy & Maji, 2007) by means of level soft sets and presented an adjustable approach to fuzzy soft set based decision making.

In order to handle fuzzy parameters (whose values could lie in a probable range), this research work seeks to use the concept of fuzzy soft set introduced by (Maji, et al., 2001) to address the limitation of (Jiang, et al., 2010).

The rest of this paper is organized as follows: the next section briefly reviewed ontologies, DLs, fuzzy set, soft set and fuzzy soft sets. In Section 3, we present our approach of fuzzy soft sets ontology and its reasoning. Finally, in Section 4, we draw conclusions and present some topics for future research.

## 2. PRELIMINARIES

### 2.1 *Ontology Engineering and Learning*

This is a subfield of knowledge engineering that focused methodologies for building ontologies, and identifying the tools and languages for supporting ontology modelling. The main purpose of this subfield is to create, represent and model knowledge in a domains. However it was established that the task of Ontology Engineering is very expensive because of its consumption of big amount of resources even when principles and methodologies are applied, also the ontology design is a difficult and burdensome task (VTempich & Simperl, 2009). For the purpose of overcoming these challenges, there is a research line that has been gaining prominence in the past two decades in the area of ontology engineering which is the extraction of domain models from text written in natural language, using Natural Language Processing (NLP) techniques. This process is known as Ontology Learning.

(Ryan, et al., 2014) proposed an approach of translating natural language knowledge into a DL ontology, there approach consist of three (3) different modules:

1. Syntactic Parsing Module: The natural language text is being analysed into lexical tagging and dependency parsing using Probabilistic Context-Free Grammars (PCFGs) for the tagging and Stanford Parser for the parsing.
2. Semantic Parsing Module: This module initiates its activities by assessing the entry

sentence and the referred result of the syntactic analysis obtained in the previous module and then starts the extraction of terms (Term Extraction) that are fit to be concepts of the ontology, concatenation of dependent terms, sentences are divided into sub-sentence when sentence breakers are found and lastly the relations between the terms are verified and validated through the verbs found in the sentences.

3. OWL DL Axioms Module: This module generates the hierarchical and non-hierarchical relations, verifies the conjunction and disjunction and detects the negations/complements within the terms extracted in the semantic parsing module.

### 2.2 *Ontology and Description Logic*

In the area of information science, ontology is considered as the term used to refer to the shared understanding or vocabulary of some domain of interest. An engineering viewpoint of ontology is given by (Uschold & Gruninger, 1996) as “an explicit account or representation of a conceptualization”. This conceptualization includes a set of concepts, their definitions and their interrelationships.

Description logics (DLs) are a family of knowledge representation languages that can be used to represent the knowledge of an application domain in a structured and formally well-understood way. They are the most recent name for a family of knowledge representation formalisms that represent the knowledge of an application domain (the world) by first defining the relevant concepts of the domain (its terminology), and then using these concepts to specify properties of objects an individual's occurring in the domain (Baader & Werner, 2002).

In a knowledge-base of description logic system, an ontology is a triple  $O = \langle RB, TB, AB \rangle$ , where RB (the Role Box or RBox) and TB (the Terminological Box or TBox) comprise the intensional knowledge, i.e., general knowledge about the world to be described (statements about roles and concepts, respectively), and AB (the Assertional Box or ABox) the extensional knowledge, i.e., particular knowledge about a specific instantiation of this world (statements about individuals in terms of concepts and roles) (Baader & Werner, 2002).

DLs are based on concepts description and roles to describe the world being modelled. Concepts, which denote the set of individuals, represent the vocabulary of an application domain. Roles denote binary relationships between individuals. Elementary descriptions are atomic concepts (unary

predicate) and atomic roles (binary predicates). Complex descriptions can be built from the atomic concepts with concept constructors (Baader & Werner, 2002). To obtain these complex descriptions, DL employs Boolean constructors such as conjunction ( $\sqcap$ ), disjunction ( $\sqcup$ ) and negation ( $\neg$ ), as well as existential restriction constructor ( $\exists R.C$ ), value restriction constructor ( $\forall R.C$ ), and number restriction constructor ( $\geq nR.C$ ), ( $\leq nR.C$ ). Where  $n$  is a cardinality,  $R$  is a role and  $C$  and  $D$  are arbitrary concepts. Below are some concept construction syntax:

- $\top$ : denotes top concept, that is, the concept that all individuals of a domain must be instance of
- $\perp$ : denote bottom concept, that is, the empty concept without no instance.
- $\neg C$ : denotes the inverse of concept  $C$ .
- $C \sqcup D$ : denotes the concept represented by the union of  $C$  and  $D$  ( $C$  'or'  $D$  logically).
- $C \sqcap D$ : denotes  $C$  intersected with  $D$ .
- $\exists R.C$ : denotes, the set of all individuals that are in relation  $R$  to at least one individual from concept  $C$ .
- $\forall R.C$ : denotes, the set of all individuals that are in relation  $R$  with individuals from concept  $C$ .

The semantics of concept descriptions is defined in terms of an interpretation  $I = (\Delta^I, \cdot^I)$ , consisting of a domain of interpretation (also known as domain of individuals)  $\Delta^I$  which is a non-empty set of individuals, and an interpretation function  $\cdot^I$  mapping every atomic concept  $A$  to a set  $A^I \subseteq \Delta^I$  and every atomic role  $r$  to a binary relation  $r^I \subseteq \Delta^I \times \Delta^I$ . The extension of  $\cdot^I$  to arbitrary concept descriptions is inductively defined as shown

in Table 1:

### 2.3 Soft Set

(Molodtsov, 1999) defined Soft Set in the following way: A pair  $(F, E)$  is called a soft set (over  $U$ ) if and only if  $F$  is a mapping of  $E$  into the set of all subsets of the set  $U$ .

In other words, a soft set is a parameterized family of subsets of the set  $U$ . Every set  $F(e), e \in E$ , from this family may be considered as the set of elements of the soft set  $(F, E)$ , or as the set of approximate elements of the soft set.

### 2.4 Fuzzy Set

(Zadeh, 1996) defined Fuzzy Set in the following way: Let  $X$  be a space of points (objects), with a generic element of  $X$  denoted by  $x$ . Thus,  $x \in X$

A fuzzy set (class)  $A$  in  $X$  is characterized by a membership (characteristics) function  $f_A(x)$  which associate each point in  $X$  a real number in the interval  $[0,1]$ , with the value of  $f(x)$  at  $x$  representing the "grade of membership" of  $x$  in  $A$ .

### 2.5 Fuzzy Soft Set

(Maji, et al., 2001) presented the concept of the fuzzy soft sets by combining the ideas of fuzzy sets and soft set, and defined it as a mapping of each element in a set to the set of all fuzzy set of the universal set.

A pair  $(F, E)$  is called a fuzzy soft set over an initial universe  $U$ , where  $F: E \rightarrow \mathcal{P}(U)$  is a mapping from  $E$  (set of parameters) into  $\mathcal{P}(U)$  (set of all fuzzy sets of  $U$ ).

## 3. TRANSLATING NATURAL LANGUAGE INTO ONTOLOGY

A natural text about disability is gotten from the official site of USA Social Security Red Book and

**Table 1: Syntax and semantics of concept descriptions (Baader & Sattler, 2001)**

Constructor	Syntax	Semantics
negation	$\neg C$	$\Delta^I \setminus C^I$
conjunction	$C \sqcap D$	$C^I \cap D^I$
disjunction	$C \sqcup D$	$C^I \cup D^I$
existential restriction	$\exists r.C$	$\{x \in \Delta^I \mid \exists y : (x, y) \in r^I \wedge y \in C^I\}$
value restriction	$\forall r.C$	$\{x \in \Delta^I \mid \forall y : (x, y) \in r^I \rightarrow y \in C^I\}$
at-least restriction	$(\geq nr.C)$	$\{x \in \Delta^I \mid \#\{y \in \Delta^I \mid (x, y) \in r^I \wedge y \in C^I\} \geq n\}$
at-most restriction	$(\leq nr.C)$	$\{x \in \Delta^I \mid \#\{y \in \Delta^I \mid (x, y) \in r^I \wedge y \in C^I\} \leq n\}$

1. Disability is not been able to engage in any substantial gainful activity (SGA) because of a medically-determinable physical or mental impairment(s):
  - That is expected to result in death, or
  - That has lasted or is expected to last for a continuous period of at least 12 months.
2. Work is “substantial” if it involves doing significant physical or mental activities or a combination of both.
3. “Gainful” work activity is:
  - Work performed for pay or profit; or
  - Work of a nature generally performed for pay or profit; or
  - Work intended for profit, whether or not a profit is realized.

translated into a DL ontology.

Source:

<https://www.ssa.gov/redbook/eng/definedisability.htm> retrieved 27/10/16.

An ontology for Disabled-Person using the three (3) modules specify by (Ryan, et al., 2014) was generated as follows:

*Module 1: Syntactic Parsing*

TAGGING

Disabled/JJ Person/NN is/VBZ a/DT person/NN that/WDT can/MD not/RB engage/VB in/IN any/DT substantial-gainful-activity/NN because/RB of/IN a/DT medically-determinable/JJ physical/NN or/CC mental/JJ impairment/NN that/WDT is/VBZ expected/VBN to/TO result/VB in/IN death/NN or/CC last/VB for/IN a/DT continuous/JJ period/NN of/IN at/IN least/JJS 12/CD months/NNS

PARSER

(ROOT  
(S  
(NP (JJ Disabled) (NN Person))  
(VP (VBZ is)  
(NP

(NP (DT a) (NN person))  
(SBAR  
(WHNP (WDT that))  
(S  
(VP (MD can) (RB not)  
(VP (VB engage)  
(PP (IN in)  
(NP  
(NP (DT any) (NN substantial-gainful-activity))  
(PP (RB because) (IN of)  
(NP  
(NP (DT a) (JJ medically-determinable)  
(ADJP (NN physical)  
(CC or)  
(JJ mental))  
(NN impairment))  
(SBAR  
(WHNP (WDT that))  
(S  
(VP (VBZ is)  
(VP (VBN expected)  
(S  
(VP (TO to)  
(VP  
(VP (VB result)  
(PP (IN in)  
(NP (NN death))))  
(CC or)  
(VP (VB last)  
(PP (IN for)  
(NP  
(NP (DT a) (JJ continuous) (NN period))  
(PP (IN of)  
(NP  
(QP (IN at) (JJS least)  
(CD 12))  
(NNS  
months))))))))))))))))))))))))))))))))))))

TAGGING

Substantial/NNP Work/NNP is/VBZ doing/VBG significant/JJ physical/NN or/CC mental/JJ activities/NNS or/CC a/DT combination/NN of/IN both/DT

PARSER

(ROOT  
(S  
(NP (NNP Substantial) (NNP Work))  
(VP (VBZ is)  
(VP (VBG doing)  
(NP  
(NP (JJ significant)  
(ADJP (NN physical)  
(CC or)  
(JJ mental))  
(NNS activities))  
(CC or)  
(NP  
(NP (DT a) (NN combination))  
(PP (IN of)  
(NP (DT both))))))

(CC or)  
(NN profit)  
(CC or)  
(NN Work))  
(VP (VBN intended)  
(PP (IN for)  
(NP (NN profit))))))  
(, ,)  
(SBAR (IN whether)  
(CC or)  
(RB not)  
(S  
(NP (DT a) (NN profit))  
(VP (VBZ is)  
(VP (VBN realized))))))

### Module 2: Semantic Parsing

TAGGING

Gainful/NNP Work/NNP is/VBZ Work/NN  
performed/VBN for/IN pay/NN or/CC profit/NN or/CC  
Work/NN of/IN a/DT nature/NN generally/RB  
performed/VBN for/IN pay/NN or/CC profit/NN or/CC  
Work/NN intended/VBN for/IN profit/NN ,/  
whether/IN or/CC not/RB a/DT profit/NN is/VBZ  
realized/VBN

### Term Extraction

Disabled/JJ Person/NN ~~is/VBZ a/DT~~ person/NN  
~~that/WDT can/MD not/RB engage/VB in/IN any/DT~~  
substantial-gainful-activity/NN ~~because/RB of/IN~~  
~~a/DT~~ medically-determinable/JJ physical/NN ~~or/CC~~  
mental/JJ impairment/NN ~~that/WDT is/VBZ~~  
~~expected/VBN to/TO result/VB in/IN~~ death/NN ~~or/CC~~  
~~last/VB for/IN a/DT~~ continuous/JJ period/NN ~~of/IN~~  
~~at/IN~~ least/JJS ~~12/CD~~ months/NNS

PARSER

(ROOT  
(S  
(NP (NNP Gainful) (NNP Work))  
(VP (VBZ is)  
(NP  
(NP  
(NP (NN Work))  
(VP (VBN performed)  
(PP (IN for)  
(NP (NN pay)  
(CC or)  
(NN profit))))))  
(CC or)  
(NP  
(NP (NN Work))  
(PP (IN of)  
(NP  
(NP (DT a) (NN nature))  
(VP  
(ADVP (RB generally))  
(VBN performed)  
(PP (IN for)  
(NP  
(NP (NN pay)

Substantial/NNP Work/NNP ~~is/VBZ doing/VBG~~  
significant/JJ physical/NN ~~or/CC~~ mental/JJ  
activities/NNS ~~or/CC a/DT~~ combination/NN ~~of/IN~~  
~~both/DT~~

Gainful/NNP Work/NNP ~~is/VBZ~~ Work/NN  
~~performed/VBN for/IN~~ pay/NN ~~or/CC~~ profit/NN ~~or/CC~~  
Work/NN ~~of/IN a/DT~~ nature/NN ~~generally/RB~~  
~~performed/VBN for/IN~~ pay/NN ~~or/CC~~ profit/NN ~~or/CC~~  
Work/NN ~~intended/VBN for/IN~~ profit/NN ~~,/~~  
~~whether/IN or/CC not/RB a/DT~~ profit/NN ~~is/VBZ~~  
~~realized/VBN~~

### Concatenation

(Disabled ↔ Person) → (NP (JJ Disabled) (NN Person))  
(continuous ↔ period) → (NP (DT a) (JJ continuous) (NN period))  
(Substantial ↔ Work) → (NP (NNP Substantial) (NNP Work))  
(Gainful ↔ Work) → (NP (NNP Gainful) (NNP Work))  
Result: **Disabled-Person | continuous-period**  
**Substantial-Work**  
**Gainful-Work**

### Break Phrases

Disabled-Person is a person | **that** cannot engage in any substantial-gainful-activity because of a medically-determinable physical | **or** mental impairment | **that** is expected to result in death | **or** last for a continuous-period of at least 12 months

Substantial-Work is doing significant-physical-activities | **or** significant-mental-activities | **or** a combination of both

Gainful-Work is Work performed for pay | **or** profit | **or** Work of a nature generally performed for pay | **or** profit | **or** Work intended for profit |, whether | **or** not a profit is realized

### Relations Extraction

Disabled-Person **is a** person | Disabled-Person **engage** in any substantial-gainful-activity because of a medically-determinable physical impairment | Disabled-Person **engage** in any substantial-gainful-activity because of a medically-determinable mental impairment | impairment **result** in death | impairment **last** for a continuous-period of at least 12 months

Substantial-Work **is doing** significant-physical-activities | Substantial-Work **is doing** significant-mental-activities | Substantial-Work **is doing** a combination of both

Gainful-Work **is** Work performed for pay | Gainful-Work **is** Work performed for profit | Gainful-Work **is** Work of a nature generally performed for pay | Gainful-Work **is** Work of a nature generally  
→ *Disabled\_Person*

$\equiv \exists \text{engage.substantial\_gainful\_activity}$   
 $\sqcap (\text{medically\_determinable\_impairment (Physical)})$   
 $\sqcup \text{medically\_determinable\_impairment (Mental)}$

Medically determinable impairment result in death **or** last for a continuous-period of at least 12 months  
→ *medically\_determinable\_impairment*  $\equiv \exists \text{result.in.death}$   
 $\sqcup \exists \text{last.continuous\_period} (\geq 12)$

Substantial\_Work **is** doing  
significant\_physical\_activities **or**  
significant\_mental\_activities  
→ *Substantial\_Work*

$\equiv \exists \text{doing.significant\_physical\_activities} \sqcup \exists \text{doing.significant\_mental\_activities}$   
 $\sqcup (\exists \text{doing.significant\_physical\_activities} \sqcap \exists \text{doing.significant\_mental\_activities})$   
 $\sqcup (\exists \text{medically\_determinable\_physical\_impairment} \sqcup \exists \text{medically\_determinable\_Mental\_impairment})$

Gainful\_Work **is** Work performed for pay or profit  
→ *Gainful\_Work*  $\equiv \text{Work} \sqcap (\exists \text{perform.Pay} \sqcup \exists \text{perform.Profit})$

performed for profit | Gainful-Work **is** Work intended for profit whether **or** not a profit is realized

### Module 3: OWL DL Axioms

Step 1: Hierarchical relations <NPs> <VP> <NPs>

*Disabled\_Person is a Person* → *Disabled\_Person*  $\sqsubseteq$  *Person*

*Gainful\_Work is Work* → *Gainful\_Work*  $\sqsubseteq$  *Work*

Step 2: Non-hierarchical relations <NPs> <VP> <NPs>

Disabled\_Person **engage** in any substantial\_gainful\_activity

→ *Disabled\_Person*  $\equiv \exists \text{engage.substantial\_gainful\_activity}$

medically determinable impairment **result** in death

→ *medically\_determinable\_impairment*  $\equiv \exists \text{result.death}$

medically determinable impairment **last** for a continuous\_period of at least 12 months

→ *medically\_determinable\_impairment*  $\equiv \exists \text{last.continuous\_period} (\geq 12)$

Substantial\_Work **is** doing significant\_physical\_activities

→ *Substantial\_Work*  $\equiv \exists \text{doing.significant\_physical\_activities}$

Substantial\_Work **is** doing significant\_mental\_activities

→ *Substantial\_Work*  $\equiv \exists \text{doing.significant\_mental\_activities}$

Gainful\_Work **is** Work performed for pay

→ *Gainful\_Work*  $\equiv \text{Work} \sqcap \exists \text{perform.Pay}$

Gainful\_Work **is** Work performed for profit

→ *Gainful\_Work*  $\equiv \text{Work} \sqcap \exists \text{perform.Profit}$

Step 3: Conjunction and disjunction <NPs> <CC> <NPs>

Disabled-Person engage in any substantial-gainful-activity because of a medically-determinable physical **or** mental impairment

Step 4: Negation detection <NPs> **is not** <NPs> **or** <NPs> **does not** <VP> <NPs>

Disabled-Person is a person that **cannot** engage in any substantial-gainful-activity because of a medically-determinable physical

→ *Disabled\_Person*  $\equiv \neg \exists \text{engage.substantial\_gainful\_activity}$

### Final Ontology

*TBox* = {

*Disabled\_Person*

$\equiv \exists \text{engage.substantial\_gainful\_activity} \sqcup \exists \text{engage.substantial\_gainful\_activity}$

$\sqcup (\exists \text{doing.significant\_physical\_activities} \sqcap \exists \text{doing.significant\_mental\_activities})$

$\sqcup (\exists \text{medically\_determinable\_physical\_impairment} \sqcup \exists \text{medically\_determinable\_Mental\_impairment})$

*Substantial\_gainful\_activity*  $\equiv \text{Substantial\_Work} \sqcap \text{Gainful\_Work}$

Substantial\_Work

$$\equiv \exists \text{doing\_significant\_physical\_activities} \sqcup \exists \text{doing\_significant\_mental\_activities}$$

$$\sqcup (\exists \text{doing\_significant\_physical\_activities} \sqcap \exists \text{doing\_significant\_mental\_activities})$$

```
Gainful_Work ≡ Work ∩ (∃perform.Pay ∪ ∃perform.Profit)
}
ABox = {
Person(MUSA)
Person(PETER)
Person(ABDUL)
Person(JOHN)
Substantial_Work(Tailor)
Substantial_Work(Teacher)
Substantial_Work(Community Service)
Gainful_Work(Tailor)
Gainful_Work(Teacher)
engage(Tailor, MUSA, 0.7)
engage(Teacher, JOHN, 0.8)
Physical_impairment(Leprosy)
Physical_impairment(Anemia)
Physical_impairment(Blindness)
Mental_impairment(Insanity)
medically_determinable(Blindness, ABDUL, 0.9)
medically_determinable(Insanity, PETER, 0.5)
}
```

4. FUZZY SOFT SET ONTOLOGY

The proposed approach for uncertainty in DL ontologies extends concepts with a fuzzification component to represent fuzzy concepts in a domain

of interest, which then resulted to a transformed ontology called Fuzzy Soft Set Ontology (FSSO) that handles uncertainty by attaching a degree of membership to vague concepts of a domain. The fuzzification component maps each assertional statement of a fuzzy concept with a membership values through a mapping function  $c:I \rightarrow [0,1]$ , where  $c \in C$  (set of concepts in the domain),  $I = \{i_1, i_2, i_3, \dots, i_n\}$  is a finite set of instance and  $[0,1]$  is a fuzzy value<sup>1</sup> in the range of 0 and 1. The FSSO together with a user query<sup>2</sup> is then passed to a reasoner. The reasoner goes through the following stages:

1. Expand the concepts used in the user query up to their terminal<sup>3</sup> concepts with the normal Tableaux expansion procedure.
2. Search in the ABox of FSSO and extract the value for the terminal concepts with regards to the instance passed in the user query or all instances in the domain if there is no any instance passed in the user query.
3. Construct a comparison table with instances in the domain as the rows and the query parameters as the columns of the table with the values extracted in (2) above as the corresponding entries of the table.

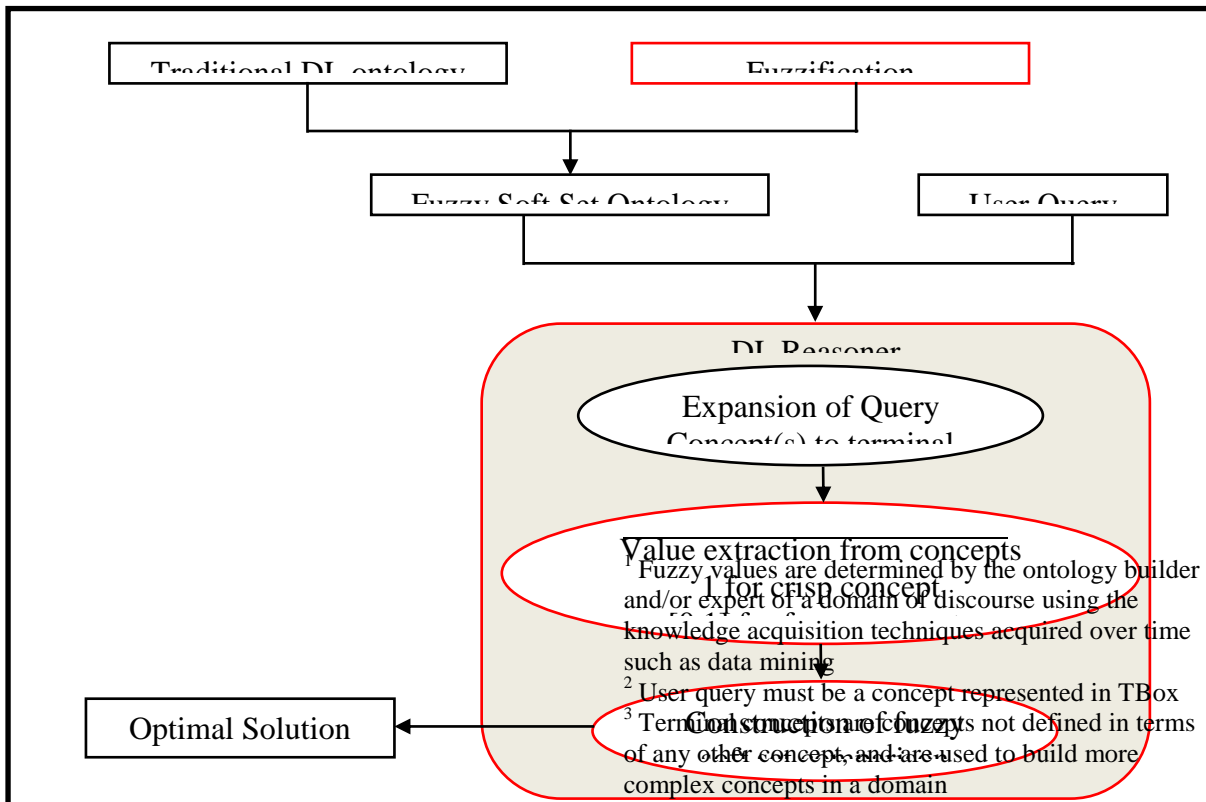


Figure 1: Proposed fuzzy soft set ontology architecture

After the construction of a comparison table, an optimal solution can be obtained by using the interpretation semantics of the constructors involved in the expansion of the query parameter. Hence the uncertainty of the concept(s) involved is resolved among all the instances in the domain. The architecture of FSSO is shown in Figure 1 below. Given a fuzzy DL-knowledge base, Fuzzy Soft Set Ontology (FSSO) can be represented as a triple  $FSSO = \langle RB_F, TB_F, AB_F \rangle$ , where  $RB_F$ ,  $TB_F$ , and  $AB_F$  are the RBox, TBox, and ABox of FSSO respectively.

The interpretation  $I$  for FSSO is obtained by combining:

1. Set of decision parameters  $M = \{m_1, m_2, \dots, m_n\}$  representing terminal concepts in the domain.
2. Domain of Interpretation  $\Delta^I$  which is the set of all instances represented in the FSSO. The combination will then results to:

$$(I, M) = \{(m_1, m_1^I), (m_2, m_2^I), \dots, (m_n, m_n^I)\}.$$

Where,  $m_i \in M, i \in \{1, 2, \dots, n\}$  are the decision parameters and  $m_i^I$  is the set of instances that satisfied the parameter  $m_i$ . Hence, the interpretation of each  $m_i$  is a proper subset of domain of interpretation ( $m_i^I \subseteq \Delta^I$ ).

**Example 1:**

A DL ontology of a medical domain was constructed using the approach of (Ryan, et al., 2014) and then extended with vague concepts to form a fuzzy soft set ontology. The  $FSSO = \langle RB_F, TB_F, AB_F \rangle$  is defined as follows:

$$RB_F = \emptyset$$

$$TB_F = \{$$

*Disabled\_Person*

$$\equiv Person \cap \neg \exists engage.substantial\_gainful\_activity$$

$$\cap (\exists medically\_determinable.Physical\_impairment$$

$$\sqcup \exists medically\_determinable.Mental\_impairment)$$

$$Substantial\_gainful\_activity \equiv Substantial\_Work \cap Gainful\_Work$$

*Substantial\_Work*

$$\equiv \exists doing.significant\_physical\_activities \sqcup \exists doing.significant\_mental\_activities$$

$$\sqcup (\exists doing.significant\_physical\_activities \cap \exists doing.significant\_mental\_activities)$$

$$Gainful\_Work \equiv Work \cap (\exists perform.Pay \sqcup \exists perform.Profit)$$

}

$$AB_F = \{$$

*Person(MUSA)*

*Person(PETER)*

*Person(ABDUL)*

*Person(JOHN)*

*Substantial\_Work(Tailoring)*

*Substantial\_Work(Teaching)*

*Substantial\_Work(Community Service)*

*Gainful\_Work(Tailoring)*

*Gainful\_Work(Teaching)*

*engage(Tailor, MUSA, 0.7)*

*engage(Teacher, JOHN, 0.8)*

*Physical\_impairment(Leprosy)*

*Physical\_impairment(Anemia)*

*Physical\_impairment(Blindness)*

*Mental\_impairment(Insanity)*

*medically\_determinable(Blindness, ABDUL, 0.9)*

*medically\_determinable(Insanity, PETER, 0.5)*

}

From the above FSSO, the set of instances is {MUSA, PETER, ABDUL, JOHN} and the set of terminal concepts is {Person, Substantial\_Work, Gainful\_Work, Physical\_impairment, Mental\_impairment}. The interpretations of the concepts are as follows:

$$Person^I = \{MUSA, PETER, ABDUL, JOHN\}$$

$$Substantial\_Work^I = \{Tailoring, Teaching\}$$

$$Gainful\_Work^I = \{Tailoring, Teaching\}$$

$$Physical\_impairment^I = \{Leprosy, Anemia, Blindness\}$$

$$Mental\_impairment^I = \{Insanity\}$$

**FSSO Reasoning Procedure**

The proposed algorithm reasons over crisp and vague concepts in a domain of interest based on fuzzy soft set theory. The reasoner takes as input a Fuzzy Soft Set Ontology and a user query, to get the degree of truthness of the query passed, to make the reasoning, the reasoner goes through two major steps:

1. Expansion step: this is where the query is expanded into its terminal concepts and the corresponding values asserted on the queried instance (or all instances if no instance passed with the query) over the terminal concepts in the  $AB_F$  will be retrieved.
2. Construction of a comparison table with the values retrieved during the expansion as entries against each instances (as rows) and the query parameters (as columns). The  $\pi_{ij}$  is the resultant fuzzy values (interpreted based on the semantics of the constructor involved) for each  $i$ , where  $i = \{1, 2, \dots, n\}$  are the instances in the domain.

Below is the FSSO reasoning algorithm:

**Input:**

A fuzzy soft set ontology

$$FSSO = \langle RB_F, TB_F, AB_F \rangle$$

A query  $Q \in FSSO$

**Output:** The degree of truthness of  $Q$

**Step 1:** Generate an interpretation for the ontology-based fuzzy soft set  $(I, M) = \{(m_1, m_1^i), (m_2, m_2^i), \dots, (m_n, m_n^i)\}$  in tabular form as follows:

- a. Traverse through the  $TB_F$  and obtain terminal concept for all the non-terminal concept in  $Q$ .
- b. Obtain the fuzzy value for each terminal concepts obtained in (a) above with respect to their corresponding instances:
  - i. Value (1) if crisply asserted.
  - ii. Value (0) if assertion not found
  - iii. Fuzzy value for assertions with fuzzy value attached in their definition

**Step 2:**

If a valid instance in the domain is attached to  $Q$ , then:

- a. Compute the resultant fuzzy value with respect to the relational operation(s) involved
- b. Return the resultant fuzzy value as the degree of truthfulness for the instance parameter against the queried concept.

Else:

- a. Construct the comparison table of the ontology-based fuzzy soft set  $(I, M)$  and compute the relational resultant for each instance  $(r_i, \forall i)$
- b. The optimal decision is to

select  $o_k$  if  $r_k = \max_i r_i$  and output  $o_k$ .

If  $k$  has more than one value then any one of  $r_k$  may be chosen.

Algorithm 1: Fuzzy Soft Set Semantic Making Decision Algorithm

The syntax and semantic of FSSO concept constructors are shown on table 2 as follow:

**Table 2: Syntax and semantic of FSSO concept constructors**

Constructor	Syntax	Semantics
Conjunction	$C \sqcap D$	$C^I \cap D^I$
Disjunction	$C \sqcup D$	$C^I \cup D^I$
Negation	$\neg C$	$1 - C^I$
Existential restriction	$\exists r. C$	$\sup_{y \in \Delta^I} \{(x, y) \in r^I \wedge y \in C^I\}$
Value restriction	$\forall r. C$	$\inf_{y \in \Delta^I} \{(x, y) \in r^I \rightarrow y \in C^I\}$
At-least restriction	$\geq nr. C$	$\{x \in \Delta^I \mid \#\{y \in \Delta^I \mid (x, y) \in r^I\} \geq n\}$
At-most restriction	$\leq nr. C$	$\{x \in \Delta^I \mid \#\{y \in \Delta^I \mid (x, y) \in r^I\} \leq n\}$

## 5. RESULT DISCUSSION

Let reason with the FSSO in Example 1 above with a user query: Disabled-Person. The query's parameters are:

- Person, which is a terminal concept itself
- $\neg \exists \text{engage.substantial\_gainful\_activity}$ , whose terminal concepts are: Substantial\_Work and Gainful\_Work
- $\exists \text{medically\_determinable.Physical\_impairment} \sqcup \exists \text{medically\_determinable.Mental\_impairment}$ , which evaluates to the terminal concepts: Physical\_impairment and Mental\_impairment.

Table 3 represent the comparison table for the query Disabled-Person using the FSSO constructors' definitions on Table 2.

**Table 3: Disabled-Person comparison table**



$i$	$Person$ $a_1$	$\neg engage.substantial\_gainful\_activity$ $a_2$	$\exists medically\_determinable.Physical\_impairment \sqcup \exists medically\_determinable.Mental\_impairment$ $a_3$	$r_i$
<b>MUSA</b>	1	0.3	0	0
<b>PETER</b>	1	0.2	0.5	0.2
<b>ABDUL</b>	1	1	0.9	0.9
<b>JOHN</b>	1	1	0	0

The columns represent the query parameters {Person,  $\neg \exists engage.substantial\_gainful\_activity$ ,  $\exists medically\_determinable.Physical\_impairment \sqcup \exists medically\_determinable.Mental\_impairment$ } for the query Disabled-Person and the rows represents the instances in the FSSO domain {MUSA,PETER,ABDUL,JOHN}. The individual entries e.g MUSA = {1, 0.3, 0} are the values gotten from the expansion of the corresponding query parameters appearing on the column header.  $r_i$  is the resultant score values for each  $i$  obtained using the semantics (as specified on Table 2) of the corresponding constructor operator connecting the decision parameters.

Consequently,  $a_1$  evaluates to 1 for each Person instance in the domain,  $a_2$  evaluates to the complement of *engage* assertion's value for each instance in the domain,  $a_3$  evaluates to the maximum value of *medically\_determinable* assertions of each instance in the domain, and finally,  $r_i$  evaluates to minimum of  $\{a_1, a_2, a_3\}$  using

the semantic of the constructor  $\sqcap$  that was used in connecting the query parameters. Hence, the following is obtained  $r_{ABDUL} = 0.9, r_{PETER} = 0.2, r_{MUSA} = 0, r_{JOHN} = 0$  from Table 3. In conclusion, the decision is maximum  $r_i$  which is ABDUL, therefore *Disabled\_Person = ABDUL (0.9)*.

### 6. CONCLUSION

In our approach, the reasoner will always output a result after a query is made into the domain regardless of the query containing a fuzzy concept or not. The reasoner extracts the membership degree attached to a fuzzy concepts in the domain of interest as represented by the ontology builder or a crispt value (1 or 0) if the concept is not fuzzy. An optimal result is established by performing the FSSO operations (shown on table 2) on the resulting fuzzy membership value of the concepts involved in the query.

This paper addresses the limitation of (Jiang, et al., 2010) of not being able to reason over fuzzy

concepts that cannot easily be specified as either true or false but better quantified in the form of degree, taking possible values of real number in the range of  $[0,1]$  and provides an efficient and flexible methods to interpret such vague concepts in an ontology in order to reason with the uncertain aspect of the world domain. This is necessary as a result of the inadequacy of traditional ontologies not been able to reason over vague concepts particularly in knowledge-based of a medical domain and also to perform inference with such vague knowledge caused by lack of having a fully and universally accepted understanding of concepts in medical domains.

In a bid to carry out the study, Ontologies are viewed as a collection of well-structured set of object whereby DL is used to represent the knowledge about the collection. In order to achieve a medical DL ontology, natural language text in a medical domain was translated into a DL ontology using the approach of (Ryan, et al., 2014). The obtained medical DL ontology was extended with fuzzy concepts to establish a Fuzzy Soft Set Ontology,  $FSSO = \langle RB_F, TB_F, AB_F \rangle$ , where  $RB_F$ ,  $TB_F$ , and  $AB_F$  are the RBox, TBox, and ABox of FSSO respectively. From this conception of ontology the notion of fuzzy soft set is used to assign a fuzzy value to non-crisp concepts in the ontology domain so as to be able to reason with the uncertain part of the world. The fuzzy soft set method of comparison for getting the optimal decision is then used as the interpretation function for both the crisp and fuzzy concepts in the domain. The approach used in this study only extends the existing models in the field of ontology modelling with inclusion of fuzzy values to vague concepts and avoidance of remodelling the whole ontology.

The results obtained shows that fuzzy value can be attached to ontological concepts, which by extension could provide efficient and flexible methods to interpret uncertainty in an ontology in order to reason with the uncertain aspect of the world domain. Thus with the approach of this study, any query involving fuzzy concept made into the proposed fuzzy soft set ontology will always be satisfiable.

This research work is suitable for medical domain where there is always uncertainty in most of the concepts used in the domain. This research can also be helpful and applicable in other real world problem such as data analysis in areas such as economics, engineering, environmental sciences, sociology, social sciences, data mining and forecasting.

The future work, one may possibly use ontology based fuzzy soft sets to address group decision making problems. An interesting topic of future research is to investigate semantic decision making using ontology-based fuzzy soft sets in a group decision making problems. It is also desirable to further explore the applications of using the ontology-based fuzzy soft set approach to solve real world problems such as data mining, forecasting, and data analysis.

## References

- Anjum, A. et al., 2007. The Requirements for Ontologies in Medical Data Integration: A Case Study. *Database Engineering and Applications Symposium, International*, pp. 308-314.
- Anon., n.d. 2016 Red Book. [Online] Available at: <https://www.ssa.gov/redbook/eng/definedisability.htm> [Accessed 27 10 2016].
- Baader, F. & Sattler, U., 2001. An Overview of Tableau Algorithms for Description Logics. *Studia Logica*, pp. 5-40.
- Baader, F. & Werner, N., 2002. Basic Description Logics. In: *The Description Logics Handbook*. Cambridge: Cambridge University Press, pp. 43-95.
- Feng, F., Jun, Y. B., Liu, X. & Li, L., 2010. An adjustable approach to fuzzy soft set based decision making. *Journal of Computational and Applied Mathematic*, pp. 10-20.
- Jiang, Y., Liu, H., Tang, Y. & Chen, Q., 2010. Semantic decision making using ontology-based soft sets. *Mathematical and Computer Modelling*, 42(11), pp. 1005-1009.
- Kana, A. F. & Akinkunmi, B. O., 2014. An Algebra of Ontologies Approximation. *Journal of Intelligence Science*, 4(2), pp. 54-64.
- Kong, Z., Gao, L. & Wang, L., 2009. Comment on "A fuzzy soft set theoretic approach to decision. *Journal of Computational and Applied Mathematics*, pp. 540-542.
- Laskey, K. J. et al., 2008. *Uncertainty Reasoning for the World Wide Web*. [Online]

- Available at: *International Handbooks on Information Systems*.  
<http://www.w3.org/2005/Incubator/urw3/XGR-urw3/>  
 [Accessed 10 January 2016]. Germany: Springer, Berlin, pp. 337-358.
- Zadeh, L. A., 1996. Fuzzy Sets. *Information and Control*, pp. 338-353.
- Maji, P. K., Biswas, R. & Roy, A. R., 2001. Fuzzy soft sets. *Journal of Fuzzy Math*, 9(3), pp. 589-602.
- Maji, P. K., Biswas, R. & Roy, A. R., 2003. Soft set theory. *Computers & Mathematics with Applications*, p. 555-562.
- Miller, G. A., 1995. A Lexical Database for English. *Communication of the ACM*, 100(1), pp. 449-462.
- Molodtsov, D., 1999. Soft set theory - First results. *Computers and Mathematics with applications*, 37(4-5), pp. 19-31.
- Richardson, R., 1994. A Semantic-based Approach to Information Processing. *School of Computer Applications, Dublin City University, Ireland*.
- Roy, A. R. & Maji, P. K., 2007. A fuzzy soft set theoretic approach to decision making problems. *Elsevier Journal of Computational and Applied Mathematics*, 203(2), pp. 412-418.
- Ryan, R. et al., 2014. Representing Knowledge in DL ALC from text. *18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014*, 35(1), p. 176 - 185.
- Sanchez, D. & Tettamanzi, A., 2006. Reasoning and quantification in fuzzy description logics. *Capturing Intelligence: Fuzzy Logic and the Semantic Web*, pp. 135-159.
- Stoilos, G., Stamou, G. Z., Tzouvaras, V. & Horrocks, I., 2007. Reasoning with Very Expressive Fuzzy Description Logics. *Journal of Artificial Intelligence Research*, pp. 273-320.
- Straccia, U., 2001. A Fuzzy Description Logic for the Semantic Web. *Artificial Intelligence Res.*, pp. 137-166.
- Uschold, M. & Gruninger, M., 1996. Ontologies: principles, methods and applications. *Knowledge Engineering Review*, pp. 93-136.
- VTempich, C. & Simperl, E., 2009. Exploring the Economical Aspects of Ontology Engineering synthetic. In: *Handbook on Ontologies*,

**13<sup>th</sup>**

**International Conference**



**Session C:**  
**E-Government, Digital Development  
and E-Readiness**

## Full Paper

## A CONCEPTUAL FRAMEWORK FOR E-CENSUS AND E-ELECTION

**A. A. Eludire**

Department of Computer Science, Joseph  
Ayo Babalola University, Ikeji Arakeji  
aaeludire@jabu.edu.ng

**ABSTRACT**

The process of official and systematic numbering or counting people is taking a census. This has led to the proposal of various census models for implementation. Some of these census models have been marred by controversies like inaccuracies, data manipulation and inflation of results. As sophisticated technology continues to impact on organizations and countries, a growing need for efficient, flexible and cost effective census system becomes paramount. The solution thus lies in employing the services of a computerized census system (e-census). This work looks at a framework for computerized census system that will enable government have first-hand information on the population of the country and tends to arrest the manual form of census taking by generating computerized database which can be updated. E-census is conceptually proposed in this work to emanate from the identification, registration and recording of all birth and death in the society. The registration of voters in electioneering processes is a major activity that can be linked with the taking of accurate census. The electorate to a larger extent has to do with a subset of the population exercising their political right within the agreed electoral framework. The identification of the qualified subset of the population to take part in the electoral process has been conceptually reduced in this work to the generation of database queries based on the selection criteria of the required population subset. The proposed framework has shown that the registration and certification of birth and death can be streamlined to improve quality of service delivery to the citizens in generating valid and reliable quantitative data on general population and voters' population. The work highlighted the needed facilities for easy registration of population for census and generation of voters' list and provision of efficient compilation of voters' register for elections.

