An Open Access Journal Available Online

Covenant Journal of Informatics & Communication Technology (CJICT)

Vol. 6 No. 2, Dec., 2018

A Bi-annual Publication of the Departments of Computer Information Science,
and Electrical & Information Engineering. Covenant University, Canaan Land,
Km 10, Idiroko Road, Ota, Ogun State, Nigeria.

Website: http//Journal.covenantuniversity.edu.ng/cjict/

It is a condition of publication in this journal that manuscripts have not been published or submitted for publication and will not be submitted or published elsewhere.

Upon the acceptance of articles to be published in this journal,the author(s) are required to transfer copyright of the article to the publisher.

ISSN - Print: 2354 – 3566
- Electronics: 2354 – 3507

Published by Covenant University Journals,
Covenant University, Canaanland, Km 10, Idiroko Road,
P.M.B. 1023, Ota, Ogun State, Nigeria

Printed by Covenant University Press

Articles

Covenant Journal of Informatics & Communication Technology. Vol. 6 No. 2, Dec., 2018

An Open Access Journal Available Online

A New Approach for the Minimization of Packet Losses in LTE Networks

Tekanyi A. M. S., Adedokun E.A, Njoku F.C & Agbon E.E

Ahmadu Bello University, Zaria,
Department of Communications Engineering Kaduna-State, Nigeria
amtekanyi@abu.edu.ng,
wale@abu.edu,
[3]franklinnjoku2008@gmail.com,
[4]eagbonehime1@gmail.com

Abstract - This research work presents a new approach for minimization of packet losses in Long Term Evolution (LTE) networks. A prominent design feature for vertical handover decision algorithm is to ensure seamless handover process between different wireless access technologies without compromising the Quality of Service (QoS) and Quality of Experience (QoE) of the users. There are scenarios in handover schemes, where due to poor handover process, frequent handover occur leading to packet losses and subsequent dissatisfaction of the users. A handover decision algorithm that incorporates the user's changing speed into a proximity model prediction technique (PMPT) in order to minimize packet losses during handover process between macrocell and femtocell networks is presented in this paper. The developed algorithm is designed to make appropriate prediction based on the established communication link to either the macrocell or femtocell network as the User Vehicle (UV) speed changes. Results obtained using MATLAB R2015b shows that the developed vertical handover algorithm (DVHA) attained a 77.07% reduction in packet loss ratio over the existing vertical handover algorithm (EVHA).

Keywords: Macrocell, Femtocell, QoS, QoE, LTE, PMPT, DVHA.

Introduction

The Long Term Evolution (LTE) network is a wireless communication standard specified for high speed data, of which the macro base stations are derived from (Deswal & Singhrova, 2017). In spite of the fact that macro base stations provide high data rates due to its operation at high bandwidth, complete coverage at indoor environment is not provided for, by the macro base stations (Chuang *et al*., 2015). As a result, femtocells are deployed to provide total network coverage. Femtocells operate within a licensed spectrum and are low powered, less expensive device with a limited range of 30 meters (Godor *et al*., 2015). Femtocells link mobile devices to the core network via the broadband connection, and improve network coverage as well as capacity (Zhang & Roche, 2010). However, only few users can be granted access simultaneously in an indoor environment (Shbat *et al*, 2012).

Users' movement between the different access networks is still a challenging issue in vertical handover process (Seth, 2013). This is due to the unpredictable nature of the user movement pattern. The conventional method of using Received Signal Strength (RSS) measurements has been deployed in the design of vertical handover algorithm to address this challenge. The results obtained, shows that there is still room for improvement in order to ensure seamless vertical handover process. The absence of a seamless vertical handover process and its resultant interruptions result in packet losses, which in turn degrades the user's call quality (Becvar & Mach, 2013).

Figure 1 shows a typical handover scenario in Long Term Evolution (LTE) networks (Wu, 2011). The handover between a macrocell and femtocell should be smooth and seamless (Wu, 2011).This can only be achieved by ensuring that the users' active voice and data sessions are maintained during the changing process of the BS (Base Station). Traditionally, handover can be of two types: Hard and soft handover. In hard handover scenario, the channel is made available at the serving BS after which the channel is engaged at the target BS, whereas for the soft handover, the channel is held at the serving BS for a certain period while been used simultaneously with the channel at the target BS. In an LTE multi-tier network, there are three handover scenarios (Gódor *et al*., 2015):

(i) **Hand-in:** A handover that occurs as the UE moves from a macrocell BS to a femtocell BS.

(ii) **Hand-out:** A handover that takes place when the UE moves from a femtocell coverage area into a macrocell zone.

(iii) **Inter-HeNB:** This is the movement of the UE from one femtocell to the other femtocell.