**Keywords:** census, election, population, registration, voters

## 1. INTRODUCTION

Over the years political leaders have been using their powers, influence and personalities, riches and positions to manipulate the results of population census for their selfish, malicious and group interest. Professor J G Ottong a social scientist at the University of Calabar, explained that population has been a sensitive and controversial issue “because of its implications for shaping regional, state and ethnic relations and balance of power” (Odunfa, 2006). In the recent past, census figures were believed to have been manipulated for political advantage while some were marred by controversies like inaccuracies, data manipulation and inflation of results. There has been substantial growth in the use of information technology to improve processes and procedures worldwide. As sophisticated technology continues to impact on organizations and countries, a growing need for efficient, flexible and cost effective census system becomes paramount. The solution thus lies in employing the services of a computerized census system (e-census). With this in mind, individuals, groups, politicians, private industries, corporate bodies and government with a long plan for the country would be able to work on real (but close to exact if not exact) figure of the population at any particular point in time. This is the justification for embarking on this work. This work therefore looks at a framework for computerized census system that will enable government have first-hand and real-time information on the population of the country and tends to arrest the manual form of census taking by converting the paper work into a database which can be updated. This process involves recording census information in a database as well as keeping track of births and death certificates to update the census figure.

A functional birth and death registration system in any country is a key towards development, by supplying the most reliable data on deaths and births and population dynamics, provides indicators for health and development; and pre-requisite data for the effective planning of health and other services, resource allocation, legal, administrative and health policy formulation, program planning, and evaluation. It is an important system in collecting accurate and useful information in a country or region (Altermann, 1969; Andersen, 2009). Birth registration data are needed to

formulate programs relating to maternal and child health including nutrition, immunization and universal education; and data from death registration provide information on the economic burden of disease, and an understanding of disease ethology (NPC, 2001; Kamen, 2005; Tobin, 2013). The right to be registered at birth is founded in article 7 of the United Nations Convention of the rights of the child (United Nations, 1990). Unfortunately, the accuracy of birth and death records, particularly in developing countries, has come under question in several studies (NPC, 2006; Helen, 2006; Wadad, 2008) with the increasing realization that each year, despite a growing awareness of the importance of this registration, and the commitment of states under international law to ensure this right, several births and deaths go uncounted.

The published studies (Redmond, 2006; Naseer, 2007) carried out to explore the reason for these small proportions show that a large percentage of the populace are aware of the registration, particularly birth registration, but practice remains poor. This work considers two areas of particular concern for top management level; data collection and system monitoring. Each and every one of those processes has equally met with stories of success and failure. In some cases the declared results have called for solid fact-finding, careful negotiating, making sure that the chances of misunderstanding are minimal, and a continuous quality assurance programme (Whitford and Reichert, 2001).

Proper record keeping or information management is a key factor in the development and productivity of any organization and serves as the backbone of the business both of organization or government sector (UN, 2008). In any country of the world the birth rate and the death rate within the country is of great importance to the government in planning and making provision for the citizens. To manage this information, it is imperative to have in place a web based birth and death registration database system that is of high capacity, flexible and user friendly to use. Such a right-to-know system can be queried at any time to deliver the statistics of birth, death and number of eligible voters within a given period. The objectives of this work and hence motivations are to assure 100% accurate registration and certification of birth and death and

to improve quality of service delivery to the citizens in an effort to generate valid and reliable quantitative data on birth and death registration. The work also provides facilities for easy registration of voters and efficient compilation of voters' register for elections. This work will be of immense importance to the National Population Commission of Nigeria and Independent National Electoral Commission (INEC), health institutions and related organisations

## 2. RELATED WORKS

Previously, there have been many less-known census innovators who have put newly discovered methods and technology to good use. Information technology has usually been on the forefront of these efforts [Akomolafe & Eludire, 2009]. Census data-processing equipment has graduated from machines just assisting in tabulation work, to indispensable tools in virtually all phases of census work. Computers are used for planning, to support mapping, in project management, in all stages of data capture, cleaning, coding, and reporting, and in demographic analysis (Dekker, 1997). Many of the recent improvements in census taking have been possible due to the ever-growing capabilities of data-processing equipment and communication networks operating on local, national, and worldwide levels. For the sake of continuity it is important that the use of newer technology is embedded into, and builds upon, existing sound methodology (United Nations, 2008).

The introduction of innovation and new technologies to census and statistical operations has brought a number of concerns. The concerns include the new technology and how to choose appropriate technology; how to maintain the integrity of existing statistical systems; how to deal with outsourcing certain tasks; and how to maintain confidentiality of data. Some technologies, such as mobile telephony, the global system of mobile communication have made person-to-person communication in the field easier, as have fax and e-mail capabilities. Bar-code technology has made management of materials more efficient. In the 2000 round of censuses (United Nations Secretariat, 2001), intelligent character recognition devices made a breakthrough in many countries, although illegible handwritten

characters and badly printed questionnaires still led to problems. In general, countries that planned carefully for the new technology and conducted pre-tests were more successful in their operations.

Census mapping has made great strides in the last few decades, from an activity requiring extensive fieldwork and manual drawing to one using remote sensing and computer-assisted map production. Geographic information system technology (United Nations Secretariat, 2001) is increasingly being used in population and housing censuses to generate maps for enumeration and for data presentation purposes. Global positioning systems (Akomolafe and Eludire, 2012) are cheap and available, and they can be used by cartographic field staff to annotate topographical maps and satellite photographs to produce excellent maps for enumerators.

Data-processing software for censuses, which was previously developed and provided by non-profit agencies, is being supplanted by commercially available software. The Internet as a tool for census data collection is still in its infancy, although several countries did allow some Internet enumeration in their most recent censuses (Keller, 2000). Generally, such data were collected from a small portion of the population on an experimental basis. Problems with this method include the need for authentication from each household; lack of coverage of households in many countries at this stage; and the fear that hackers could compromise the integrity of the census. Moreover, data collected via the Internet would have to be integrated into other data streams, including mail-back questionnaires and telephone responses.

In analysing voter registration practices worldwide (Maley, 2000; Pintor and Gratschew, 2002) highlighted Universality, Equality, Direct Access and Secrecy as elements to be safeguarded by a voter register. The registration of voters is crucial for participation in democratic political context. An all-inclusive clean voter register should be considered a safeguard to the integrity of the suffrage and therefore an essential condition for the legitimacy of democracy as well as for the political stability of the country (Rial, 1999; Smith, 2001).

3. MATERIALS AND METHODS

In this work two conceptual approaches are investigated, the first approach is to link the process of census enumeration to efficient and reliable birth and death registration while the second approach links voters' registration and compilation of voter's list linked to accurate census.

3.1. Design Assumptions

The population census of any community at a locality; at the local, state or national level can be conceptual presented as in Fig. 1. In this representation t - is the total population at a given period p - year interval of enumeration (usually fixed at 10 years)

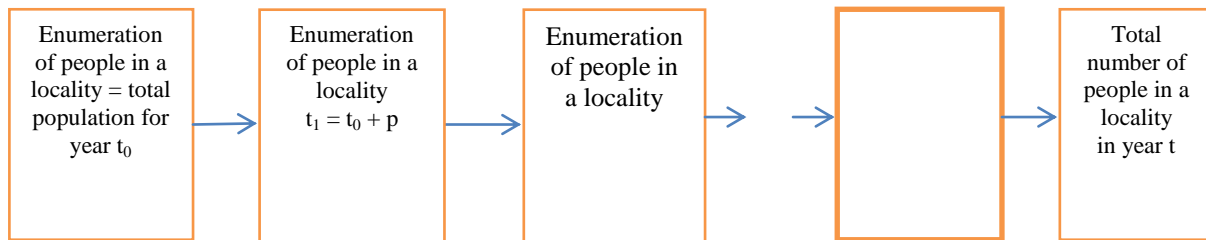


Figure 1: Classic Census Concept of fixed year interval

The graphical representation in Fig. 1 assumed that the modality for the enumeration is constant, but in reality, this cycle of census can be graphically summarised as a recurrent event that can be mathematically depicted in Fig.2. The modalities

differ but the processes are the same, so Figure 1 and Figure 2 can be taken to be the same.

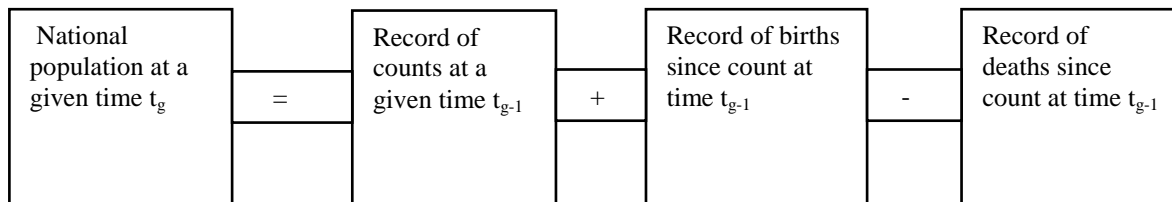


Figure 2: Mathematical representation of national population census capturing

In conducting a census, it is assumed that a starting year (year<sub>0</sub> or year<sub>1</sub>) is identified in which the physical counting or enumeration is carried out. The result of this exercise is documented and can be analysed to determine various societal/social benchmarks or variables. At the lapse of a given period (year) usually 10 or 15 years, another count or enumeration is carried out and so on.

Critically examining the process of life that has happened within the period (time span) between year<sub>0</sub> and year<sub>9</sub> (or year<sub>1</sub> and year<sub>10</sub>) in case of ten-

year period; that is the period between one census and the other, it can be summarised that two principal events would have impacted on the results of year<sub>0</sub> and year<sub>9</sub> counts. The events are number of births and deaths from the initial count of year<sub>0</sub> to year<sub>9</sub>. This means that the results of year<sub>9</sub> (Result<sub>9</sub>) differs from results of year<sub>0</sub> (Result<sub>0</sub>) by the difference in the total number of births and deaths recorded since Result<sub>0</sub>. This can be represented mathematically as shown (1):

$$\begin{aligned}
 Result_1 &= Result_0 + noOfBirths_1 - noOfDeaths_1 \\
 Result_2 &= Result_1 + noOfBirths_2 - noOfDeaths_2 \\
 Result_3 &= Result_2 + noOfBirths_3 - noOfDeaths_3
 \end{aligned}
 \tag{1}$$



$$Result_n = Result_{n-1} + noOfBirths_n - noOfDeaths_n$$

From equation (1) it can be summarised that

$$Result_n = Result_0 + \sum_1^{n-1} noOfBirths - \sum_1^{n-1} noOfDeaths \tag{2}$$

The number of deaths (noOfDeaths) would be classified to include physical death and number of emigration from the locality or other activities having reducing impact on the population. The number of birth (noOfBirths) would include physical live births and possible immigration into the locality.

The error to be introduced to final count would have to be determined to give a reliable estimation.

census of the locality over any period (5years, 10, years or 15 years). In this case, the process of population census can be reduced to the question of maintaining an initial database of locality's records, plus regular update of births and deaths records. This means that conceptually, population census can be seen as a proper maintenance of electronic records in a database as presented in Fig.3.

This work is premised on the assumptions that if a reliable and acceptable database could be established in year to record all the count then subsequent recording of occurrence of noOfBirths and noOfDeaths in the database with birth being recorded as addition and death/migration recorded as subtraction/ deletion would give the population

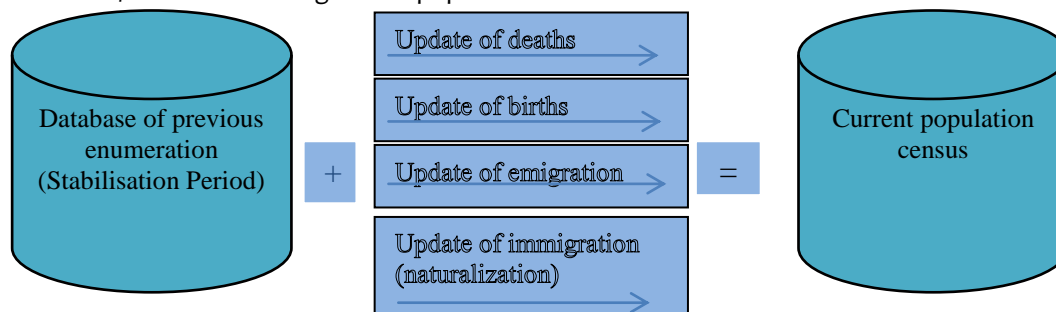


Figure 3: Enhanced census concept

For the record of birth to be accurately taken and record of death accurately taken, there must be properly controlled sources of records of birth, legal birth certificates, and registered death certificate requiring the enforcement of registration of birth and properly controlled sources of records of death with enforced registration of death. A possible element in registration will be that of issuing a National Identification number which would be available from the registration of birth.

This also calls for raising awareness, improving inter-agency cooperation and associated human rights issues amongst the identified stakeholders. Initially some stakeholders at the implementation level would include, the general populace, the local government authority, government census department, government electoral office, government employment department and foreign embassies. Conceptually, the process of obtaining and maintaining data in this system is shown in Fig.4.

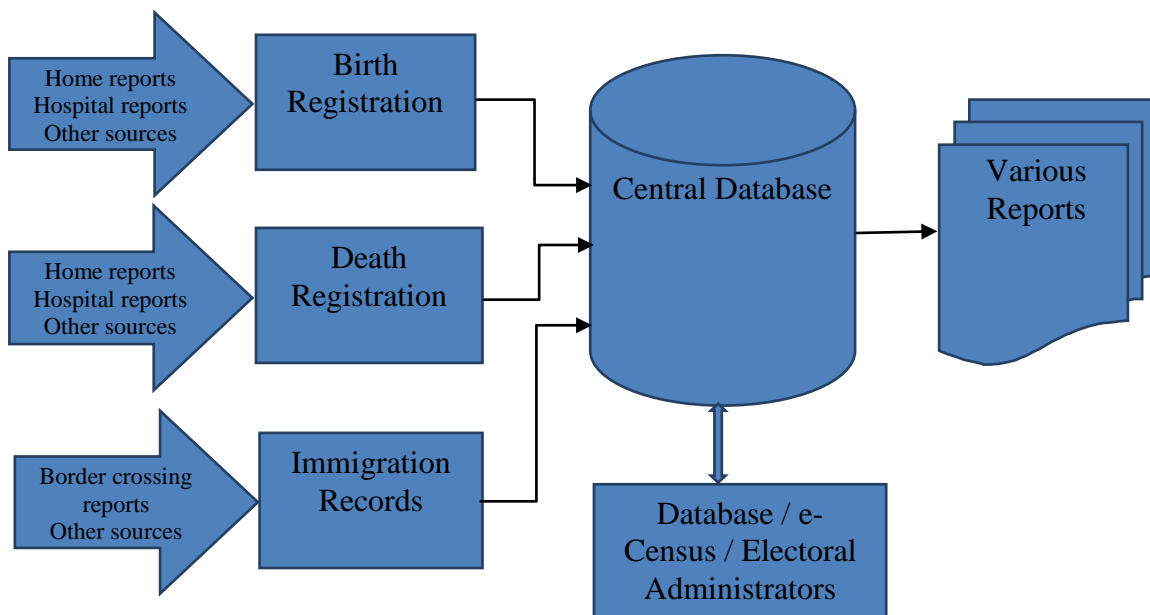


Figure 4: Framework Concept of E-Census

### 3.2. Conceptual View of Electoral Registration

Along the same thought pattern, it has been observed that registration for elections is carried out for a subset of the population. This legally depends on the age of universal suffrage ranging from 18 to 25 years. (Azinge, 1994). From the foregoing, mathematically, it can be inferred that if elections were to be conducted in year  $n$  then the list of eligible voters  $P_v$  can be generated as a subset of  $Result_n$  where age equal to or greater than 18 (taking 18 years as the universal suffrage). Mathematically, the process of registering for any election can be reduced to the determination of population subset with  $Age \geq M$  where  $M$  is the universal suffrage.

$$P_v = QuerySubset(Age \geq M) \quad (3)$$

The relationship function (3) in any given year is a number that will be a constant. The controls here are such that  $subset_n M < Result_n$  for any given year.

Bellare et al (2000) opined that it is natural for citizenry to be jittery about the polls because of certain security challenges such as the argument between pro-zoning and anti-zoning group, political thuggery, kidnapping, bomb blast and other politically motivated violence which forms the

challenges to credible elections. Challenges of free fair election in any society could best be viewed against the structural setting of such society emanating primarily from the inefficiency associated with registration of prospective voters (Caarls, 2004; Haya, 2007).

The need for building an efficient Voter Registration System is expressed in the fact that citizens are often blamed for their apathy towards the democratic process. It is said that they neglect their duty to get their name added to the voter list subsequently resulting in low percentage of voting (Saltmann, 2005). When citizens change their residence they do not transfer their name to the voter list of the new place. This leads to low voter turnout since most are not able to vote as they live far away from the place where they had registered

as Voters and sometimes bogus voting could be done in their name.

This problem is especially serious in cities of developing countries that are urbanising rapidly. Here, millions of young educated and hardworking people are migrating to cities every year, but only a small percentage of this group votes. This is noted as a great loss to the nation since if the educated and politically aware citizens vote, then they would elect leaders with progressive, development oriented policies and such leaders can quickly guide the country to a prosperous future.

The classic method for registering voters for election is depicted in Fig. 5 where registration records from different locations are pulled into a common database for the generation of voters' list to be dispatched to the different polling stations.

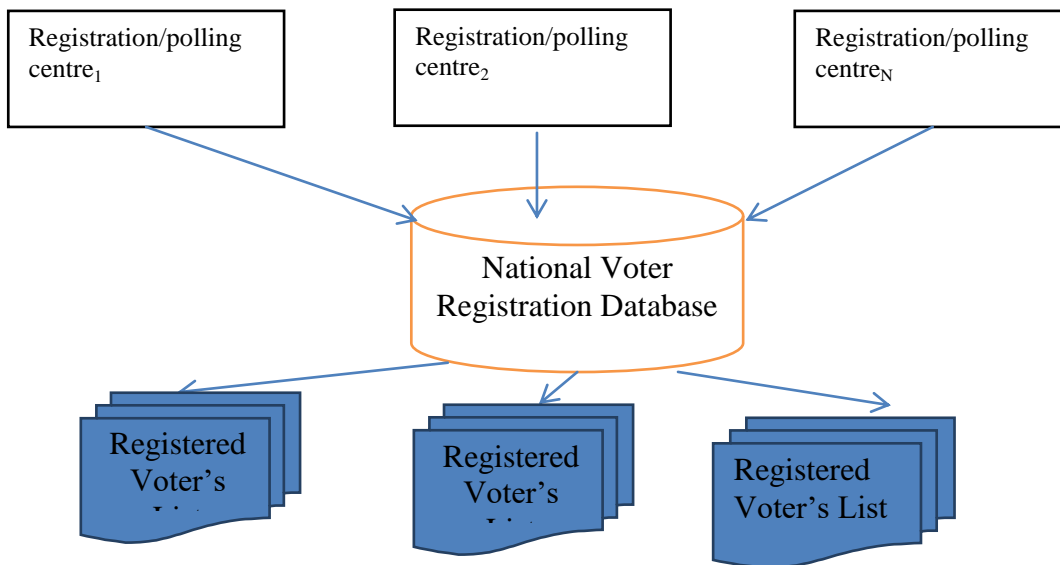


Figure 5: Classic registration of voters for elections

An enhanced approach to voters' registration for electoral process is shown in Fig. 6 where the registration records are processed for voters' list

generation based on the registered voters in the locality.

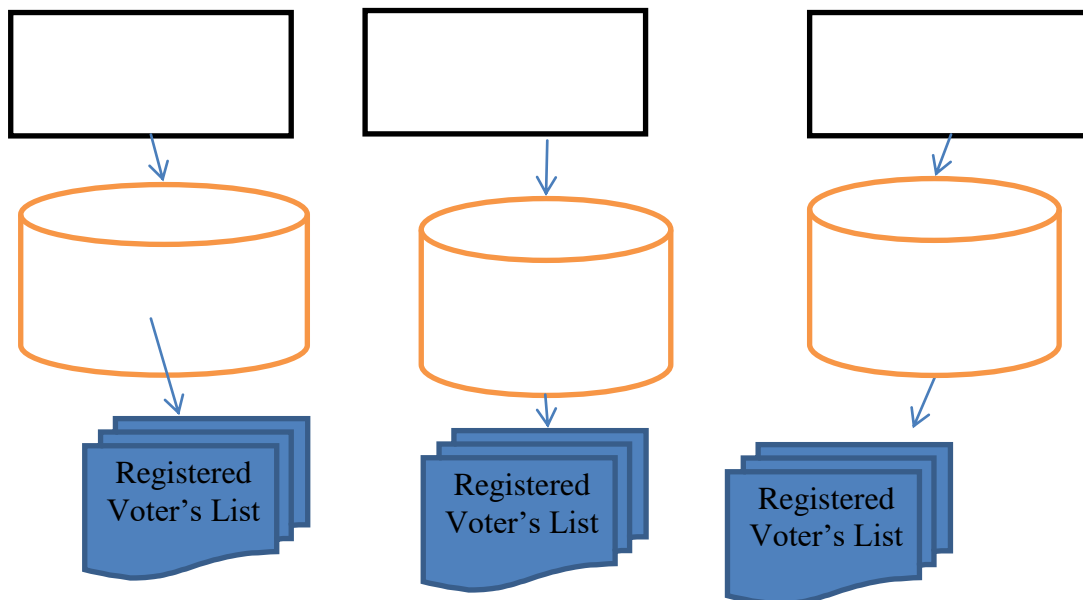


Figure 6: Enhanced registration of voters for elections

This work proposed a census based approach to the registration of voters for elections. The diagrammatic representation of this concept is shown in Fig.7. The registered voters' list is the output of a query generated on the database. The

generated list will be seen to contain the number of voters which is less than or equal to the population subset for year to determined in accordance with equation (3).

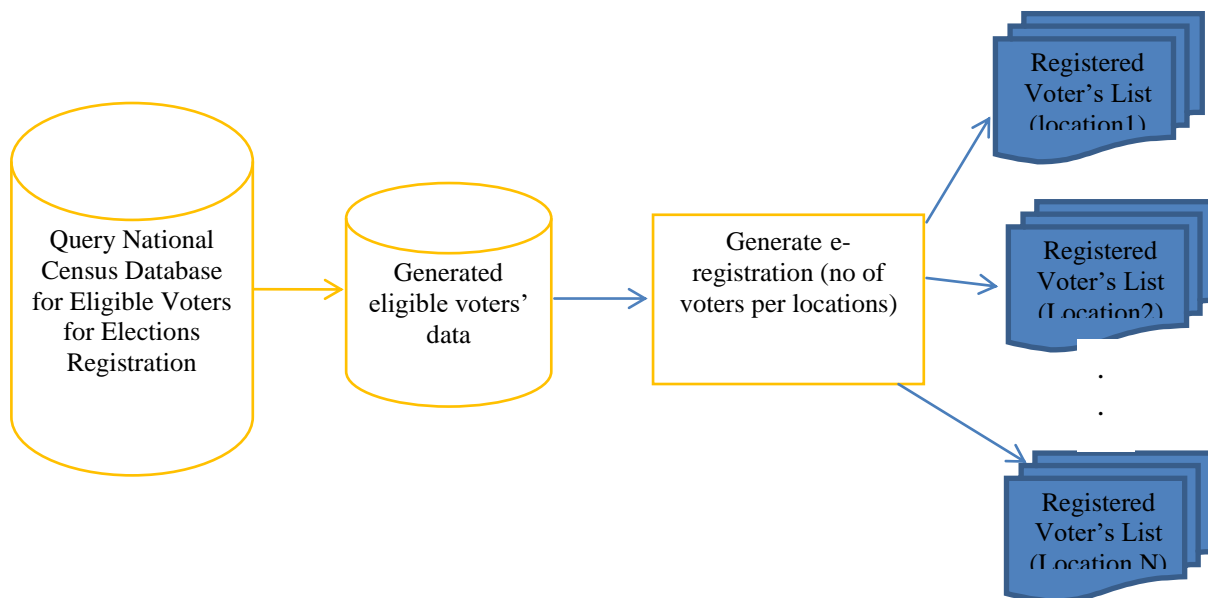


Figure 7: Proposed Census-based Electoral Registration Process

**3.3. Population/Voter Registration Systems Design Criteria**

We have adapted the voter systems design criteria specified in (Shamos, 2000; Umonbong, 2006; Grim, 2010) to census based voter registration system design and therefore identify the following design criteria:

- Authentication: Only authorized/eligible voters should be able to register in the database.
- Uniqueness: No voter should be able to register more than once.
- Accuracy: Voter registration systems should record the registration records correctly.

- Integrity: Registration records should not be modified without detection.
- Verifiability: Should be possible to verify that registration records are correctly accounted for in the main population database.
- Auditability: There should be reliable and demonstrably authentic registration records.
- Reliability: Systems should work robustly, even in the face of numerous failures.
- Flexibility: Equipment should allow for a variety of free form report generation formats.
- Convenience: Generation of voters' registration should be possible with minimal equipment and skills.

**Certifiability:** Systems should be testable against essential criteria.

**Transparency:** Voters should be able to possess a general understanding of the whole process.

**Cost-Effectiveness:** Systems should be affordable and efficient.

The Requirement analysis to guarantee the design criteria would take care of functional and non-functional requirements. The functional

requirements include Data insertion module, User registration module, Searching module using different criteria and Downloadable birth and death certificate. A Ucase diagram for the functional requirement is shown in Fig.8. Three categories of actors and their responsibilities are depicted and these are, the Universal Admin, User and Common user.

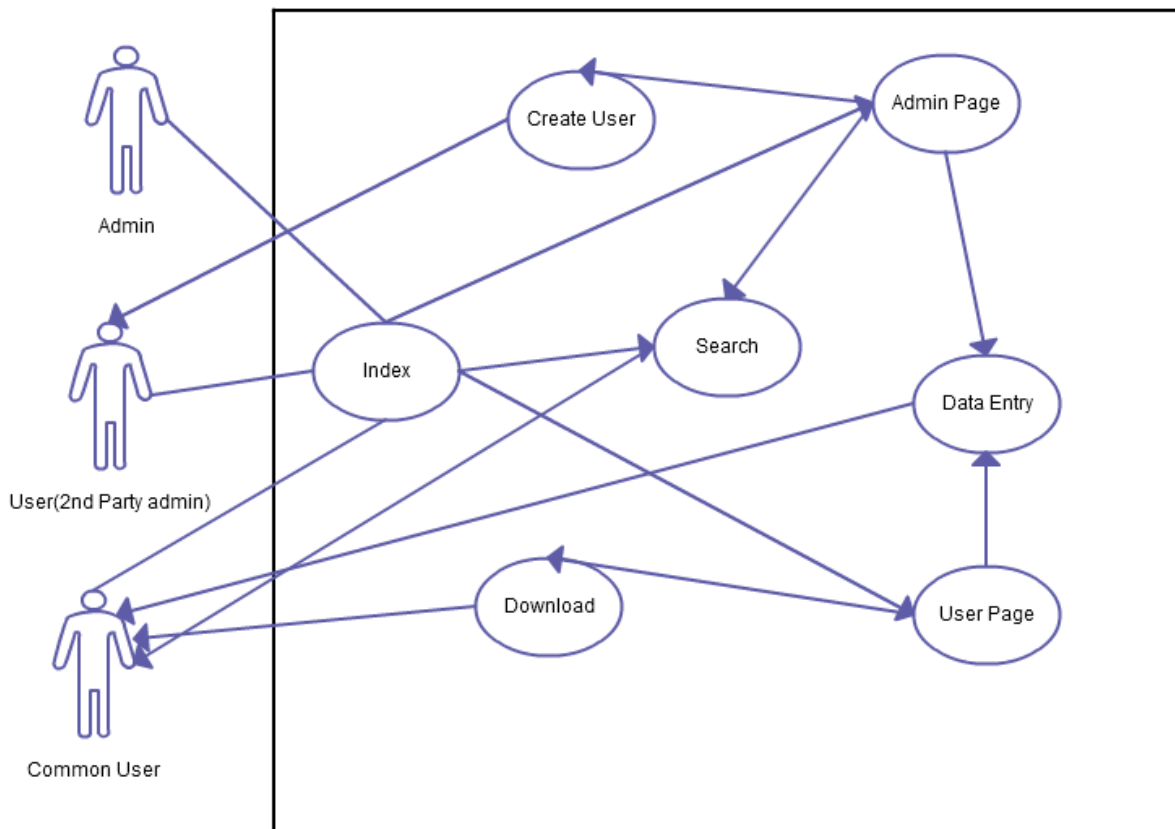


Figure 8: Use case diagram for Census Registration System

**3.4. System Database Modelling**

The database modelling was done to achieve high level of elicitation that guarantees the underlisted

- ☐ Easy accessibility to the system
- ☐ Handling data redundancy
- ☐ Reliable data preservation
- ☐ Searching option
- ☐ Efficient data handling

- ☐ Providing flexibility yet preserving the reliability of the information provided.

In the light of this, the entity structure will in the minimum contain some identified tables and associated fields. An entity structure diagram for possible data structure is shown in Table 1.

Table 1: Census Registration Database Entities

PersonalInfo
--------------

birthID	Surname	firstname	othernames	gender
---------	---------	-----------	------------	--------

birthInfo

birthID	birthDate	birthPlace	birthTime	referredby
---------	-----------	------------	-----------	------------

parentInfo

birthID	fatherName	fatherNtnalID	fatherNationality	MotherName	motherNationalID	motherNationality
---------	------------	---------------	-------------------	------------	------------------	-------------------

Address

birthID	houseNo	roadStreet	townCity	lga	state
---------	---------	------------	----------	-----	-------

deathInfo

deathID	deceasedID	dateDied	timeDied	placeDied	referredby
---------	------------	----------	----------	-----------	------------

adminInfo

adminName	adminPassword	adminType	securitylevel
-----------	---------------	-----------	---------------

#### 4. SYSTEM IMPLEMENTATION STRATEGIES AND DISCUSSIONS

There are a number of cost cutting options that must be avoided while creating this critical system to effectively realise the non-functional requirements such as:

- ☐ Advocacy team for waking awareness about the census registration
- ☐ Trained work force for data entry
- ☐ Web hosting server and Internet connected PCs
- ☐ Maintenance stuffs and mass access to the Internet

An approach for implementation to reduce costs sometimes to half is by using private contractors. One of the major items kicking against the use of private contractors instead of hiring employees for the population and election department. However in view of the significant amount of wrong doings that might occur in this critical activity if it is done by low cost private contractors, this approach is not strongly recommended. Private contractors usually employ low paid staff, who switch jobs in short time spans and no practical action can be taken against such a work force. With the security and audit trail embedded in the system, if any wrong doing is committed in this critical work then the IT System will allow us to easily pin point the operator who has committed it and appropriate strict action like suspension or dismissal can be taken against the errant person.

- In developing the IT System for e-election/voter registration, in-house IT team should be avoided since most in-house teams do not have expertise and experience to take up such a complex project.

- The most modern technology and expertise should be utilized to create a system that is very robust and can handle the load of thousands of people using it.

- The competent top IT companies in the world should be invited to bid for developing the IT system for this Census Registration System.

- The Census Registration System can be developed using Public Private Partnership involving Banks, Telephone/Mobile and Internet Service Providers to raise the needed huge amount of funds to maintain a first class system.

#### 5. CONCLUSION

In this research work, it was discovered that census results play a great role in national politics and decision making. A scroll through archives of census information and results collected shows that there exists low level of data integrity. The electoral processes starting with voter's registration is also found to be deficient. Computerized aggregate data describing the characteristics of small geographic areas for the entire period are obtainable through a computerised census system can assist in building a centralised database on recent information to be collected from the populace. The registration of deaths should also be taken into account as much

as that of births. It is unnecessary and a waste of time and resources to count people before, during and after elections following the current hand recording method and house to house census data collection systems.

This work has addressed a number of issues that could be dealt with to obtain a reliable and timely census in the country which can be integrated to provide services for electronic registration of voters using the possibility made available through the use of information technology. The proposed centralised database can be made accessible to other stakeholders such as Banks, service providers like – Telephone / Mobile Connection Providers, Internet Connection Providers etc. requiring updating of their customer's address.

The implementation cost is not addressed in this paper, however it is quite likely that banks will be happy to finance the cost of setting up and maintaining this system. The possibility that citizen can get his/her address updated in the database as soon as he/she relocates to a new address will lead to 100% population / voter registration. This is a conceptual framework that needs further works in area of data collection, refinement of implementation requirements and upgrades where more features may be included and can be implemented in real life.

## 6. REFERENCES

- Akomolafe D. T. and Eludire A. A. (2009) Prospects and Challenges of Information and Communication Technology in Achieving the Millennium Development Goals. *Journal of Science, Osun State College of Education, Ilesa*,
- Akomolafe D. T., Eludire A. A., Ofere A. F. (2012) Using Geospatial Technology to obtain Spatial Data of Nigerian Road Networks. *JABU Journal of Science and Technology* 2, 5-9
- Alterman, H. (1969). *Counting People* Harcourt: Brace Company.
- Anderson, M. (2009). *Census Bolton*: Bolton Press.
- Avi Rubin (2000): Security considerations for remote electronic voting over the internet, 2000. Available at <http://avirubin.com/e-votingsecurity.pdf>
- Azinge Epiphany (1994). *The Right to Vote in Nigeria: A Critical Commentary on the Open Ballot System* Journal of African Law Vol. 38, No. 2, pp. 173-180
- Bartolini, S. 2000. Franchise Expansion. In *International Encyclopedia of Elections*, edited by R. Rose. London: Macmillan.
- Grim J. H. (2010). *Statistical Model of the Czech Census for Interactive Presentation*. *Journal of Official Statistics*, vol. 26, no. 4.
- Helen V. (2006). *Census Results Resurrect North South Rivalry*. Paris: Agence France Press.
- Kamen, C. S. (February, 2005). *The Israel Integrated Census of Population and Housing*: Israel: Israel Publisher.
- Keller, Wouter, and Ad Willeboordse (2000). *Statistical Processing in the Internet Era: the Dutch View*. Conference on Network of Statistics for Better European Compliance and Quality of Operation, Radenci, Slovenia, 13-15 November 2000. (This paper can be retrieved from the web site of the Statistical Office of Slovenia at <http://www.sigov.si/zrs>)
- Maley, M. 2000. Administration of Elections. In *International Encyclopedia of Elections*, edited by R. Rose. London: Macmillan.
- Michael Shamos (2000): "Electronic voting/evaluating the threat". Presented at CFP '93. Available at <http://www.cpsr.org/conferences/cfp93/shamos.html>
- Nasser, Haya. (2007). *Papers show Census role. WWII camps*: USA Today.
- Okop Umonbong (2006): A paper on the voting system in Nigeria presented by, at the area seminar held in Blackpool, England.
- Odufa, Sola (2006) "Nigeria's counting controversy". *bbc.co.uk* (BBC News, 14 December 2005). Retrieved 2016-02-19 <http://news.bbc.co.uk/2/hi/africa/4512240.stm>
- National Population Commission (NPC), 1991 *Population Census of the Federal Republic of Nigeria* (Abuja, Nigeria: NPC, 1998) Page 6.
- Rafael López Pintor and Maria Gratschew *Voter Registration and Inclusive Democracy: Analysing Registration Practices Worldwide*
- Redmond, W. A. (2006). *Population Census New York*: Mc Grew.



- Report of Nigeria's National Population Commission on the 2006 Census *Population and Development Review* 33, no. 1 (2007) 209.
- Rial, J. 1999. El Registro Electoral como Herramienta para la Consolidacion Democratica. In *Seminario Internacional sobre Legislacion y Organizacion Electoral: Una Vision Comparada*. Lima: Organizacion de Estados Americanos/Transparencia
- Roy G. Saltman (2005): Accuracy, integrity, and security in computerized vote-tallying.
- Smith, A.S. (2001). *U.S Census Bureau Census Special Reports Series*. United State: Iwrin Publisher.
- Susanne Caarls (2004): E-voting handbook Key steps in the implementation of e-enabled elections
- Symposium on Global Review of 2000 Round of Population and Housing Censuses: Mid-Decade Assessment and Future Prospects. Statistics Division, Department of Economic and Social Affairs, United Nations Secretariat, New York, 7-10 August 2001
- Tobin EA, Obi AI, Isah EC (2013). Status of birth and death registration and associated factors in the South-south region of Nigeria. *Ann Nigerian Med* 2013; 7:3-7.
- VoteHere Inc., Network Voting Systems Standards, Public Draft 2, USA, April 2002. California
- Truesdell, L. E. (1965). *The Development of Punch Card Tabulation in the Bureau of the Census* New York: New York Press
- Wadād, A. (2008). *Population Census and Land Surveys under the Umayyads* Chicago: Adventure Press.
- Whitford, David, and Jennifer Reichert (2001). Quality Assurance Challenges in the United States' Census 2000. Q2001 - International Conference on Quality in Official Statistics, Organized by Statistics Sweden and Eurostat, Stockholm, Sweden, 14-15 May 2001.
- United Nations (2008) *Principles and Recommendations for Population and Housing Censuses* Statistical Papers: Series M No.67/Rev.2
- United Nations Secretariat (2001) Symposium on Global Review of 2000 Round of Population and Housing Censuses: Mid-Decade Assessment and Future Prospects. Statistics Division, Department of Economic and Social Affairs, United Nations Secretariat, New York, 7-10.

<http://www.nigerianstat.gov.ng>

<http://www.prb.org/Articles/2007/ObjectionsOverNigerianCensus.aspx>

<http://www.census.gov>

<https://www.gov.uk/government/collections/individual-electoral-registration-2014>

[https://en.wikipedia.org/wiki/Individual\\_Electoral\\_Registration#References](https://en.wikipedia.org/wiki/Individual_Electoral_Registration#References).

**13<sup>th</sup>**

**International Conference**



**Session D:**

**Cybersecurity, Infrastructure  
Protection and Digital Privacy**

## Full Paper

# A CHAOS BASED IMAGE ENCRYPTION ALGORITHM USING SHIMIZU-MORIOKA SYSTEM

**H. J. Yakubu**

Department of Mathematics/  
Statistics/Computer Science,  
University of Maiduguri, Maiduguri, Borno  
State, Nigeria.  
thejoe\_gdf@yahoo.com

**T. Aboiyar**

Department of Mathematics/  
Statistics/Computer Science,  
University of Agriculture, Makurdi,  
Benue State, Nigeria.  
taboiyar@gmail.com

**ABSTRACT**

Recent research on image encryption schemes has focused on chaotic systems in order to meet the demand for real-time secure image transmission over the Internet. In this paper, we propose a new image encryption scheme based on Shimizu-Morioka chaotic system. The scheme consists of two stages: the confusion (mixing) stage and the diffusion stage. In the confusion stage, we utilized the rich chaotic properties of the Shimizu-Morioka chaotic system by solving the system  $N$  time's steps using Euler's method and scrambled the positions of the pixel values of the image using the randomness of the solutions obtained from the chaotic system. In the diffusion stage, we generate  $N$  (where  $N$  is the size of image per colour) random integer numbers that is non-periodic and performed MOD and bitXOR operations on the shuffled image using the random numbers to obtain encrypted (diffused) image. The proposed algorithm is tested on a standard RGB image that is of size  $256 \times 256$  and is stored with TIFF file format. Performance analysis on the proposed scheme such as the statistical analysis and the sensitivity analysis show that the proposed encryption scheme is reliable and strong enough to withstand different attacks.