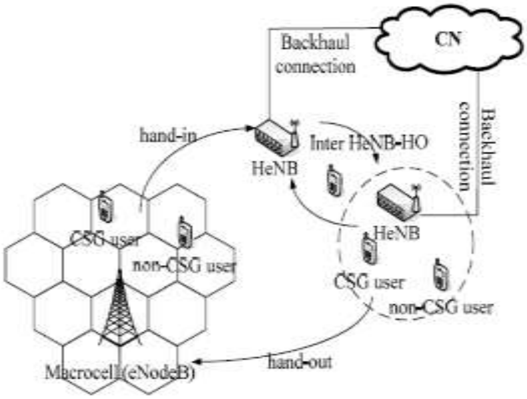


Figure 1: Handover Scenario in LTE Network (Wu, 2011)

This paper proposes a New Approach for the Minimization of Packet Losses in LTE Network. This approach introduces the concept of incorporating the changing speed of the users into a prediction technique in order to efficiently reduce the rate at which packets are lost in LTE networks. The first section of this work contains the introduction and the second section is the review of related works which gives an insight on the state of act of works done on LTE networks. The third section highlights the methodology applied in the minimization of packet losses. The fourth section shows the parameters used for the simulation. And the fifth section discusses the validation and comparison of the developed work.

2. Review of Related Work:

(Ben Cheikh et al, 2013) proposed an optimized handover algorithm with an efficient call admission control. The proposed algorithm was basically designed to reduce the number of unnecessary handovers and to maintain the communication quality during the handover. In the authors approach, the choice of the target femtocell took into account the direction of the mobile user, its velocity, and the quality of the signal in terms of Signal-to-Interference Ratio (SINR). The results showed that the proposed algorithm minimized the number of hand-in and drop rate handovers when compared to the traditional handover procedure which considered only signal strength. However the changing speed of the user was not considered in the design of the algorithm, as this could have improved the results obtained in terms of reducing the packet losses.

(Kalbkhani et al, 2014) developed a handover decision algorithm that relied on the prediction of the Received Signal Strength (RSS) in order to have a better throughput and also minimize the ping pong handover encountered by the users. In the proposed work, base stations that fulfilled the conditions of having a higher RSS than the set threshold as well as also been greater than the RSS of the serving base station with a hysteresis margin were known. The future RSS samples of the known base station and the serving base station were predicted using adaptive Recursive Least Square (RLS) algorithm. These approximated samples were used for future SINR samples. In conclusion, the candidate base stations list were reduced on the basis of the approximated SINR and predicted RSS, after which the base station with the highest throughput was selected. The proposed algorithm was validated based on the performance metrics of: Ping Pong Rate (PPR), number of handover (NHO), throughput, Outage Probability (OP) of the UE, and error due to the RSS prediction. The major drawback of the work was that the authors did not consider the varying speed of the User Equipment (UE) towards the mobile stations, which could have improved the throughput thereby leading in an enhanced result.

(Al-Shahin, 2015) presented an inter-femtocell handover scheme for dense femtocell networks. In the proposed work, two parameters namely; the movement direction of the mobile user and the location of the neighbor FAPs were considered in the minimization of the neighbor cell list for the purpose of reducing the need to scan a large cell list . Two algorithms were used to address the movement direction of the user; the first was used to predict the user mobility pattern, while the other was for generating mobility rules. To determine the location of the FAPs, a user mobility analysis server was used. Results obtained showed that as the mobile user direction was predicted towards (west-south) direction, 65% reduction in the FAPs list was achieved and an 85% reduction was obtained when the prediction was made in the (east-south) direction. The problem with the handover scheme was that no parameter was set to aid in the selection of a suitable FAP for handover which could result in increased dropped packets and eventual degradation of the user’s QoS.

(Habibzadeh et al, 2015) analyzed a handover decision algorithm that was based on received signal strength and channel holding time. In their work, the UE connects to the femtocell when the mean signal power from the macrocell was less than the set threshold and the mean signal power from the femtocell was higher by a fixed hysteresis margin. Furthermore, connection to the femtocell was made when the time spent in the femtocell was greater than the holding time. This was done to obtain an increased rate of handover to the femtocell while also reducing the number of unnecessary handover. The major challenge of the work was the delay that could occur as a result of the algorithm, when applied in a highly mobile environment which would result in signaling overhead during the process of selecting the target base station and this could be a major limitation of the work.

Xenakis *et al.*, (2016) designed a novel handover decision algorithm which utilizes measurement from candidate cells to optimize two Handover Hysteresis Margins (HHMs). The first HHM is used to avoid cells that can compromise service continuity, while the second HHM is used to identify the cell with the minimum required UE transmit power. When the handover event is triggered, the serving cell acquires the maximum transmit power, the cell interference and the downlink RS transmit power for all candidate cells by using the private mechanism for non-standard use. The two adaptive HHMs are subsequently evaluated for all candidate cells and the subset of cells that sustain service continuity is identified. The cell that requires the minimum UE transmit power is subsequently selected and the handover execution phase is initiated. However, in using the candidate cell list as an input to the algorithm, there are delays associated with the handover decision since the list is not optimized.

Alhabo & Zhang (2017) proposed a handover decision algorithm that makes use of the actual distance between the UE and the SCs (small cells) and the UE angle of movement, for the purpose of creating a shortened candidate list which assisted in minimizing the energy resource that would have been dissipated in the process of scanning for the candidate cells (macrocell or small cells), thereby reducing the number of unnecessary HOs (Handovers) while increasing the network throughput of the system. By the introduction of the shortened SC list, the performance of the algorithm was improved upon, as the number of unnecessary HOs were avoided because less number of target SCs were selected and the cell with the highest SNR (Signal to Noise Ratio) was selected as a candidate HO target. However, the delay that could eventually arise from the scanning process could affect the performance of the HO process, especially in a scenario where the UE moves at varying speeds.