**KEYWORDS:** IMAGE ENCRYPTION, CHAOS, CRYPTOSYSTEM, SHIMIZU-MORIOKA SYSTEM, EQUILIBRIUM POINT.

## 1. INTRODUCTION

We are in an age where information is an asset that has value like any other asset. Information dissemination has continued to be much easier than before owing to the rapid advancement in communication technology. Today, huge amount of information (in form of text, image, audio or video) are transferred across the world over a public network called the Internet, though efficient is highly insecure, and therefore exposed to various threats (Abd El-Samie *et al.*, 2014). The need to protect sensitive images from unauthorized person wanting to have access to them becomes necessary. Image security is based on cryptography, which is the technique that transforms information to be transmitted into an unreadable and unintelligent form by encryption process so that only authorized persons can correctly recover the information by decryption process and is generally acknowledged as the best method of information protection and image security (Abraham and Daniel, 2013, Mishkovski and Kocarev, 2011).

Traditional encryption methods which include Advanced Encryption Standard (AES), Data Encryption Standard (DES), International Data Encryption Algorithm (IDEA), Rivest-Shamir-Adleman (RSA) algorithm, ElGamal algorithm have been effective solutions to the information security problems (Cao, 2013, Ye, 2013). They are still being used heavily in different forms of information security. However, they are primarily designed for text and though can be used for image encryption, are usually found not suitable due to the following three reasons: (i) Image size is always very large and therefore needs more time to encrypt it with the traditional methods, (ii) A decrypted image need not be exactly the same as the original image, since decrypted image with small distortion is usually acceptable due to human perception property and the high redundancy of image data, (iii) Digital image contents are strongly correlated and this feature is not used by the traditional methods thereby affecting their encryption efficiency (Cao, 2013, Ramadan *et al.*, 2016, Mishra *et al.*, 2012).

To improve efficiency and security of image encryption, numerous image encryption and

hiding schemes were proposed. Among these schemes, the chaos based encryption schemes turn out to be most attractive to many researchers because of its interesting properties which includes sensitivity to initial condition and control parameters, deterministic and the ergodicity (Cao, 2013, Ramadan *et al.*, 2016). These properties of chaos have much potential for application in cryptography as it is hard to make long-term predictions on chaotic systems and that means the scheme utilizing these properties will be strong against the statistical, the differential, and the brute-force attacks (Abd El-Samie *et al.*, 2014). The application of chaos to encryption of digital images started in 1997 by Fridrich and since then, many researchers applied chaos to different fields of image security (Wu *et al.*, 2012). In this paper, a three-dimensional Shimizu-Morioka chaotic system is used in developing and implementing an image encryption scheme.

## 2. SHIMIZU-MORIOKA SYSTEM

The Shimizu-Morioka system is a classical three-dimensional chaotic system first studied by Shimizu and Morioka in 1980 as a simplified model for studying the dynamics of the well-known Lorenz system for large Rayleigh number. The Shimizu-Morioka system is defined by the following nonlinear equations.

$$\begin{aligned}\dot{x} &= y, \\ \dot{y} &= -xz + x - \beta y, \\ \dot{z} &= x^2 - \alpha z.\end{aligned}\quad (1)$$

where  $(x, y, z) \in \mathbb{R}^3$  are state variables, the dot ( $\dot{\cdot}$ ) on a variable indicates the derivative of the variable with respect to time  $t$ , while  $\alpha$  and  $\beta$  are positive parameters (Köse, 2015). In this system, stable symmetric and asymmetric periodic motions as well as stochastic behaviour of trajectories, were discovered by Shimizu and Morioka through a computer simulation. Shil'nikov, (1991) presented the following observations:

- (i). As in the Lorenz model, the Shimizu-Morioka system is invariant with respect to the substitution  $(x, y, z) \rightarrow (-x, -y, z)$ .

- (ii). System (1) has three equilibrium states:  $(0,0,0)$ ,  $(\sqrt{\alpha},0,1)$  and  $(-\sqrt{\alpha}, 0,1)$ .

2.1 Stability Analysis of the Equilibrium Points of System (1)

Salih, (2011) presented the following observations and their proofs:

- a) If  $\alpha \geq 0$  then system (1) has three isolated equilibrium points:  $P_0(0,0,0)$ ,  $P_1(\sqrt{\alpha},0,1)$  and  $P_2(-\sqrt{\alpha}, 0,1)$  and for  $\alpha < 0$ , it has only one isolated equilibrium point  $P_0(0,0,0)$ .
- b) The equilibrium point  $P_0(0,0,0)$  is unstable for all  $\alpha \in \mathbb{R}$
- c) The equilibrium point  $P_1(\sqrt{\alpha},0,1)$  is asymptotically stable if and only if  $\alpha > \alpha_0 = \frac{2-\beta^2}{\beta}$  where  $\beta \in (0, \sqrt{2})$ .
- d) The equilibrium point  $P_1(\sqrt{\alpha},0,1)$  is unstable if and only if  $\alpha < \alpha_0 = \frac{2-\beta^2}{\beta}$  where  $\beta \in (0, \sqrt{2})$ .

2.2 Phase Portrait of the Shimizu-Morioka Chaotic System

The Shimizu-Morioka chaotic system is described by

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 1-z & -\beta & 0 \\ x & 0 & -\alpha \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 1-z & -0.91 & 0 \\ x & 0 & -0.365 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

where we defined our control parameters value as  $\beta = 0.91$  and  $\alpha = 0.365$ . Using a MATLAB/Simulink model, version 7.10.0 (R2010a) the phase portraits of the Shimizu-Morioka chaotic system in the xy, xz, yz, and xyz phase space were obtained as shown in Figure 1 by a, b, c, and d respectively when initial conditions are chosen as  $x_0 = 0.1$ ,  $y_0 = 0.1$  and  $z_0 = 0.1$

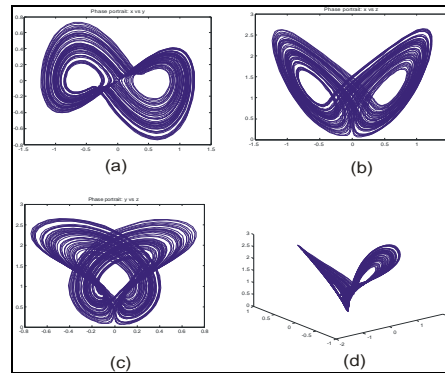


Figure 1: Phase portrait of the Shimizu-Morioka chaotic system in the (a) xy, (b) xz, (c) yz, (d) xyz phase space.

3. PROPOSED CRYPTOSYSTEM

The proposed encryption scheme consists basically, of two stages. The first stage is the *confusion* (mixing) stage and the second stage is the *diffusion* stage. In the confusion stage, we utilized the rich chaotic properties of the Shimizu-Morioka chaotic system to shuffle the image using initial conditions and control parameters as the key. In the diffusion stage, we generated a set of N (size of image per colour) random integer numbers that has irregularity and non-periodicity properties and performed MOD and bitXOR operations on the shuffled image using the random numbers in order to change the pixel values of the confused image. The resulting image is the cipher (encrypted) image. The decrypted image is obtained by applying the same operations carried out in the encryption process using the same initial conditions and control parameters but in the reverse order. The detail algorithm for encryption and decryption processes is presented below

3.1 Encryption Algorithm

- i. Read RGB image from a file as I,

- ii. Obtain the image dimension  $m \times n \times 3$ ,
  - iii. Compute number of pixels per colour ( $N = m \times n$ ),
  - iv. Enter the parameters value for  $\alpha$ ,  $\beta$ ,  $x_0$ ,  $y_0$ ,  $z_0$ ,  $h$  ( $h$  is the step size),
  - v. Solve the Shimizu-Morioka chaotic system  $N$  times steps using the Euler's method to obtain solutions in vector form as  $X$ ,  $Y$ ,  $Z$ ,
  - vi. Add confusion to the solution using round function,
  - vii. Sort the vectors  $X$ ,  $Y$ , and  $Z$  to obtain  $X_1$ ,  $Y_1$ , and  $Z_1$  with their list of indices  $l_x$ ,  $l_y$ , and  $l_z$ .
  - viii. Define  $A$ ,  $B$ , and  $C$  to be matrices for red, green and blue intensities respectively.
  - ix. Convert the  $A$ ,  $B$  and  $C$  matrices to double to obtain  $A_1$ ,  $B_1$ , and  $C_1$ .
  - x. Reshape  $A_1$ ,  $B_1$ , and  $C_1$  into row vectors as  $A_2$ ,  $B_2$ , and  $C_2$ .
  - xi. Use the indices of the sorted solution of the Shimizu-Morioka chaotic system to scramble the row vectors  $A_2$ ,  $B_2$ , and  $C_2$  and obtain new row vectors as  $A_3$ ,  $B_3$ , and  $C_3$ ,
  - xii. Generate a set of  $N$  random integer numbers that has irregularity and non-periodicity properties.
  - xiii. Perform MOD and bitXOR operations on  $A_3$ ,  $B_3$ , and  $C_3$  and the random numbers to obtain our encrypted image as  $A_4$ ,  $B_4$ , and  $C_4$ .
  - xiv. Reshape  $A_4$ ,  $B_4$ , and  $C_4$  into  $m \times n$  matrices to obtained  $A_5$ ,  $B_5$  and  $C_5$ .
  - xv. Form the encrypted image as  $I_1$  by merging  $A_5$ ,  $B_5$  and  $C_5$ .
  - xvi. Convert the image  $I_1$  to uint8.
  - xvii. Display the scrambled image  $I_1$ .
  - xviii. Save the encrypted image  $I_1$ .
- vii. Reposition the entries in  $A_9$ ,  $B_9$ , and  $C_9$  with the indices  $l_x$ ,  $l_y$ , and  $l_z$  to obtain new row vectors  $A_{10}$ ,  $B_{10}$ , and  $C_{10}$ ,
  - viii. Reshape  $A_{10}$ ,  $B_{10}$ , and  $C_{10}$  into square matrices to obtain  $A_{11}$ ,  $B_{11}$ , and  $C_{11}$ .
  - ix. Form the decrypted image as  $I_2$  by merging the  $A_{11}$ ,  $B_{11}$ , and  $C_{11}$ .
  - x. Convert the image  $I_2$  to uint8.
  - xi. Display the decrypted image  $I_2$ .
  - xii. Save the decrypted image  $I_2$  in a file

#### 4. RESULTS AND DISCUSSION

##### 4.1 Implementation

In carrying out the practical aspect of this work, we used a standard digital colour image that is of size  $256 \times 256$ , stored with TIF file format (Lena\_colour.tif) as our input data for encryption as shown in Figure 2. We implemented the code on MATLAB version 7.10.0 (R2010a) to simulate the proposed encryption algorithm.

##### 4.2 Results Obtained

After applying the proposed algorithm to the plain image in Figure 2 using initial conditions and control parameters as the key, the following results were obtained during the encryption processes as shown in Figures 3 and 4 below:

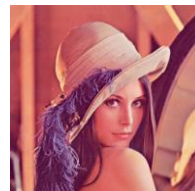







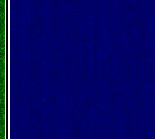
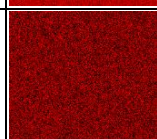
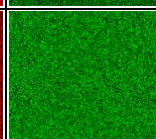
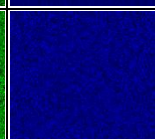
Figure 2: Original-image (Plain-image)

##### 3.2 Decryption Algorithm

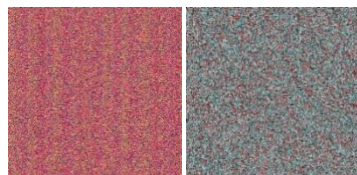
- i. Read the encrypted image  $I_1$ ,
- ii. Convert the image to double,
- iii. Define  $A_6$ ,  $B_6$ , and  $C_6$  to be matrices for the red, green and blue respectively for  $I_1$ .
- iv. Convert  $A_6$ ,  $B_6$ , and  $C_6$  to double as  $A_7$ ,  $B_7$ , and  $C_7$ .
- v. Reshape  $A_7$ ,  $B_7$ , and  $C_7$  into row vectors to obtain  $A_8$ ,  $B_8$ , and  $C_8$ ,
- vi. Perform MOD and bitXOR operations using the set of  $N$  random integer numbers on  $A_8$ ,  $B_8$ , and  $C_8$  to obtain the scrambled image as  $A_9$ ,  $B_9$ , and  $C_9$ .



Figure 3: Red, Green and Blue channel of the plain, scrambled and cipher image.

Image Type	Red Channel	Green Channel	Blue Channel
Plain			
Scrambled			
Cipher			

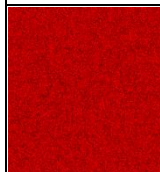
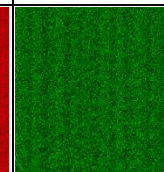
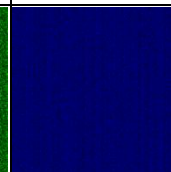
resulted in a decrypted image as shown in Figure 7.



(a) (b)

Figure 4: (a) Scrambled (Confused) image, (b) Cipher image

The decryption processes began with the cipher image- Figure 4b as input data. The cipher image is first split into red, green and blue channels as shown in Figure 3. We then performed the bitXOR and MOD operations using same set of random integer numbers to obtained undiffused image along the red, green and blue channel as shown in Figure 5. These Undiffused channels are then scrambled using the solutions obtained from the Shimizu-Morioka chaotic system with

Red Channel	Green Channel	Blue Channel
		

same initial conditions and control parameters that were used as encryption key during the encryption process are used as decryption key. Figure 6 shows the unconfused image in the red, green and blue channel. Merging these channels

Figure 5: Red, Green and Blue channels of the undiffused image.

Figure 6: Red, Green and Blue channels of the unconfused image

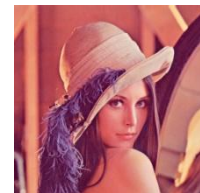
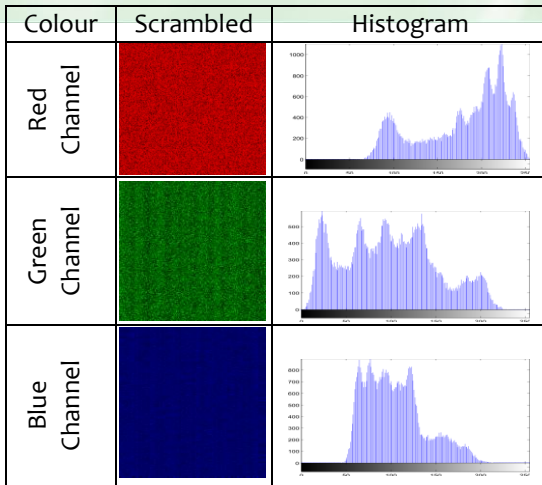


Figure 7: Decrypted image

5. SECURITY ANALYSIS

With the application of an encryption algorithm to an image, it is expected that its pixel values change when compared with the original image. A good encryption algorithm must make these changes in an irregular manner and maximize the difference in pixel values between the original and the encrypted images. Also, to obtain a good encrypted image, it must be composed of totally random patterns that do not reveal any of the features of the original image (Abd El-Samie et al., 2014). To test the robustness of the proposed scheme, security analysis such as the statistical analysis (which include histogram uniformity analysis and the correlation coefficient analysis) and the differential analysis (which include the



indicating that the attacker cannot find any hint about the plain image from the cipher image.

Figure 8: Histogram for Red, Green and Blue channel of the plain image

Figure 9: Histogram for Red, Green and Blue Channel of the scrambled (confused) image

Number of Pixel Change Rate-NPCR and Unified Average Changing Intensity-UACI) was performed.

5.1 Histogram Uniformity Analysis

For image encryption algorithm to be considered worthy of use, the histogram of the encrypted image should satisfy these two properties (Abd

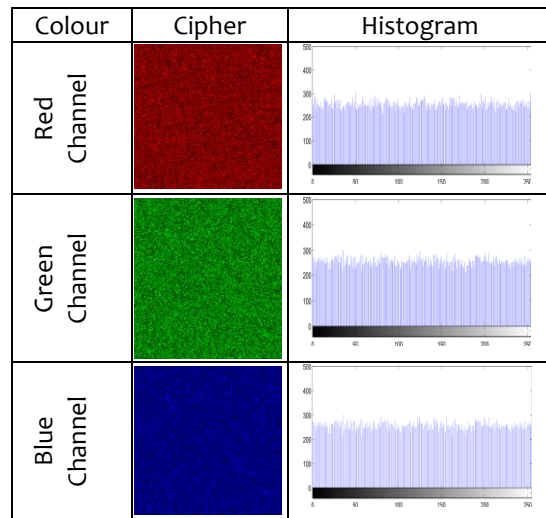
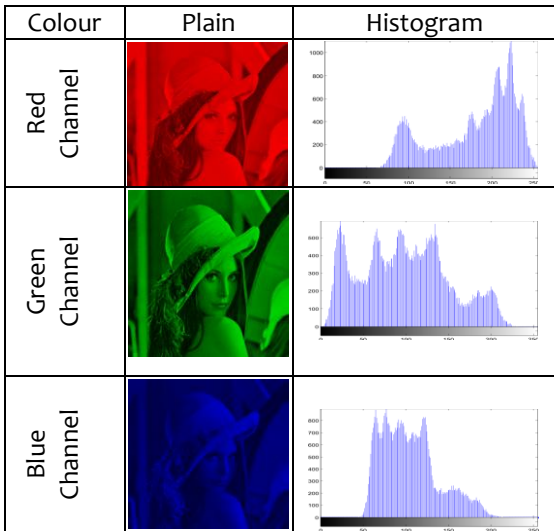


Figure 10: Histogram for Red, Green and Blue Channel of the encrypted (diffused) image



El-Samie et al., 2014):

1. It must be totally different from the histogram of the original image.
2. It must have a uniform distribution, which means that the probability of occurrence of any gray scale value is the same.

Looking at the histogram of the encrypted image-Figure 10 and that of the original image-Figure 2, the proposed scheme satisfied the two conditions of histogram uniformity analysis


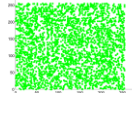
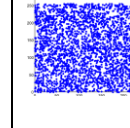
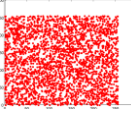

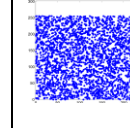
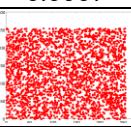
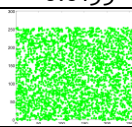
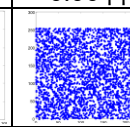
5.2 Correlation Coefficient Analysis

A useful metric to assess the encryption quality of any image encryption algorithm is the correlation coefficient between adjacent pixels of the cipher-image. In our proposed encryption algorithm, we analyzed the correlation between two vertically adjacent pixels, two horizontally adjacent pixels and two diagonally adjacent pixels in the cipher-image. We also obtained the same correlation coefficients in the plain-image for comparison purposes. This metric is calculated as follows:

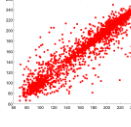
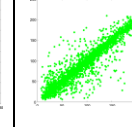
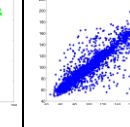
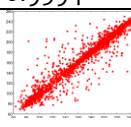
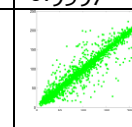
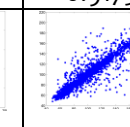
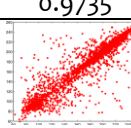
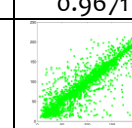
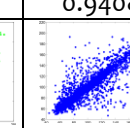
$$r_{xy} = \frac{Cov(x, y)}{\sqrt{D(x)}\sqrt{D(y)}}$$

where x and y are the values of two adjacent pixels in the cipher-image. In numerical computations, the following discrete formulas can be used:



	Red Channel	Green Channel	Blue Channel
Horizontal			
	-0.0043	-0.0014	-0.0240
Vertical			
	-0.0061	-0.0155	0.0044
Diagonal			
	-0.0018	0.0345	0.0215

$$E(x) = \frac{1}{L} \sum_{l=1}^L x_l, \quad D(x) = \frac{1}{L} \sum_{l=1}^L (x_l - E(x))^2,$$

	Red Channel	Green Channel	Blue Channel
Horizontal			
	0.9594	0.9397	0.9175
Vertical			
	0.9735	0.9671	0.9408
Diagonal			
	0.9333	0.9139	0.8808

and 
$$Cov(x, y) = \frac{1}{L} \sum_{l=1}^L (x_l - E(x))(y_l - E(y))$$

where  $L$  is the number of pixels involved in the calculations. The closer the value of  $r_{xy}$  to zero, the better the quality of the encryption algorithm will be (Abd El-Samie et al., 2014, Sathishkumar et al., 2011, Ye, 2013).

The correlation coefficient analysis in the plain and cipher image of Lena are shown in Figures 11

and 12 respectively. From Figure 11, we can see that the plain image is strongly correlated with an average of about 0.94 while in Figure 12, we see that there is almost no correlation among the adjacent pixels in the cipher images as these can be seen clearly from their respective correlation values which is almost zero in all the three directions. This indicates that the attacker cannot find any information regarding the plain image from the cipher image.

**Figure 11:** Correlation Coefficient Analysis of the Shimizu-Morioka chaotic image encryption scheme on plain image.

**Figure 12:** Correlation Coefficient Analysis of the Shimizu-Morioka chaotic image encryption scheme on cipher image.

### 5.3 Sensitivity Analysis

For an image encryption scheme to be able to resist the differential attack efficiently, it must be sensitive to small changes in the original image. That is, one small change in the plain image must cause a significant change in the cipher image. To test the influence of only one-pixel change in the plain-image over the whole cipher-image, we used two common measures: The Number of Pixel Change Rate (NPCR) and the Unified Average Changing Intensity (UACI). The NPCR measures the percentage of different pixels' numbers between the two cipher-images whose plain-images only have one-pixel difference, whereas, the UACI measures the average intensity of differences between the two cipher-images. They indicate the sensitivity of the cipher-images to the minor change of plain-image. The formula for evaluating NPCR and UACI are as follows:

$$NPCR = \frac{\sum_{i,j} D(i,j)}{W \times H} \times 100\% \quad \text{and}$$

$$UACI = \frac{1}{W \times H} \left[ \sum_{i,j} \frac{|C_1(i,j) - C_2(i,j)|}{255} \right] \times 100\%$$

where  $C_1$  and  $C_2$  denote the two ciphered images whose corresponding plain-images have only one-pixel difference, the  $C_1(i,j)$  and  $C_2(i,j)$  represent the gray scale values of the pixels at

grid  $(i,j)$  in the  $C_1$  and  $C_2$  respectively, the  $D(i,j)$  is a binary matrix with the same size as the images  $C_1$  and  $C_2$  whose entries is determined from  $C_1(i,j)$  and  $C_2(i,j)$  by the following: if  $C_1(i,j) = C_2(i,j)$ , then  $D(i,j) = 0$ , otherwise,  $D(i,j) = 1$ . The  $W$  and  $H$  are the width and height of the image (Ramhrishnan et al., 2014, Ramadan et al., 2016, Wu et al., 2011, Wu et al., 2012).

Although these two tests are compactly defined and are easy to calculate, test scores are difficult to interpret in the sense of whether the performance is good enough. Wu et al. (2011) made some findings on the acceptable NPCR and UACI scores for an image encryption scheme to be considered secured. Theoretical values of NPCR and UACI scores of binary and gray images were evaluated at 0.05-level, 0.01-level and 0.001-level. Their results show that the type and size of image used have significant influence on the NPCR and UACI scores. An NPCR score is acceptable if the experimental score is equals to or greater than the theoretical NPCR score. Also, for UACI score, the experimental UACI score should be on or within the theoretical UACI critical scores.

Table 1 below present our experimental NPCR and UACI scores. The theoretical NPCR scores for gray images with size 256 x 256 at 0.05-level, 0.01-level and 0.001-level are 99.5693%, 99.5527% and 99.5341% respectively (Wu et al., 2011). Looking at our experimental NPCR scores in Table 1, the proposed scheme has satisfied the requirement. The theoretical UACI critical values for gray images with size 256 x 256 at 0.05-level, 0.01-level, and 0.001-level are 33.2824% - 33.6447%, 33.2255% - 33.7016%, and 33.1594% - 33.7677% respectively (Wu et al., 2011). Also, our experimental UACI score in Table 1 has passed the requirement. Thus, our proposed scheme can withstand any differential attack.

**Table 1:** The NPCR and the UACI Scores for the Proposed Scheme on the Image-Lena.

NPCR (%)	UACI (%)
99.57	33.43

**6. CONCLUSION**

To improve the security of image transmission, we proposed in this paper, a new confusion-diffusion cryptosystem which we achieved by utilizing the rich chaotic properties of the 3-D Shimizu-Morioka chaotic system to shuffled the image. The encrypted image is obtained by performing bitXOR and MOD operations on the shuffled image using a set of generated random integer numbers that is non-periodic. The proposed scheme is tested on a standard colour image Lena.Tif. We also performed security analysis such as the histogram uniformity analysis, the correlation coefficient analysis, the NPCR and the UACI on the proposed scheme. From the experimental results obtained, the proposed scheme is highly secured and strong against the statistical, the differential and the brute-force attacks.

**7. REFERENCES**

Abd El-Samie, E. F., Ahmed, H. E. H., Elashry, F. I, Shahieen, H. M., Faragallah, S.O., El-Rabaie, M. E., and Alshebeili, A. S. (2014). Image Encryption- A Communication Perspective. 1<sup>st</sup> Edition. CRC Press, London, pp 1-86.

Abraham, L., and Daniel, N. (2013). Secure Image Encryption Algorithms: A Review. *International Journal of Scientific and Technology Research*, Vol. 2, No. 4, pp 186 – 189.

Cao, Y. (2013). A New Hybrid Chaotic Map and its Application on Image Encryption and Hiding. *Mathematical Problems in Engineering*. 728375: 13pp.

Köse, E. (2015). Controller Design by Using Sliding Mode and Passive Control Methods for Continuous Time Non-linear Shimizu-Morioka Chaotic System. *International Journal of Engineering Innovation and Research*. Vol. 4, No. 6, pp 895-902.

Mishkovski, I. and Kocarev, L. (2011). Chaos-Based Public-key Cryptography. Springer-Verlag Berlin Heidelberg. SCI 354, pp 27-65.

Mishra, M., Mishra, P., Adhikary, M. C. and Kumar, S. (2012). Image Encryption Using Fibonacci-Lucas Transformation. *International Journal on Cryptography and Information Security*. Vol. 2, No. 3, pp 131-141.

- Ramadan, N., Ahmed, H. H., Elkhamy, S. E., Abd Abd El-Samie, F. E. (2016). Chaos-Based Image Encryption Using an Improved Quadratic Chaotic Map. *American Journal of Signal Processing*, Vol. 6, No. 1, pp 1-13.
- Ramahrishnan, S., Elakkiya, B., Geetha, R., and Vasuki, P. (2014). Image Encryption Using Chaotic Maps in Hybrid Domain. *International Journal of Communication and Computer Technologies*, Vol. 2, No. 5, pp 44 – 48.
- Sallih, H. R. (2011). The Stability Analysis of the Shimizu-Morioka System with Hopf Bifurcation. *Journal of Kirkuk University-Scientific Studies*, Vol. 6, No. 2, pp 184-200.
- Sathishkumar, G. A., Bagan, K. B., and Sriraam, N. (2011). Image Encryption Based on Diffusion and Multiple Chaotic Maps. *International Journal of Network Security and its Applications*, Vol. 3, No. 2, pp 181 – 194.
- Shil'nikov, A. L. (1991). Bifurcation and Chaos in the Shimizu-Morioka System. *Selecta Mathematica Sovietica*, Vol. 10, No. 2, pp 105-117.
- Wu, Y., Noonan, J. P., and Agaian, S. (2011). NPCR and UACI Randomness Tests for Image Encryption. *Cyber Journals: Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Telecommunications*, pp 31-38.
- Wu, Y., Yang, G., Jin, H., and Noonan, J. P. (2012). Image Encryption Using the Two-dimensional Logistics Chaotic Map. *Journal of Electronic Imaging*, Vol. 21, No.1, 28pp.
- Ye, R. (2013). A Highly Secure Image Encryption Scheme Using Compound Chaotic Maps. *Journal of Emerging Trends in Computing and Information Sciences*, Vol. 4, No. 6, pp 532 – 544.

## Full Paper

**AN IMPROVED RSA ALGORITHM BASED ON RESIDUE  
NUMBER SYSTEM****Y.K. Saheed**

Department of Physical Sciences,  
Al-Hikmah University, Ilorin  
yksaheed@alhikmah.edu.ng

**K.A. Gbolagade**

Department of Computer Science, Kwara  
State University, Malete  
kazeem.gbolagade@kwasu.edu.ng

**ABSTRACT**

Cryptography is the science of writing in secrets. In cryptography, public key algorithms are known to be slower than symmetric key alternatives because of their basis in modular arithmetic. The modular arithmetic is computationally heavy and time consuming. In view of this, it has become a great challenge to implement RSA in a faster way. The encryption operation in the RSA cryptosystem is  $c = me \text{ mod } n$ . This seems to be an expensive computation, involving  $e-1$  multiplications by  $m$  with increasingly large intermediate results, followed by a division by  $n$ . In this work, we provide two optimizations to make the operation easy: firstly, multiplying by an appropriate sequence of previous intermediate values, rather than only by  $m$ , can reduce the number of multiplications to no more than the twice the size of  $e$  in binary. And secondly, dividing and taking the remainder after each multiplication keeps the intermediate results the same size as  $n$ . Hereafter, we introduce a conversion from binary to residue number system using traditional moduli set  $\{2n-1, 2n, 2n+1\}$  for the further encryption of the message and to confuse a cryptanalyst. This moduli set provides faster speed and required little hardware. The key length is also increased as the moduli set are used as part of the private key. The residue number system promises the proposed algorithms to be highly parallelizable, well adapted to parallel architecture and well suited to hardware implementations. Theoretically, our results provide a faster way of encryption operation of Rivest Shamir Adleman (RSA) algorithm while also improving the security.

**KEYWORDS:** RSA algorithm, moduli set, Cryptanalyst, Encryption, Cryptography.

## 1. INTRODUCTION

Cryptography provides techniques for keeping information secret, for determining that information has not been tampered with, and for determining who authored pieces of information. Cryptography is an important technique used on various applications, especially for internet and business transactions. With the help of cryptography most of the communication applications provide security, authentication, privacy, integrity, and non-repudiation (Damrudi and Norafida, 2013). Cryptography is important whenever sensitive information is to be communicated in network. Computer security has an important role in securing information (Damrudi and Norafida, 2013).

Most of the cryptography algorithms can be classified as private key cryptography and public key cryptography. In private key cryptography, same key is used for encrypting and decrypting the message, whereas different keys are used for encrypting the message and decrypting the message in public key cryptography. Public key cryptography provides more security compared to private key cryptography (Abu, Deepthi and Puskar, 2015).

The problem with the number theoretic cryptosystems (e.g., RSA) is that they require a lot of computational power for providing a high level of security and most likely a low level of efficiency (Saraiva, 2009). The main bottleneck of Public key algorithms is that they are slower compared to symmetric key alternatives because of their basis in modular arithmetic. Hence, how to make a more efficient and faster implementation of public key algorithms is a great concern to researchers in the field of cryptography. For solving the mentioned problem, we proposed a better and faster way to increase the speed of RSA cryptosystem in this paper.

The optimization of the exponentiation computation is our focus in this paper. The problem of the RSA algorithm is that executing the exponentiation for large numbers takes lots of time, and needs advanced technologies. In this paper, we solved this problem via decomposing the exponentiation into smaller parts. The speedup result from theoretical analysis confirms that using this approach leads to better results.

### 1.1 Background and related work

Recently, Abu et al., (2015) presented a parallel algorithm to manage the RSA decryption complexity by exploring the impact of compute unified device architecture and Pthread on decryption in RSA. Mohammed, Heba, and Jawad (2016) presented an acceleration of the RSA Processes based on Parallel Decomposition and Chinese Remainder theorem. This work proposes variant decompositions to gain extra speed up. Saxena and Kapoor (2014) proposed the new Parallel RSA algorithm based on repeated square-and multiply method. Also, a few researchers have worked on parallel RSA such as Fan et al., (2010) Li et al., (2010), and Qing et al., (2010). Fewer multiplications and less delay in a new parallel exponentiation algorithm for RSA were presented in Sepahvandi et al., (2009).

Computer systems have a significant role in securing important information from bank accounts to medical records, personal e-mails, and websites, confidential records and much sensitive data stored digitally (Damrudi and Norafida, 2013). Due to the growth of the internet, much of this information needs to be secured where stored and while transferring daily from one station to another. Security has become an increasingly important feature of daily life with the growth of electronic communication (Alkar and Sonmez, 2004). At first, cryptography was just used by government or military and then larger businesses started to use it.

Nowadays, by increasing the users of the internet and e-mail, ATM (automated teller machine), and smart cards, cryptography is used by almost all people. In the near future, the consumers of the internet and e-mail services, ATM and finally, digital world services will look for more security. In the internet era, a great deal of data is being encrypted and decrypted for each net user, and so it is extremely important to increase the encryption and decryption speed of algorithms (Bielecki and Burak, 2007). This growth will require a new approach to cryptography.

### 2.1 RSA ALGORITHM

The RSA cryptosystem, invented by Ron Rivest, Adi Shamir, and Leonard Adleman (1978), was first publicized in the August 1977 issue of Scientific American. The RSA algorithm can be used for public key encryption and digital signatures. Figure 1 shows the public key operations. The cryptosystem

is most commonly used for providing privacy and ensuring authenticity of digital data. These days RSA is deployed in many commercial systems. It is used by web servers and browsers to secure web traffic, it is used to ensure privacy and authenticity of Email, it is used to secure remote login sessions, and it is at the heart of electronic credit-card payment systems. In short, RSA is frequently used in applications where security of digital data is a concern.

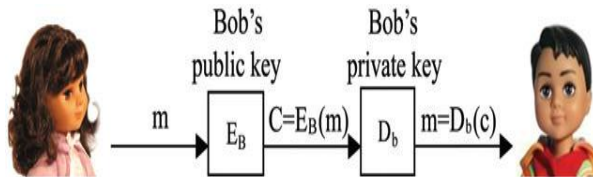


Figure.1 Public key cryptography scheme

## 2.2 Key generation Algorithm for RSA public-key encryption

Each entity creates an RSA public key and a corresponding private key (Menezes, Van, and Vanstone, 1996). Each entity A should do the following:

- i. Generate two large random (and distinct) primes  $p$  and  $q$ , each roughly the same size.
- ii. Compute  $n = pq$  and  $\theta = (p - 1)(q - 1)$ .
- iii. Select a random integer  $e$ ,  $1 < e < \theta$ , such that  $\gcd(e; \theta) = 1$ .
- iv. Use the extended Euclidean algorithm to compute the unique integer  $d$ ,  $1 < d < \theta$ , such that  $ed \equiv 1 \pmod{\theta}$ .
- v. A's public key is  $(n; e)$ ; A's private key is  $d$ .

**Definition** The integers  $e$  and  $d$  in RSA key generation are called the encryption exponent and the decryption exponent, respectively, while  $n$  is called the modulus.

## 2.3 RSA Algorithm

SUMMARY: B encrypts a message  $m$  for A, which A decrypts.

1. Encryption. B should do the following:

- (a) Obtain A's authentic public key  $(n; e)$ .
- (b) Represent the message as an integer  $m$  in the interval  $[0; n - 1]$ .
- (c) Compute  $c = m^e \pmod{n}$ .
- (d) Send the ciphertext  $c$  to A.

2. Decryption. To recover plaintext  $m$  from  $c$ , A should do the following:

- (a) Use the private key  $d$  to recover  $m = c^d \pmod{n}$

## 3.1 RESIDUE NUMBER SYSTEM

A residue number system is defined in terms of a relatively-prime moduli set  $\{P_1, P_2, \dots, P_n\}$  that is  $\gcd(P_i, P_j) = 1$  for  $i \neq j$ . A weighted binary number  $X$  can be represented as  $X = (x_1, x_2, \dots, x_n)$ , where  $x_i = X \pmod{P_i}$ ,  $0 \leq x_i < P_i$  (1)

Such a representation is unique for any integer  $X$  in the range  $[0, M-1]$ , where  $M = P_1 P_2 \dots P_n$  is the dynamic range of the moduli set  $\{P_1, P_2, \dots, P_n\}$  (Jenkins and Leon, 1977). Addition, subtraction and multiplication on residues can be performed in parallel without carry propagation. Hence, by converting the arithmetic of large numbers to a set of the parallel arithmetic of smaller numbers, RNS representation yields significant speed up (Mi, 2004). Binary to residue conversion is very simple and can be implemented with modular adders (Guan and Jones, 1988). When binary to residue conversion of the needed operands had finished, arithmetic operations on RNS numbers are performed in parallel without carry-propagation between residue digits. Hence, RNS leads to carry-free, parallel and high-speed arithmetic. It should be noted that each modulo of the moduli set has its own arithmetic processor which consists of a modulo adder, a modulo subtractor and a modulo multiplier (Molahosseini and Navi, 2007; Wang, 1998). In order to use the result of arithmetic operations in outside of RNS, the resulted RNS number must be converted into its equivalent weighted binary number.

## 3.2 PROPOSED RSA CRYPTOSYSTEM

The exponentiation is optimized as follow. To compute  $m^3 \pmod{n}$ , we compute  $m^2 \pmod{n}$  with one modular squaring, then  $m^3 \pmod{n}$  with a modular multiplication by  $m$ . Then, we convert the encrypted message into residue number system using three moduli set  $\{2^n - 1, 2^n, 2^n + 1\}$ . The decryption is done similarly: one first computes  $c^2 \pmod{n}$ , then  $c^3 \pmod{n}$ ,  $c^6 \pmod{n}$ , and  $c^7 \pmod{n}$  by alternating modular squaring and modular multiplication.

### Key Pair

Public key:  $n = 55, e = 3$       Private key:  $n = 55, d = 7$

### Key Pair Generation

Primes:  $p = 5, q = 11$   
 modulus:  $n = pq = 55$   
 Public exponent:  $e = 3$   
 Private exponent:  $d = 3^{-1} \pmod{20} = 7$

Message	Encryption	Ciphertext (C <sub>1</sub> )	Ciphertext (C <sub>2</sub> )	Decryption			
M	$m^2 \bmod n$	$m^3 \bmod n$	$\{2^n - 1, 2^n, 2^n + 1\}$	$M = c^7 \bmod n$	$C^2 \bmod n$	$c^6 \bmod n$	$c^7 \bmod n$
0	0	0	{0,0,0}	0	0	0	0
1	1	1	{1,1,1}	1	1	1	1
2	4	8	{2,0,3}	9	17	14	2
3	9	27	{0,3,2}	14	48	49	3
4	16	9	{0,1,4}	26	14	31	4
5	25	15	{0,3,0}	5	20	15	5
6	36	51	{0,3,1}	16	46	26	6
7	49	13	{1,1,3}	4	52	9	7
8	9	17	{2,1,2}	14	18	49	8
9	26	14	{2,2,4}	31	49	36	9

The main operation of modular exponentiation is multiplication and we have tried to reduce the number of multiplication in our approach and also introduced a residue number system based on three traditional moduli set  $\{2^n - 1, 2^n, 2^n + 1\}$  which was proposed in (Saheed and Gbolagade, 2016) for further transformation of the ciphertext so as to create more confusion to cryptanalyst. The reduction in the size of the power makes the computation of the proposed approach to be faster.

#### 4. CONCLUSION

Public key algorithms has there basis in modular arithmetic. The modular arithmetic in RSA is computationally expensive. Hence, public key algorithms become slower. In this work, we suggest a better way to enhance the speed of RSA algorithm by splitting and reducing the power operation. Theoretically, our proposed approach optimizes the exponentiation of the RSA algorithm compared to the primitive RSA encryption and decryption operation. In future, we intend to look at the implementation and parallelism concept to improve the speed of public key algorithms, since parallelism is known to be a technique to accelerate various applications.