Khan *et al.*, (2017) presented a handover algorithm for hand-in procedure. In the work, a multi-step handover scheme was used for the purpose of reducing the number of unnecessary handover, while also selecting the appropriate femtocell access point (FAP) for the incoming UE. Three filtering phases were proposed to achieve it. The first of the filtering phases for the proposed scheme was to measure the power of the available candidate femtocells. The next phase was to filter out the femtocells which cannot support an unregistered UE. The final filtering phase was to determine whether the UE was registered in the closed subscriber group (CSG) or not. All three phases in the proposed scheme were set as input parameters to make the handover decision, for the purpose of minimizing unnecessary handover and improving the throughput of the UE. The proposed handover algorithm showed better performance than already existing hand-in algorithms. However the work did not consider the delays that could arise from having to check, select and measure for the available and target femtocell, which could lead to degradation of the UE's QoS.

(Deswal & Singhrova, 2017) designed a handover decision algorithm for an integrated macrocell and femtocell network in which the macrocell network was overlayed by several femtocell networks using equivalent received signal strength (RSS_{eq}) and a dynamic hysteresis margin. In the proposed work, the RSS_{eq} was calculated by normalizing the unequal transmit powers of the femtocell and macrocell with respect to the distance of their respective base stations. The distance was to ensure that the UV (User Vehicle) performs handover to a femtocell whenever a femtocell was in the boundary of a macrocell as this guarantees successful packet delivery. The major problem of the work was that the changing speeds of the users were not considered as input parameters for deciding where and when to perform the handover, since both networks have dissimilar wireless access technologies. This could lead to degradation in quality of service. Also, no mobility prediction technique was used in order to accurately describe the user's movement towards the target cell, which could have improved the results obtained.

From the literatures reviewed, degradation of the user's quality of service caused by frequent handovers has been a major drawback in achieving a seamless handover process. This proposed work improved on the work of Deswal & Singhrova, (2017) where the scenario of the changing speed of the UV was not considered in deciding where and when to handover, as this could result in degradation in both QoS and QoE. A developed vertical handover decision algorithm (DVHA) that incorporates the changing speed of users into a proximity model is proposed in this paper.

3. Improved Algorithm: Dvha

This work proposes a new approach for minimizing packet losses in LTE networks. To address the issue of packet losses especially as the users speed changes, a mobility prediction technique that considers the changing speed of the user is presented. The following steps were used in the development of the DVHA.

1. For simulation purpose, an LTE macrocell-femtocell network architecture made up of one macrocell and 60 femtocells was used.
2. The Received Signal Strength (RSS) of macrocell and femtocell at all UV positions was obtained.
3. Random movement mobility model to model the movement pattern of the UV at varying speed was obtained.
4. Path loss between the UV and the FAPs was generated. The proximity prediction technique was used to mitigate the occurrence of packet losses during the handover process.
5. The conditions considered for the generation of the RSS for both the macrocell and femtocell are given as (Deswal & Singhrova, 2017):

   NCL = ($RSS_f \geq RSS_{fth}$) or ($RSS_{fe} \geq (RSS_m + \delta)$) and NCL = ($RSS_f \leq RSS_{fth}$) and ($RSS_m \geq (RSS_{fe} + \delta)$)

   Where: NCL= Neighbor Cell List

   RSS_f = Received Signal Strength of the femtocell

   RSS_{fth}= Received Signal Strength of the threshold femtocell

   RSS_{fe} = Equivalent Received Signal Strength

   RSS_m= Received Signal Strength of the macrocell

   δ = Hysteresis margin

6. The packet loss ratio is now calculated using equation 1(Deswal & Singhrova,

$$PLR = \frac{\left(\sum_{i=1}^{n} P_{r_i} - \sum_{i=1}^{n} P_{s_i}\right)}{\sum_{i=1}^{n} P_{s_i}} \qquad (1)$$

The equation used to describe the rate of handover during the packet delivery process is given as (Singh et al, 2005):

$$\lambda_H(h) = \frac{KD}{320}\left(1 - exp\left(-\frac{b}{h^a}\right)\right) \qquad (2)$$

The variables in the equation 2 are defined as follows:

K = adaptive parameter

D = base station separation

h = hysteresis margin

a = ratio of the path loss exponent (γ) to the standard deviation (σ) b = ratio of the correlation distance (d_o) to the averaging distance (d_{av})

The total link availability of the Proximity model is given as (McDonald & Znati, 1999). thus:

$$A_{m,n}^{T}(t) = [A_{m,n}^{i}(t)]S_{uv} \qquad (3)$$

The variables in the equation 3 are defined as follows:

$A_{m,n}^{T}(t)$ = Total link availability between two nodes m and n

$A_{m,n}^{i}(t)$ = Link availability when the mobility is random

S_{uv} = Speed of the user vehicle

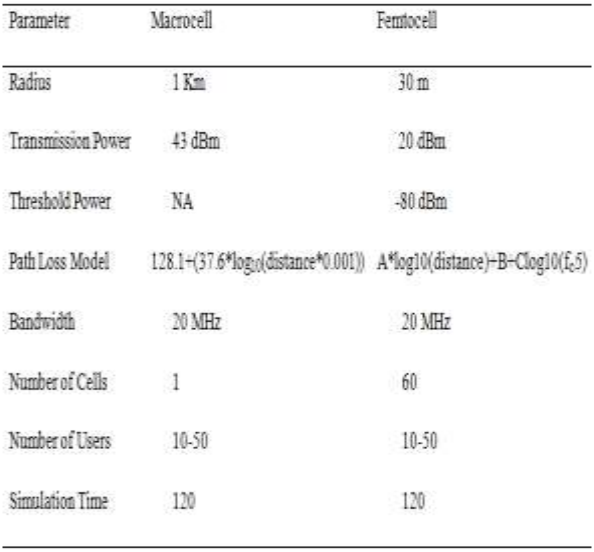
m = node of the UV

n = node of the target femtocell or macrocell

4. Simulation Environment

MATLAB R2015b version was used in the simulation of this work. MATLAB R2015b was used to compare the performance of the Improved Vertical Handover Algorithm (IVHA) and the Existing Vertical Handover Algorithm (EVHA). The simulation parameters used in this work is as shown in table 1

Table 1: Simulation Parameters



Parameter	Macrocell	Femtocell
Radius	1 Km	30 m
Transmission Power	43 dBm	20 dBm
Threshold Power	NA	-80 dBm
Path Loss Model	$128.1+(37.6*\log_{10}(distance*0.001))$	$A*\log10(distance)+B+C\log10(f_c/5)$
Bandwidth	20 MHz	20 MHz
Number of Cells	1	60
Number of Users	10-50	10-50
Simulation Time	120	120

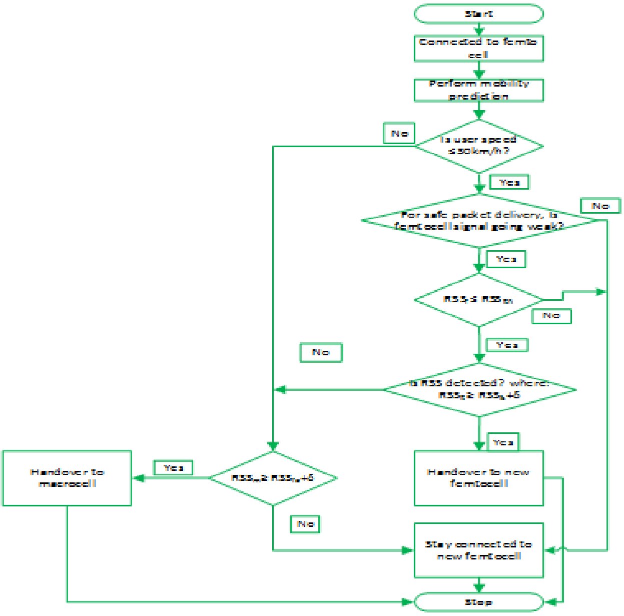

Start
Connected to femto cell
Perform mobility prediction
Is user speed ≤ 30km/h?
No
Yes
For safe packet delivery, is femtocell signal going weak?
No
Yes
RSS_f ≤ RSS_th
No
Yes
No
Is RSS detected? where: RSS_t ≥ RSS_s+δ
Yes
Handover to macrocell
Yes
RSS_m ≥ RSS_s+δ
No
Handover to new femtocell
Stay connected to new femtocell
Stop

fig 2: Flowchart of the Ivha when the users were connected to Femtocell.

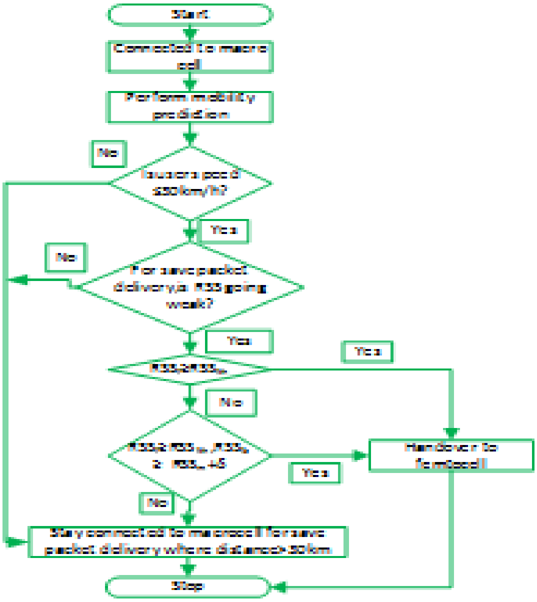

Start
Connected to macro cell
Perform mobility prediction
No
Is users speed ≤30km/h?
Yes
No
For save packet delivery, is RSS going weak?
Yes
Yes
RSS_f≥RSS_m
No
RSS_f≥RSS_th ,RSS_f ≥ RSS_m+5
Yes
Handover to femtocell
No
Stay connected to macrocell for save packet delivery where distance>30km
Stop

Figure 3: Flowchart of the DVHA when the users were connected to Macrocell

5. Results and Discussion

Figure 4 shows the result of the performance of EVHA and DVHA. The packet loss ratio is calculated using equation 1. The packet loss was simulated for both low and high speed users ranging from 30-120kmph. A Markovian model for random motion was used for the network simulation. From the plot, it can be observed that the developed vertical handover scheme displayed less packet loss ratio count when compared to the existing vertical handover scheme. This was achieved as a result of the effect of the mobility prediction scheme deployed, as well as the dwell time that was incorporated. Furthermore, the fluctuations as evident in the plot are due to the varying speed of the users as they move towards either the femtocell or macrocell.

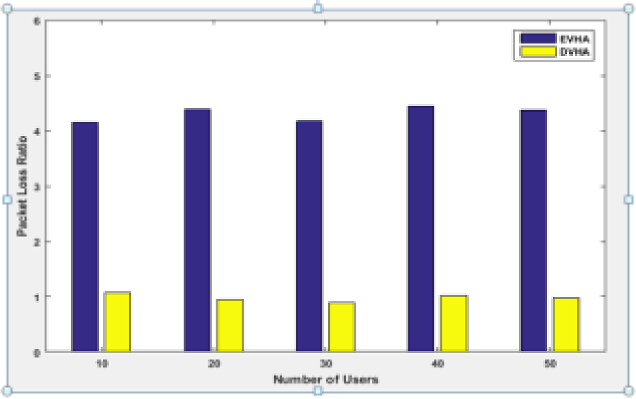


Fig 4: Plot of Packet Loss Ratio to number of Users

6. Conclusion

This paper presents the development of a new approach for the minimization of packet losses in LTE networks. After simulation, the performance of the DVHA was shown to perform better than the EVHA. The DVHA achieved this by using a PM prediction technique that incorporates the changing speed of the UV. The DVHA reduced the packet loss ratio by 77.07% with respect to the number of users over the EVHA.

References

Alhabo, M. D. J., Zhang, L., & Oguejiofor, O. (2017, June). Inbound Handover Interference-Based Margin for Load Balancing in Heterogeneous Networks. In *2017 International Symposium on Wireless Communication Systems (ISWCS)*. IEEE.

Al-Shahin F.A. (2015). Femtocell-to-Femtocell Management in Dense Femtocellular Networks. *International Journal of Computer and Communication Engineering*. (Vol 4, pp 346-353). doi: 10.17706/ijcce.2015.4.5.346-353.

Becvar, Z., & Mach, P. (2013). Mitigation of redundant handovers to femtocells by estimation of throughput gain. *Mobile Information Systems, 9*(4), 315-330.

Chuang, T. H., Chen, G. H., Tsai, M. H., & Lin, C. L. (2015). Alleviating Interference through Cognitive Radio for LTE-Advanced Network. *International Journal of Electrical and Computer Engineering*, *5*(3), 539.

Deswal, S., &Singhrova, A. (2017). A Vertical Handover Algorithm in Integrated Macrocell Femtocell Networks. *International Journal of Electrical and Computer Engineering*, *7*(1), 299..

Gódor, G., Jakó, Z., Knapp, Á,& Imre, S. (2015). A survey of handover management in LTE-based multi-tier femtocell networks: Requirements, challenges and solutions. *Computer Networks, 76*, 17-41.

Habibzadeh, A., Moghaddam, S. S.,

Razavizadeh, S., & Shirvanimoghaddam, M. (2015). A novel handover decision-making algorithm for HetNets. Paper presented at the *2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. 438-442

Kalbkhani, H., Yousefi, S., &Shayesteh, M. G. (2014). Adaptive handover algorithm in heterogeneous femtocellular networks based on received signal strength and signal-to-interference-plus-noise ratio prediction. *IET Communications*, *8*(17), 3061-3071. doi: 10.1049/iet-com.2014.0230

Khan, M., Ashraf, M., Zafar, H., & Ahmad, T. (2017). Enhanced Handover Mechanism in Long Term Evolution (LTE) Networks. *International Journal of Communication Networks and Information Security*, *9*(1), 40-47.

McDonald, A. B., &Znati, T. F. (1999). A mobility-based framework for adaptive clustering in wireless ad hoc networks. *IEEE Journal on Selected Areas in communications*, *17*(8), 1466-1487

Seth, A. (2013). Vertical Handoff Decision Algorithms for Next Generation Wireless Networks: Some Issues. *International Journal of Advanced Research in IT and Engineering*, *2*(8).

Singh, B., Aggarwal, K. K., & Kumar, S. (2005). A New Empirical Formula for Handover Rate in Microcellular Systems. *International Journal of Wireless Information Networks*, *13*(3), 253-260.

Shbat, M. S., &Tuzlukov, V. (2012). Handover technique between femtocells in LTE network using collaborative approach. Paper presented at the *18th Asia-Pacific Conference on Communications (APCC)*, Jeju Island. 61-66

Wu, S. J. (2011). A new handover strategy between femtocell and macrocell for LTE-based network. In *Ubi-Media Computing (U-Media), 2011 4th International Conference on* 203-208. IEEE.

Xenakis, D., Passas, N., Merakos, L., & Verikoukis, C. (2014). Mobility management for femtocells in LTE-advanced: key aspects and survey of handover decision algorithms. *IEEE Communications Surveys & Tutorials*, *16*(1), 64-91.