#### REFERENCES

- Masumeh Damrudi & Norafida Ithnin. 2013. Parallel RSA encryption based on tree architecture. *Journal of the Chinese Institute of Engineers*, 36:5, 658-666.
- Abu Asaduzzaman, Deepthi Gummadi, and Puskar Waichal. 2015. A Promising Parallel Algorithm to Manage the RSA Decryption Complexity. *In: Proceedings of the IEEE Southeast Con 2015*, April 9 - 12, 2015 - Fort Lauderdale, Florida
- P. Saraiva, 2009. Openssl acceleration using graphics processing units, <https://fenix.tecnico.ulisboa.pt/downloadFile/395145839854/Resumo.pdf>, 2009. Retrieved on 15 April, 2017.
- Mohammed Issam Younis, HebaMohammed Fadhil, Zainab Nadhim Jawad 2016. Acceleration of the RSA Processes based on Parallel Decomposition and Chinese Remainder Theorem. *International Journal of Application or Innovation in Engineering & Management*. Volume 5, Issue 1, January 2016. Pg12-23
- S. Saxena, B. Kapoor.2014.An efficient Parallel Algorithm for Secured Data Communications using RSA Public Key Cryptography Method. *IACC, 4th IEEE International Advance Computing Conference*.
- Fan, W., Chen, X., and Li, X., 2010. Parallelization of RSA algorithm based on compute unified device architecture. *In: 9th international conference on grid and cooperative computing (GCC)*, 1-5 November 2010, Nanjing, China. Washington, DC: IEEE Computer Society, 174-178.
- Li, Y., Liu, Q., and Li, T.,2010. Design and implementation of an improved RSA algorithm. *In: International conference on e-health networking, digital ecosystems and technologies (EDT)*, 17-18 April 2010, Shenzhen, China. Piscataway, NJ: IEEE Computer Society, 390-393.
- Qing, L., Yunfei, L., and Lin, H., 2010. On the design and implementation of an efficient RSA variant. *In: 3rd international conference on advanced computer theory and engineering (ICACTE)*, 20-22 August 2010, China. Piscataway, NJ: IEEE Computer Society, Vol. 3, 533-536.
- Sepahvandi, S., et al., 2009. An improved exponentiation algorithm for RSA cryptosystem. *In: International conference on research challenges in computer science, 2009. ICRCCS '09*, 28-29 December 2009, Shanghai. Piscataway,



- NJ: IEEE Computer Society, 128–132.
- Alkar, A.Z. and Sonmez, R., 2004. A hardware version of the RSA using the montgomery's algorithm with systolic arrays. *Integration, the VLSI journal*, 38 (2), 299–307.
- Bielecki, W. and Burak, D., 2007. Parallelization method of encryption algorithms. New York, NY: Springer.
- Rivest, R.L., Shamir, A., and Adleman, L. 1978. A method for obtaining digital signatures and public key cryptosystems. *Commun. of the ACM*, 21:120 -126, 1978.
- Saheed, Y.K. and Gbolagade, K.A., 2016. Efficient Image Encryption Scheme Based on the Moduli Set  $\{2^n - 1, 2^n, 2^n + 1\}$ . *Al-Hikmah Journal of Pure & Applied Sciences* Vol.3 (2016): 15-21.
- Menezes A., Van Oorschot, and S. Vanstone 1996. *A Handbook of Applied Cryptography*. CRC Press.
- Jenkins W.K. and Leon, B.J. 1977. The use of residue number systems in the design of finite impulse response digital filters. *IEEE Trans. Circuits System*, Vol. 24, pp. 191–201.
- Guan, B. and Jones, E.V. 1988. Fast conversion between binary and residue numbers. *Electronics Letters*, Vol. 24, No. 19, pp. 1195–1197.
- Molahosseini, A.S. and Navi, K. 2007. New Arithmetic Residue to Binary Converters. *International Journal of Computer Sciences and Engineering Systems*, Vol.1, No.4, pp. 291-295.
- Mi, L. 2004. *Arithmetic and Logic in Computer Systems*. John Wiley & Sons, Inc., Hoboken, New Jersey.
- Wang, Y. 1998. New Chinese Remainder Theorem, In: *Proceedings of Asilomar Conference, USA*.

## Full Paper

**ARITHMETIC OPERATIONS IN DETERMINISTIC P SYSTEMS  
BASED ON THE WEAK RULE PRIORITY****C. M. Peter**

Department of Mathematical Sciences and  
Information Technology,  
Federal University, Dutsin-ma – Nigeria.  
macpee3@yahoo.com

**D. Singh**

Department of Mathematics,  
Ahmadu Bello University, Zaria – Nigeria  
(Former Professor, Indian Institute of  
Technology, Bombay)  
mathdss@yahoo.com

**ABSTRACT**

Membrane computing, otherwise known as P system, is a recently introduced area of distributed parallel computing of a biochemical type. Several variants have been considered in the literature. In this paper, a variant of P systems for arithmetic operations on non-negative integers based on weak priorities for rule application is considered. Consequently, we obtain deterministic P systems. Two membranes suffice. There are four objects for multiplication and five objects for division. These objects are acted upon by six rules for each of the two binary operations. Therefore, the model is simple – without complicated moves and does not face the task of deciding which rule to apply. Moreover, there exist potentials for possible extensions of the P system model to accommodate negative integers and rational numbers.

**Keywords:** Membrane computing, binary operation, determinism, weak rule priority

## 1. INTRODUCTION

The concept of membrane computing was introduced by Gheorghe Păun a 1998 report as a computability model which abstracts its structures and functions from the biological cell. The main ingredient of membrane computing is the notion of a membrane structure, which consists of several cell-like membranes recurrently placed inside a unique skin membrane. It is usually represented in the form of a Venn diagram without intersecting sets and with a unique superset. The number of membranes in a membrane structure is called its *degree*. To each membrane there exists a region which it encloses.

Multisets of objects are placed inside the regions. Such objects correspond to the molecules swimming in the solution in the compartments of a biological cell. The multiplicity of a chosen object in a region corresponds to a positive integer. The input region contains multisets of objects which encode the input data. The absence of such an object corresponds to zero. The encoding takes the form which is given in Alhazov (2006).

The reaction rules placed inside the regions, which are cooperative in some cases, are responsible for the objects to evolve and be transferred from one membrane region to another. They are similar to the chemical reactions taking place in a biological cell. If the objects placed inside the regions are able to evolve, we obtain a computing device called a *P system*. The evolution takes place in stages called *configurations*. If the system does not halt then no result can be said to have been computed. Therefore, a computation terminates if the system halts at a configuration which is thereafter called the *final configuration*. The result of the computation is encoded by the multisets of pre-determined objects remaining in the output membrane. In other words, the multiplicity of a chosen object in the output membrane represents the output of the computation. See Păun (2000) and Păun (2006) for details.

*Deterministic P systems* and *P systems with priorities* are some of the basic variants currently under investigation. Some binary operations have been modeled with *P systems*; see Alhazov (2006), Atanasiu (2000), Chen *et. al.* (2014), Guo *et. al.* (2013) and Zeng *et. al.* (2012). These operations are basically the elementary mathematical operations – addition, subtraction, multiplication and division. Addition and subtraction are trivial with particular

references to Guo and Chen (2008) and Guo and Zhang (2008); however, we explain them here in subsequent sections. A variant of multiplication appears in Guo and Zhang (2008) with single membrane. It uses eleven rules and fourteen objects probably as a trade off to its single membrane. Moreover, division uses twelve rules and thirteen objects. There is a need to minimize the number of rules and objects. In this paper, an attempt is made to present *P systems* with at most five rules and two membranes for arithmetic operations on non-negative integers by exploiting determinism and the *weak priority* relation for rule application.

## 2. A JUSTIFICATION OF THE MODEL

Two applications of membrane computing, namely, static sorting and circuits simulations were studied (Păun and Thierrin, 2001). Number sorting has been extensively studied in computer science. Membrane computing allows one to simulate such Boolean circuits. It becomes necessary to model *P systems* for the basic binary operations.

An important feature of *P systems* that provides an edge over digital computing is their use of parallelism. Objects that have access to a rule should use such rule (with the restriction imposed by the priority relation). Moreover, all membranes work in parallel. Though the effect of these two levels of parallelism on the complexity of computations performed by *P systems* is not clarified as of yet (Păun, 2000).

So far, the weak rule priority has not been considered seriously by authors of *P systems*, particularly in the modeling of the basic binary operations. Instead, the strong rule priority relation has been in use probably to consider energy accounting and other resources in their model. Nevertheless, the concluding remark:

*Of course also, the weak interpretation of the priority is of interest: a rule is always used when objects exist which were not used by a rule of a higher priority.*

presented in Păun (2000) provides a motivation to explore applications of weak rule priority relation in modeling *P systems* for arithmetic operations efficiently. Further motivation arises from the fact that the maximality of the parallelism achieved by

the weak rule priority exceeds that achieved by the strong rule priority. This is because in a weak rule priority system, rules of lower priorities will not have to wait for a subsequent iteration when there exist objects upon which they can be applied. Moreover, the objects that the rules are applied to may also have to wait for a subsequent iteration in a strong rule priority system. Therefore, the maximality of parallelism is achieved in both the object level of parallelism and the rule application level of parallelism by exploiting the weak rule priority system.

3. SOME PRELIMINARY CONCEPTS

3.1 Determinism

A P system is said to be *deterministic* if in each step of a computation there is no more than one choice of rules to be applied (that is, either the computation has reached its halting configuration, or the next configuration is uniquely determined). On the other hand, a system is non-deterministic if there exists at least a step in the computation such that the system can assume any one of two or more states. In a membrane system, a rule whose objects exist in a region is applied on the objects. Possibilities that there are at least two rules in a region exist. These rules may or may not compete for objects placed within the region. If the rules compete for objects then the P system will be deterministic. However, if *priorities* are assigned to the rules then the P system could be non-deterministic. (Freund and Păun, 2000, Ibarra, 2005 and Păun, 2000).

3.2 Weak versus strong rule priority relations

As mentioned earlier, in order to obtain determinism rules are sometimes assigned priorities to determine which rule is applied first and which is applied next. So far, there are two types of priority relations in membrane computing, namely, strong and weak priority relations. In a strong priority relation, when a rule with higher priority has been applied, the rule with lower priority is not applied even if the two rules do not compete for objects. For instance, if  $a \rightarrow c > b \rightarrow d$  are rules with the priority relation '>', meaning that  $a \rightarrow c$  has priority over  $b \rightarrow d$  and both  $a$  and  $b$  are available, then only the first rule is used, even though it has nothing to do with the object  $b$ . The relevance of this type of priority relation corresponds with the way of using priorities in

*ordered grammars* in the regulated rewriting. It also has biochemical relevance. This relevance emanates from the idea that the rules could consume not only objects, but also a common resource such as energy. If a rule with higher priority is used, then no energy would be left for a rule with lower priority. On the other hand, in a weak priority relation, a rule with lower priority is applied after a rule with higher priority has been applied or cannot be applied, provided there are objects available for the rule with lower priority. In other words, a rule of lower priority is always used where a rule of higher priority has or has not been used provided the former can be applied. Both interpretation of rule priority are of interest (Păun, 2000, for further details).

4. VALIDITY VERIFICATION OF THE P SYSTEM MODEL ON BASIC ARITHMETIC OPERATIONS

We present the P systems for the basic binary operations. There are two rules for addition and one rule for subtraction. No priority exists among the rules for addition – they are applied in parallel. On the other hand, there exist priorities among the rules in membrane 2 for both multiplication and division.

4.1 Addition P system

An addition P system is a construct of the form

$$\Pi^+ = (V, \mu, w_1, (R_1, \rho_1), i_0)$$

where

$V = \{a, b, c\}$  is a set of objects.

$\mu = [1]_1$  is a membrane structure of degree 1.

$w_1 = a^m b^n$  is the initial multiset of objects in the membrane.

$R_1 = \{a \rightarrow c, b \rightarrow c\}$  is a set of rules.

$i_0 = 1$  is the label for the output membrane.

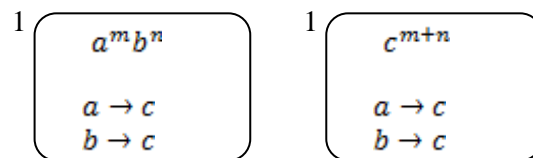


Figure 4.1: Addition P system

There are  $m$  copies of an object  $a$  and  $n$  copies of an object  $b$  in membrane 1 which is both an input

and an output membrane. The  $m$  copies of  $a$  and  $n$  copies of  $b$  encode the numbers to be added. There are two rules necessary for the addition to take place, namely  $a \rightarrow c$  and  $b \rightarrow c$ . The rule  $a \rightarrow c$  consumes the  $m$  copies of  $a$  and produces  $m$  copies of  $c$  while the rule  $b \rightarrow c$  consumes the  $n$  copies of  $b$  and produces  $n$  copies of  $c$ . The rules are applied in parallel in a single transition. Afterwards, no copies of  $a$  or  $b$  exists within the membrane. Consequently, none of the rules can be applied. Therefore, the computation halts after the first transition. The result of the computation is encoded by the  $m + n$  copies of  $c$ .

#### 4.2 Subtraction P system

A subtraction P system is a construct of the form

$$\Pi^- = (V, \mu, w_1, (R_1, \rho_1), i_0)$$

Where

$V = \{a, b, c\}$  is a set of objects.

$\mu = [1]_1$  is a membrane structure of degree 1.

$w_1 = a^m b^n$  is the initial multiset of objects in the membrane.

$R_1 = \{ab \rightarrow (c, out)\}$  is a set of rules.

$i_0 = 1$  is the label for the output membrane.

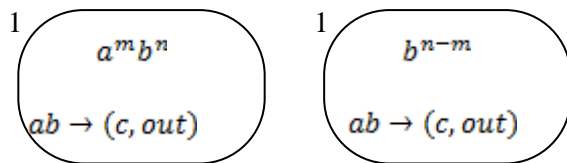


Figure 4.2: Subtraction P system

The computation works as follows:  $m$  copies of an object  $a$  and  $n$  copies of an object  $b$  are placed in the only membrane labeled 1 which is both an input and output membrane. We assume  $m \leq n$ . The  $m$  and  $n$  encode the two numbers involved in the binary operation of subtraction where  $m$  is to be subtracted from  $n$ . There is only one rule in the membrane, namely,  $ab \rightarrow (c, out)$ . The rule consumes  $m$  copies of the pair  $ab$  of  $a$  and  $b$  and produces  $m$  copies of a new object  $c$  which it sends out of the membrane. Afterwards, no copy of the pair  $ab$  exists in the membrane. Therefore, the rule cannot be applied again. The computation halts after the first transition. Consequently, there are  $n - m$  copies of  $b$  in the output membrane. The result of the computation is determined by the  $n - m$  copies of  $b$ .

#### 4.3 Multiplication P system

Let  $m$  and  $n$  be non-negative integers. A multiplication P system for the product  $m \times n$  is a construct of the form

$$\Pi^* = (V, \mu, w_1, w_2, (R_1, \rho_1), (R_2, \rho_2), i_0)$$

where

$V = \{a, b, c, e\}$  is a set of objects.

$\mu = [1[2]_2]_1$  is a membrane structure of degree 2.

$w_1 = \emptyset$  is the empty multiset (no object present in membrane 1 at the initial configuration).

$w_2 = a^m b^n e$  is the initial multiset of objects in membrane 2.

$R_1 = \{r_1: a \rightarrow (a, in_2), r_2: e \rightarrow (e, in_2), r_6: a \rightarrow (a, out)\}$

is the set of rules in membrane 1.

$\rho_1 = \{r_1 > r_6\}$  is the priority relation of the rules in membrane 1.

$R_2 = \{r_3: be \rightarrow (e, out), r_4: e \rightarrow \delta, r_5: a \rightarrow (ac, out)\}$  is the set of rules in membrane 2.

$\rho_2 = \{r_3 > r_4 > r_5\}$  is the priority relation of the rules in membrane 2.

$i_0 = 1$  is the label for the output membrane.

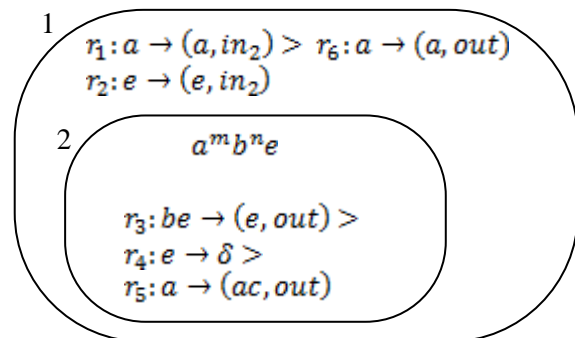


Figure 4.3: Arbitrary initial configuration for a multiplication P system

At the initial configuration of the system, there are  $m$  copies of object  $a$ ,  $n$  copies of object  $b$  in membrane 2 which encode the numbers to be multiplied and one copy of object  $e$  in membrane 2, the input membrane. Membrane 1 has three rules  $r_1, r_2$  and  $r_6$  where  $r_1 > r_6$  is the weak rule priority relation which determines the precedence in selecting the rules. Membrane 2 also has three

rules,  $r_3$ ,  $r_4$  and  $r_5$  and they have been assigned the weak rule priority relations  $r_3 > r_4 > r_5$ .

Since there are no objects in membrane 1 at the initial configuration, none of the rules in membrane 1 can be applied. Since  $r_3$  has priority over  $r_4$  in membrane 2,  $r_3$  will be applied. It consumes one copy of  $b$  and sends the only copy of  $e$  to membrane 1, leaving  $m$  copies of  $a$  and  $n - 1$  copies of  $b$  behind. Next, even though  $r_4$  has priority over  $r_5$ ,  $r_4$  cannot be applied since no object  $e$  is available in membrane 2 after  $r_3$  has been applied, therefore,  $r_5$  will be applied in parallel with  $r_3$  owing to the weak priority relation on the rules. It produces  $m$  copies of  $c$  according as there are  $m$  copies of  $a$  in membrane 2. The  $m$  copies of  $c$  and  $a$  are sent to membrane 1. Afterwards, there will be  $n - 1$  copies of  $b$  in membrane 2. This is the first transition.

Next, the copies of  $a$  and the only copy of  $e$  are sent back to membrane 2 by  $r_1$  and  $r_2$ , respectively. Each time that the use of  $r_3$  decreases the copies of  $b$  by one in membrane 2 and at the same time that  $r_4$  cannot be applied,  $m$  copies of  $c$  are produced and sent to membrane 1 by  $r_5$ . The process continues until there are no copies of  $b$  left in membrane 2. Consequently,  $r_3$  will no longer be applied. However,  $r_4$  being the next in priority to  $r_3$  will be applied owing to the availability of  $e$  in the membrane. It dissolves membrane 2. Eventually,  $r_1$  and  $r_2$  cannot be applied since they point to membrane 2. This will cause  $r_6$  to be applied. It sends the copies of  $a$  out of the membrane. At this point, no rule can be applied and the computation halts. The result of the computation is determined by the multiplicity of  $c$  available in membrane 1, which is the output membrane.

#### 4.4 Multiplication P system exemplified

The following example demonstrates the computation of  $5 \times 7$  by the multiplication P system. At the initial configuration there are seven copies of  $a$ , five copies of  $b$  and one copy of  $e$  in membrane 2, the input membrane. The input integers for the computation are encoded by the multiplicities of  $a$  and  $b$ . No rule can be applied in membrane 1 at this moment as it has no objects in it. Meanwhile,  $r_3$  has priority over  $r_4$  in membrane 2, therefore,  $r_3$  will be applied. It consumes one copy of  $b$  and sends the only copy of  $e$  to membrane 1. The rule  $r_4$  is next in priority, however, it will not be applied since no object  $e$  exists in membrane 2

having applied  $r_3$ . Rather,  $r_5$  will be applied. It produces seven copies of  $c$  which it sends to membrane 1 together with the seven copies of  $a$ .

There are now seven copies of each of  $a$  and  $c$  and one copy of  $e$  in membrane 1. Since  $r_1$  has priority over  $r_6$  then  $r_1$  and  $r_2$  will be applied. The rule  $r_1$  sends the seven copies of  $a$  to membrane 2 while  $r_2$  sends the only copy of  $e$  to membrane 2. Thus, membrane 1 contains seven copies of  $c$  while membrane 2 contains seven copies of  $a$ , four copies of  $b$  and one copy of  $e$ .

**Table 4.1:** An illustration of the product of 5 and 7 with the multiplication P system

Membrane 1		Membrane 2	
Rule	Object	Rule	Object
$r_1: a \rightarrow (a, in_2)$ $r_6: a \rightarrow (a, out)$ $r_2: e \rightarrow (e, in_2)$	(empty)	$r_3: be \rightarrow (e, out)$ $r_4: e \rightarrow \delta >$ $r_5: a \rightarrow (ac, out)$	$a^7 b^5 e$
$r_1: a \rightarrow (a, in_2)$ $r_6: a \rightarrow (a, out)$ $r_2: e \rightarrow (e, in_2)$	$a^7 c^7 e$	$r_3: be \rightarrow (e, out)$ $r_4: e \rightarrow \delta >$ $r_5: a \rightarrow (ac, out)$	$b^4$
$r_1: a \rightarrow (a, in_2)$ $r_6: a \rightarrow (a, out)$ $r_2: e \rightarrow (e, in_2)$	$c^7$	$r_3: be \rightarrow (e, out)$ $r_4: e \rightarrow \delta >$ $r_5: a \rightarrow (ac, out)$	$a^7 b^4 e$
$r_1: a \rightarrow (a, in_2)$ $r_6: a \rightarrow (a, out)$ $r_2: e \rightarrow (e, in_2)$	$a^7 c^{14} e$	$r_3: be \rightarrow (e, out)$ $r_4: e \rightarrow \delta >$ $r_5: a \rightarrow (ac, out)$	$b^3$
$r_1: a \rightarrow (a, in_2)$ $r_6: a \rightarrow (a, out)$ $r_2: e \rightarrow (e, in_2)$	$c^{14}$	$r_3: be \rightarrow (e, out)$ $r_4: e \rightarrow \delta >$ $r_5: a \rightarrow (ac, out)$	$a^7 b^3 e$
$r_1: a \rightarrow (a, in_2)$ $r_6: a \rightarrow (a, out)$ $r_2: e \rightarrow (e, in_2)$	$a^7 c^{21} e$	$r_3: be \rightarrow (e, out)$ $r_4: e \rightarrow \delta >$ $r_5: a \rightarrow (ac, out)$	$b^2$
$r_1: a \rightarrow (a, in_2)$ $r_6: a \rightarrow (a, out)$ $r_2: e \rightarrow (e, in_2)$	$c^{21}$	$r_3: be \rightarrow (e, out)$ $r_4: e \rightarrow \delta >$ $r_5: a \rightarrow (ac, out)$	$a^7 b^2 e$
$r_1: a \rightarrow (a, in_2)$ $r_6: a \rightarrow (a, out)$ $r_2: e \rightarrow (e, in_2)$	$a^7 c^{28} e$	$r_3: be \rightarrow (e, out)$ $r_4: e \rightarrow \delta >$ $r_5: a \rightarrow (ac, out)$	$b$
$r_1: a \rightarrow (a, in_2)$ $r_6: a \rightarrow (a, out)$ $r_2: e \rightarrow (e, in_2)$	$c^{28}$	$r_3: be \rightarrow (e, out)$ $r_4: e \rightarrow \delta >$ $r_5: a \rightarrow (ac, out)$	$a^7 b e$
$r_1: a \rightarrow (a, in_2)$ $r_6: a \rightarrow (a, out)$ $r_2: e \rightarrow (e, in_2)$	$a^7 c^{35} e$	$r_3: be \rightarrow (e, out)$ $r_4: e \rightarrow \delta >$ $r_5: a \rightarrow (ac, out)$	(empty)
$r_1: a \rightarrow (a, in_2)$ $r_6: a \rightarrow (a, out)$ $r_2: e \rightarrow (e, in_2)$	$c^{35}$	$r_1: a \rightarrow (a, in_2) >$ $r_6: a \rightarrow (a, out)$ $r_2: e \rightarrow (e, in_2)$	$a^7 e$

$r_1: a \rightarrow (a, in_2)$ $r_6: a \rightarrow (a, out)$ $r_2: e \rightarrow (e, in_2)$	$a^7 c^{35}$	(dissolves)	
$r_1: a \rightarrow (a, in_2)$ $r_6: a \rightarrow (a, out)$ $r_2: e \rightarrow (e, in_2)$	$c^{35}$	(dissolved)	

In membrane 2,  $r_3$  consumes one copy of  $b$  and sends one copy of  $e$  to membrane 1, at the same time  $r_5$  produces seven copies of  $c$  and sends them to membrane 1 together with the seven copies of  $a$ . Thereafter, there would be seven copies of  $a$ , fourteen copies of  $c$  and one copy of  $e$  in membrane 1.

In membrane 1,  $r_1$  sends seven copies of  $a$  to membrane 2 while at the same time  $r_2$  sends one copy of  $e$  to membrane 2. In membrane 2,  $r_3$  consumes one copy of  $b$  and sends the only copy of  $e$  to membrane 1, at the same time  $r_5$  produces seven copies of  $c$  and sends them to membrane 1 together with the seven copies of  $a$ . There are now seven copies of  $a$ , twenty-one copies of  $c$  and one copy of  $e$  in membrane 1 while there are two copies of  $b$  in membrane 2.

Again in membrane 1,  $r_1$  sends seven copies of  $a$  to membrane 2 while  $r_2$  sends the only copy of  $e$  to membrane 2. In membrane 2,  $r_3$  consumes one copy of  $b$  and sends one copy of  $e$  to membrane 1, while again  $r_5$  produces seven copies of  $c$  and sends them to membrane 1 together with the seven copies of  $a$ . There are now seven copies of  $a$ , twenty-eight copies of  $c$  and one copy of  $e$  in membrane 1, while only one copy of  $b$  is left in membrane 2.

When the rules  $r_1$  and  $r_2$  are used again, the seven copies of  $a$  and the only copy of  $e$  are sent to membrane 2. Again in membrane 2,  $r_3$  sends the only copy each of  $b$  and  $e$  to membrane 1, while again  $r_5$  produces seven copies of  $c$  and sends them to membrane 1, together with the seven copies of  $a$ . There would then be seven copies of  $a$ , thirty-five copies of  $c$  and one copy of  $e$  in membrane 1. Finally, when  $r_1$  and  $r_2$  send, respectively, the seven copies of  $a$  and the only copy of  $e$  to membrane 2, there would be thirty-five copies of  $c$  in membrane 1 while membrane 2 would contain seven copies of  $a$  and one copy of  $e$ .

The availability of a copy of  $e$  without any copy of  $b$  in membrane 2 makes possible the use of  $r_4$  for the first time. It dissolves membrane 2, consuming the only copy of  $e$ . Consequently, rules  $r_1$  and  $r_2$  cannot be applied as their target membrane has

been dissolved. Rule  $r_6$  will be applied. It sends the seven copies of  $a$  out of membrane 1, after which the computation halts. There are now only thirty-five copies of  $c$  which encode the result of the computation in membrane 1, the output membrane.

#### 4.5 Division P system

Consider two non-negative integers  $m$  and  $n$  where  $m \leq n$ . The P system to compute  $n/m$  is a construct of the form

$$\Pi' = (V, \mu, w_1, w_2, (R_1, \rho_1), (R_2, \rho_2), i_0)$$

where

$V = \{a, b, c, d, e\}$  is a set of objects.

$\mu = [1[2]_2]_1$  is a membrane structure of degree 2.

$w_1 = \emptyset$  is the empty multiset (no object present in membrane 1 at the initial configuration).

$w_2 = a^m b^n d$  is the initial multiset of objects in membrane 2.

$$R_1 = \{r_1: c \rightarrow (a, in_2), r_2: d \rightarrow (d, in_2), r_6: d \rightarrow (d, out)\}$$

is the set of rules in membrane 1.

$\rho_1 = \{r_2 > r_6\}$  is the set of priority relation of the rules in membrane 1.

$$R_2 = \{r_3: ab \rightarrow (c, out), r_4: a \rightarrow \delta, r_5: d \rightarrow (de, out)\}$$

is the set of rules in membrane 2.

$\rho_2 = \{r_3 > r_4 > r_5\}$  is the set of priority relations of the rules in membrane 2.

$i_0 = 1$  is the label for the output membrane.

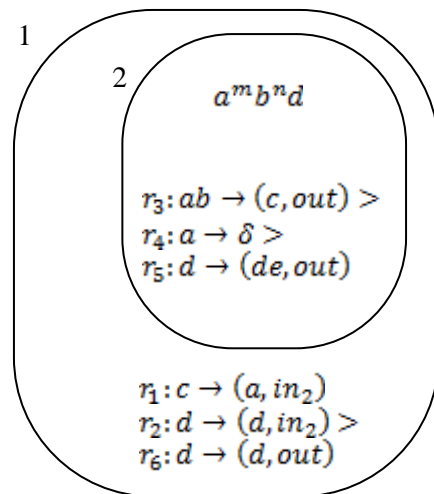


Figure 4.4: Arbitrary initial configuration for a division P system

Membrane 2 has three rules  $r_3, r_4$  and  $r_5$  where  $r_3 > r_4 > r_5$  are the weak rule priority relations among the rules. Membrane 1 has three rules, namely,  $r_1, r_2$  and  $r_6$  where  $r_2 > r_6$ . At the initial configuration, there are  $m$  copies of object  $a$  and  $n$  copies of object  $b$  in membrane 2 which is the input membrane, where  $m$  and  $n$  (assuming that  $m \leq n$ ) encode the input integers for the division.

Since there are no objects in membrane 1 at the moment, none of the rules can be applied within. Since  $r_3$  has priority over  $r_4$  in membrane 2,  $r_3$  will be applied. It consumes  $m$  copies of the pair of  $a$  and  $b$  and produces  $m$  copies of  $c$  and sends them to membrane 1. Since no object  $a$  exists in membrane 2 after  $r_3$  has been applied, a rule with lesser priority is to be applied, which is  $r_4$ . But then it will not be applied since no object  $a$  is present in membrane 2. Therefore,  $r_5$  will be applied in parallel with  $r_3$ . It produces one copy of a new object  $e$  and sends it to membrane 1 together with the only copy of  $d$ . Thus,  $n - m$  copies of  $b$  are left in membrane 2, while  $m$  copies of  $c$ , one copy of  $d$  and one copy of  $e$  are present in membrane 1.

In membrane 1,  $r_6$  will not be applied since it competes for the object  $d$  with  $r_2$  which has a higher priority. Therefore, both  $r_1$  and  $r_2$  will be applied. The rule  $r_1$  consumes the object  $c$  and produces the object  $a$  which it sends to membrane 2 while  $r_2$  sends  $d$  back to membrane 2. As the iteration continues, the multiplicity of  $b$  is being reduced by  $m$  until it is less than  $m$  in membrane 2. Two cases to consider are where the multiplicity of  $b$  is zero and where it is greater than zero.

If the multiplicity of  $b$  is zero, then  $r_3$  cannot be used again rather  $r_4$  will be used for the first time. It dissolves membrane 2 and consumes the object  $a$ , at the same time  $r_5$  sends  $d$  to membrane 1 for the last time. Thus,  $r_1$  and  $r_2$  cannot be used again as their target membrane has been dissolved. Therefore,  $r_6$  will be used for the first time. It sends

out of membrane 1. No rule can be applied at this point and the computation halts. Thus, there are only copies of  $e$  in membrane 1 and they encode the quotient in the division.

On the other hand, if the multiplicity of  $b$  is greater than zero then  $r_3$  will be used for the last time. This time however, it is copies of  $a$  and not  $b$  that are left in membrane 2, together with the only copy of  $d$ . The rule  $r_4$  consumes the remaining

copies of  $a$  and dissolves membrane 2 while  $r_5$  produces one copy of  $e$  and sends it to membrane 1 together with the only copy of  $d$ . Thus, copies of  $c$  and  $e$  and one copy of  $d$  are in membrane 1. The dissolution of membrane 2 renders  $r_1$  and  $r_2$  unusable, hence the use of  $r_6$  for the first time. It sends the only copy of  $d$  out of the membrane. At this point, no rule can be applied and the computation halts. The copies of  $e$  encode the quotient in the division while the copies of  $c$  encode the remainder.

#### 4.6 Division P system exemplified

In the following example, we compute  $11/4$ . At the initial configuration of the system, four copies of an objects  $a$ , eleven copies of an objects  $b$  and one copy of an object  $d$  are put in membrane 2. Any of the rules in membrane 2 can be applied, however, due to the priority among the rules,  $r_3$  will be applied first. It consumes four copies of the pair of  $a$  and  $b$  and produces four copies of  $c$  which it sends to membrane 1. Now no object  $a$  exists in membrane 2 and so  $r_5$  will be applied in parallel with  $r_3$ . It produces one copy of a new object  $e$  and sends it to membrane 1 together with the only copy of object  $d$ . No other rules can be applied in membrane 2 at this point. There are now four copies of  $c$ , one copy of  $d$  and one copy of  $e$  in membrane 1.

**Table 4.2:** An illustration of the division of 11 by 4 with the division P system (division with remainder)

Membrane 1		Membrane 2	
Rule	Object	Rule	Object
$r_1: c \rightarrow (a, in_2)$ $r_2: d \rightarrow (d, in_2)$ $r_6: d \rightarrow (d, out)$		$r_3: ab \rightarrow (c, out) >$ $r_4: a \rightarrow \delta >$ $r_5: d \rightarrow (de, out)$	$a^4 b^{11} d$
$r_1: c \rightarrow (a, in_2)$ $r_2: d \rightarrow (d, in_2)$ $r_6: d \rightarrow (d, out)$	$c^4 de$	$r_3: ab \rightarrow (c, out) >$ $r_4: a \rightarrow \delta >$ $r_5: d \rightarrow (de, out)$	$b^7$
$r_1: c \rightarrow (a, in_2)$ $r_2: d \rightarrow (d, in_2)$ $r_6: d \rightarrow (d, out)$	$e$	$r_3: ab \rightarrow (c, out) >$ $r_4: a \rightarrow \delta >$ $r_5: d \rightarrow (de, out)$	$a^4 b^7 d$
$r_1: c \rightarrow (a, in_2)$ $r_2: d \rightarrow (d, in_2)$ $r_6: d \rightarrow (d, out)$	$c^4 de^2$	$r_3: ab \rightarrow (c, out) >$ $r_4: a \rightarrow \delta >$ $r_5: d \rightarrow (de, out)$	$b^3$
$r_1: c \rightarrow (a, in_2)$ $r_2: d \rightarrow (d, in_2)$ $r_6: d \rightarrow (d, out)$	$e^2$	$r_3: ab \rightarrow (c, out) >$ $r_4: a \rightarrow \delta >$ $r_5: d \rightarrow (de, out)$	$a^4 b^3 d$



$r_1: c \rightarrow (a, in_2)$ $r_2: d \rightarrow (d, in_2)$ $r_6: d \rightarrow (d, out)$	$c^3 d e^2$	(dissolves)	
$r_1: c \rightarrow (a, in_2)$ $r_2: d \rightarrow (d, in_2)$ $r_6: d \rightarrow (d, out)$	$c^3 e^2$	(dissolved)	

In membrane 1, both  $r_1$  and  $r_2$  can be applied. While  $r_1$  consumes the four copies of  $c$  and produces four copies of  $a$  which it sends to membrane 2,  $r_2$  sends the only copy of  $d$  to membrane 2. One copy of  $e$  is left in membrane 1. There are now four copies of  $a$ , seven copies of  $b$  and one copy of  $d$  in membrane 2. The rules  $r_3$  and  $r_5$  will be applied. While  $r_3$  consumes four copies of the pair of  $a$  and  $b$  and sends four copies of  $c$  to membrane 1,  $r_5$  sends the only copy of  $d$  to membrane 1 and produces a new copy of  $e$  which it sends to membrane 1. There are now four copies of  $c$ , one copy of  $d$  and two copies of  $e$  in membrane 1.

Again,  $r_1$  consumes the four copies of  $c$  and produces four copies of  $a$  and sends them to membrane 2 while  $r_2$  sends  $d$  to membrane 2. Consequently, all the rules in membrane two are lost and the object  $d$  falls back to membrane 1. Moreover, rules  $r_1$  and  $r_2$  cannot be applied in membrane 1 since their target membrane has been dissolved, hence,  $r_6$  will be applied. It sends  $d$  out of the membrane.

The computation halts. There are now three copies of  $c$  and two copies of  $e$  in membrane 1 which is the output membrane. The result of the computation is encoded by the two copies of  $e$  which represent the dividend and the three copies of  $c$  which represent the remainder.

We now examine the nature of the result in this example if the task was to divide 12 by 4 starting from the configuration before the dissolution of membrane 2, where there are four copies of each of  $a$  and  $b$  and one copy of  $d$  in membrane 1 and two copies of  $e$  in membrane 2. The rule  $r_3$  would consume the four copies of each of  $a$  and  $b$  and produce four copies of  $c$  which it would send to membrane 1. At the same time  $r_5$  would produce a copy of  $e$  and send it to membrane 1 together with the only copy of  $d$ . There would be four copies of  $c$ , three copies of  $e$  and one copy of  $d$  in membrane 1, while membrane 2 would contain no objects.

Next,  $r_1$  would consume the four copies of  $c$  and produce four copies of  $a$  which it would send to membrane 2 while  $r_2$  would send the only copy

of  $d$  to membrane 2. In membrane 1,  $r_3$  would be applied for the first time. It would consume the object  $a$  and at the same time dissolve membrane 2, allowing  $d$  to return to membrane 1. Again,  $r_1$  and  $r_2$  could not be applied since they both point to the dissolved membrane 2. Therefore,  $r_6$  would be applied for the first time. It sends the only copy of  $d$  out of the membrane. The computation would then halt. There would be three copies of  $e$  which encode the quotient. The non availability of  $c$  is an indication that there exists no remainder from the division.

**Table 4.3:** An illustration of the division of 12 by 4 with the division P system (division without remainder)

Membrane 1		Membrane 2	
Rule	Object	Rule	Object
$r_1: c \rightarrow (a, in_2)$ $r_2: d \rightarrow (d, in_2)$ $r_6: d \rightarrow (d, out)$		$r_3: ab \rightarrow (c, out) >$ $r_4: a \rightarrow \delta >$ $r_5: d \rightarrow (de, out)$	$a^4 b^{11} d$
$r_1: c \rightarrow (a, in_2)$ $r_2: d \rightarrow (d, in_2)$ $r_6: d \rightarrow (d, out)$	$c^4 de$	$r_3: ab \rightarrow (c, out) >$ $r_4: a \rightarrow \delta >$ $r_5: d \rightarrow (de, out)$	$b^7$
$r_1: c \rightarrow (a, in_2)$ $r_2: d \rightarrow (d, in_2)$ $r_6: d \rightarrow (d, out)$	$e$	$r_3: ab \rightarrow (c, out) >$ $r_4: a \rightarrow \delta >$ $r_5: d \rightarrow (de, out)$	$a^4 b^7 d$
$r_1: c \rightarrow (a, in_2)$ $r_2: d \rightarrow (d, in_2)$ $r_6: d \rightarrow (d, out)$	$c^4 de^2$	$r_3: ab \rightarrow (c, out) >$ $r_4: a \rightarrow \delta >$ $r_5: d \rightarrow (de, out)$	$b^3$
$r_1: c \rightarrow (a, in_2)$ $r_2: d \rightarrow (d, in_2)$ $r_6: d \rightarrow (d, out)$	$e^2$	$r_3: ab \rightarrow (c, out) >$ $r_4: a \rightarrow \delta >$ $r_5: d \rightarrow (de, out)$	$a^4 b^4 d$
$r_1: c \rightarrow (a, in_2)$ $r_2: d \rightarrow (d, in_2)$ $r_6: d \rightarrow (d, out)$	$c^4 de^3$	$r_3: ab \rightarrow (c, out) >$ $r_4: a \rightarrow \delta >$ $r_5: d \rightarrow (de, out)$	
$r_1: c \rightarrow (a, in_2)$ $r_2: d \rightarrow (d, in_2)$ $r_6: d \rightarrow (d, out)$	$e^3$	$r_3: ab \rightarrow (c, out) >$ $r_4: a \rightarrow \delta >$ $r_5: d \rightarrow (de, out)$	$a^4 d$
$r_1: c \rightarrow (a, in_2)$ $r_2: d \rightarrow (d, in_2)$ $r_6: d \rightarrow (d, out)$	$de^3$	(dissolves)	
$r_1: c \rightarrow (a, in_2)$ $r_2: d \rightarrow (d, in_2)$ $r_6: d \rightarrow (d, out)$	$e^3$	(dissolved)	

5. CONCLUSION

A simple deterministic P system for carrying out the four basic arithmetic operations of addition, subtraction, multiplication and division have been constructed based on the weak rule priority relations. The numbers of rules are six for each of multiplication and division while the highest number of membranes remains two. There is a strong possibility of extending the idea to negative integers and subsequently to real numbers.