Zhang, J. & De la Roche, G. (2010). Femtocells: technologies and deployment: Wiley Online Library. Xvii.

Covenant Journal of Informatics & Communication Technology. Vol. 6 No. 2, Dec., 2018

An Open Access Journal Available Online

A Review on Web Page Classification

Ayodeji Osanyin, Olufunke Oladipupo & Ibukun Afolabi

[1] Department of Computer and Information sciences,
Covenant University, Ota, Nigeria
osanyindeji@gmail.com,
funke.oladipupo @covenantuniverity.edu.ng,
ibukun.fatudimu@covenantuniversity.edu.ng

Abstract—With the increase in digital documents on the World Wide Web and an increase in the number of webpages and blogs which are common sources for providing users with news about current events, aggregating and categorizing information from these sources seems to be a daunting task as the volume of digital documents available online is growing exponentially. Although several benefits can accrue from the accurate classification of such documents into their respective categories such as providing tools that help people to find, filter and analyze digital information on the web amongst others. Accurate classification of these documents into their respective categories is dependent on the quality of training dataset which is dependent on the preprocessing techniques. Existing literature in this area of web page classification identified that better document representation techniques would reduce the training and testing time, improve the classification accuracy, precision and recall of classifier. In this paper, we give an overview of web page classification with an in-depth study of the web classification process, while at the same time making awareness of the need for an adequate document representation technique as this helps capture the semantics of document and-also contribute to reduce the problem of high dimensionality.

Keywords/Index Terms— Classification, Document representation, TF-IDF, Web Page classification, Word2Vec

1. Introduction

Aggregating and categorizing information from these sources seems to be a daunting task as the information on the World Wide Web is increasing every second at a very high rate due to the influx of internet usage (Raj *et al.*, 2016). Automatic web classification / categorization is the main technology to achieve this.

Web page categorization which is also referred to Web Page Classification (WPC) is the process of assigning a web resource to one category or the other (Deri *et al.*, 2015). WPC problem can be sub-divided into two categories: the traditional manual method and the automated method of web page categorization. The traditional manual method is typically performed by experts who assigns web pages manually to the correct category, but this is impossible nowadays because of the influx of digital documents which will take a great deal of effort and time (Dey Sarkar *et al.*, 2014). While the former uses humans to achieve the categorization, the automatic method of WPC uses classification algorithms to determine the correct category in which the web pages belongs to automatically (Shibu *et al.*, 2010).

The former is tedious and time consuming, the latter reduces the large number of manpower, time needed for the classification, as well as resources (Dixit & Gupta, 2015).

Web classification is different from the standard text classification in some aspects: Traditional text classification is typically performed on structured documents which are stored in structured data stores such as relational databases and written with consistent styles which web collections do not possess (Qi & Davison, 2010; Abdelbadie *et al.*, 2013; AbdulHussien, 2017).

Web documents are semi-structured, formatted with the web markup language (HTML) which increases the rendering of the web pages to users. Also, web pages are linked together by hypertext within the same page or from one document to another (Qi & Davison, 2010). Several benefits can accrue from the accurate classification of documents into their respective categories such as providing resources that help the users to locate and retrieve the pertinent information amidst the vast resources on the web. Also news filtering, document routing and personalization of information on the web are additional advantages that can be harvested from web page classification.

According to (Mangai *et al.*, 2012), the commercial applications of web page categorization are as follows: most web directories owned by I.T giants such as Google, Yahoo and Microsoft bing are built, retained and extended by advanced WPC technologies (Huang *et al.*, 2004; Mangai *et al.*, 2012). Web page categorization are used to produce better search results from a search query. Searching for a particular resource proceeds by entering a keyword, and the search engine provides results related to the keyword, with WPC the search engine provides relevant and increased search results (Tsukada *et al.*, 2001). Advanced WPC techniques are used to improve the answers from a search result in a question and answering system (Cui *et al.*, 2004). Also, another very important application of WPC is web content filtering (Hammami *et al.*, 2003). Many WPC system have been proposed by several authors over the years, in which different approach have been formulated to tackle the problem of the classifier performance (Kato & Goto, 2016). Amongst the notable machine learning algorithm which have been proposed by several authors in literature for web page classification include Naive Bayes, KNN, Support Vector Machine (SVM), Artificial Neural Network (ANN), Decision trees (DT) (Fatima & Srinivasu, 2017). The classification result of the web page classification system in achieving high result is dependent on making the pre-processed document represent as much information as contained in the original document i.e. the pre-processing stage determines the quality of the results of the web classifier (Wang *et al.,* 2016). Also, the accuracy of most classification algorithms relies on the quality and size of training data which is dependent on the document representation technique (Dey Sarkar *et al.*, 2014).

This paper is a review paper which is intended in exploring the research question in the net section in a bid to achieve a systematic review of web page classification process, evaluate the document representation techniques and methodology used in web page classification.

2. Research Questions

The aim of this research is to answer the following research questions (RQs):

RQ1: What is the state of the art on WPC process?

The motivation for this question is to identify the current stages involved in the WPC process

RQ2: What are the Corpuses Used in web classification systems?