## 6. RECOMMENDATION FOR FUTURE RESEARCH

It seems promising that the deterministic P systems based on the weak rule priority presented in this paper can be extended to accommodate negative integers as well as real numbers. For instance, in order to encode negative integers, one designates a separate membrane 3 within the skin membrane. A rule is placed within the membrane which sends objects to membrane 2 by first converting the objects to another object prior to entering the membrane; such objects are regarded as “borrowed objects” in membrane 2. It is easy to see that one can encode negative integers with less effort in this manner. In the case of real numbers, rational and irrational components need to be identified. It is easy to encode positive rational numbers – the P system for every rational number is a division P system for two integers where the P system of the denominator does not represent zero. However, encoding irrational numbers is more challenging.

## 7. ACKNOWLEDGEMENTS

Our acknowledgment goes to Gh. Păun, the Pioneer author of membrane computing. We also acknowledge the authors of the various texts we have consulted, and other individuals who have contributed directly or indirectly to the success of this paper. My special thanks to the anonymous reviewers who have helped to put the current paper to a better shape.

## 8. REFERENCES

Alhazov, A., Bonchis, C., Ciobanu, G., & Izbasa, C., 2006. Encodings and arithmetic operations in P systems. *Proceedings of the Fourth Brainstorming Week on Membrane Computing*, Vol. 1, 1-27.

- Păun, Gh., 2000. Computing with membranes. *Journal of Computer and System Sciences*, Vol. 61(1), 108-143.
- Păun, Gh., 2006. Introduction to membrane computing. In *Applications of Membrane Computing*. pp. 1-42 Springer Berlin Heidelberg.
- Atanasiu, A., 2000. Arithmetic with membranes. Pre-proc. In *Workshop on Multiset Processing, Curtea de Arges, Romania, TR Vol. 140*, pp. 1-17.
- Chen, Y., Zhang, G., Wang, T. & Huang, X., 2014. Automatic design of a P system for basic arithmetic operations. *Chinese Journal of Electronics*, Vol. 23, no. 2, 302-304.
- Guo, P., Zhang, H., Chen, H. Z., & Chen, J. X. Fraction arithmetic operations performed by P systems. *Chinese Journal of Electronics*, Vol. 22, no. 4. 2013, 689-694.
- Zeng, X. X., Song, T., Zhang, X. Y. & Pan, L. Q., 2012. Performing four arithmetic operations with spiking neural P systems, *IEEE Transaction on NanoBioscience*, Vol.11, No.4, pp.366-374.
- Guo P. & Chen, J., 2008. Arithmetic operation in membrane system. In *proceedings of the 2008 international conference on BioMedical Engineering and informatics*, pp. 13-39.
- Guo, P. & Zhang, H., 2008. Arithmetic operation in single membrane. In *Computer Science and Software Engineering, 2008 International Conference on*, Vol. 3, pp. 532-535.
- Păun Gh. & Thierrin, G., 2001. Multiset processing by means of systems of finite state transducers. In *Automata Implementation*. Springer, pp. 140-157.
- Freund, R. & Păun, Gh., 2003. On deterministic P systems. See *P Systems Web Page at <http://psystems.disco.unimib.it>*.
- Ibarra, O. H., 2005. On determinism versus nondeterminism in P systems. *Theoretical Computer Science*. Vol. 344, no. 2, , 120-133

**13<sup>th</sup>**

**International Conference**



**Session E:**

**Sustainable Infrastructure for  
Innovation, Research and  
Development**

## Full Paper

**AN AUTOGENERATED APPROACH OF STOP WORDS USING  
AGGREGATED ANALYSIS****Tijani O.D.**

Department of Computer Science,  
Federal University of Agriculture,  
Abeokuta, Nigeria  
dantjoo7@yahoo.com

**Akinwale A.T.**

Department of Computer Science,  
Federal University of Agriculture,  
Abeokuta, Nigeria  
akinwale@funaab.edu.ng

**Onashoga S.A.**

Department of Computer Science,  
Federal University of Agriculture,  
Abeokuta, Nigeria  
onashogasa@funaab.edu.ng

**Adeleke E.O.**

Department of Mathematics,  
Federal University of Agriculture,  
Abeokuta, Nigeria.  
adelekeeo@funaab.edu.ng

**ABSTRACT**

The importance of stop words list generation helps in the elimination of stop words, which contribute to reduce the size of the vector space of the corpus and indexing structure considerably to obtain a high compression rate, speed up calculation and increase the accuracy of information retrieval systems. The proposed system focuses on the different characteristics of stop words which distinctively identify stop words based on a carefully adopted aggregated approach of Frequency Analysis, Word Distribution Analysis and Word Entropy Measure. Each of the method generates its own stop words list after passing through a thorough text preprocessing stage including the diacritization of the Yoruba corpus, which is then aggregated using set theory to redefine stop words as those with high occurrences, stable distribution and less informative. The system uses a novel approach of machine learning using Multinomial Naïve Bayes to make the system perform the automatic generation of stop words list and keep the list updated in the event of an evolving new word. When applied to Yoruba language corpus, the output produced a standardized automatically generated stop words list, which outperforms the existing stop words lists that were mostly produced using frequency measure. New words in comparison to existing list were also identified. The system which outperforms existing ones generated 255 stop words with a text compression rate of 63% when the stop words are removed from text document.

**Keywords:** Natural Language Processing, Zipf's law, Stop words, Entropy, Variance, Diacritization

### 1.0 Introduction

Stop Words are words in a body of text that does not have any informative value and make up a large fraction of a document. They are found to always dominate the corpus without any contextual contribution. In fact, a common feature of these words is that they carry no significant information to the document; instead, they are used just because of grammar Zou et al. (2006).

Stop-words are language-specific functional words that are frequent words and carry no information according to Sharma et al. (2015). Medhat et al. (2016) corroborated this, when they described Stop words are more typical words used in many sentences and have no significant semantic relation to the context in which they exist.

Stop words also tend to have very little contribution to query as they rarely make up index or query words in natural language processing. It is imperative to note that stop words cannot be used as index terms.

Stop words can be described as evenly distribute-d noise signal that should be filtered from a text, as these words have no discriminative values. According to Lo et al. (2005), stop words are described as words in a document that are frequently occurring but meaningless in terms of Information Retrieval (IR).

### 1.1 Word Distribution Method

Stop words are homogenously distributed across documents. Existing methods used in measuring the distribution of words in a document such as mean probability, variance etc. Figure 1 shows word distribution characteristics and frequency measures.

Also, Chekima et al. (2016) in reported that, the ten most commonly used words in English normally account 20 to 30 percent of the words in a document. Therefore, it is usually worth to eliminate all stop words terms when indexing a document or processing queries.

From the angle of parts of speech, stop words fall into the category of articles, prepositions, adverbs, conjunctions, interrogative words, negative words, exclamations, also they include all the pronouns, demonstratives, subject and object pronouns, some numbers, additions and verbs. Stop-words may be separate or attached ones in a form of prefixes or suffixes Cook et al. (2008).

### 1.2 Characteristics of Stop Words

1. These words are said to have a very low discrimination value.
  2. The amount of information carried by these words is negligible.
  3. They have high frequency of occurrence
  4. Stop word have a stable/high distribution
  5. They are evenly and homogenously distributed words.
  6. Stop words are poor index terms and hence cannot be used as index terms.
  7. They are never or rarely used as query terms/search words.
  8. They are words like articles, pronouns, adverbs, prepositions, conjunction, injunction etc.
  9. They are general words and not used specifically in a certain field.
  10. They are necessary for the construction of the language.
- (Sadeghi et al., 2008 & Yousef et al., 2016)

### 1.3 Application/Importance of Stop Words List

Stop words are widely used in text mining,

information retrieval, text summarizer, indexing, social media platforms, clustering, supervised machine learning, SEOs, search and retrieval systems, Natural Language Processing etc.

A stop words list refers a set of terms or words that have no inherent useful information. Stop words create problems in identification of key concepts and words from textual sources when they are not removed due to their overwhelming presence both in terms of frequency as well as occurrence in textual sources according to Sadeghi et al. (2014).

The importance of stop words list generation helps in the elimination of stop words, which could contribute to reduce the size of the vector space of the corpus and indexing structure considerably and obtain a compression of more than 40%. On the other hand, as highlighted by Saif et al. (2015) the removal of stop words also help to speed up calculation and increase the accuracy of information retrieval systems as reported.

Numerous researches have been carried out in the generation of English stop word lists such as the popular Brown corpus which consists of 421 stop words as derived by Fox as quoted by Savoy et al. (1999), which contains obvious stop words like: the, for, is, and, it etc, the Van stop word list which has 250 stop words. This list is generally known as the standard or classic stop words (Chekima et al., 2016). Also recent works from Choy 2012, Cook et al. 2008 and Saif et al. 2015 among others are based on stop word list generation for English language.

Due to its role as a vital foundational work in text mining and natural language processing (NLP), lots of researches are being attracted

into other languages as regards stop word list generation. Lot of other researches have also been carried out in other languages including French by Savoy (1999), Chinese (Zou et al. 2006 & 2008) and Arabic (Alajmi et al. 2012 & Sadeghi et al. 2014). Chekima et al. 2016 worked on Malay language, while a stop words list in Punjabi language was developed by Puri et al. (2013) and Medhat et al. (2016) did a work on Egyptian stopword list.

However in sharp contrast to its European and Asian counterpart very few works has been done on stop word list generation in African languages. In specific, only one research work by Asubiaro (2013 & 2015) on stop word generation in Yoruba language was found. This drew our curiosity and as our own contribution to the development of local languages, this research work will be specifically dealing with the construction of stop word list in Yoruba language of Nigeria.

Even though Yoruba language is spoken by more than 43 million people in West African countries such as Nigeria, Republic of Benin, Togo, Ghana, and Ivory Coast, and even in Europe and North America, Asubiaro (2015) stated that: Yoruba language like most African languages is a technologically resource-scarce language. Resource scarce languages lack necessary language technologies and this will lead to these languages going into extinction, because the official language in Nigeria for instance is English language and this is killing our mother-tongue, Yoruba language unlike what obtains in China where Chinese language is the official language of the country.

The motivation for this research work is based on the urgent need to generate stop words list in Yoruba language which is imperative and the foundation for other

important Information Retrieval and Natural Language Processing (NLP) for the further development of the language. In the process of doing so, this will lead to generating resources for technologically resource-scarce languages like Yoruba.

Also, looking at the level and height of NLP automation achieved in other languages e.g. English; we intend to build up the foundation of NLP on our own Yoruba Language to stimulate further researches. The need to domesticate available technological methodologies on Yoruba language also drew our interest in order to project and save Yoruba language from extinction.

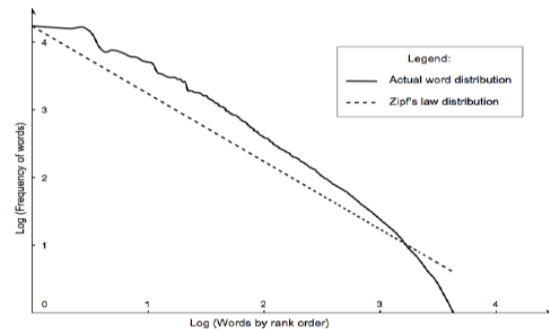
According to Asubiaro (2013 & 2015), the challenge thrown-up by Yoruba language been listed as one of the technologically resource-scarce languages needs to be tackled by conducting researches especially in the areas of Natural Language Processing.

Furthermore, Yoruba language being a language spoken by over 43 million people in Nigeria and abroad needs to be saved from extinction by developing sophisticated systems that will ensure it can be well-documented and preserved with the use of Information and Communication Technology tools for easy learning and understanding by the generations of the modern digital age. Any language that is not moving at the speed of light at which technology is evolving has a high chance of dying a natural death. The use of other evolving NLP methodologies to automatically generate stop words list in Yoruba language also drew our curiosity.

## 2.0 Related Work

Even though, a wide range of stop word lists have been generated in different languages and can be easily sourced for use, the

methodology used in the creation of the list is very important. Choy (2012) argued that stop words lists are rarely investigated and validated compared to the results of the mining process or mining algorithm.



**Figure 1:** A graph showing the actual and zipf's law distributions plotted in a log-log scale (Blanchard, 2007).

Many authors in order to improve the quality of textual data have proposed different methods to extract an effective stop word list for a particular corpus. A very common approach used in the generation of a stop word is to manually identify frequently occurring words from a body of texts. This approach has been proven to be generally applicable to some situations; however the difficulty, time consumption and somewhat ineffectiveness of this herculean task can easily be imagined and noticed.

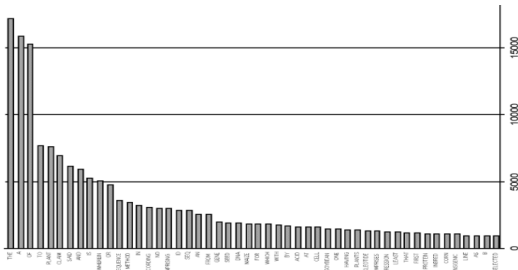
To conquer this limitation, other researchers have proposed several methods to automatically generate stop words list. The various methodologies used can however be streamlined and classified into two, namely:

- (1) the statistical approach and
- (2) the semantic approach.

The statistical approach is a measure of how frequent a word appears. Such frequency

measure (statistical) methods include: term frequency, document frequency or inverse document frequency [6], methods based on zipf's law (Chekima et al. 2016 & Lo et al. 2005) etc.

On the other hand, the semantic approach is based on the information gain or measure that a word carries to determine if it is a stop word. Stops words are known to carry very little information in comparison to non-stop words. Entropy measure, which calculates the probability of a word being a stop word just like Kullback-Leibler divergence measure and Maximum Likelihood Estimation measure that determines the amount of information a word contains, are methods used to determine the information measure or semantic content of a word.



**Figure 2:** showing the distribution of the top 50 frequently occurring words in the Syngenta patent portfolio in agricultural biotechnologies (Blanchard, 2007).

Most early works done on stop words list generation were based on the frequency count method i.e. how often a word appears in a document.

W. Nelson Francis and Henry Kucera (1985) generated list of common (stop) words in English language using frequency analysis from 500 samples, each approximately 2000 words long divided into 15 genres, with the following top 20 stop words appearing conspicuously across the samples.

**Table 1:** Top 20 stops words (Francis, 1985)

1.	The	8.	Is	15.	his
2.	Of	9.	Was	16.	on
3.	And	10.	He	17.	be
4.	To	11.	For	18.	at
5.	A	12.	It	19.	by
6.	In	13.	With	20.	I
7.	That	14.	As		

Zou et al. (2006) in the table below shows 40 out of the generated stop words in English which was extracted from TIME magazine articles using the traditional method of **accumulated frequency**.

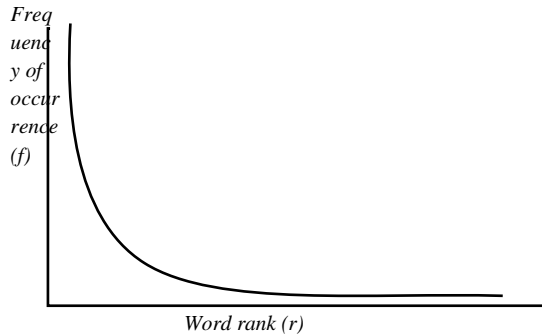
**Table 2:** Top 40 words with highest frequencies from 423 short TIME magazine articles (245,412 word occurrences, 1.6 MB) - (Zou et al., 2006)

Wo	Fre	Wo	Fr	Wo	Fr	Wor	Fr
rd	q.	rd	eq.	rd	eq.	d	eq.
Th	158	his	181	U	95	Wer	84
e	61		5		5	e	8
Of	72	Is	181	Ha	94	Thei	81
	39		o	d	o	r	5
To	63	He	17	Las	93	Are	81
	31		oo	t	o		2
A	58	As	15	Be	91	One	811
	78		81		5		
An	561	On	155	ha	91	We	79
d	4		1	ve	4	ek	3
In	52	By	14	wh	89	The	69
	94		67	o	4	y	7
Th	25	At	133	No	88	Gov	68
at	07		3	t	2	ern	7
For	222	It	12	Ha	88	All	67
	8		90	s	o		2
Wa	214	fro	122	An	87	Yea	67
s	9	m	8		3	r	2
Wit	183	but	113	As	86	Its	62
h	9		8		5		0

The frequency count method employs the word frequency in documents by taking into account the terms and their corresponding frequencies, similar to Lo et al.'s (2005) work motivated by zipf's law. Zipf's law states that



given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. (Wikipedia, 2016). Zipf's law is easily represented by a log-log graph which is similar to an elbow as shown in Fig. 1 below:



**Figure 3:** A typical log-log graph showing Zipf's law relationship

Based on the weight Chi square method, Choy (2012) proposed a new approach called Combinatorial Counts. The technique as opposed to measuring the information value of words to establish the stop words focuses on the extreme number of combinations that most non-meaningful words display to establish stop words. The result shows that the proposed technique outperforms most of the other techniques by a fair margin.

Chekima et al. (2016) used an aggregated method to construct a Stop word list in Malay language by combining three different lists generated using the set theory of union and intersection. Having studied existing works which was based on either the use of one or two techniques, Chekima et al. (2016) used a multifaceted approach to generate a standard stop words list in Malay language. The statistical approach was used to determine word frequency of occurrences against their rank as inspired by zipf's law

while the distribution of words across documents was checked by the variance measure to confirm the high distribution rate of stops words. To measure the informative content of a word, the authors used entropy measure to decide the less-informative words. The overlap of the words in the three differently generated lists was high.

A threshold of 300 words been a reasonable average of the number of stop words produced in each of the list by the three methods was chosen by Chekima et al. (2016). The three lists were aggregated using the union and intersection method in set theory and a final stop words list of 502 malay words were generated. The combination of the three lists redefined stop words as those with high occurrences, stable distribution and less informative.

According to the authors, comparison could not be carried out due to the absence of standard malay stop words list, hence a manual evaluation was carried out by language expert and this further ensured the list was pruned down to 339 stop words. The strength of the research lies in the fact that in comparison to using a single method, the aggregation method ensured that a stronger and more robust final list is produced based on the various characteristics of stop words.

Meanwhile, a Punjabi stop-words list was generated using the combination of the classification of words based on frequency and the statistical approach based on word distribution by Puri et al. (2013). The purpose of the work was to find a suitable, automated method for identification of stop words in Punjabi Text which results in reduction of the overall vector space, thereby leading to performance improvements in terms of

execution speed and the relevance of results. To achieve their aim Puri et al. (2013) used two approaches of frequency count to generate a quick list of stop words by finding the most frequent words in the document and probability distribution which is used to identify words with stable distribution which are stop words.

The combination of the two lists generated by the approaches was achieved using Borda's rule and this changed the order of the output. However, the weakness of the work was that no evaluation was carried out.

In the research to investigate the effect of removing Egyptian Dialect (ED) stopwords on the Sentiment Analysis (SA) task, Medhat et al. (2016) extracted Egyptian dialect stop word list by calculating the frequency to determine words with high occurrence after preprocessing had been carried out on the text corpora extracted from the online social network (OSN) and a cut-off threshold of 200 most frequent words was set to validate the stop words. The preprocessing includes removal of usernames, URLs and non-arabic words. The data used was collected from Twitter, Facebook, and a movie review site on the same topic in ED. The authors went ahead to validate the generated list by diacritising the words, comparing it with the Modern Standard Arabic (MSA) stop words list and when not found, they went further to check for its equivalent in English and for additional scrutiny. Thereafter, where needed all possible prefixes and suffixes were added. According to Medhat et al. (2016), the result revealed that the methodology showed better performance. Evaluation was carried out using accuracy and F-measure. In the results, the author concluded that for the type of corpora used, decision tree classifier

performs better than Naïve Bayes classifier.

Saif et al. (2015) proposed an interesting and novel approach by using contextual semantic and sentiments analysis of words to automatically identify and remove stopwords from Twitter data. Unlike existing approaches that generate context insensitive stop words list, the authors captured the contextual semantic and sentiments of word using SentiCircle - a semantic representation model to automatically identify stop words. The SentiCircle model extracts the contextual semantics of a word from its co-occurrences with other words in a given tweet corpus. Using this model, stop words were identified based on their weak semantics and sentiments within the context which they occur by calculating the SentiMedian and check whether it falls within the stop word region or not. In evaluating the approach, the binary sentiment classification into positive and negative classification of tweets were performed using the Maximum Entropy classifier to compare their list with the chosen baseline i.e. van stoplists, the system outperforms the classic method.

Another experimental approach to automatically generate stop words list reviewed by Blanchard (2007) is the mapping tool concept. The aim of the work is to provide a good understanding of how mapping tools work, with an emphasis on stopwords and to open perspectives with automated stopword list construction.

Even though the author considered the stop words generation feature of commercial mapping software such as AnaVist, OmniViz and Aureka, the concept appears to be interesting, the blackbox system involved in the software may not allow thorough

scrutiny of the methodologies employed. However, the flaws of the mapping approach include the ability of the user to manipulate the system output, opacity and redundancy (if combined with other scoring algorithm such as tf-idf). The author however proposes the algorithmic methods of automatically generating stop words such as evolutionary (generic) algorithm and term-based approach.

Zou et al. (2006) in order to save the time and release the burden of manual stop word selection proposed an automatic aggregated methodology based on statistical and information models for extraction of a stop word list in Chinese language. Despite the peculiar challenge of absence of word boundary in the Chinese words of not been separated by spaces using bigram and boundary detection segmentation, Zou et al. (2006) are able to create a Chinese stop word list based on an aggregated methodology of statistical and information models.

Based on the statistical method, they took a statistics of the distribution of word frequencies in different documents and observed that stop words are ranked at the top of the list with much larger frequency than the other words. On the other hand, stop words are also those words with quite a stable distribution in different documents. Mean and variance were used to measure the distribution criteria. A combination of these two observations redefines the stop words as those words with stable and high frequency in documents.

In comparison to the general English stop word list, result analysis showed that the Chinese stop list is comparable and much more general and outperform other Chinese

stop lists. The stop word extraction algorithm used saves the time for manual generation and constructs a standard. Zou et al. (2006) however achieved the aggregation by applying the Borda's rule used in the social choice theory of voters.

In a subsequent work, Zou et al. (2008) carried out an evaluation of stop word lists in Chinese Language. A proposed novel segmentation algorithm was used for the evaluation of the stop word lists. The researchers posited that the research became expedient in order to have not only a standardized but an effective Chinese stop word list. The authors' argued that their approach produced a great improvement to the original lists used and segmentation improves directly with an improvement in the stop word list generation.

Lo et al. (2005) published an algorithmic approach to generate an automated stopword list. It achieves both good performance and minimal effort by using iterative sampling and computing. A term is randomly selected in the corpus and all documents containing this term are retrieved. The weight of each term is calculated using the Kullback-Leibler divergence measure and normalized by the maximum weight as to be comprised between 0 and 1. All terms are ranked in ascending order of their associated weight and the least informative words are then extracted.

The method introduced by Lo et al. (2005) was compared to four baseline approaches of Term Frequency, Normalized Term Frequency, Inverse Document Frequency and the Normalised Inverse Document Frequency using four different standard TREC collections. Out of the baseline approaches

considered the Normalized Inverse Document Frequency emerged as the best variant of all the Zipf's law inspired approaches aforementioned. Unfortunately, according to their result, the proposed method failed to outperform the baseline approaches but successfully reduce computational overhead. Also, Lo et al. (2005) discovered that when the stop list generated by the best approach is merged with the Fox list, the result is much desirable.

In generating what appears to be the first stop words list for Yoruba language, Asubiaro (2013) based on the methodology used in an earlier work by Savoy for French stop words applied Entropy measure in generating Yoruba stop word list. One important note from Asubiaro's work is that diacritics as far as Yoruba language is concerned are very important as it not only provides morphological and lexical information, ignoring it will lead to loss of information and specificity. Asubiaro generated the stop words list in Yoruba language by calculating the entropy of each word in the dataset, the resulting list was ordered by ascending entropy to reveal the words that have a greater probability of being noise words, since stopwords carry little information they are high entropy words.

Using this approach, a list of 256 stopwords was drawn from the diacritized texts and for the undiacritized version of the texts, 189 stopwords were drawn. The result by the author upon comparing the stoplist with English stoplist shows that only 69.1% of the stopwords are present in the English stoplist. Asubiaro further discovered that some of the (Yoruba) stopwords have no corresponding words in English. He therefore warned that adapting English stopwords list for Yoruba

will not work optimally for the language. Also, to further evaluate the effect of the generated stop words list, the full text was reduced by 65.91% and 67.46% when the stopwords were removed from the full diacritized and undiacritized text respectively.

However, one limitation in Asubiaro's work is that generating stop words based on only one method of entropy measure that specifies the information-gain of a word may not be very effective. Additionally, from the 189 diacritized and 251 undiacritized generated Yoruba stop words respectively, there is need to properly evaluate the list generated by investigating and comparing it with other list generated in Yoruba language, with a view to standardize it. In comparison to the English language, not much extensive work has been done for generating stop words in Yoruba language.

### 3.0 Materials and Methodology

For this research work, we sourced for Yoruba corpus (both diacritized and undiacritized) from different sources and domains to ensure the generality of the output of stop words across the domains. The domains noteworthy of mention among others are literature, religion, education, news, science, sports, business, technology, arts etc.

Data collected are merged to produce a general corpus. This was subjected to pre-processing as stated in 3.1 to normalize the data and filter noise (i.e. unwanted characters) such as symbols, urls, foreign words, digits among others. As a result, a total of 1,294,001 tokens consisting of 9,345 distinct words as size of our corpus.

This research work are based on the following methodologies starting from text preprocessing to ensure that we will be working with clean and processed corpus

devoid of external characters such as other languages, undiacritized texts and special characters such as urls, numeric, alpha-numeric and news tag etc. This will be followed immediately by the three main proposed methodologies including Frequency Analysis, Word Distribution Analysis and Word Entropy Measure. These three methods will produce their respective stop words list which will be eventually aggregated using the intersection of the set theory to yield a standardized stop word list in Yoruba language that will basically not just one but all the features of the characteristics of stops words as discussed in chapter one will improve to redefine the stop words list as words having high occurrence, wide distribution and are less informative (i.e. high entropy).

As a contribution to knowledge, we propose an adaptive mechanism by using Naïve Bayes Algorithm to ensure the system as a machine learns all the processes involved in the automatic generation of the stop words list and keep the list updated peradventure a new word which does not exist before this work is introduced – as this is a possible phenomenon in languages. The system should be able tell us if such a new word is a stop word and regenerate the list and include such word as a stop word.

### 3.1 TEXT PREPROCESSING

#### 3.1.1 Diacritization

Diacritization is the inclusion of a mark above or below a printed letter that indicates a change in the way it is to be pronounced or stressed. These sub-dots and tone marks are appended to base or American Standard Institute (ANSI) characters.

According to Can et al. (2008), Diacritics are appended on base characters to represent some speech sounds that are beyond the scope of ANSI conventional codes for writing which is based on Latin encoding system. Hence, diacritics extend the functionality of these base characters, therefore new

characters are formed by appending diacritic mark(s) on a base character.

In some languages like Yorùbá, tonality is represented with the tone marks; high tone (̀) and low tone (/) which are applied on its vowels and nasal consonant. Yorùbá also cater for speech sounds that are not represented in the 26 alphabets of Latin encoding from which it inherited its writing style.

Like Yorùbá, some African and European languages such as Hausa, Igbo, French, German, Italian and Finnish use diacritics on some base characters. While diacritics carry morphological information in some of these languages, in others, diacritics do not.

In Yorùbá, German and Finnish for instance, the use of diacritics provide morphological and lexical information.

Ògùn (charm), Ogún (inheritance), Ọgun (war) for instance, are different Yorùbá words derived by appending diacritical marks on “ogun”. Each has a distinct meaning which differs from others derived from the same base characters. Unlike in Yoruba, Italian and French languages also use diacritics, but the use of diacritics bear insignificant morphological or lexical information.

#### 3.1.2. Tokenization

Tokenization is the process of segmenting running text into words and sentences.

Electronic text is a linear sequence of symbols (characters or words or phrases). Naturally, before any real text processing is to be done, text needs to be segmented into linguistic units such as words, punctuation, numbers, alpha-numeric, etc. This process is called tokenization.

In Yoruba language just like in English, words are often separated from each other by blanks (white space), but not all white space is equal. Both “Los Angeles” and “rock 'n' roll” are individual thoughts despite the fact

that they contain multiple words and spaces. We may also need to separate single words like “I’m” into separate words “I” and “am”.

Tokenization is a kind of pre-processing in a sense; an identification of basic units to be processed. It is conventional to concentrate on pure analysis or generation while taking basic units for granted. Yet without these basic units clearly segregated it is impossible to carry out any analysis or generation.

The identification of units that do not need to be further decomposed for subsequent processing is an extremely important one. Errors made at this stage are very likely to induce more errors at later stages of text processing and are therefore very dangerous.

### 3.1.3. Word Normalization

This is a preprocessing technique to remove foreign words or other languages, numeric, alpha-numeric, special characters and news tag from the corpus. This is to ensure that the corpus is entirely language based and free of foreign text bodies which can hamper the effectiveness of the system.

## 3.2 FREQUENCY ANALYSIS METHOD

One outstanding feature of stop words is that they have high frequency of occurrence. This method utilizes the word frequency in documents by taking into account the terms and their corresponding frequencies, similar to Lo et al. (2015) inspired by zipf’s law.

Named after American linguist, George Kingsley Zipf, According to Wikipedia (2016), Zipf’s law states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table.

$$f = \frac{1}{r^\alpha} \quad (1)$$

Where f is the frequency of occurrence and r is the rank of the word and  $\alpha \approx 1$

Zipf’s law is easily represented by a log-log graph which is similar to an elbow as shown in Fig. 1.

Distinct words from the preprocessed Yoruba corpus with their corresponding frequencies of occurrence are extracted. Words with highest frequencies are ranked in the descending order (highest to lowest) with the most occurring word taking rank 1 and so on.

We make use of word probability against a document for normalization instead of using word’s frequencies as the gap between the frequencies of one word and another will be too wide. The probability of a word can be derived by dividing the summation of word’s frequency in a particular document by the number of tokens n in the document.

Suppose there are m distinct words in a document and there are n documents all together. We denote each word as  $w_x$  ( $x=1, \dots, m$ ) and each document as  $D_y$  ( $y=1, \dots, n$ ). For each word  $w_{x,y}$  we calculate its frequency in document  $D_y$  denoted as  $f_{x,y}$ .

However, the document has different lengths. In order to normalize the document length, we calculate the probability  $P_{x,y}$  of the word  $w_x$  in document  $D_i$  which is its frequency in the document  $D_y$  divided by the total number of words (tokens) in document  $D_i$ .

$$Pr_{x,y} = \frac{\sum f_{x,y}(w_{x,y})}{\sum N} \quad (2)$$

Where  $f_{x,y}$  is the frequency of word  $w_{x,y}$  and N is the total number of tokenized words.

**Table 3:** Showing top 20 words in Yoruba language with highest frequencies

Word	Translation	Frequency	Rank	%
yíí	this	6,345	1	0.4903
sí	to	5,888	2	0.4550
ó	he/she/it	5,799	3	0.4481
àtí	and	5,634	4	0.4354
tó	that	5,222	5	0.4036
ní	in	5,061	6	0.3911
fún	for	4,969	7	0.3840
pẹ̀lú	with	4,335	8	0.3350
é	you	4,130	9	0.3192
tàbí	or	4,006	10	0.3096
bí	as	3,564	11	0.2754
nígbà	while	3,489	12	0.2696
tí	if	3,344	13	0.2584
awón	they	3,204	14	0.2476
àwà	we	2,901	15	0.2242
nítorí	because	2,345	16	0.1812
náa	too	2,109	17	0.1630
lè	can	1,987	18	0.1536
báwo	how	1,788	19	0.1382
wọ̀n yí	these	1,390	20	0.1074

Some words that had low frequencies were excluded as they are not considered as stop words.

**Table 4:** Stop words in Yoruba with low frequencies

Word	Translation	Frequency
òpin	End	34
jùlọ	above	37
odindin	Whole	58
síwájú	Firstly	66
̀nkan	Something	92
lòótó	Truly	94
enikan	Someone	123
tètè	Quickly	234
lẹhin	After	256
miiran	Another	406

**3.3 WORD DISTRIBUTION ANALYSIS METHOD**

The other characteristic of a stop word been considered is that it is homogenously distributed across documents. The higher a word is spread across documents, the higher the chances of it being a stop word.

The distribution analysis was carried out using variance of probability to extract stop words, since it is an important measurement of a distribution.

The variance of probability (VP) is defined by the standard formula:

$$VP(w_x) = \frac{\sum_{1 \leq x \leq n} (P(w_{xy}) - \bar{P}(w_{xy}))^2}{n} \quad (3)$$

Where  $P(w_{xy})$  is the frequency of word,  $w_{xy}$  in document D and n is the number of distinct words in the document.

The  $\bar{f}(w_{xy})$  represents the mean value which is calculated by:

$$\bar{P}(w_{xy}) = \frac{\sum f_{x,y}(w_{xy})}{\sum n} \quad (4)$$

With all these values, a descending ordered list was generated. Those ranked in the top will have a larger chance to be considered as stop words in this model. The table below

shows the variance of the top 20 words calculated.

**Table 5:** Showing corresponding variance of the top 20 words

Word	Translation	Variance
Yíí	this	4122.10
Sí	To	3537.41
ó	he/she/it	3428.74
Àti	And	3231.77
tó	That	2765.36
Ní	In	2592.97
fún	For	2496.95
pèlú	With	1884.52
é	You	1704.90
Tàbí	Or	1600.62
Bí	As	1255.67
Nígbà	While	1201.29
Tì	If	1099.56
Awon	They	1005.62
Àwa	We	816.65
Nítórí	Because	521.00
Náa	Too	415.52
Lè	Can	365.66
Báwo	How	291.17
wònyí	These	167.61

### 3.4 ENTROPY MEASURE METHOD

Entropy offers us an opportunity to determine stop words based on the amount of information that a word carries. According to information theory, stop words are words that carry little information.

As defined by Claude Shannon [1948] in his paper, entropy is a measure of randomness. Words with very high randomness will have low entropy; however on the other hand, stop words have low randomness and high entropy.

The basic concept of entropy in information theory is a measure to count that how much randomness is in a signal or in a random event. An alternative way to look at this is to talk about how much information is carried by the signal. As an example, consider some

English text, encoded as a string of letters, spaces and punctuation (so our signal is a string of characters). Thus we measure the information value of the word  $w_j$  by its entropy.

As calculated earlier, the probability  $P_{x,y}$  is its frequency in the document  $D_i$  divided by the total number of words in document  $D_i$ . We calculate the entropy value ( $H$ ) for word  $w_j$  as following:

$$H(w_j) = \sum_{i=1}^n P_{x,y} \log\left(\frac{1}{P_{x,y}}\right) \quad (5)$$

The higher the entropy a word has, the lower the information value of such a word. Therefore, the words with lower entropy were extracted as candidates of stop words. From the analysis, the entropy value is calculated and the top 20 words with their corresponding entropy are given below:

**Table 6:** Showing entropy value of top 20 words

Word	Translation	Entropy
fún	For	0.1596
Ní	In	0.1595
pèlú	With	0.1591
tó	That	0.1590
é	You	0.1583
Tàbí	Or	0.1576
Àti	And	0.1572
ó	he/she/it	0.1562
Sí	To	0.1556
Bí	As	0.1542
Nígbà	While	0.1535
Tì	If	0.1519
Yíí	this	0.1518
Awon	They	0.1501
Àwa	We	0.1456
Nítórí	because	0.1344
Náa	Too	0.1284
Lè	can	0.1250
Báwo	How	0.1188
wònyí	These	0.1041



3.5 AGGREGATION METHOD

Three stop words list is generated through the frequency analysis, distribution analysis and entropy. These stop words list namely frequency analysed list (FAL), distribution analysed list (DAL) and entropy analyzed list (EAL) was aggregated using the intersection rule of the set theory to generate the final Yoruba stop words list.

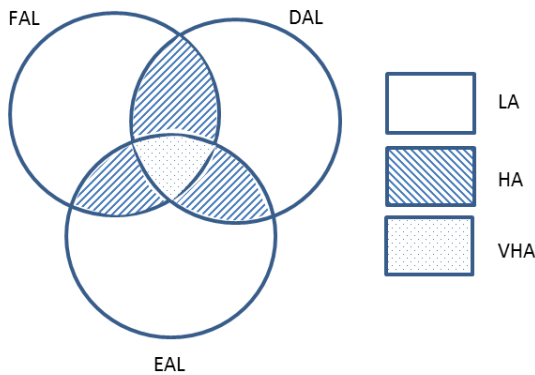


Figure 4: Showing the Venn diagram of the intersection (set theory) of the 3 methods used

A threshold of 300 stop words per list was used. The intersection of the three different stop word lists, namely: low aggregation (LA), high aggregation (HA) and very high aggregation (VHA) improved and redefined the stop words list as having high occurrences, widely distribution and are less informative. The output will produce a standardized automatically generated Yoruba stop words list.

3.6 ADAPTIVE INPUT

As our own contribution to knowledge, we introduced a learning approach to the system by ensuring that all the approaches used by the generation processes as stated above are learnable by the computer i.e. machine learning. This is to give the system an opportunity to be able to perform the automatic generation of stop words list and keep the list updated peradventure a new word which does not exist before is

introduced in the language. The system should be able tell us if such a word is a stop word and therefore include such word in the list of stop words.

3.6.1 Multinomial Naive Bayes Classification

We adopted the Multinomial Naïve Bayes Classifier (MultinomialNB) as the learning algorithm for the stopwords generation system. MultinomialNB implements the naive Bayes algorithm for multinomially distributed data, and is one of the two classic naive Bayes variants used in text classification where the data are typically represented as word vector counts.

Based on the aforementioned methodologies employed to identify a stop word as having high frequency, wide distribution and high entropy, the Multinomial Naïve Bayes classifier classified words as a stop word or a non-stop word. This algorithm helps in constantly checking and updating the stop words list to ensure consistency and prevent it from being outdated.

Although using multinomial distribution, this algorithm can be applied for case of text mining by changing the text data into a form that can be calculated with the nominal value of the integer.

Generally, the Multinomial Naïve Bayes algorithm for text classification cases where the probability of a document d being in class c is computed as

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c) \tag{6}$$

where  $P(t_k|c)$  is the conditional probability of term  $t_k$  occurring in a document in class c.

3.7 Architecture for Generation of Yoruba Stop Words

All our approaches to generate an automatic stop words list in Yoruba language is presented as an overview in the architecture below:

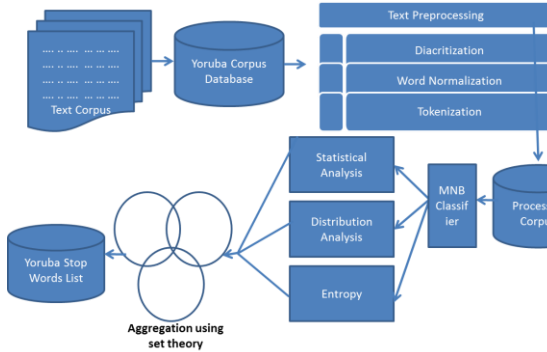


Figure 5: Architecture of the Automatic Generation of Stop Words

4.0 Results and Conclusion

Our system based on the different characteristics of stop words distinctively identifies stop words centered on our carefully adopted approaches. The output produced a standardized automatically generated 255 Yoruba stop words.

The list of Yoruba stop words generated from aggregated methods outperforms the existing stop words lists which were mostly produced using frequency measure. New words in comparison to existing list were also identified.

Our list apart from being thoroughly evaluated by language experts was also a generalized list which cut across different domains. Also a text compression rate of 63% was achieved after the removal of the stop words generated by our system.