The purpose of this question is to discover the recent corpus or training data set used for web page classification

RQ3: What are current document representation techniques utilized in web classification systems?

The purpose of this question is to identify the gaps in the DR technique (semantic matching) utilized in web page classification

RQ4: What feature of the web page is used for web classification systems?

The purpose of this question is to identify the main part of web page that is used for building the web page classification system

RQ5: What kind of methods are used for web page classification

The motivation for this question is to discover trends applied methodologies used in web page classification and thereby establish the state of the art methodologies

3. Methodology

The research questions are structured to express content of literature review particularly following the approach of (Webster & Watson, 2002) and of (Kitchenham, 2004). Scopus, IEEExplore, CiteSeerx, ACM library, Google Scholar were the main source for the publication due to their richness and relevance in content as regards to web page classification publications. The initial search keyword in Google scholar was “web page classification” in the search bar, sorted by year and relevance. Then the search keyword was refined to consist of the following: “web page categorization” “feature selection techniques for web classification” “document representation techniques for web page classification” “web page classification process” “semantic matching” “Word2Vec for web page classification” “automatic text categorization” “topic models for web page classification" "comparison of document representation techniques" in the article title. 80% of the papers used were found in the Google Scholar database.

Inclusion criteria:

(1) Web page classification, Web page categorization, Document representation techniques, web page classification process, topic models for web page classification, LDA for web page classification are used to arrive at the search criteria and the major topics of the publications, (2) In a situation where several articles have reports that are similar, the latest publication is selected.

Exclusion criteria:

(1) Web classification using ontology based publications were excluded, because this research is focused on statistical techniques used in web page classification. (2) Publications that focused on general web classification using the multimedia content were excluded. (3) The contents of some online journal publications that could not be retrieved were also removed.

The necessary relevant criteria's highlighted above was the reference point for the articles and abstracts of the journal publications. In situations when details of the title and the abstract of the article don't match with the set of criteria, the whole content of the journal publication is examined, after which a decision for choice for either inclusion or exclusion is the made. The above highlighted procedure resulted in to 70 publications which was included in the next stages in the research process. These 70 publications were selected from a total of 85 which was retrieved before applying the inclusion criteria. The year of the publications selected ranged from 1999 to 2018.

3.1 RQ1: What is the state of the art on web page classification process?

This template was designed for two affiliations. According to (Fatima & Srinivasu, 2017), the web page classification system is divided in to several components as shown in Figure 1 below. The stages of the Web Page classification process includes: Creating a corpus of web pages, pre-processing / document representation, organization of the pre-processed pages, building the WPC model, obtaining a trained classifier, evaluating the classifier.

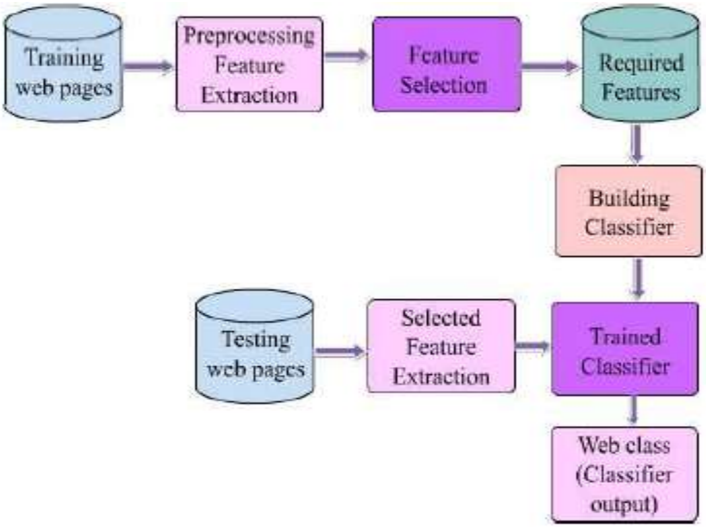


Figure 1. Web page classification process

A. Corpus or Web Pages Training Dataset

The first stage in the web classification process proceeds with extracting the main contents of the webpage along with other web page elements such as Internal and external hyperlinks, Metadata, Flash animation, Java script, Video Clips, Embedded objects, advertisement, Google ad-sense (Deri, Martinelli, Sartiano & Sideri, 2015). The extracted web contents are used in creating a corpus of labeled web pages i.e. training web pages which would be utilized by the classifier to building the learning system (Qi & Davison, 2009).

B. Pre-processing/Document Representation

The next stage in the web page classification process is the pre-processing stage also known as Document Representation (DR) or dimensionality reduction in this context (Mangai, Kothari & Kumar, 2012). This stage can also be further sub-divided into Feature Extraction (FE) and Feature Selection (FS) (Baharudin, Lee & Khan, 2010). FE process begins by extracting the raw content of the pages and discard HTML tags and other WWW contents. Web page document are characterized by high dimensionality, the first technique to reduce this high dimensionality is FE (Shibu, Vishwakarma & Bhargava, 2010; Raj, Francis & Benadit, 2016).