Table 7: An alphabetically ordered list of some of the Yoruba stops words generated

A	á	àbí	Afi	àgàgà
Án	ara	àsìkò	àtàwọn	àtí
àwa	àwọn	bá	bákan	báyíí
bẹẹ	bẹrẹ	bí	Bíí	bó
bọ	bóyá	dá	dáadá	dé

			a	
déédé	di	díẹ	Dúró	e
é	èmi	ẹni	ẹnikan	ẹnikẹn
				i
eré	èyí	ẹyin	fáwọn	fẹ
fẹgbẹ	fi	fún	Fúnra	gan
gba	gbà	gbé	gbọ	gbọdọ
gbogbo	gégé	hàn	l	í
o				
lbi	ìdí	ìgbà	Ín	inú
lrú	ìṣé	ìyẹn	Já	jáde
jẹ	jé	jọ	jù	jùlọ
ká	kaàkir	kan	kàn	kanka
	i			n
káwọn	kẹta	Kí	kì	kín
kínní	kiri	Kó	kò	kọ
kọjá	kọjú	kọkọ	kú	kúrò
la	lààrin	lái pé	lára	lásán
látí	làwọn	Le	lè	lẹ
lélẹ	lẹnu	lẹyin	ló	lọ
lódún	lójọ	lójú	lọnà	lọọ
lòótọ	lórí	lọsẹ	loun	lọwọ
má	máa	márún	méjèèj	Méji
			ì	
méjilá	mẹrin	mẹta	mi	Miíràn
mo	mọ	mọ	mú	Múra
n	ń	ná	nàà	Ni
ní	níbẹ	níbi	nídíí	Nígbà
nìkan	nílẹ	ṅnílùú	nínú	Nípa
níṣé	nítórí	nìyẹn	nìyí	ńkọ
ńlá	ǹnkan	O	ó	Ò
yọ	Yóò	Náa	odindi	lẹhin
			n	
ọ	odidi	òdọ	ohun	òhún
ọjọ	ọkan	omọ	on	onà
òpin	òpọ	òpọlọp	òrọ	Òun
		ọ		
owọ	pa	pàápàá	padà	Pàdé
parí	pé	peléke	pẹlú	Péré
pọ	rán	Rára	rẹ	rẹ
rí	rò	Sá	san	Saré
ṣe	ṣẹbọ	ṣeé	ṣẹlẹ	Sí
sì	síbẹ	síbi	sílẹ	Sínú
ṣiṣé	síwáj	sọ	sódọ	sọrọ
	ú			
ṣùgbọ	sùn	Ta	tàbí	Tán
n				
tàwọn	tẹ	tẹlẹ	tètè	Ti

tí	tiẹ	tilẹ	títí	Tó
tọ	Tóó	torí	tòun	Tún
tuntun	Ú	Un	Ún	Wa
wá	Wà	wáá	wàhál à	Wáyé
wẹwẹ	Wí	wo	Wò	wọlé
wọn	Wọn	wọnyí	Yá	Yàrá
yẹ	yẹn	yí	yíí	Yín

The novel approach using Multinomial Naïve Bayes is performing well but still under observation for classification and evaluation.

However, the challenges experienced by using a resource-scarce Yoruba language corpus including scarcity of corpus, diacritization of words, system word format were well tackled but subject to further improvement.

For future work, we propose the use of semantic analysis and mapping approach for the generation.

## References

Alajmi A., Saad E. M., Darwish R.R. 2012. Toward an Arabic Stop-Words List Generation. *International Journal of Computer Applications (0975 – 8887)*, Volume 46– No.8, May 2012.

Asubiaro T. 2015. Statistical Patterns of Diacritized and Undiacritized Yorùbá Texts

Asubiaro T.V. 2013. Entropy-Based Generic Stopwords List for Yoruba Texts. *International Journal of Computer and Information Technology (ISSN: 2279 – 0764)*, Volume 02– Issue 05, September 2013

Aurangzeb K., Baharum B., Lam H.L., and Khairullah K. 2010. A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances In Information Technology*, Vol. 1, No. 1, February 2010.

Barquin J. (2016) Yoruba Translation Web - Free online translation service instantly Yorùbá. <http://yorubatranslation.com/> sighted 8<sup>th</sup> June, 2017

Blanchard A. 200. Understanding and customizing stopword lists for enhanced patent mapping.

Can, F. , Kocberber, S., Balcik, E., Kaynak, C., Ocalan, H., C., Vursavas, O. M. Information retrieval on Turkish Texts. *Journal of the American Society for Information Science and Technology*. Vol. 59, No. 3 (February 2008), pp. 407-421.

Chekima K. and Rayner A. 2016. An Automatic Construction of Malay Stop Words Based on Aggregation Method. *Second International Conference, SCDS 2016, Kuala Lumpur, Malaysia, September 21–22, 2016*, pp. 180-189.

Choy M. 2012. Effective Listings of Function Stop words for Twitter. *International Journal of Advanced Computer Science and Applications*, Vol. 3, No. 6, 2012

Cook N., Gillam L. 2008. Distributional Lexical Semantics for Stop Lists.

Corpora Preparation and Stopword List Generation for Arabic data in Social Network

Fox. C. 1992. Lexical analysis and stoplists. *Information Retrieval - Data Structures & Algorithms*, pages 102-130. Prentice-Hall.

Francis W.N. and Kučera H. 1982. Frequency Analysis of English Usage- Lexicon and Grammar.

Lo R.T., He B., Ounis I. 2005. Automatically Building a Stopword List for an Information Retrieval System.

Onashoga S.A., Abayomi-Alli A., Idowu O., Okesola, J.O. 2016. A Hybrid Approach for Detecting Malicious Web Pages Using Decision Tree and Naïve Bayes Algorithms. *Georgian Electronic Scientific Journal*:

*Computer Science and Telecommunications* 2016|No.2(48)

Rajeev P., Bedi R.P.S., Goyal V. 2013. Automated Stopwords Identification in Punjabi Documents. *An International Journal of Engineering Sciences, Issue June 2013, Vol. 8*

Sadeghi M. et. al. 2014. Automatic identification of light stop words for Persian information retrieval systems

Saif H. et. al. 2015. Automatic Stopword Generation using Contextual Semantics for Sentiment Analysis of Twitter.

Saif H., Fernandez M. and Alani H. 2015. Automatic Stopword Generation using Contextual Semantics for Sentiment Analysis of Twitter.

Savoy J. 1999. A Stemming Procedure and Stopword List for General French Corpora. *Journal of the American Society for Information Science, 50(10), 1999, 944-952*

Sharma D. and Jain S. 2015. Evaluation of Stemming and Stop Word Techniques on Classification Problem. *International Journal of Scientific Research in Computer Science and Engineering, 2015, Volume-3, Issue-2 ISSN: 2320-7639.*

Trim C. 2013. The Art of Tokenization. *Language Processing. Cited on website on Jan. 23, 2013.*

Walaa M., Ahmed H., Hoda K. 2016. Egyptian Dialect Stopword List Generation from Social Network Data.

Wang Q.A. 2008. Probability distribution and entropy as a measure of uncertainty.

Wilbur W.J. and Sirotkin K. 1991. The automatic identification of stop words.

Zaman A. N. K., Matsakis P., Brown C. 2011. Evaluation of Stop Word Lists in Text Retrieval Using Latent Semantic Indexing. *978-1-4577-1539-6 (2011) IEEE.*

Zipf H. 1949. Human Behaviours and the Principle of Least Effort. *Addison-Wesley, Cambridge, MA.*

Zipf's Law 2016. Cited on Wikipedia website (15<sup>th</sup> September, 2016). Url: [https://en.wikipedia.org/wiki/Zipf's\\_Law](https://en.wikipedia.org/wiki/Zipf's_Law)

Zou F., Wang F.L., Deng X., Han S. 2008. Evaluation of Stop Word Lists in Chinese Language.

1. Zou F., Wang F.L., Deng X., Han S., Wang L.S. 2006. Automatic Construction of Chinese Stop Word List. *Proceedings of the 5th WSEAS International Conference on Applied Computer Science, Hangzhou, China, April*

---

## Full Paper

# INFORMATION TECHNOLOGY AS A TOOL FOR ATTAINING FOOD SECURITY AND SUSTAINABLE DEVELOPMENT IN NIGERIA

---

**F. O. Okorodudu**

School of Applied Sciences,  
Department of Computer Science,  
Delta State Polytechnics,  
Otefe-Oghara,  
Delta State, Nigeria.  
okoroblackx4@yahoo.co.uk

**G. O. Eloho**

School of Applied Sciences,  
Department of Computer Science,  
Delta State Polytechnics,  
Otefe-Oghara,  
Delta State, Nigeria.  
goodluckeloho@yahoo.com

**B. Ossai**

School of Applied Sciences,  
Department of Computer Science,  
Delta State Polytechnics,  
Otefe-Oghara,  
Delta State, Nigeria.  
ossaiblessing18@yahoo.com

**ABSTRACT**

In the face of rising pressure from climate change, rising populations and decreasing crop yields, Nigeria is faced with the task of confronting the critical challenge of efficiently delivering sustainable and healthy diets for her citizens. However, the poor performance of Nigeria's agricultural sector has been attributed to the insufficient application of Information and Communication Technology (ICT) which has resulted to poor food availability and poor access and utilization problems at the household and national levels. ICT is required to successfully combine secure sustainable food systems through the application of innovative digital technologies that take into account the interactions between food and agricultural systems with broader industrial systems. This paper therefore outlines the possible roles of ICT as a tool for attaining food security and sustainable development of Nigeria. The paper advocates for a rapid action to harness the contributions of ICT towards the achievement of the Sustainable Development Goals (SDGs) as a precursor to achieving food security.

**Keywords:** Food security, Food availability; Information technology, Sustainable development, Crop yield, Nigeria.

## 1. INTRODUCTION

Agriculture accounts for 17.8% of Nigeria's Gross Domestic Product (GDP) in 2015 and supports 70% of the total population hence it is a major source of livelihood in Nigeria (NBS, 2015)

The performance of the agricultural sector especially as it relates to food security has been abysmal despite the fact that it is a growing sector in Nigeria's economy with a mean annual growth rate of 6.3% (Olaniyi *et al*, 2016). Currently, the world is facing increasing populations, limited crop yields and large amounts of waste both within the food supply chain and by consumers in western economies. It is expected that world population would have reached 9.1 billion in 2050 by what time an estimated food production increase of 70% to feed an additional over 2.2 billion people (UN/DESA, 2016).

However, contrary to popular beliefs and expectations, merely increasing production of existing crops alone is unlikely sufficient to achieve overall food security. Previous solutions to these issues involved increasing agricultural productivity. Unfortunately again, the approach of increasing agricultural productivity has led to a mirage of other issues such as the extension of monocultures and a significant loss of agro biodiversity and increased washing away of the top soil.. These human interventions in the natural environment has thus created and combined a complex interaction of climatic change, constraining the productivity of existing agricultural techniques and leading to poor crop yield and consequently, the challenge of food security (Guardian newspapers, 2014).

Hunger and poverty have been twin issues that have ravaged the world leading to the adoption of the Millennium Development Goals (MDGs) in the year 2000 by the then 189 member states of the United Nations. The International Telecommunication Union (ITU), being the leading agency for ICT, has, overtime, sought to

promote the use of ICT to address emergency situations and food security as increased access to and use of ICTs to address urgent issues as it relates to food security as it plays a key role in the agricultural industry (Olivier, 2014)

However, efforts by these bodies and agencies to employ these tools have not been sufficiently widespread enough to adequately tackle the need of the populace and ensure adequacy of food. Key factors such as legal framework, technology, markets and policy formulations etc have to be considered to adequately address food security. Available statistics on the MDGs show that the 15-year life span of the objectives did not fully achieve its targets hence the need to reformulate the Sustainable Development Goals (SDGs) which came into force early last year.

Three pillars, according to a World Bank report, are fundamental to food security. These are: food availability, food accessibility and food utilization. By extension, any country whose food production level is unable to satisfy these three fundamentals are food insecure. In most developing countries like Nigeria, ICT use has only been limited to one aspect which is food production that leads to food availability, but for the other dimension, a clause exist. ICT can be used to avail effective and relevant information concerning food accessibility and absorption to attain a satisfying state of food security and agricultural sustainability if well planned, (FAO, Brussels, 2000). The trust of this paper therefore is to correlate food security and information technology as necessary agents for attaining the new sustainable development goals as put forward by the United Nations in November 2015 so as to attain the status of a secure country.

## II. FOOD SECURITY AND SUSTAINABLE DEVELOPMENT OVERVIEW

**a. Food Security**

Two issues of profound importance lie at the heart of current thinking about the development of global societies: the challenge of food security, and the potential of attaining food security through sustainable development. Food is the basic need and necessity of life that must be satisfied before any other developmental issue. Inadequate nutrition is considered as a measure of poverty in many societies or a substitute to poverty (Datt *et al*, 2000). According to (Helen, 2002), food security helps to make for a stable political society and ensures that citizens are law abiding while food inadequacy or food insecurity results not only to poor health but also to social unrest and political instability. Hence, the world food summit defines food security as ‘a situation when all people at all times have physical and economic access to sufficient, safe and nutritious food to meet their dietary needs and food preferences for a healthy and active life’ (World Food Summit, 2003).

Further to the above, according to the United Nations Food and Agricultural Organization (FAO), "food security exists when all people, at all times, have physical and economic access to sufficient, safe and nutritious food to meet their dietary needs and food preferences for an active and healthy life." (FAO, Rome, 2006). Implicit in this definition is the recognition of the fact that food security is multi-dimensional. In recognition of the multi-dimensional view point of food security, the Committee on World Food Security identified four main dimensions or pillars’ of food security (FAO Sweden, 2014).

**Availability:** This can be ensured with adequate production of food made available to final consumers.

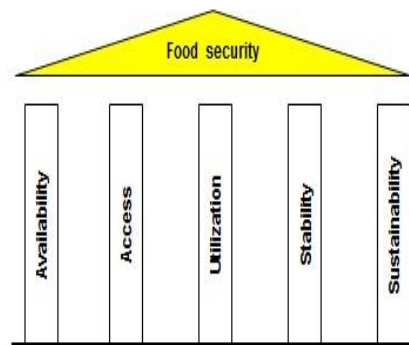
**Access:** This can be assured if all individuals and households have sufficient resources to buy food when they need them to ensure adequate nutrition.

**Utilization:** This can be assured when the human body is able to ingest and metabolize food. Nutritious and safe diets, an adequate biological and social environment, and a proper health care to avoid diseases help achieve adequate utilization of food.

**Stability** can be assured when the three other pillars are maintained over time.

Experts have noted the need for a pillar on environmental sustainability, where food production and consumption patterns do not deplete natural resources or the ability of the agricultural system to provide sufficient food security (Craig, 2013). Therefore, for the purpose of this paper, five pillars of food security is as identified in the figure 1 below.

Figure 1. Pillars of food security



WORLD RESOURCES INSTITUTE

Figure 1: pillars of food security (Craig, 2013)

**b. Sustainable Development**

The World Commission on Environment and Development (WCED) defined sustainable development as “development that meets the needs of the present, without compromising the

ability of future generations to meet their own needs.” This definition contains within it two key concepts: one is the concept of ‘needs,’ in particular the essential needs of the world’s poor, to which overriding priority should be given; and the other is the idea of limitations imposed by the state of technology and social organization, on the environment’s ability to meet present and future needs.

Although the definition of sustainable development emerged forms an international enquiry into the relationship between environment and development, it is not concerned primarily with the environment but with the sustainability of the overall developmental context. While the definition of sustainable development emerged from an international enquiry into the relationship between environment and development, it is not concerned primarily with the environment but with the sustainability of the overall developmental context.

On the 25<sup>th</sup> of September 2015, 193 Member States of the United Nations came together and adopted the SDGs as a direct replacement for the MDGs which phased out after the 15 year period elapsed. The SDGs is a set of 17 objectives and 169 targets which is expected to guide actions of governments, international agencies, civil societies and other institutions over the next 15 years (2016- 2030). The ambitious 2030 Agenda is a global vision for people, for the planet and for long-term prosperity. The SDGs is a plan to make the future better unto a sustainable and resilient course that will lead to a transformation in the standard of living and chart a course for a more inclusive and dynamic pathways to development. It has been argued and rightly so by scholars that the new SDGs are more ambitious than the MDGs. While the MDGs helped to halve extreme poverty, the SDGs aims to end it. While MDGs focused on the poorest countries, the SDGs engage all nations in a

shared, universal agenda. The MDGs prioritized prosperity and inclusion while the SDGs takes cursory look at environmental sustainability as a fundamental pillar of global wellbeing. These goals are achievable, but they require a breakthrough in both the speed and degree of progress.

### III. AGRICULTURAL PRODUCTIVITY AND FOOD SECURITY

Agricultural productivity is a measure of the amount of agricultural output produced for a given amount of inputs, such as an index of multiple outputs divided by an index of multiple inputs, (FAO, Sweden, 2014)

Agricultural technologies that can increase food security in developing countries have existed for decades yet these countries continue to be live in hunger and poverty. The reason for this is not far-fetched. Farmers are faced with a number of challenges yet these problems have workable solutions, but the global difficulty is getting the appropriate information to farmers which is pertinent in ensuring food security in Nigeria. To benefit from knowledge, one should be able to acquire existing knowledge, produce new knowledge and apply this knowledge to foster development. In the light of the above, ICT must meet the needs of the local people in sharing the indigenous and the acquired knowledge. This will solve the challenges and lead to increased agricultural productivity, increased awareness and sharing of information that will eventually ensure food security for all in the country.

Tackling the issue of hunger and malnutrition is not only about boosting agricultural productivity but also has to do with increasing income and creating a resilient food system that will strengthen markets so that people can safely access safe and nutritious food to feed the increasing hunger and starvation that is ravaging the human race.



#### IV. INFORMATION TECHNOLOGY AND FOOD SECURITY IN RURAL AREAS

ICT is an existing and widely deployed technology that can be mobilized to step up the pace and scale of transformation in the rural areas. ICT can play a key role in delivering an innovative and integrated cross sectorial sustainable development outcomes. According to ITU, the transformational potential of ICT can be used as a means of implementing the SDGs using the UN Broadband Commission for sustainable development, (Broadband Commission, 2014). ICT can also deliver innovation, connectivity, productivity and efficiency across all sectors in remote areas and strengthen the resilience of critical infrastructure in the process

Scholars have however argued that the traditional ways of communication have been monologue in nature and have not allowed much interaction with users. An important area of innovation in ICT is by combining different ICTs in order to deliver a more complete communication package. New ICTs have to be linked to traditional communication forms in order to meet identified needs and reach out to specific groups who have been shut across from information access that could increase their agricultural productivity. The realization of the opportunities offered by ICTs for rural development, agricultural sustainability and food security requires a culture of information and new skills, (FAO Rome, 2016).

A critical factor in meeting the challenge of ensuring food security among rural households in developing countries like Nigeria is human resource development through knowledge building and information sharing (Murithii, 2009). According to (Rafea, 2009), there is a general lack of relevant and accurate information on production practices, farm

management, prices of agricultural products and food security dimensions for agricultural products that can better the lots of farmers.

#### V. INTEGRATING INFORMATION TECHNOLOGY IN FOOD SECURITY AND SUSTAINABLE DEVELOPMENT

The abilities of ICTs on food security and sustainable development can be linked with improving communication between research systems, farmers, information regarding inputs and introducing technologies to provide more rapid accessibility to high quality information. This will help to ensure that accurate information are optimized for agricultural product sales to increase agricultural products and decrease agricultural product losses (Temu *et al*, 2004).

ICT can dramatically boost the uptake of SDGs in five major ways:

##### 1. Speed and Scale of ICT Uptake

ICT itself diffuses with remarkable speed and at a global scale; the digital transformation has already begun. Mobile subscriptions went from a few tens of thousands in 1980 to over 7.4 billion subscriptions in 2015. Facebook users skyrocketed from a few users 2004 when the platform was launched to 1.5 billion users in mid-2015. According to projections, mobile broadband will cover more than 90 percent of the world's population and go from almost one billion subscribers in 2010 to 7.7 billion subscriptions by the year 2021. Smartphones use is expected to grow from near-zero subscriptions in 1999 to around 6.4 billion subscriptions by 2021

##### 2 Reduced Deployment Costs

ICT can also help to reduce the cost of deploying new services. ICT also enables students to access quality online teaching even when no qualified teachers are locally available. Online

finance allows individuals to obtain banking services even in regions where no banks are present. This shows that ICT is introducing vital services for low income countries like Nigeria. A continuation in this direction will make more farmers to get the needed information and increase productivity.

### 3 Growth of public awareness

In the past, information on new technologies spread by word of mouth, local demonstration, and gradual scale-up of government programs and services. Now, with torrents of information flowing in real time through the internet, social media, mobile communications and other e-channels, information travels instantly around the globe in the shortest of times. News, music, fashion and new technologies revolve around the world in days, not decades, making it easier to reach more people in a shorter timeframe.

### 4 Rapid upgrade rate

Global information flows are enhanced and technology developers are much more attuned to advances in other parts of the world. There is a trend towards many ICT applications becoming open-source—or at least interoperable—which enables gains made by a developer in one part of the world to be picked up and built on by others on the other side of the globe, accelerating the whole process of technology upgrade. The growing speed of the global innovation cycle is shortening the duration of each technology generation, especially for ICT-based solutions, meaning progress happens faster.

### 5. Low-cost digital training

Another major way that ICT can accelerate technology diffusion is by providing low-cost online platforms for training workers, students and others in these new technologies. Special training materials are also being delivered conveniently over smart phones, tablets, laptops and other devices. Deploying multiple

channels for training materials makes it easier to provide workers with real-time, in-service training that does not disrupt work schedules but integrates training into the work itself. In this way, ICT-hosted training modules and courses provide a means to train millions of workers, especially young and under-employed workers, in the uses of new ICT applications for SDG-oriented service delivery.

## VI. INFORMATION TECHNOLOGY IN FOOD SECURITY PROMOTION AND SUSTAINABILITY

### a. Local Radio and Mobile Phones

The use of community radio is an old form of communication channel that has been in use for decades. Local radio helps rural communities to access information that they need to improve their farm yield. Hence local community radios have been recommended over time for use in promoting all food security and agricultural sustainability dimensions in rural communities. Also, mobile phones can aid in the delivering of complete and current information about market prices. With the use of ICT, information can be delivered in a timely fashion.

### b. Extension Agents

Extension agents play a critical role in ensuring that needed information about current trends in food production gets to farmers in an efficient and timely manner. Extension agents also provide a forum through which farmers can access information concerning food production, planning and food marketing.

### c. The Internet

Internet use in accessing information concerning; crop varieties, increased yields, prices and markets, food production and other agricultural issues is key to ensuring food security. If farmers are able to use the internet in sharing information, it will lead to improved production, marketing and economy growth.

#### d. Training and Retraining of Local Farmers

With the right information, local farmers are as civilized and up to date as urban dwellers. Thus, governments, through its agents such as the FADAMA operators must ensure that constant training and retraining of local farmers is carried out to enlighten them on the current trends in agricultural farming. This will go a long way in bringing us nearer to attaining ample food production that will lead to food sustainability and eradicating hunger and poverty.

#### e. Grants

A key challenge to rural farmers is the challenge faced in securing loans from money agencies. These farmers have little or nothing to tender in return for a loan which makes it herculean for them to get grants to farm on a large scale basis. If the world is to be hunger free by 2030 as anticipated by the sustainable development goals, then, effort must be made to ensure that rural farmers are able to access loans without much ease.

### VII. CONCLUSION

We have submitted, from the above, that ICT and food security are essential platforms for the attainment of the SDGs. While this is recognized in the 2030 dateline Agenda for sustainable development, ICT is neither systematically nor adequately reflected in the individual goals and subsequent targets. Rapid action is thus needed to harness the contribution that ICT can make toward the achievement of the global goals.

Fully embracing the potential of ICT is therefore a key ingredient to achieving food security and subsequently the SDGs so as to fast track the attainment of the setout goals by 2030, and possibly to even accelerate their achievement. Beyond goals and targets, efforts must be made to address the means of implementation, monitoring and financing of means of boosting food security in the country. Researches in the

area of boosting food security through the use of ICT must be given top priority so as to fast track and support the continued development of SDGs as Nigeria move to formulate national targets and strategies for its full implementation.

### References

- Broadband Commission, 2014. Means of Transformation, ITU/UNESCO <http://www.fao.org/docs/eims/upload/295346/Rafea%20Managing%20Agriculture%20Knowledge.pdf> 4pp. Sourced online on 30<sup>th</sup> January, 2017
- Craig Hanson, 2013. Food Security, Inclusive Growth, Sustainability, and the Post-2015 Development Agenda. *Research Background paper submitted to the High Level Panel on the Post-2015 Development Agenda*. Geneva, Switzerland
- Datt, G., Simler, K; Mukherjee, S and Dava, G. 2000. .Determinants of Poverty in Mozambique 1996-97 (FCND Discussion Paper. No.78. *International Food Policy Research Institute*: Washington, DC,USA.
- Department of Economic and Social Affairs of the United Nations Secretariat (UN/ DESA).
- FAO, 2006. Food Security. Policy Brief. FAO Agricultural and Development Economics Division with support from FAO Netherland Partnership Programme and the EC-FAO Food Security Programme, FAO, Rome, Italy
- FAO, 2014. Food Security in the Sustainable Development Goals: Where is the process heading? *Swedish International Agriculture Network Initiative*. Stockholm, Sweden
- Helen, H. J. 2002. Food Insecurity and the Food Stamp Programme. *American Journal of Agricultural Economics* 84(5): 1215-1218.
- Murithii, O, 2009. Information Technology for Agriculture and rural development in South Africa: Experiences from Kenya. *Paper presented at the conference on International research on*

*Food Security, Natural Resources Management and Rural development, Tropentag: University of Hamburg Nzirasanga Commission on Education, Germany*

National Bureau of Statistics (NBS), 2015. Nigerian Gross Domestic Product Report Q2 2015". Retrieved 20 January 2017. From <http://www.nigerianstat.gov.ng/pages/download/312>

Olaniyi, Olumuyiwa Akin O.A. and Ismaila, Kayode O., 2016. Information and Communication Technologies (ICTs) Usage and Household Food Security Status of Maize Crop Farmers in Ondo State Nigeria: Implications for Sustainable Development. *Library Philosophy and Practice (e-journal)*. Paper 1446.

Olivier De Schutter, 2014. Report of the Special Rapporteur on the right to food, sourced online on 24 January 2014

Rafea, A., 2009. Managing Agriculture Knowledge: Role of information and communication Technology. FAO publications, Rome, Italy

Temu, A. and Msuya, E., 2004. Capacity Human Building in Information and Communication Management Towards Food Security., *CTA Seminar on the Role of Information Tools in Food and Nutrition Security*, Maputo, Mozambique, pp. 8-12

The Guardian Newspapers, (2014): <http://www.theguardian.com/environment/2014/mar/31/climate-change-food-supply>. Sourced online on February, 2, 2017.

The following definitions are paraphrased from Gross, R., H. Schoeneberger, H. Pfeifer, and H-J A. Preuss, *The Four Dimensions of Food Security: Definitions and Concepts*, European Union, Internationale Weiterbildung und Entwicklung gGmbH (InWEnt), and FAO, Brussels, 2000.

World Food Summit 2003. Agriculture and Sustainable Development, Rome, Italy.

---

Full Paper

**USING THE ADJUSTED WEIGHTING FUNCTION TO BRIDGE  
THE NETWORKED READINESS DIGITAL DIVIDE**

---

**P. K. Oriogun**

Department of Computer Science  
Lead City University,  
Ibadan,  
Oyo State, Nigeria  
p.oriogun@lcu.edu.ng;  
p.oriogun@gmail.com

**ABSTRACT**

P. K. Oriogun

This paper critically examines the current state of the Networked Readiness Index (NRI) framework as administered and published by the World Economic Forum through its yearly Global Information Technology Report (GITR) series since its inception in 2001. The newly published Adjusted Weighting Function (AWF) is a model specifically designed to overcome the inadequacies reported by a number of authors concerning the existing NRI framework, with the goal of bridging the digital divide of the developed and developing economies. In this paper, the AWF model was implemented using case studies from three world economic regions (sub-Saharan Africa, ASEAN and Nordic countries) comprising a sample size of 18 countries. The initial results are promising in terms of reducing the gap between the developed and developing economies. Consequently, this paper argues that using the AWF model for computing Networked Readiness Index within the existing framework of the World Economic Forum will go a long way in bridging this digital divide. The paper further claims that the AWF model is an unbiased measure of a country's Networked Readiness Index, and will be more acceptable to developed and developing economies.

**Keywords:** ASEAN (Association of Southeast Asian Nations), NRI (Networked Readiness Index), ICTs (Information and Communication Technologies), ITU (International Communication Union), WEF (World Economic Forum)

## 1. INTRODUCTION

This is a follow-on paper to a recently published paper (Oriogun, 2017) proposing a weighting function for adjusting the Global Information Technology Report (GITR) Networked Readiness Index (NRI) Framework on the basis that, a number of authors have suggested that the credibility of the NRI is called into question by the non-transparent manner in which the authors report the sources of the data and the methodology followed to collect the raw data. In support of this new weighting function, the paper explores the legitimacy of the link between a country's Networked Readiness Index and economic development on the basis that economic development is a context-specific process which is entangled with indigenous politics and historically based institutions. In order to support the claim that this new *Adjusted Weighted Function* is a much fairer and unbiased measure of a country's Networked Readiness Index, an in-depth investigation of case studies from three world regional economies based on continuous primary data over the past 5 years from the World Economic Forum was conducted. The investigation specifically compared the developed and developing economies from the same baseline. A total sample size of 18 world economies were examined, they include sub-Saharan Africa (Botswana, Kenya, Mauritius, Namibia, Nigeria and South Africa), ASEAN (Cambodia, Indonesia, Malaysia, Philippine, Singapore, Thailand and Viet Nam – Association of Southeast Asian Nations) and the Nordic (Finland, Sweden, Norway, Denmark and Iceland –Northern European countries).

## 2. LITERATURE REVIEW

A number of learned authors have been very critical of the NRI framework as administered by the World Economic Forum through its yearly Global Information Technology Report series. When Avgerou (2003) examined the validity of the relationship between ICTs and economic development that has been constructed in the discourse of some influential publications, namely, The Global Information Technology Report (Kirkman et al., 2002), Competitive Report (Porter et al., 2002), The World Bank (2002) and United Nations Development Programme (2001), she discovered that tool-and-effect association suggested in such discourse was not only misleading, it was also dubious. Avgerou (2003) vehemently disagree with the position of these four influential publications, and argue that such a

link is based on distorted economic viewpoint, and that economic development is situated, context-specific process that is entangled with indigenous politics and historically based institutions. In similar vein, a cautionary note was given by Dutta and Jain (2003) stating that: *'countries ranked together can show very small variation in the index... Additionally small differences in the index may be outside the limits of statistical significance due to the fact that some missing observations were estimated using analytic techniques such as regression and clustering'* (5). The same report explicitly outlined the research challenges of computing the Networked Readiness Index as *'Absence of key usage matrices; selection of countries; ensuring statistical significance; data estimation and calculating the NRI'*. furthermore that computation of the *'Networked Readiness is a complex phenomenon, and measuring countries NRI remains a significant challenge, and any framework or model representing NRI is a simplified representation at best, a simplified version of reality'*.

According to (Goswami, 2006) a number of extraneous variables have been included that do not shed any light on ICT environment, readiness or usage, while others that may have added greater robustness to the measure, are missing. Additionally, that *'the credibility of the NRI is called into question by the non-transparent manner in which the authors report the sources of the data and the methodology that was followed to collect the raw data'*. The Economic and Social Commission for Western Asia (ESCWA, 2011), suggested that some of the disadvantages of the NRI include but not limited to the following: *'may send distorted policy messages if it is incorrectly compiled or misinterpreted; it may not highlight weaknesses faced by the country ICT infrastructure; it may disguise serious failings in some dimensions and increase the difficulty of identifying appropriate remedial actions, if the construction process is not transparent, and, it may also create friction between different countries'*. In the context of cyber warfare, Watkins and Hurley (2015) reported that the NRI model is based on qualitative factors that do not necessarily transition well into predictability of future trends and behaviour (p.386). When Malisuwan et al (2016) investigated Thailand's position in the Networked Readiness Index, they concluded that *'... it is clear that there is no fixed formula for the economic policy that suits each individual country, but various widespread procedures normally share some common characteristics'* (p. 408). The *Adjusted Weighted*

Function (Oriogun, 2017) is a way of addressing all the inadequacies of the current NRI framework as reported by the Global Information Technology Reports from its inception to date.

3. METHODOLOGY

Oriogun’s (2017) *Adjusted Weighting Function* (AWF) is based on the current NRI methodology. It is assumed that the initial membership rating will normalize all the anomalies and inefficiencies of the current framework as alluded to in this paper. The AWF model offers definitions for what is required for a country to be at a ‘**ready state**’ and ‘**minimum rating**’ for membership of the NRI community. In the existing unadjusted NRI framework, the minimum NRI rating a country is able to score is 1.0 from a maximum value of 7.0 (although the 2005 and 2006 report differs greatly from this norm –without any reason to support the

sudden change). In the context of AWF model, a country that is deemed to be important enough with some initial basic infrastructure in place to be a member of the NRI community, the ‘**minimum rating**’ that should be awarded must be 3.5 out of the maximum of 7.0 (already 50% of the ratings). This ‘**minimum rating**’ is what the AWF model recognize as the ‘**ready state**’ for membership of the NRI community. It now depends on individual country/economy to prove itself in terms of moving from 50% rating up to 100%, which according to this paper should be possible, or at least the opportunity is available. The AWF model was derived initially through Mathematical Induction, however, when it transpired that this approach was not ideal or forthcoming, an Experimentation approach was employed instead. This led to a Simple Linear Regression equation description of the general form below:

General Form:

$$\text{Oriogun NRI Adjusted Weighting Function } -f(\text{awf\_nri})$$

$$f(\text{awf\_nri}) = aN + b$$

Where

$$N = (\text{gitr\_nri})/7; \quad 0.0 \leq \text{gitr\_nri} \leq 7.0$$

$$a + b = 7.0; \quad 3.5 \leq a \leq 6.0; \quad 1.0 \leq b \leq 3.5$$

4. EVALUATION

The AWF model was implemented to normalize and improve the existing World Economic Forum framework of NRI ratings for three separate world regional economies. The reason for selecting these economies is because continuous primary data was available through the Global Information Technology Report In the past 5 years. Furthermore, for a number of these economies in the same region there are developed and developing economies. The continents that were captured for examination and analysis include sub-Saharan Africa, ASEAN

and the Nordic economies were captured for examination and analysis.

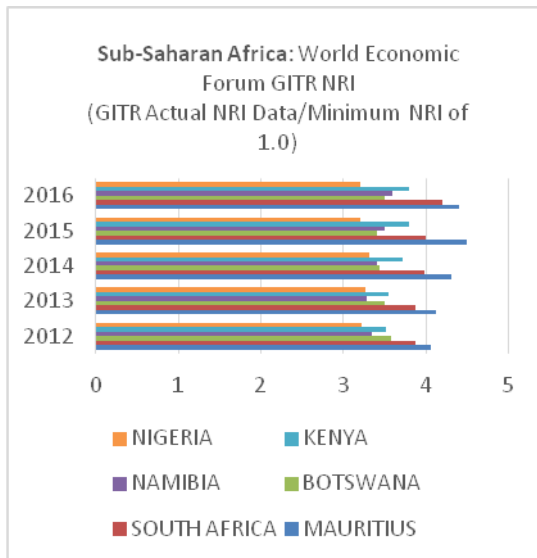
Table 1, Table 3 and Table 5 shows the computed NRI to date as computed by the World Economic Forum on the basis of the assumptions made in this paper that, the baseline rating for all the countries is currently set to 1.0 (except from 2005 and 2006). Table 2, Table 4 and Table 6 shows the implementation of Oriogun AWF model for a representative sample of sub-Saharan Africa, ASEAN and Nordic economies.

Table 1: Sub-Saharan Africa: World Economic Forum GITR NRI (Actual NRI Data) 2012 – 2016

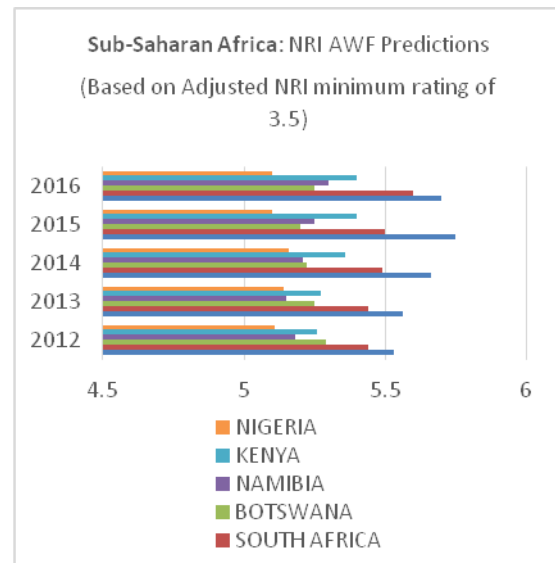
	MAURITIUS			4.06	4.12	4.31	4.5	4.4
	SOUTH AFRICA			3.87	3.87	3.98	4	4.2
	BOTSWANA			3.58	3.5	3.43	3.4	3.5
	NAMIBIA			3.35	3.29	3.41	3.5	3.6
	KENYA			3.51	3.54	3.71	3.8	3.8
	NIGERIA			3.22	3.27	3.31	3.2	3.2

**Table 2** Sub-Saharan Africa: NRI AWF Predictions (based on Actual NRI/Minimum NRI of 1.0)

MAURITIUS	5.53	5.56	5.66	5.75	5.7	
SOUTH AFRICA	5.44	5.44	5.49	5.5	5.6	
BOTSWANA	5.29	5.25	5.22	5.2	5.25	
NAMIBIA	5.18	5.15	5.21	5.25	5.3	
KENYA	5.26	5.27	5.36	5.4	5.4	
NIGERIA	5.11	5.14	5.16	5.1	5.1	



**Figure 1** Graphical Representation of Table 1



**Figure 2** Graphical Representation of Table 2

**Table 3** ASEAN: World Economic Forum GITR NRI (GITR Actual NRI Data)

	2012	2013	2014	2015	2016
Singapore	5.86	5.96	5.97	6	6
Malaysia	4.8	4.82	4.83	4.9	4.9
Thailand	3.78	3.86	4.01	4	4.2
Indonesia	3.75	3.84	4.04	3.9	4
Philippine	3.64	3.73	3.89	4	4
Viet Nam	3.7	3.74	3.84	3.9	3.9

**Table 4** ASEAN: Oriogun NRI AWF Predictions (based on Actual NRI/Minimum NRI of 1.0)

	2012	2013	2014	2015	2016
Singapore	6.43	6.48	6.49	6.5	6.5
Malaysia	5.9	5.91	5.92	5.95	5.95
Thailand	5.39	5.43	5.51	5.5	5.6
Indonesia	5.38	5.42	5.52	5.45	5.5
Philippine	5.32	5.37	5.45	5.5	5.5
Viet Nam	5.35	5.37	5.42	5.45	5.45



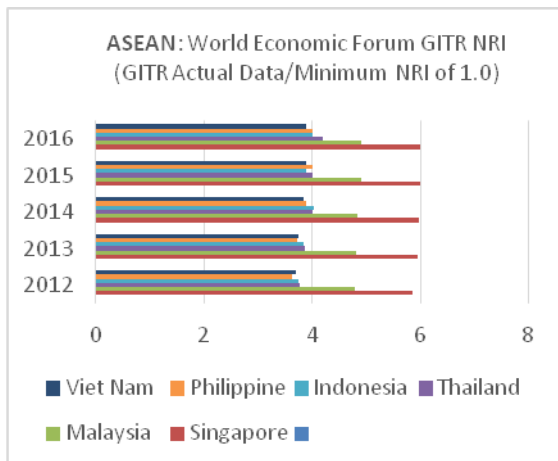


Figure 3 Graphical Representation of Table 3

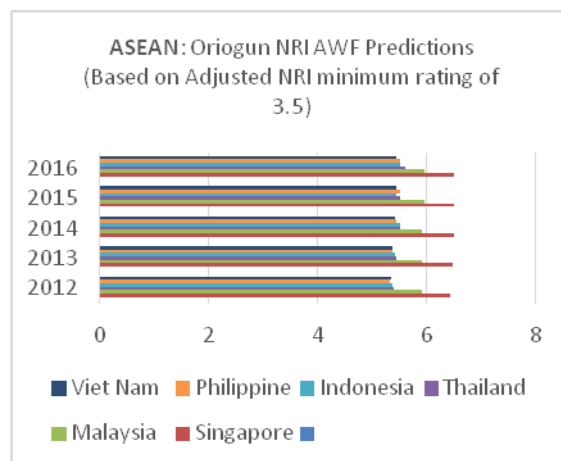


Figure 4 Graphical Representation of Table 4

Table 5 Nordic: World Economic Forum GITR NRI (GITR Actual NRI Data)

	2012	2013	2014	2015	2016
FINLAND	5.81	5.98	6.04	6	6
SWEDEN	5.94	5.91	5.93	5.8	5.8
NORWAY	5.59	5.66	5.7	5.8	5.8
DENMARK	5.7	5.58	5.5	5.5	5.6
ICELAND	5.33	5.31	5.3	5.4	5.5

Table 6 Nordic: Oriogun NRI AWF Predictions (based on Actual NRI/Minimum NRI of 1.0)

	2012	2013	2014	2015	2016
FINLAND	6.41	6.49	6.52	6.5	6.5
SWEDEN	6.47	6.46	6.47	6.4	6.4
NORWAY	6.3	6.33	6.35	6.4	6.4
DENMARK	6.35	6.29	6.25	6.25	6.3
ICELAND	6.17	6.16	6.15	6.2	6.25

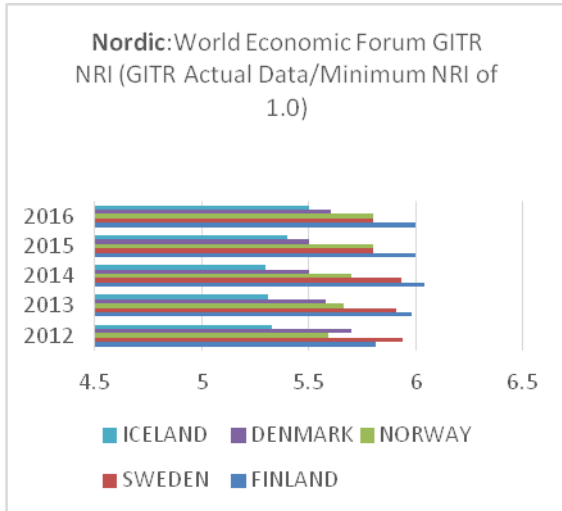


Figure 5 Graphical Representation of Table 5

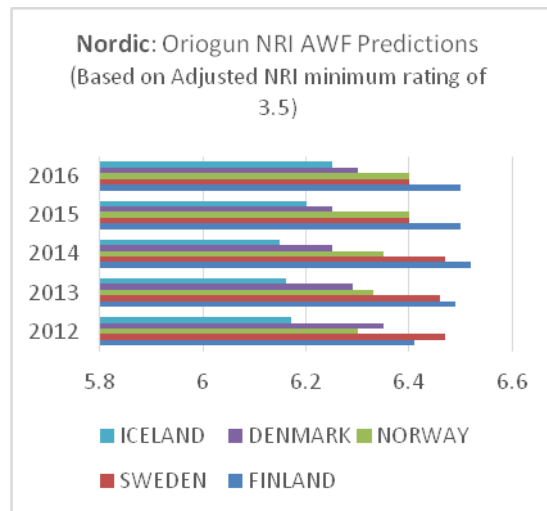
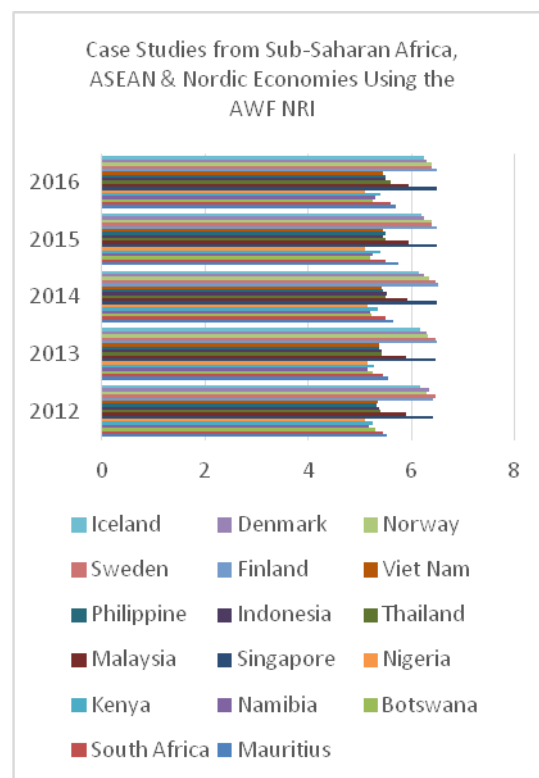
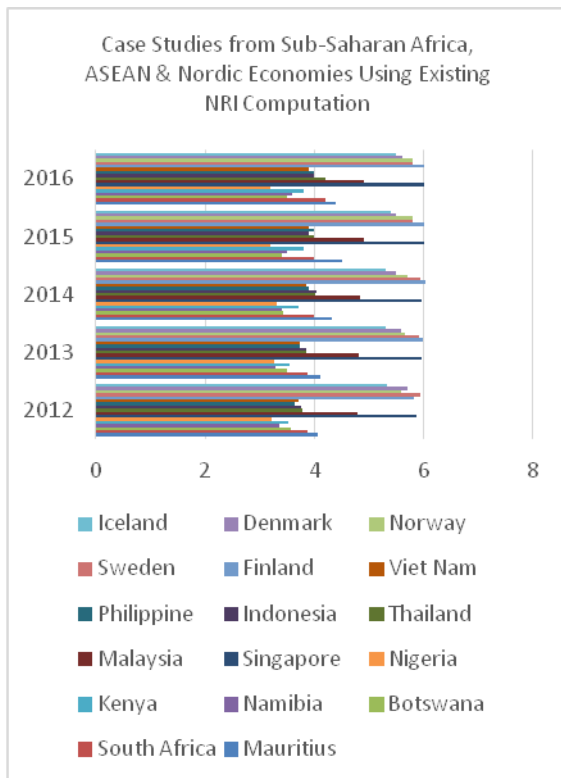


Figure 6 Graphical Representation of Table 6

4.1 Reducing the Digital Divide

It is evident from Figure 7 and Figure 8 that using Oriogun’s AWF model for the recalibration of existing NRI will reduce the so called *digital divide* immensely. Given that the World Economic Forum has invested heavily on developing the framework for the current

Networked Readiness Index, it is the suggestion in this paper that if the current framework is adjusted as specified in this article, all the economies deemed capable of being part of the NRI community would readily embrace and accept the existing framework as being adequate and robust indication of their individual performance.



**Figure 7:** Sub-Saharan Africa, ASEAN & Nordic Economies using Existing NRI Computation

## 5. DISCUSSION

In terms of sub-Saharan Africa, the NRI is still very low compared to developed economies such as the Nordic countries that tend to be within the top 5% of the NRI rankings over the past 5 years. Developing economies are still lagging behind the developed economies. This digital divide according to International Telecommunication Union (ITU) is largely based on uptake of wireless-broadband and fixed broadband services. Furthermore, many developing economies remain at very low levels. The World Bank (2016) also noted that there is digital divide within developing nations, noting that those with good education have better connectivity and are more resourceful in the context of the Usage with respect to Networked Readiness. This could possibly explain the reason why sub-Saharan Africa economies, and to some extent, other developing economies remain at the bottom half of the NRI ranking. The intention of Oriogun's AWF model is to make sure that before being accepted into the NRI community, every country deemed ready for membership would have had in place the basic ICT infrastructure to flourish as a full member of the NRI community.

It is evident from the ASEAN economies, Singapore economy in terms of NRI ranking over the past 5 years have been the highest, scoring average rating of 5.96 out of a maximum of 7.0 under the unadjusted NRI data (as shown in row 1 of Table 3). However, when the same computation is performed with adjusted NRI AWF the average rating increased to 6.48 (as shown in row 1 Table 4) out of maximum of 7.0 rating on the basis of the current NRI framework. Malaysia on average over the same 5 years have an rating of 4.85 using the unadjusted NRI data (as shown in row 2 Table 3). In contrast, a better result was obtained using Oriogun's AWF model of 5.93 (as shown in row 2 Table 4). It is not

**Figure 8:** Sub-Saharan Africa, ASEAN & Nordic Economies using the AWF NRI

surprising that Malaysia is closely following Singapore because when Philippine Institute for Development Studies examined trends in ICT statistics, on how Philippines fair in ICT (Albert, Serafica, and Lumbera 2016), they discovered that the percentage of households with computer and percentage of household with internet access are: Singapore (85% household with computer, 84% household with internet access) and Malaysia (65% household with computer, 64.5% household with internet access). In contrast, Thailand (29.1% household with computer, 23.2% household with internet access 2013 study), Indonesia (15.6% household with computer, 5.7% household with internet access 2013 study), Viet Nam (16% household with computer 211 study, 12.5% household with internet access 2010 study) and Philippines (13.1% household with computer, 10.1% household with internet access 2010 study). In essence, except from Singapore and Malaysia that have grown steadily over the past 5 years in aspiring to obtaining the status of developed economies, the rest of the ASEAN economies are still operating as developing economies, very much the same as sub-Saharan Africa economies with the exception of Mauritius with similar NRI rating as Thailand.

The 2013 report (Bilbao-Osorio, B, et al., 2013) on NRI suggest that four (Finland, Sweden, Norway, and Denmark) of the five Nordic economies continue to feature in the top 10, and that Iceland is not too far behind. The 2015 report (Di Battista et al., 2015) further highlighted the fact that the Nordic countries continue to perform well. The five Nordic countries have featured in the top 20 of every edition of the Global Information Technology Report since 2012. The same 2015 report claims that Norway has the best digital infrastructure in the world, and that Iceland has the best access to the internet (95%). Dutta and Mia (2011) reported that Sweden came

first in the NRI ratings due to the country's favourable climate for technological innovation adoption and penetration of new technologies. According to van Marion and Hovland (2015),

average of 70% in 2011 and about equal to the Nordic average.

## 6. CONCLUSION

This paper offers a weighting function for adjusting the current NRI final computation on the basis of the current World Economic Forum framework.

## 7. REFERENCES

- Albert, J. A. G., Serafica, R. B., and Lumbea B. T., 2016. *Examining Trends in ICT Statistics: How Does the Philippines Fare in ICT?* Philippine Institute for Development Studies, Discussion Paper Series No: 2016-16, May 2016. [Available Online] Accessed 14<sup>th</sup> May 2017 <http://dirp3.pids.gov.ph/websitecms/CD/PUBLICATIONS/pidsdps1616.pdf>
- Avgerou, C., 2003. *The link between ICT and economic growth in the discourse of development*, In: Korpela, Mikko; Montealegre, Ramiro and Poulymenakou, Angeliki (eds) (2003) *Organizational information systems in the context of globalization*. New York, USA: Springer, pp. 373-386, ISBN 978 1402074882
- Bilbao-Osorio, B., S. Dutta, T. Geiger, and B. Lanvin., 2013. "The Networked Readiness Index 2013: Benchmarking ICT Uptake and Support for Growth and Jobs in a Hyperconnected World." *The Global Information Technology Report*. B. Bilbao-Osorio, S. Dutta, and B. Lanvin, editors. Geneva: World Economic Forum.
- Di Battista A., S. Dutta., T. Geiger., and B. Lanvin., 2015. *The Networked Readiness Index: Taking the Pulse of the ICT Revolution*, *The Global Information technology Report 2015*. World Economic Forum 2015, ISBN 978-92-95044-48-7, pp 3-30. Available at [www.weforum.org/gitr](http://www.weforum.org/gitr) (accessed 16th May 2017)
- Dutta, S. and A. Jain., 2003. *The Networked Readiness of Nations*, *The Global Information Technology Report 2002-2003*. New York: Oxford University Press. 2-25.
- Denmark is one of the most advanced countries in terms of internet usage, with 89% of Danes using the internet regularly in 2011. This was reported to be above the European Union (EU)
- The author claims that computing of the NRI rankings based on Oriogun's AWF model will minimize the so called 'digital divide' alluded to
- Dutta, S., and I. Mia., 2011. *The Global Information Technology Report 2010 – 2011 Transformation 2.0*. World Economic Forum 2011. Available at [www.weforum.org/gitr](http://www.weforum.org/gitr) (accessed 16th May 2017)
- ESCWA., 2011. *Measuring the Information Society International Benchmark Models*, [online] Available at: <https://www.unescwa.org/events/standardizing-information-society-measurement-models-escwa-region>, [accessed 13th May 2017].
- Goswami, D., 2016. *A Review of the Network Readiness Index*. Lyngby: LIRNE.NET.
- Kirkman, G. S., C.A. Osorio and J.D. Sachs., 2002. "The Networked Readiness Index: Measuring The Preparedness of Nations for the Networked World," *The Global Information Technology Report 2001 – 2002*, March 2002.
- Malisuwan, S., W. Kaewphanuekrungsi, N. Tiamnara and N. Suriyakrai., 2016. *Thailand's Position in the Network Readiness Index (NRI): Analysis and Recommendations*, *Journal of Economics, Business and*

Management, Vol. 4, No. 5, May 2016

van Marion L., and Hovland, J. H., 2015. *The Nordic Digital Ecosystem Actors, Strategies, Opportunities*, Nordic Innovation Publication, Nordic Innovation, Oslo 2015, ISBN 978-82-8277-080-4 (Print),

December 2015.

Oriogun P. K., 2017. *Proposing a Weighting Function for Adjusting the Global Information Technology Report Networked Readiness Index Framework*, To Appear in the Proceedings of the British Computer Society Software Quality Management (SQM 2017) Conference, Southampton Solent University, UK, 10th April 2017.

Porter, M. E., Sachs, J. D., Cornelius, P. K., McArthur 1. W., and Schwab K., 2002. *Executive Summary: Competitiveness and Stages of Economic Development*, The Global Competitiveness Report 2001-2002, M.

E. Porter, 1. D. Sachs, P. K. Cornelius, 1. W. McArthur and K. Schwab, New York: Oxford University Press, 2002, pp. 16-25.

Watkins, L and J. Hurley., 2015. *Cyber Maturity as Measured by Scientific Risk-Based Metrics*, In the Proceedings of the International Conference on Cyber Warfare and Security (ICCSWS), March 2015.

World Bank., 2016. *World Development Report 2016: Digital Dividends*, Washington, DC: World Bank. doi: 10.1596/978-1-4648-0671

**13<sup>th</sup>**

# **International Conference**



**13<sup>th</sup>**

**International Conference**



**Session F:**

**Digital Economies: Capacity  
Building, Start-ups and Youth  
Innovation**

---

Full Paper

**A NEURO-FUZZY SYSTEM FOR CHARACTERISING SAKI  
(FULANI WEAR) IN WOVEN FABRICS**

---

**R. R. Madaki**

Department of Computer Science,  
Northwest University, Kano  
rafeeahmadaki@yahoo.com

**Y. Baguda**

Faculty of Information Science,  
King Abdul Aziz University,  
Jeddah  
baguday@yahoo.com

**L. Abdulwahab**

Faculty of Computer Science and  
information Technology,  
Bayero University, Kano  
abd\_wahhb@yahoo.com

**ABSTRACT**

The need for pattern recognition system cannot be over emphasized as it cuts across many fields. Majority of the work on fabric pattern recognition focuses on determining the nature of the wefts and wraps in a given cloth. Others are more concerned with defect detection of hand woven fabrics and a few consider recognizing some African fabrics. However, there are limited studies on Saki Pattern recognition. Saki is a traditional Fulani hand woven material worn as everyday cloth by the Fulani Clan of West and Central Africa. This research aims to recognize Saki Pattern in woven fabrics using neuro-fuzzy system. A total of 600 images from four (4) different samples of Saki are collected and pre-processed to extract relevant features. The images are trained using Backpropagation algorithm (BP) Neural Network in Matlab environment. Fuzzy inference rules are then used for classification. The experimental results obtained showed that all four (4) Saki samples were predicted accurately with an average of 80% similarity. Thus, providing a lot of information on Saki which will help preserve the Fulani cultural heritage and boost the Saki textile industry.

**Keywords:** Fuzzy inference, Image processing, Neural network, Pattern recognition, Saki



## 1. INTRODUCTION

Over the years, Machine learning has gained a lot of significance in the IT world which led to the production of many intelligent systems (Kasabov 1996). The need for pattern recognition is in every field and cannot be over-estimated; the medical breakthrough of heartbeat pattern recognition systems, the security importance of face recognition in surveillance, the relevance of image processing and classification etc. (Li 2002). Pattern recognition in image processing is eminent as systems are trained to be intelligent; to learn and master a certain pattern be it a face, picture, finger print or fabric etc. and then recognize it through a series of tests. Woven fabrics have certain unique patterns and are numerous in the African region as all most every tribe has its own unique locally woven cloth. *Saki* is one example of such cloth used by the Fulani Clan of West and Central Africa. It is a thick cloth made from cotton and embroidered with colorful thread designs differing according to sex, region or intent of use. Figure 1 and Figure 2 show a picture of *Saki* cloth used in Nigeria.



Figure 1: Saki Sample A



Figure 2: Saki Sample B

Many Strategies were suggested by different individuals in achieving accurate detection and recognition of different fabric patterns, most employ the use of Neural Networks, Gray level Co-occurrence Matrix (GLCM), Fast-Fourier transform, Data Mining etc. These methods might have work best for some fabrics and not so good for others however, there is a dart in literature regarding recognition of *Saki* pattern. Hence the need for this system that aims at detecting *Saki* pattern in woven fabrics using neuro-fuzzy approach with the aim at providing a lot of information about the fabric.

## 2. RELATED WORKS

There are many works on Fabric pattern recognition, the distinction is based on the type and nature of pattern and mechanism for recognising the pattern.

A search system proposed by (Nara 1994), used Genetic algorithms and a Neural Network. The representation of entries in the memory is distributed ("an auto associative neural network") and the problem was to find an attractor under a given access information where the uniqueness or even existence of a solution was not always guaranteed (an ill-posed problem). Search takes a lot of time before results are obtained. Other works on fabric pattern recognition focuses on determining the nature of the wefts and wraps in a given cloth (Jing 2012) (Jain&Duin 2004) (Kang et al. 2013). A system proposed by (Kumar 2003), presented a new approach for the segmentation of local textile defects using feed-forward neural network. Every fabric defect alters the grey-level arrangement of neighbouring pixels, and this change was used to segment the defects. However, Small sized defects in low-resolution images cannot be detected by this method. Other works on defect detection of hand woven fabrics are (Dobrea & Blaga 2007) (Patel et al. 2013).

Few systems have considered recognition of African Fabric patterns; (Olawale & Ajayi 2013) developed a Model for African Fabric Analysis Recognition. The system analyse and distinguish traditional African fabric pattern (*Adire*, *Ewe Aso oke*, *Amafu*). The model was a valuable tool in image retrieval systems. Selected African fabric patterns were analysed using image processing and wavelet analysis techniques to extract relevant features for the recognition purpose. A pattern recognition system for Nigerian Fabrics developed by (Babatunde et al. 2013), acquired and analysed Nigeria fabric patterns (*Adire*, *Aso oke*). The system provided a recognition model for the patterns and implemented a prototype of the model for mobile devices. This was with a view to addressing the problem of information failure constraining the development of commerce and business in Nigeria. Collected Nigeria fabric patterns were analysed using image processing and wavelet analysis techniques to extract the relevant features for recognition purposes. The system did not truly capture all Nigerian Fabrics as it is limited to recognising very few fabrics amidst of the diverse nature of the Nigerian populace. Due to the dart in literature on *Saki* Fabric, there is the need for this new system that takes as input images of *Saki*, pre-process the images, train and characterize *Saki* pattern.

### 3. METHODOLOGY

A system description of the proposed system is given in Figure 3.

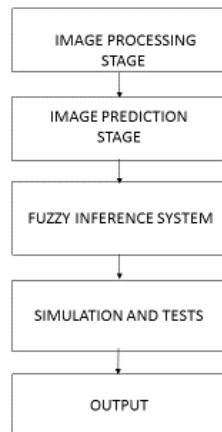


Figure 3: System block diagram

#### 3.1 IMAGE PROCESSING

Image processing is a method to convert an image into digital form and perform some operations on it, in order to get an enhanced image or to extract some useful information from it (Baguda. 2014). The image of different Saki fabric was acquired from local markets dealing in sales of traditional artefacts and clothing in northern Nigeria. The images were captured using a Sony Cyber-Shot DSC-TX1 camera with high resolution and back illuminated sensor enabling more light to be collected in producing sharp images. The image samples were read and displayed in Matlab 2014a environment for processing where the captured images are being read, turned to gray scale and histogram of each image is obtained to calculate threshold value of the grayscale image. Then images are being compared and Structural Similarity Index SSIM of each image is obtained A SSIM map is generated showing the level of similarity between images after rotation. A SSIM value of 0 depicts no similarity at all and that of 1 shows the two images compared are alike and the same. A mathematical model of the SSIM/angle relationship is derived from the SSIM/Angle plot to aid in understanding the behaviour of the relationship. The image processing stage determines the characteristics of Saki from the

sample images and sends the information to the Neural Network.

#### 3.2 IMAGE PREDICTION USING ARTIFICIAL NEURAL NETWORK

This is the process of training an Artificial Neural Network to recognize and classify an image (fed to it during learning) from a sequence of images (Serdaroglu et al. n.d.). Artificial neural networks are composed of many simple elements operating in parallel. These elements are inspired by biological nervous systems. The neural network consists of many different artificial neurons. The training is to be carried out using a multilayer back propagation algorithm to the minimize error function. This is because the BP algorithm is a gradient descent algorithm which aims at finding the local minima/maxima by iteratively moving towards the negative of the slope of the function to be maximized or minimized. The sample data is divided into training data and test data, with 140 sample used for training and 10 samples for testing. The activation function used is the Unipolar Sigmoid function as it is semi linear, differentiable and it produce a value between 0 and 1 which enables the BP algorithm to adapt to lower layers of weights in the Neural Network. The value of the activation parameter is (0.5). Maximum number of epochs is 300 with a learning rate of  $\eta$  (0.001) for better convergence.

#### 3.3 FUZZY INFERENCE

Fuzzy logic (FL) is a problem solving control system methodology connected to the degree to which events occur rather than the likelihood of their occurrence. It is way of making machines more intelligent enabling them to reason like human (Baguda S 2014). A set of fuzzy rules defines the characterisation process, these rules are based on a given range of Average SSIM values.

Average SSIM (Structural Similarity Index) is a similarity comparison between predicted image and actual image as we saw earlier in this chapter, with a scale between 0 and 1, a result of 0 depicting no similarity at all and 1 depicting strong similarity. This system is aimed and reducing time taken for training and faster convergence. The groupings are as follows;

- Group1 Not Similar: 0.0 – 0.2**
- Group2 Fairly Similar 0.2 – 0.4**
- Group 3 Similar: 0.4 – 0.6**
- Group 4 Very Similar: 0.6 – 0.8**
- Group 5 Most Similar: 0.8 – 1.0**

4. RESULTS

The Neuro-Fuzzy characterization was carried out at the training stage, where image sets were tested compared and characterized based on a set of predefined rules. This reduces the number of iterations as the first set of image is trained and tested and compared to the second image if results are the same the images are classified in the same group else a third image is compared to the two previous images and results obtained determine the group of the image. This process leads to fast convergence and reduces the number of epoch as an image does not have to be blindly tested for all the available groups. Figure 4 shows a graph of the RMS (Root Mean Square) error against the number of epoch.

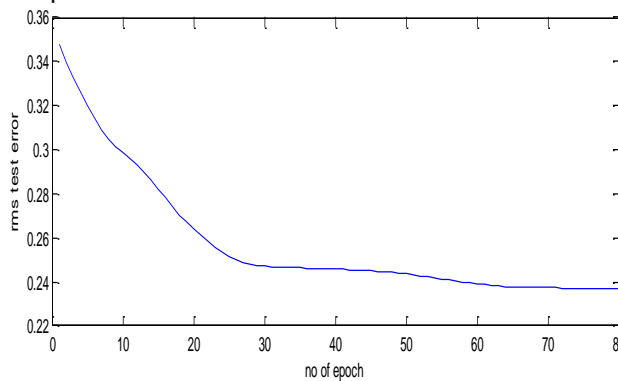


Figure 4: RMS error / Epoch graph

Figure 5 and 6 shows the actual image used in training and predicted image obtained respectively.



Figure 5: Original Image (RGB mode)

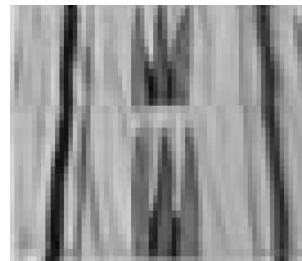


Figure 6: predicted image (Gray mode)

5. CONCLUSION

In this research we describe the application of Neural Networks, Fuzzy inference and image prediction in characterizing *Saki* patterns in woven fabrics. The results obtained for the four different types of *Saki* were accurate and acceptable with a success rate of more than 80%. Reduced dimensionality and quality of the image aid in fast and accurate prediction. The similarity index obtained are good in almost all cases. The accuracy is high in gray level mode of *Saki* images. The Neuro-Fuzzy characterization classifies all the images at once in the training process given a class range which makes it almost impossible to have varying results. Results were obtained earlier usually below the 100<sup>th</sup> epoch with some as low as 50 epoch. This makes the Neuro-Fuzzy technique to be faster, accurate and more efficient. Thus Neuro-fuzzy system proved to be an effective tool for characterization.

6. REFERENCES

Babatunde, J. et al., 2013. Development Of Pattern Recognition System For Nigeria Fabric. *International Journal of Scientific & Engineering Research*, Volume 4, Issue 11, November-2013 ISSN 2229-551 , 4(11), pp.1405–1424.

Bushman, F. et al., 1996. *Pattern-Oriented Software Architecture: A System of Patterns*. John ‘Wiley & Sons Ltd. Bans Lane. Chi Chester.

West Sussex PO19 1UD. England.

Dobrea, D. & Blaga, M., 2007. Genetic Algorithm For Textile Pattern Recognition. *ATC'06 – 34th Aachen Textile Conference*, November 29-30, 2007.

Garage, E., Introduction to Image Processing. Available at: <http://www.engineersgarage.com/articles/image-processing-tutorial-applications> [Accessed April 23, 2016].

Jing, J. et al., 2012. Automatic Recognition of Woven Fabric by Using Gray Level Co-occurrence Matrix 2-D Wavelet Transform Gray Level Co-occurrence Matrix. *journal of information & computational Science* 9:11, pp.3181–3188.

Kang, X., Wang, J. & Jing, J., 2013. The Recognition of Woven Fabric Based on 2-D Wavelet Transform and. *Advances in information sciences and service sciences (AISS)*, 5, pp.984–993.

Kasabov, N.K., *Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering*, The MIT Press. Cambridge, Massachusetts London, England.

Kumar, A., 2003. Neural network based detection of local textile defects. , 36, pp.1645–1659.

Kang X., et al., 2015. Automatic Classification of Woven Fabric Structure Based on Computer Vision Techniques \* 2-D Wavelet Transform. *Journal of Fiber Bioengineering and informatics* 8:1 (2015) 69–79

Li Y., Hu, C. & Yao, X. Innovative batik design with an interactive evolutionary art system. *Journal of Computer Science and Technology* 24:6 1035-1047

Li R, et al., 2002. A fuzzy neural network for pattern classification and feature selection. *Elsevier Science B.V.* 130, pp101-108.

Nayak M., et al., 2013. Pattern Classification Using Neuro Fuzzy and Support Vector Machine (SVM) - A Comparative Study. *International journal of Advanced Research in computer and Communication Engineering* 2(5) 2301-2306.

Nara, S., & Wolfgang Banzhaf., 1994. Pattern Search Using Genetic Algorithms and a Neural Network Model. *Complex Systems* 8 (1994) 295- 309

Olawale, J.B. & Ajayi, A., 2013. A Model for African Fabrics Analysis and Recognition. *International Journal of Computer Applications* (0975-8887), 81(15), pp.38–43.

Patel, J., Jain, M. & Dutta, P., 2013. Detection of Faults Using Digital Image Processing Technique. *Asian Journal of Engineering and Applied Technology*, 2(1), pp.36–39. Available at: [www.trp.org.in](http://www.trp.org.in).

Serdaroglu, A., Ertuzun, A. & A, E., *Defect Detection in Textile Fabric Images Using Subband Domian Subspace Analysis*.

---

## Full Paper

# A HYBRID DIMENSIONALITY REDUCTION MODEL FOR CLASSIFICATION OF MICROARRAY DATASET

---

**M.O. Arowolo**

Department of Computer Science, College of Information and Communication technology,  
Kwara State University,  
Malete, Nigeria  
olliray2002@yahoo.com

**S.O. Abdulsalam**

Department of Computer Science, College of Information and Communication Technology,  
Kwara State University,  
Malete, Nigeria  
abdulsalamny@gmail.com

**R.M. Isiaka**

Department of Computer Science, College of Information and Communication technology,  
Kwara State University,  
Malete, Nigeria  
imabdulrafiu@yahoo.com

**K. Gbolagade**

Department of Computer Science, College of Information and Communication technology,  
Kwara State University,  
Malete, Nigeria  
Kazeem.gbolagade@kwasu.edu.ng

**ABSTRACT**

In this paper, a combination of dimensionality reduction technique, to address the problems of highly correlated data and selection of significant variables out of set of features, by assessing important and significant dimensionality reduction techniques contributing to efficient classification of genes is proposed. One-Way-ANOVA is employed for feature selection to obtain an optimal number of genes, Principal Component Analysis (PCA) as well as Partial Least Squares (PLS) are employed as feature extraction methods separately, to reduce the selected features from microarray dataset. An experimental result on colon cancer dataset uses Support Vector Machine (SVM) as a classification method. Combining feature selection and feature extraction into a generalized model, a robust and efficient dimensional space is obtained. In this approach, redundant and irrelevant features are removed at each step; classification presents an efficient performance of accuracy of about 98% over the state of art.

**Keywords:** Dimensionality Reduction, Feature Selection, Feature Extraction, Classification

## 1. INTRODUCTION

Recently, developments in data possession ability, data compression and improvement of database as well as data warehousing knowledge have shown the way to the emergence of high dimensional dataset. Data are often irrelevant and redundant, giving rise to increase in the search space size and giving rise to complexity of processing the data. This curse of high dimensionality is a key problem in machine learning. Hence dimensionality reduction is a dynamic research area in the area of microarray gene analysis, machine learning, data mining and statistics (Veerabhadrapa, and Lalitha, 2010).

Dimensionality reduction as a pre-processing approach, it helps in removing redundant or irrelevant features from high dimension microarray dataset. Diagnosing and classifying cancer diseases based on innovative gene expression information is a challenge (Austin, Chia, and Chih, 2013), microarray-based gene expression has become a realistic method in the prediction of classification and prognosis outcomes of diseases (Chen, and Lee, 2011).

Several algorithms have been proposed for dimensionality reduction in literature, this study develops a hybrid dimensionality reduction model in order to overcome the drawback of very high dimensional data, by imploring feature selection algorithm (One-Way-ANOVA) on colon cancer dataset to identify the most relevant features for classification (Jazzar, and Muhammed, 2013), (Shen, Diao, and Su, 2011), (Han, and Kamber, 2006), it reduces the dataset from 2001 to 416 attributes. The 416 attributes are passed to feature extraction algorithm using PCA and PLS separately to project the reduced datasets, into a low-dimensional space and create a unique dimension for the interaction of the dataset, it achieved 10 components and 20 components respectively. It trims down the data to a significant relevant quantity, and the complexity on supplementary processing, in other to enhance learning accuracy, and get a better result clarity. Classification is carried out using Support Vector Machine (SVM) and accuracy of PLS based reaches 98% and outperforms PCA based approach.

## 2. RELATED WORKS

Several studies have been proposed for dimensionality reduction of microarray data, using different techniques.

Zhang, and Deng, 2007, applied a reduced dataset of genes by carrying out gene pre-selection with a univariate principle task, and subsequently estimated the upper bound of errors in Bayes, to sort unneeded genes resulting from pre-selection step. To prove their scheme K-Nearest Neighbor (KNN) and SVM classifiers were used on five datasets (Zhang, and Deng, 2007).

Abeer, and Basma, 2014, worked the innovation of Differentially Expressed Genes (DEGs) in microarray data, by building a precise and cost efficient classifier. T-Test feature selection method with KNN classifier was used on Lymphoma dataset to build the DEGs, the classifier accuracy was recorded.

Vaidya and Kulkarni, 2014 proposed cluster elimination and dimension reduction for cancer classification. ANOVA, PCA, Recursive Cluster Elimination (RCE) as a classification algorithm were implored by employing an innovative gene selection method. It reduces gene expression data into minimal number of gene subset.

Zena and Dunca, 2015, proposed variety forms of dimensionality reduction approach on a high-dimensional microarray data. Several feature selection and feature extraction processes seeking to eliminate redundant and irrelevant features, for innovative instances of classification, which can be accurate, were established.

Song and Sejong, 2016, proposed a feature selection based dimensionality reduction model on microarray biological data to improve the pre-evaluation information, the seature selection subset improved the data by using the top ranking pairs for features in the selection process, KNN and SVM were used in testing the classification results.

Yang, et al, 2017 proposed and implemented a dimension reduction method for microarray

datasets, using k-means clustering algorithms Laplacian eigen map and isomap methods were used to achieve clustering accuracy, the overall result showed that Laplacian eigenmap provides an improvement performance of redundancy than isomap.

Research is still on for innovative techniques to select unique features for classification improvement. According to previous studies, there are several limitations as regards to the problem of developing efficient and effective classification models. This paper combines Feature selection using one-way ANOVA with feature extraction using PCA and PLS separately to enhance classification using SVM.

### 3. METHODOLOGY

Different methods have been proposed using several dimensionality reduction algorithms. The main objective of dimension reduction is to enhance and improve the efficiency of the data and discover strong relationship among gene expression by finding the hidden original data.

#### A. Hybrid Approach

The propose system consists of a hybrid module; Feature selection and feature extraction for Classification to evaluate the performance of classification. Colon cancer dataset (Alon, Barkai, Notterman, Gish, Ybarra, Mark and Levine, 2001) loads as an input to the feature selection algorithm (One-Way ANOVA) algorithm. The input from the feature selection module is passed to the feature extraction and classified by using the SVM and the result is displayed in the result module.

The methodology used in this paper is as follows:

- Employ one-way ANOVA feature selection with PCA features extraction and SVM.
- Use one-way ANOVA feature selection with PLS feature extraction and SVM classification.

Compare the performances in terms of; accuracy, sensitivity, specificity, precision and time, to improve classification performance.

Figure 1 shows a framework of dimension reduction for analyzing microarray data in this study. The colon cancer dataset (Alon, et al,

2001) is passed into One-Way-ANOVA algorithm as a pre-processing process; the result is passed into the PCA and PLS as feature extraction algorithms, in other to fetch for the latent components before classifying using SVM model.

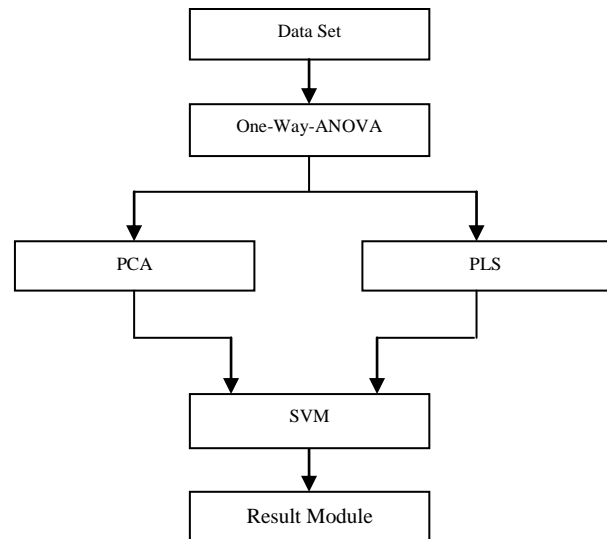


Figure 1: Technique Workflow

#### B. Experimental Dataset and Setup

A public available Colon cancer dataset by (Alon, et al, 2001) was used for the experiment.

The system configuration used for the paper; Processor: iCore 2, RAM size: 4GB, Speed: 2.13 GHz, System: 64-bit, Operating System: Windows 8 Pro, and Implementation Tool: MATLAB R2015a

#### C. Feature Selection

In step 1, the gene dataset is computed with the use of One-Way-Analysis of Variance (ANOVA) as a feature selection method, it is a frequently used approach analyzing data and drawing interesting information based on P-Value, it is a robust technique; it presume all sample of a data to be distributed in general, having equal variance and independent (Vaidya, and Kulkarni, 2014). The motivation is to carry out a one-way ANOVA feature by ranking the significant features through small values as 0.05 p-values

and the classed numbers of features are further processed for selection of responsive data. The following features apply:

$N_j$  = The number of cases with  $Y = j$

$X_j$  = The sample mean of predictor  $X$  for target class  $Y = j$

$S_j^2$  = The sample variance of predictor  $X$  for target class  $Y = j$ :

$$S_j^2 = \sum_{i=1}^{N_j} (X_{ij} - \frac{X_j}{N_j-1})^2 \quad (1)$$

$\bar{X}$  : The grand mean of predictor  $X$ :

$$\bar{X} = \frac{\sum_{j=1}^J N_j X_j}{N} \quad (2)$$

The notations above are base on non-missing pairs of the sample and attribute of the dataset used in terms of  $(X, Y)$ .

Calculating the p-value; Prob  $\{F (J-1, N-J) > F\}$ :

Where,

$$F = \frac{\sum_{j=1}^J N_j (x_j - \bar{x})^2 / (J-1)}{\sum_{j=1}^J (N_j - 1) x_j^2 / (N-1)} \quad (3)$$

$F (J-1, N-1)$  is an indiscriminate variable which works with an F distribution with level of freedom  $J-1$  and  $N-J$ . When denominator for a predictor is zero, position the p-value as 0.5. Predictor is classed by sorting the p-value in ascending order. If there is tie, sort F in descending order and if it still ties, sort N in descending order. Classification of features shows that 416 features were the most significant features correlated to the microarray data analysis out of 2001 features.

#### D. Feature Extraction

In step 2, the feature extraction module uses PCA and PLS separately on the microarray colon cancer datasets after passing through the One-Way-ANOVA Feature selection to paper the variation of efficiency performance.

Partial Least Square (PLS) is a procedure in modeling associations linking large piece of experimental variables using latent variable, it finds uncorrelated linear modification (latent components) of the selected predictor variables which comprises of response variables of high covariance (Nadir, Othman, and Ahmed, 2014). PLS fetches the linear connection linking the response and descriptive variables  $y$  and  $X$ :

$$X = TP^T + E_x \quad (4)$$

$$y = TC^T + E_y \quad (5)$$

$T$  signifies the scores (latent variables)  $P$  and  $C$  are loadings, and  $E_x$  and  $E_y$  are the outstanding matrices achieve by the original  $X$  and  $y$  variables.

Principal Component Analysis is suitable when there are measures achieved on a number of observed values; it is a procedure to determine the key variables in a multidimensional dataset explaining variations in the observations. It is very helpful for analysis visualization and simplification of high dimensional datasets (Nadir, Othman, and Ahmed, 2014). The general formula for calculating the score of weight of extracted components is;

$$C_1 = b_{11}(X_1) + b_{12}(X_2) + \dots + b_{1p}(X_p) \quad (6)$$

Where;

$C_1$  = the principal component 1 on subject's score (the first component extracted)

$b_{1p}$  = the regression coefficient (or weight) for experimental variable  $p$ , used in generating principal component 1

$X_p$  = the subject's score on experimental variable  
Feature selection preserves data characteristics for interpretability, but discriminates power, with a lower and shorter training time as well as reducing overfitting. While, feature extraction has higher discriminating power and it controls overfitting when it is unsupervised, but losses data interpretability and also its transformation may be expensive.

#### E. Classification

In step 3, the results for classification were computed using Support Vector Machine (SVM). SVM is a statistical knowledge theory for learning constructive procedure (Xue-Qiang, and Guo-Zheng, 2014), it is used for classification tasks, and it uses linear models in implementing non-linear class boundaries by transforming input space using a non-linear mapping into a new space. SVM produces an accurate classifier with less over fitting and it is robust to noise.

Assuming  $\{(x_1, y_1), \dots, (x_n, y_n)\}$  be a training set with  $x_{1i} \in R^d$  and  $y_i$  is the corresponding target class. SVM can be reformulated as:



Maximize:

$$J = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j (x_i^T, x_j) \quad (7)$$

Subject to;

$$\sum_{i=1}^n \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0, i = 1, 2, \dots, n \quad (8)$$

A weighted average of the training features.  $\alpha_i$  is an optimization task of Lagrange multiplier and  $\alpha_i$  is a rank label.  $\alpha_i$ 's are non zero values for every points in the margin and on the accurate plane of the classifier.

#### 4. RESULTS AND DISCUSSION

The performances of the proposed methods are studied, in which the dataset is restructured after the evaluation of each steps. The details of restructured dataset are shown in Table I

TABLE I: Result Evaluations

Dataset	Features Selected (one-way ANOVA)	Feature Extracted (PCA)	Feature Extracted (PLS)
Colon Cancer (2001x62)	416	10 Component 5	20 Component 5

In the paper, the methods used obtained reduced results shown in Table I above, it presents confusion matrices for the paper in terms of accuracy, specificity, prediction, training time and error for justification, which are illustrated to determine the performance in the figures below.

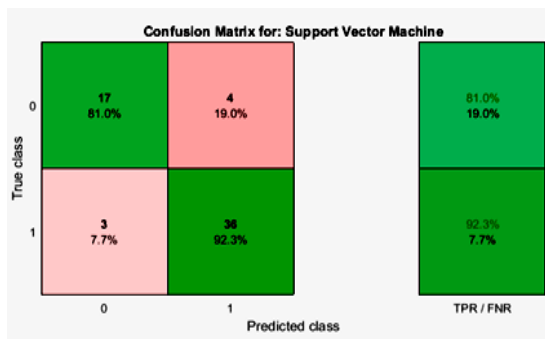


Figure 2: Confusion Matrix of Proposed Classification, using One-Way-ANOVA-PCA-Based Classification

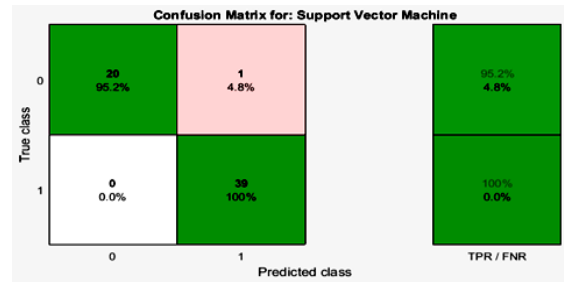


Figure 3: Confusion Matrix of Proposed Classification, using One-Way-ANOVA-PLS-Based Classification

Figure 2 and Figure 3 demonstrates the confusion matrices of the proposed paper, One-Way-ANOVA-PLS-Based and One-Way-ANOVA-PCA-Based methods achieved for the paper. The performance metrics are illustrated based on the confusion matrices and the reliability of the performances is discussed. The adopted terms are defined below (Zainal, 2009):

$$\text{Sensitivity} = TP / (TP + FN) \% \quad (9)$$

$$\text{Specificity} = TN / (TN + FP) \% \quad (10)$$

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \% \quad (11)$$

$$\text{Precision} = TP / (TP + FP) \% \quad (12)$$

Colon (Alon, et al, 2001) developed a classification method based on microarray dataset gene expression. To evaluate the performance of the proposed approach, one-way ANOVA uses p-value 0.05 as the hold out validation procedure to select relevant features. Feature extraction (PCA and PLS) methods that reduces dimensionality of the preprocessed data were compared using the SVM classification method.

TABLE 2: PERFORMANCE MEASURES OF PROPOSED, PCA AND PLS BASED METHOD.

S/N	Performance Metrics	PCA- Based Method	PLS-Based Method
1	Training Time	2.973	0.28958
2	Accuracy (%)	83.33	98.33
3	Sensitivity (%)	92.31	100
4	Specificity (%)	80.95	95.60
5	Precision (%)	90	97.5
6	Area Under Curve	0.893773	1
7	Error (%)	11.7	1.67

**Table 2:** illustrates a comparative chart between the two methods used in terms of performance measures. The One-Way-ANOVA-PLS-Based method achieves necessary higher value in the dataset when compared to the One-Way-ANOVA-PCA method.

## 5. CONCLUSION

This paper studied the performance of dimensionality reduction in microarray gene classification technique, using Colon Cancer datasets. The learning gritty on classification performance measures such as training time, accuracy, sensitivity, specificity, prediction, Receiver Operating Curve and Overall Error. PLS Based method showed a better performance than PCA-based method with 98.33% to 83.33% accuracy. Hence PLS based dimensionality reduction scheme is suitable for microarray gene classification as it extracts relevant and a reduced amount of information from the feature selection based technique. In future studies PLS can be compared with another feature extraction method with the aforementioned criteria. Another dataset will be a good avenue for further research of dimensionality reduction.

## 7. REFERENCES

Veerabhadrapa, P., and Lalitha, R., 2010. Bi-level dimensionality reduction methods using feature selection and feature extraction. *IJCA* vol. 4, pp. 33-38.

Austin, H.C., Chia, H.L., and Chih, H.C., 2013. New Approaches to Improve the Performance of Disease Classification Using Nested-Random Forest and Nested-Support Vector Machine Classifiers. *RNIS*. Vol. 14. pp.105.

Chen, A.H., and Lee, M., 2011. Novel Approaches for the Prediction of Cancer Classification. *IJACT*, vol. 3, pp. 30-39.

Jazzar, M.M., and Muhammad, G., 2013. Feature Selection Based Verification /Identification System Using Fingerprints and Palm Print. *Arabian Journal for Science and Engineering*. Vol. 38, pp.849-857.

Shen, Q., Diao, R., and Su, P., 2011. Feature Selection Ensemble. In: proceedings of Computing. Springer-Verlag, pp. 289-306.

Han, J.W., and Kamber, M., 2006. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers.

Zhang, J. and Deng, H., 2007. Gene selection for classification of microarray data based on the Bayes error. *BMC Bioinformatics*, Vol. 8, No. 1, pp. 370.

Abeer, M., and Basma, A., 2014. A Hybrid Reduction Approach for Enhancing Cancer Classification of Microarray Data. *IJARAI*, Vol. 3.

Zena, M and Duncan, F.G., 2015, A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data.

Vaidya, M., and Kulkarni, P.S., 2014. Innovative Technique for Gene Selection in Microarray Based on Recursive Cluster Elimination and Dimension Reduction for Cancer Classification. *IJIRAE*, pp.209-213.

Alon, U., Barkai, N., Notterman, D.A., Gish, K. Ybarra, S., Mack, D., and Levine, A.J., 2001. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceeding. Nat. Acad. Sci. USA*, Vol. 96, pp. 6745-6750.

Nadir, O.E., Othman, I., and Ahmed, H.O., 2014. A Novel Feature Selection Based on One-Way ANOVA F-Test for E-Mail Spam Classification. *Research Journal of Applied Sciences, Engineering and Technology*, Vol. 7, No. 3, pp. 625-638.

Xue-Qiang, Z., and Guo-Zheng, L., 2014. Dimension Reduction for p53 Protein

- Recognition by Using Incremental Partial Least Squares. IEEE, Vol. 13, No. 4, pp. 73-79.
- Yang, J., Wang, H., Ding, H., Ning, A., and Gil, A., 2017. Nonlinear Dimensionality Reduction for Synthetic Biology Biobrick Visualization. BMC Bioinformatics, Vol. 17, No. 4, pp. 1-10.
- Songlu, L., and Sejong, O., 2016. Improving Feature Selection Performance Using pairwise Pre-evaluation. BMC Bioinformatics, Vol. 17, pp. 1-13
- Zainal, A., 2009. Relation between eye movement and fatigue: Classification of morning and afternoon measurement based on Fuzzy rule. International Conference on Instrumentation Communication Information Technology and Biomedical Engineering, pp. 1-6.

## Full Paper

**PROFIT MAXIMIZATION FOR A PRODUCT-MIX PROBLEM IN  
SMALL AND MEDIUM ENTERPRISE****Ademola O. Adesina**

Olabisi Onabanjo University (OOU),  
Ago-Iwoye, Nigeria  
inadesina@gmail.com

**David O. Iyanda**

Department of Mathematics  
National Open University Nigeria (NOUN),  
Ibadan, Nigeria  
expertsolutionconcepts@gmail.com

**ABSTRACT**

Operations research or mathematical optimization is the selection of a best element (with regard to some criteria) from some sets of available alternatives. An optimization problem consists of maximizing or minimizing a real function by systematically choosing input values from within an allowed sets and computing the value of the function. Linear programming (LP) is a technique for determining the optimum schedule of the interdependent activities in view of the available resources as it applies to the optimization models in which objectives and constraint functions are strictly linear. It is therefore imperative to apply optimization model in the cost of production and resource management in order to manage the use of raw materials and resources for effective optimal production which will result into maximum profit or low cost of production. It may be observed that optimization concept is rarely applied in the small and medium enterprise; in this era where many government policies encourage entrepreneurship. This study thus embarks on the method of finding and arriving at the optimal product-mix of an indigenous manufacturing company (ESC-Souvenir) that is engaged in leather production. The challenge encountered in getting the right proportion of raw materials needed during production was modelled as linear programming problem. The result shows that for the firm to obtain an optimal percentage production and an average monthly profit of 20%, thirty-nine thousand, six-hundred naira (N39,600 i.e. USD 396) will be needed for the production and supply of Types A, B and C custom-designed bags. The firm must vary the input values as a combination of thirty (30) of Type B and ninety (90) of Type C for the three products under consideration. The result also revealed that (for an optimal production level) the raw materials, leather/lining, parking foam and leather tape must be 54.4%, 25% and 32.5% in appropriate proportion respectively.

**Keywords:** Linear programming, Mathematical optimization, Operations research

## 1. INTRODUCTION

Planning is required on various occasions in day-to-day activities, especially when the resources meant for attaining certain objectives are limited in supply. The best strategy is the one that gives a maximum output from a minimum input. The objective (which is in form of output) may be to get the maximum profit, minimum cost of production or minimum inventory cost with a limited input of raw materials, manpower, and machine capacity. Such problems are referred to as problems of constrained optimizations. *Programming* is another word for 'planning' which refers to the process of determining a particular path of action from amongst several alternatives i.e. taking decisions systematically (Soni, 1996). In subject areas like Mathematics (Mathematics is considered as the queen of Sciences and number theory as the queen of Mathematics (Gray, 2008, Gray, 2006, von Waltershausen, 1856), Computer science and Operations research, Mathematical optimization (alternatively, optimization or mathematical programming) is the selection of a best element (with regard to some criteria) from some sets of available alternatives. In the simplest case, an optimization problem consists of maximizing or minimizing a real function by systematically choosing input values from within an allowed sets and computing the value of the function. The generalization of optimization theory and techniques to other formulations comprises of a large area of Applied Mathematics. More generally, optimization includes finding *best available* values of some objective functions given a defined domain (or a set of constraints), including a variety of different types of objective functions and domains.

Linear programming (LP) is a mathematical method for determining a way to achieve the best outcome (such as maximum profit or lowest production cost) in a given mathematical model for some list of requirements represented as linear equations. The word *linear* stands for indication that all relationships involved in a particular problem are of degree one, that is, the relationships handled are those represented by straight lines, i.e. the relationships are of the form  $y = a + bx$ . Thus, linear programming is a decision making technique under given constraints and based on the assumption that the relationships amongst the variables representing different phenomena happen to be linear. LP is therefore a technique for determining an optimum schedule of interdependent activities in view of the available resources. LP applies to

optimization models in which objectives and constraint functions are strictly linear.

Linear programming (LP) is a powerful analytical tool that can be used to determine an optimal solution that satisfies the constraints and requirements of the current situation (Better, 1988, Afful et al., 2016). A linear programming problem consists of three parts: first, there is an objective function which is to be either maximized or minimized; second, there is a set of linear constraints which contains the technical specifications of the problems in relation to the given resources or requirements; and third, there is a set of non-negativity constraints since negative production has no physical counterparts (Chinneck, 2004, Lim et al., 2014). In formulating the linear programming problem, the assumption is that a series of linear (or approximately linear) relationships involving the decision variables exist over the range of alternatives being considered in the problem (Chinneck, 2004, Lim et al., 2014, Horst, 2013). The obligation to meet infinite needs with restricted resources is one of the biggest challenges encountered in the market today Ozsan, Simsir, & Pamukcu, (2010). The LP output not only provides an optimal solution, it also provides sensitivity analysis. Sensitivity analysis evaluates how changes in the objective function coefficients affect the optimal solution of a linear programming model. It could examine how well the changes of objective function coefficients and the right hand side values could affect the optimal solution (Anderson et al., 2016). More formally, linear programming is a technique for the optimization of a linear objective function, subject to linear equality and linear inequality constraints. Given a polytope and a real-valued affine function defined on this polytope, a linear programming method will find a point on the polytope where this function has the smallest (or largest) value if such point exists, by searching through the polytope vertices.

This study demonstrates the use of computational linear programming technique for solving product-mix problem of three parts (objective function, linear constraints and non-negativity constraints), using computational and analytical approaches. These approaches in combination with the use of *Excel Solver* are applied to solve the production problem in *ESC-Souvenir* firm so as to determine the product-mix supply for three (3) products (bags).

## 2. LITERATURE REVIEW

Linear programming (LP) technique is used in a wide range of applications, including Agriculture, Banking, Industry, Transportation, Economics, Health systems, Behavioural and Social sciences and Military. It also boasts efficient computational algorithms for problems with thousands of constraints and variables. LP is a major innovation since World War II in the field of Business, decision making, particularly under conditions of certainty. Schrijver (1998) confirmed the approach of LP formulation of a problem and its corresponding solution using general linear programming. This approach was developed during the World War II as a method used in solving the plan expenditures and returns. LP approach made again to the army inventory costs i.e. reduced army cost but increases the losses in the enemy's camp. Indeed, because of its tremendous computational efficiency, linear programming forms a backbone of the solution algorithm for other operative research models, including integer, stochastic and non-linear programming. The solution to solving a system of linear inequalities dates back to 1827 when Fourier published a method for solving them, (Sierksma, 2001, Balinski and Tucker, 1969), and after whom the method of Fourier–Motzkin elimination is named.

LP was introduced to optimise the crude blending and refining operations (Hassan et al., 2011). This operation includes crude evaluation, selection, and scheduling and product logistics planning. The objective of the study considered developing a mathematical programming for solving a blending problem of refinery and maximizing *Naphtha* production. A LP software package was used as a black box model by the users for the refinery planning and optimization. The model developed in that work was proved to be highly effective at the level of solving the blending problem. The study yielded better overall *Naphtha* productivity for the case of the oil refinery studied, as compared to results obtained by the commercial software.

Saudi Public Transport Company (SAPTCO) intercity commercial bus transportation operation engaged 338 busses across 250 cities and villages within a list of 382 major trips per day. The transport company operates Mercedes 404 SHD and Mercedes 404 RI-IL fleet types for the intercity trips. The two brands of Mercedes 404 apply fleet assignment model which was adapted and applied to a sample of the intercity bus schedule. The results showed a substantial saving of 29% in the total number of

needed buses. This encouraged the decision makers at SAPTCO to use only Mercedes 404 SHD fleet type. The fleet assignment model was later modified to incorporate only one fleet type and applied to the same sample. With the increase in the problem size, the fleet assignment model was decomposed by stations. Finally, the modified decomposed model was applied to the whole schedule. The model results showed a saving of 16.5% in the total number of needed buses of Mercedes 404 SHD. Further study using a predefined minimum connection time for model efficiency was considered in connection with time modification for 11 stations and it resulted into saving about 14 buses. With these observations and recommendations, 27.4% (90) buses was the expected saving of the total number of the needed buses. A net saving of 16.44 million Saudi Riyals (USD 4.4million) was yielded annually by SAPTCO in addition to contributing to the growth of the air traffic and better coordination of hiring new employees. The revenue analysis showed that these 90 surplus buses would yield about USD 20,744,000 additional revenue yearly (Hasan and Al Hammad, 2010).

Naifer, Al-Rawahy & Zekri (2011) reported that 112 farmers were divided into three (3) groups based on the soil salinity levels i.e. low, medium and high. The research purpose was to know the extent at which farmers sustain an economically viable agricultural production in the salt-prone areas of Omar. Linear programming was used to maximize each type of farm's gross margin under water, land and labour constraints. The economic losses incurred by farmers due to salinity were estimated by comparing the profitability of the medium and high salinity farms to the low salinity farm's gross margin. The results showed that when salinity increased from low to medium, the salinity level damage was USD1,604 / hectare and USD2,748 / hectare if it increased from medium to high salinity level. Introduction of salt-tolerant crops in the cropping systems showed that the improvement in gross margin was substantial thus attractive enough for medium salinity farmers to adopt the new crops and/or varieties to mitigate the effect of water salinity.

A linear programming technique was employed to determine the most efficient way of combining the locally available ingredients or feedstuffs to formulate least cost rations for broilers. Nutrient requirements of the broilers, nutrient composition

of the available ingredients and any other restriction factor of the available ingredient for the formulation were taken into consideration before Mathematical models were applied to investigate, analyse and formulate the ration. After applying LP, the result of the study for the least costs are grouped into two for starter and finisher rations, as presented in Table 1.

**Table 1: Ingredients to formulate feedstuffs**

Ration	Yellow corn %	Soya bean %	Wheat bran %	Fish meal %	Calcium di-phosphate %	Lysine %	Methionine %	Limestone %	NaCl %	Ready premix %	Soya oil %	vitamins and mineral mix %
Starter	68.0	25.07	4	0.5	0.5	0.1	0.32	0.3	0.3	0.5	0.4	0.01
Finisher	67.5	20.45	5	0.25	1.5	0.25	0.35	0.3	0.5	3	0.75	0.15

Wambugu, Okello, Nyikal, & Shiferaw (2009) and Muthini, Nyikal, & Otieno (2017) researched on financing small-holder farming in Kenya. It has been noticed that huge amount was incurred by government through credit facilities released by the government. The study sought to find the best way to fund smallholder agriculture, it became necessary to analyse and document smallholders' effective demand for credit. The comparison of the existing production plans and production plans under strictly profit maximization were carried out. Linear programming model was used to formalize the observed plans and determine those under profit maximization. The two activities and the values of outputs under different objectives were analysed and compared. Farm Investment Analysis was undertaken to determine the suitability of releasing fund for farm activities through credit. A typical small holder area was undertaken in the selected zones of Muranga and Kisumu districts. Sample farmers were visited and structured questionnaires administered to cover farm events and physical resources of short and long rains for 1995 and 1996 respectively. This formed a basis of formulating the farm plans. Ten years later, the objectives of smallholders have not changed as it was observed during outreach programs. The results were grouped into three: (1) the observed plans and those under profit maximization were different for the farmers' activities; (2) the observed plans had significantly lower profit than those under profit maximization; and (3) meeting constraints through credit was only feasible when the objective was profit maximization. The small holder agriculture, characterized by subsistence production, does not exhibit effective demand for credit and funding. It therefore requires means other than the competitive market.

Mullan (2008) used linear programming to solve the ice cream mix calculations and provided a proof of calculation showing that the mass of the mix sums to the correct value and all the components. Eleven (11) *Excel spreadsheets* were developed as the calculator and these covered many of the ice-cream formulation challenges that commercial manufacturers may encounter. The metabolic pathway and the enzymes properties involved in the Citric acid biosynthesis in the mould *Aspergillusniger* were well known. This fact, together with the availability of new theoretical frameworks aimed at quantitative analyses of control and dynamics in the metabolic systems. A mathematical model of the Carbohydrate metabolism in *Aspergillusniger* under conditions of Citric acid accumulation was introduced. The model makes use of the 'S-system representation' of biochemical systems. This system made the use of linear programming possible to optimize the process. It was found that maintaining the metabolite pools within narrow physiological limits (20% around the basal steady-state level) and allowing the enzyme concentrations to vary within a range of 0.1 to 50 times their basal values. It was possible to triple the *glycolytic flux* while maintaining 100% yield of substrate transformation. To achieve these improvements, it was necessary to modulate seven or more enzymes simultaneously.

A municipal water supply system over a 15-year planning period with initial infrastructure and possibility of construction and expansion during the first and sixth year on the planning horizon was considered (Chung et al., 2009, Naderi and Pishvae, 2017). Correlated uncertainties in water demand and supply were applied on the form of the robust optimization approach of Bertsimas and Sim to design a reliable water supply system. Robust optimization aims to find a solution that remains feasible under data uncertainty. It was found that the robust optimization approach addressed parameter uncertainty without excessively affecting the system. While they applied their methodology to hypothetical conditions, extensions to real-world systems with similar structure were straight forward. Therefore, their study showed that this approach was a useful tool in water supply system design that prevented system failure at a certain level of risk. On the other way, an extensive study of water supply optimization at Jatimlerek covering an irrigation area of 1236 hectares was studied by Hoesein &

Limantara (2010). The irrigation scheme was designed to serve more than one district. LP was introduced as the method for the optimization of the water supply. The results contained were used as the guidance in cropping pattern and allocating water supply for irrigation at the area.

Al-Sitt (2004) developed four (4) models of optimal water allocation with deficit irrigation in order to determine the optimal cropping plan for a variety of scenarios. The first model, Dynamic programming (DP) model allocated a given amount of water optimally over the different growth stages to maximize the yield per hectare for a given crop, accounting for the sensitivity of the crop growth stages to water stress. The second model, Single Crop Model tried to find the best allocation of the available water both in time and space in order to maximize the total expected yield of a given crop. The third model, Multi crop Model was an optimization model that determined the optimal allocation of land and water for different crops. It showed the importance of several factors in producing an optimal cropping plan. The output of the models was prepared in a readable form to the normal user by the fourth model, Irrigation Schedule Model.

A multi-period linear programming model to identify the optimal size of fingerling to under-stock and maximize multi-period returns on a catfish grow-out farm was developed (Bouras and Engle, 2007, Kumar and Engle, 2014). Grow-out production alternatives included under-stocking three different sizes (7.6 cm, 12.7 cm, and 17.8 cm) of fingerlings in multiple-batch production at 15,000 fingerlings per hectare. Fingerlings were produced either with or without thinning at different stocking densities. The results showed that the optimal size of fingerling to under-stock was 12.7 cm. On-farm production of fingerlings was optimal across all farm sizes but the fingerling production technique selected varied with farm size. Models of larger farm sizes indicated that it was optimal to thin fingerlings, while for smaller farm sizes, producing fingerlings without thinning was optimal. When farm size was treated as an endogenous variable in the farmer's profit-maximization decisions, the optimal size of a catfish farm was 404 water-acres. Sensitivity analysis suggested that the net returns were sensitive to changes in the key parameters of the model (such as interest rates, feed conversion ratios, survival rates, catfish prices, harvesting costs, and the availability of operating capital), whereas the optimal size of fingerlings to under-

stock was robust to variations in the model's parameters.

Carlson (1988) and Soares Machado & Gassenferth (2015) presented a practical proposition for the application of the LP quantitative method in order to assist planning and control of customer circuit delivery activities in telecommunications companies working with the corporative market. Based upon data provided for by a telecom company operating in Brazil, the LP method was employed for one of the classical problems of determining the optimum mix of production quantities for a set of five products of that company: Private Telephone Network, Internet Network, Intranet Network, Low Speed Data Network, and High Speed Data Network, in face of several limitations of the productive resources, seeking to maximize the company's monthly revenues. By fitting the production data available into a primary model, observation was made as to what number of monthly activations for each product would be mostly optimized in order to achieve maximum revenues in the company. The final delivery of a complete network was not observed but the delivery of the circuits that made it up, and that was a limiting factor for the study herein, which, however, brought an innovative proposition for the planning of private telecommunications networks.

Matthews (2004) evaluated and optimized the utility of the nurse personnel at the Internal Medicine Outpatient Clinic of Wake Forest University Baptist Medical Centre. Linear programming was employed to determine the effective combination of nurses that would allow for all weekly clinic tasks to be covered while providing the lowest possible cost to the department. A specific sensitivity analysis was performed to assess just how sensitive the outcome was to the stress of adding or deleting a nurse to or from the payroll. The nurse employee cost structure in this study consists of five (5) certified nurse assistants (CNA), three (3) licensed practicing nurses (LPN), and five (5) registered nurses (RN). The LP revealed that the outpatient clinic should staff four (4) RNs, three (3) LPNs, and four (4) CNAs with 95 percent confidence of covering nurses' demand on the floor.

Kuo, Schroeder, Mahaffey, & Bollinger (2003) stated that from the period of December 1, 2000, to July 31, 2002, the following individualized data were obtained for the Division of General Surgery at



Duke University Medical Centre: allocated operation time (hours), case mix as determined by current procedural terminology (CPT) codes, total OR time used and normalized professional charges and receipts. In-patient, Outpatient, and Emergency cases were included. The Solver linear programming routine in *Microsoft Excel (Microsoft Corp.)* was used to determine the optimal mix of surgical OR time allocation to maximize professional receipts. Their model of optimized OR allocation maximized weekly professional revenues at 237,523 USD, a potential increase of 15% over the historical value of 207,700 USD or an annualized increase of approximately 1.5 million USD. Their results suggested that mathematical modelling techniques used in operations research, management science, or decision science might rationally optimize OR allocation to maximize revenue or to minimize costs. These techniques may optimize allocation of scarce resources in the context of the goals specific to individual academic departments of surgery.

### 3. METHODOLOGY

As in the real world, theoretically, maximum profit is pursued by all manufacturers and producers and maximising profit is one way to reach the goal. In this study, Linear programming model of solving product-mix problems is used to model the problem and the method used in solving the problem is Simplex method of solving LP which involves Computation Linear Programming Technique using *Microsoft Excel Solver*. Simplex method has been the standard technique for solving a linear program. It involves simple simplex algorithm, which is an iterative procedure that examines the vertices of the feasible region to determine the optimal value of the objective function. It usually starts at the corner that represents doing nothing. It moves to the neighbouring corner that best improves the solution. It does this over and over again by improving the objective function each time until the optimal solution is found at the most attractive corner.

The method passes from vertex to vertex on the boundary of the feasible polyhedron, repeatedly increasing the objective function until either an optimal solution is found, or it is established that no solution exists. In principle, the time required might be an exponential function of the number of variables, and this can happen in some contrived cases. In practice, however, the method is highly efficient, typically requiring a number of steps

which is just a small multiple of the number of variables (Schrijver, 1998, Tunçel, 2016, Al-Deseit, 2009, Hassan et al., 2011) Linear programs in thousands or even millions of variables are routinely solved using the *simplex* method on modern computers (Kuo et al., 2003). Efficiently, highly sophisticated implementations are available in the form of computer software packages such as *Microsoft Excel Solver* used in the study. This method is applied as an analytical tool to solve ESC-Souvenir product-mix supply problem in order to know right combination of the bags to produce and to know the exact quantities of each material involved in the production of all types of the bags which ultimately results in optimal profit.

The standard problem is formulated as maximum or minimum LP. The standard maximum LP is formulated as follows:

$$\text{Maximize: } z = c_1 x_1 + c_2 x_2 + \dots + c_n x_n$$

Subject to the following constraints:

$$a_{11} x_1 + a_{12} x_2 + \dots + a_{1n} x_n \leq b_1$$

$$a_{21} x_1 + a_{22} x_2 + \dots + a_{2n} x_n \leq b_2$$

$$a_{m1} x_1 + a_{m2} x_2 + \dots + a_{mn} x_n \leq b_m$$

and

$$x_j \geq 0 (j = 1, 2, \dots, n)$$

The standard minimum Linear Programme is formulated as follows:

$$\text{Minimize: } z = c_1 x_1 + c_2 x_2 + \dots + c_n x_n$$

Subject to the following constraints:

$$a_{11} x_1 + a_{12} x_2 + \dots + a_{1n} x_n \geq b_1$$

$$a_{21} x_1 + a_{22} x_2 + \dots + a_{2n} x_n \geq b_2$$

$$a_{m1} x_1 + a_{m2} x_2 + \dots + a_{mn} x_n \geq b_m$$

and

$$x_j \geq 0 (j = 1, 2, \dots, n)$$

#### 3.1 Data source

Data source used in solving the ESC-Souvenir product-mix supply problem was collected as an extract from the records of ESC-Souvenir firm between June 2011 and May 2016. The firm is a subsidiary of Expert Solution Concepts Limited, Ibadan, Nigeria from her main product line: custom-designed bags and three (3) different custom-designed bags made to be supplied to *Phototech Studio*, Ibadan, Nigeria. ESC-Souvenir was contracted by *Phototech Studio* to supply one hundred and twenty (120) bags per month. The bags are of three (3) different types (A, B and C)

and different dimensions: 12” by 12”, 12” by 14” and 14” by 16” respectively. The firm makes a profit of N250 (USD 0.25) on Type A, N300 (USD 0.33) on Type B and N340 (USD 0.34) on Type C. The raw materials used in the production are Leather, Lining, Parking Form and Leather Tape and the proportion of these materials are shown in Table 2. A tannery supplies the firm with at least 18 sets of Leather, 18 sets of Lining, 10 sets of Parking foam and not more than 4 rolls of Leather tapes (100 yards per roll).

The firm has the capacity to produce 90 bags of Type C which has the highest demand per month. It is important for ESC-Souvenir to determine the proportion of each bag to produce per month in order to maximize profit. The adequacy, inadequacy and surplus of each of these materials supplied is an important factor in order to meet its target of one hundred and twenty (120) bags per month.

Table 2: Dimension of raw materials

S/N	Materials	Dimension (inch)	Size (inch <sup>2</sup> )
1	Leather	36” by 58”	2088
2	Lining	36” by 58”	2088
3	Parking Foam	55” by 77”	4235
4	Leather Tape (Length)	36” by 1”	36

3.2 Problem evaluation

To evaluate the problem, mathematical model of the product-mix problem is formulated by expressing the objective function and constraints in canonical form and converting the expressions from canonical form to standard form with the introduction of slack or surplus variables and artificial variables where necessary.

3.2.1 Mathematical model formulation of the problem

Using data in Table 1 expressing the objective function and constraints in canonical form:

Let Type A =  $X_1$ , Type B =  $X_2$ , and Type C =  $X_3$

The profit on Type A is N250, Type B is N300 and Type C is N340

∴ Maximise objective function (Profit):

$$P = 250X_1 + 300X_2 + 340X_3 \tag{1}$$

Subject to (raw materials and other constraints):

$$X_3 \leq 90 \text{ Number of Type C} \tag{2}$$

$$X_1 + X_2 + X_3 = 120 \text{ Number of Contracted Item} \tag{3}$$

The dimension of leather and lining material to make Type A is 12” by 25”, Type B is 12” by 29” and Type C is 16” by 29” tailoring allowance inclusive. From this, it was deduced that  $\frac{1}{6}$  of these materials is needed to make Types A and B respectively and  $\frac{1}{4}$  for Type C. For the Parking foam,  $\frac{1}{12}$  of the material is needed to make Types A and B as well as  $\frac{1}{9}$  for Type C.

$$0.1667 X_1 + 0.1667 X_2 + 0.25 X_3 \geq 18 \text{ Leather} \tag{4}$$

$$0.1667 X_1 + 0.1667 X_2 + 0.25 X_3 \geq 18 \text{ Lining} \tag{5}$$

$$0.0833 X_1 + 0.0833 X_2 + 0.1111 X_3 \geq 10 \tag{6}$$

$$\text{Parking foam} \tag{6}$$

$$1.8888 X_1 + 2.0 X_2 + 2.3333 X_3 \leq 400 \tag{7}$$

$$\text{Leather tape} \tag{7}$$

3.2.2 Convert the problem from canonical to standard form

Converting the expressions from canonical form to standard form with introduction of slack or surplus variables in equations 2 and 7, since the inequality sign is “less than” ( $\leq$ ) the slack variables must be added to change the sign to “equal to” (=). Equation 3 is “equal to” (=) but when  $X_1, X_2$  and  $X_3 = 0, 0 \neq 120$  therefore slack variable must be added to balance the equation. In Equations 4, 5 and 6 respectively, the inequality sign is “greater than” ( $\geq$ ). Hence, surplus variable must be subtracted to change the sign to “equal to”(=).

After conversion, the constraints now become:

$$X_3 + S_1 = 90 \text{ Number of Type C} \tag{2}$$

$$X_1 + X_2 + X_3 + S_2 = 120 \tag{3}$$

$$\text{Number of Contracted Items} \tag{3}$$

$$0.1667 X_1 + 0.1667 X_2 + 0.25 X_3 - S_3 = 18 \tag{4}$$

$$\text{Leather} \tag{4}$$

$$0.1667 X_1 + 0.1667 X_2 + 0.25 X_3 - S_4 = 18 \tag{5}$$

$$\text{Lining} \tag{5}$$

$$0.0833 X_1 + 0.0833 X_2 + 0.1111 X_3 - S_5 = 10 \tag{6}$$

$$\text{Parking foam} \tag{6}$$

$$1.8889 X_1 + 2.0 X_2 + 2.3333 X_3 + S_6 = 400 \tag{7}$$

$$\text{Leather tape} \tag{7}$$

By introducing the slack or surplus variable the original objective function in equation (1) now becomes:

$$P = 250X_1 + 300X_2 + 340X_3 + 0S_1 + 0S_2 - 0S_3$$

$$-0S_4 - 0S_5 + 0S_6$$

### 3.2.3 The initial simplex tableau

The initial simplex tableau that is constructed from these equations and inputted into *Microsoft Excel* is shown in Appendix I. In this initial starting point, the *Basis* column shows the variables that have values corresponding to the *Slack* column; these are:

$$S_1 = 90, S_2 = 120, S_3 = 18, S_4 = 18, S_5 = 10, S_6 = 400$$

and the initial basic solution i.e. *Profit* = 0. All other variables are zero (0) i.e.  $X_1 = X_2 = X_3 = 0$ .

The data in the initial simplex tableau in Appendix I will be inputted into the *Microsoft Excel Solver* to solve the problem and find the optimal solution.

## 4. DISCUSSION OF RESULTS

The result of the first optimisation is shown in Appendix II and the summary are  $S_1 = 18, S_2 = 48, S_3 = 2, S_6 = 232.002, X_3 = 72$  and  $P = 24,480$  and the basic solution. All other variables are zero (0). If these values are substituted into the equations, the Left Hand Side (LHS) of the equations satisfy the Right Hand Side (RHS). This is not the optimal solution because there is a shortage of 48 bags in the supply as indicated by  $A_1 = 48$ . The raw materials i.e. leather and lining are used completely, two (2) of the parking foams and 232.002 yards of leather tape are in surplus. Similar results are obtained from second to the fifth optimisation (as shown in Appendices III– IV).

Appendix V shows the result of the sixth optimisation which is the optimal solution. Summary of the result are  $S_3 = 9.5, S_4 = 9.5, S_5 = 2.5, S_6 = 130.003, X_2 = 30, X_3 = 90$  and the basic solution  $P = 39,600$ . All other variables are zero (0).  $X_3 = 90$  is the number of Type C bags this satisfy equation (2),  $X_2 = 30$  is the number of Type B hence the combination of Types B and C satisfy equation (3) i.e. the one hundred and twenty (120) bags supply. It is this combination that resulted to optimal value (profit) of thirty-nine thousand, six hundred naira (N39,600 i.e. USD 39.6) only; the leather and lining as indicated by  $S_3$  and  $S_4$  is inadequate in its need for additional 9.5; the parking form as indicated by  $S_5$  is also inadequate in its need for additional 2.5 and finally leather tape is in surplus of 130 yards.

Note: The completed data analysis and presentation findings can be found in Appendices I - VII

## 5. CONCLUSION

The data collected from *ESC-Souvenir* firm was modelled into Linear Programming Problem. An objective function (profit maximisation) and constraints (resources, production and supply) were then developed out of the Linear Programming Problem. The initial simplex tableau was constructed out of the mathematical equations that modelled the Linear Programming Problem. Data in the initial simplex tableau was inputted into *Microsoft Excel Solver* for optimisation. The results of the optimisation are presented in Appendices I – V where various basic solutions are enumerated and the optimal solution was identified. The optimal value (profit) of thirty-nine thousand six hundred naira (N39,600) only (USD 39.6) was resulted from production and supply of combination of thirty (30) Type B bags and ninety (90) Type C bags. The results also revealed that 54.4% (9.5 pieces of leather), 54.4% (9.5 pieces of lining) and 25% (2.5 pieces of parking form) are needed in meeting the contracted supply sum of one hundred and twenty (120) bags. It also revealed that 32.5% (130 yards of leather tape) are in excess.

## 5. RECOMMENDATIONS

*ESC-Souvenir* product-mix supply as aforementioned in the problem statement has not been performing well in resources and production management. These problems have contributed immensely in *ESC-Souvenir* firm's profit maximisation, raw materials minimisation and meeting the supply/demands. The study has come at an opportune time in addressing the product-mix supply of *ESC-Souvenir* firm. This reason stems from the fact that getting exact number of each of the raw materials involves in production and the exact product-mix of types of bags to produce has been major challenge for the firm and can now be addressed by this study. Based on the findings and results from the study it is therefore recommended as follows:

1. *ESC-Souvenir* firm should look into other elements involved in its production such as labour and overhead cost when incorporated into the LP model.
2. The combination of Types B and C bags have positive effect in maximising the profit and
3. Inadequate supply of the raw materials can have effect on production capacity.

## 6. REFERENCES

- Afful, A. A. et al, 2016. Optimising a bank's credit portfolio. *International Journal of Applied Management Science*, 8, 68-82.
- Al-Deseit, B. 2009. Least-cost broiler ration formulation using linear programming technique. *Journal of Animal and Veterinary Advances*, 8, 1274-1278.
- Al-Sitt, E. K. 2004. *Modeling Optimal Water Allocation Under Deficit Irrigation*. MSc., College of Engineering, King Saud University Riyadh.
- Anderson, D. R. et al, 2016. *Statistics for business & economics*, Nelson Education.
- Balinski, M. and Tucker, A. 1969. Duality theory of linear programs: A constructive approach with applications. *Siam Review*, 11, 347-377.
- Bettters, D. R. 1988. Planning optimal economic strategies for agroforestry systems. *Agroforestry systems*, 7, 17-31.
- Bouras, D. and Engle, C. R. 2007. Optimal size of fingerling to understock in catfish grow-out ponds: an application of a multi-period integer programming model. *Aquaculture Economics & Management*, 11, 195-210.
- Carlson, C. K. 1988. Information management approach and support to decision-making. *Information & Management*, 15, 135-149.
- Chinneck, J. W. 2004. Chapter 13: Binary and mixed-integer programming. *Practical optimization: A gentle introduction*. Ottawa, Ontario, Canada.
- Chung, G. et al, 2009. Reliable water supply system design under uncertainty. *Environmental Modelling & Software*, 24, 449-462.
- Gray, J. 2006. Gauss and non-Euclidean geometry. *Non-Euclidean Geometries*, 61-80.
- Gray, J. 2008. *Linear differential equations and group theory from Riemann to Poincaré*, Springer Science & Business Media.
- Hasan, M. K. and Al Hammad, A. A. 2010. Intercity bus scheduling for the Saudi Public Transport Company to maximize profit and yield additional revenue. *Journal of Service Science and Management*, 3, 373.
- Hassan, M. et al, 2011. Improving oil refinery productivity through enhanced crude blending using linear programming modeling. *Asian journal of scientific research*, 4, 95-113.
- Hoesein, A. A. and Limantara, L. M. 2010. Linear programming model for optimization for water irrigation area at Jatimlerek of East Java. *International Journal of Academic Research*, 2.
- Horst, J. 2013. *Hierarchically integrating the production planning and scheduling to optimize the production planning process of a beverage compan*. University of Twente.
- Kumar, G. and Engle, C. R. 2014. Optimizing catfish feeding and stocking strategies over a two-year planning horizon. *Aquaculture Economics & Management*, 18, 169-188.
- Kuo, P. C. et al, 2003. Optimization of operating room allocation using linear programming techniques. *Journal of the American College of Surgeons*, 197, 889-895.
- Lim, N. et al, 2014. Engineering resource management middleware for optimizing the performance of clouds processing mapreduce jobs with deadlines. *Proceedings of the 5th ACM/SPEC international conference on Performance engineering*. ACM.
- Matthews, C. H. 2004. Using linear programming to minimize the cost of nurse personnel. *Journal of health care finance*, 32, 37-49.
- Mullan, W. 2008. Dairy science and food technology improving your writing using a readability calculator.
- Muthini, D. N. et al, 2017. Determinants of small-scale mango farmers market channel choices in Kenya: An application of the two-step Craggs estimation procedure. *Journal of Development and Agricultural Economics*, 9, 111-120.
- Naderi, M. J. and Pishvae, M. S. 2017. A stochastic programming approach to integrated water supply and wastewater collection network design problem. *Computers & Chemical Engineering*, 104, 107-127.
- Naifer, A. et al, 2011. Economic Impact of Salinity: The Case of Al-Batinah in Oman. *International Journal of Agricultural Research*, 6, 134-142.
- Ozsan, O. et al, 2010. Application of linear programming in production planning at marble processing plants. *Journal of mining science*, 46.
- Schrijver, A. 1998. *Theory of linear and integer programming*, John Wiley & Sons.
- Sierksma, G. 2001. *Linear and integer programming: theory and practice*, CRC Press.
- Soares Machado, M. A. and Gassenferth, W. 2015. An application for efficient telecommunication networks provisioning

- using linear programming. *Independent Journal of Management & Production*, 6.
- Soni, R. 1996. *Business Mathematics with Applications in Business and Economics*, Pitambar Publishing.
- Tunçel, L. 2016. *Polyhedral and semidefinite programming methods in combinatorial optimization*, American Mathematical Soc.
- Von Waltershausen, W. S. 1856. *Gauss zum Gedächtniss*, S. Hirzel.
- Wambugu, S. N. et al, 2009. Effect of social capital on performance of smallholder producer organizations: the case of groundnut growers in Western Kenya. *International Association of Agricultural Economists Conference*.

**13<sup>TH</sup>**

# **INTERNATIONAL** *Conference*



**Theme**

## **Information Technology Innovation for Sustainable Development**

**Conference Proceedings**  
**Volume 28**

Edited by:  
**Professor Adesola ADEROUNMU**  
**Professor Adesina SODIYA**

ISSN: 2141-9663