Then FE process continues by breaking down text into small chunks known as token which can either be a phrase, word or symbols in a process known as tokenization. After tokenization, then the tokens are reduced to their root or inflectional words know as stemming or lemmatization. Then lowercase conversion and filtering out of stop words (They are generally regarded as 'functional words' which do not carry meaning such as “the”, “a”, “and’’) (Fatima & Srinivasu, 2017).

The feature selection stage precedes after feature extraction. This stage involves constructing a vector matrix of the web document which is aimed at improving the accuracy of the web classifier. The basic aim of feature selection is to select the most important features that would represent the whole document (Alamelu Mangai *et al.,* 2010). Also with the inherent characteristics of web document which is high dimensional datasets, FS is used to reduce the space the original high dimensional space to a lower dimensional space which helps to increase the overall accuracy of the classifier and efficiency. Feature selection approaches can be broadly classified as filter, wrapper, and embedded.

The most generic of all the approaches is the filter approach which is the independent of the classification being utilized (Dey Sarkar, Goswami, Agarwal & Aktar, 2014). The filter approach uses metrics such as mutual information, correlation, entropy and so on, which analyzes general the general structure of the dataset and selects the optimal feature set (Talavera, 2005). The filter approach is a straight forward method and easier to work out than the other (embedded and wrapper) approaches (Kojadinovic & Wottka, 2000).

However, it is to be noted that wrapper and embedded methods often outperform filter in real data scenarios (Alamelu Mangai *et al.,* 2010). But in former (embedded approach), the algorithm is designed to embed the FS together with the objective function. Examples of embedded approach are DT, LASSO, 1-N SVM and so on. While in the later (wrapper approach), works on the basis of several combinations of the whole data for training and testing, which is usually an exhaustive search for the target function that learns the best feature set for the dataset. Metrics such as classification accuracy are used for selecting the optimal feature set. A major drawback of this approach is that, it is computationally expensive because of the brute force approach (Dey Sarkar *et al.,* 2014).

In contrast to the approaches discussed earlier, that selects the optimal feature set from the set of features, other techniques try to transform the original high dimensional feature matrix in to a lower dimensional matrix. This effectively helps to determine the semantics of a document and also, the main concepts in the document (Said, 2007; Qi & Davison, 2009; Li, Xia, Zong & Huang, 2009). Also, the above approaches cannot infer the inter or intra document statistical structure of the corpus (Biro, Benczur, Szabo, Maguitman, 2008). Such methods include: bag of words model TF-IDF (Ayyasamy *et al.,* 2010; Wang *et al.,* 2013; Sartiano & Sideri, 2015; Weiping & Chunxia, 2015; Moiseev, 2016; Raj *et al.,* 2016; Deri *et al.,* 2015; Fatima and Srinivasan, 2017), Latent Semantic Indexing (LSI) (Deerwester *et al.,* 1990; Chen & Hsieh, 2006; Biro *et al.,* 2008), Probabilistic Latent Semantic Indexing (PLSI) [34], Word2Vec (Lilleberg *et al.,* 2015; Wang *et al.,* 2016), Latent Dirichlet Allocation (LDA) (Sriurai *et al.,* 2010; Špeh *et al.,* 2013; Wang *et al.,* 2016).

Each technique has its own pros and cons. Lots of discussions are ongoing in the pre-processing and document representation stage of the WPC system. Document representation is very crucial stage in the web page classification process as irrelevant and noisy features in the data set will impact badly on the performance of the classifier in terms of its accuracy, speed and reducing overfitting issues (Alamelu Mangai *et al.,* 2010).

Also this stage has gain more attention recently than any other component of the WPC as good dimensionality reduction will improve the learning capabilities of the classifier and good storage capabilities (Ayyasamy *et al.,* 2010; Azam & Yao, 2012; Dey Sarkar *et al.,* 2014; Lilleberg *et al.,* 2015).

C. Obtaining the Required Features

The next stage after pre-processing stage is to gather the required feature set for classification which is usually achieved by creating matrix representation of the document vectors which would be fed to the classifier (Alamelu Mangai *et al.,* 2010).

D. Building the WPC Model

After gathering the required features, the next stage is to build the WPC model using a classification algorithm with the selected features as the input data set. Several machine learning algorithm have been used for the building the model of the WPC system systems such as KNN (Miao *et al.,* 2009; Bang *et al.,* 2010; Karima *et al.,* 2012), Support Vector Machine (SVM) (Chen & Hsieh, 2006; Sriurai, Meesad & Haruechaiyasak, 2010; Patil & Pawar, 2012; Wang, Chen, Jia & Zhou, 2013; Lilleberg, Zhu & Zhang, 2015; Fatima & Srinivasu, 2017), Naïve Bayes (Dey Sarkar et, al., 2014; Raj, Francis & Benadit, 2016), Decision trees (DT) (Kim *et al.,* 2001), Deep Learning (Kato & Goto, 2016), Weighted Voting of Feature Intervals known (WVFI) (Mangai *et al.,* 2012a; Mangai *et al.,* 2012b), Artificial Neural Network (ANN) (Ruiz & Srinivasan, 1998; Yu *et al.,* 2008) and so on. After training the classifier, the model obtained is thereafter utilized to automatically categorize new web resources to the appropriate category.

Several authors have argued about the best ML technique for web page classification but literature has shown that the accuracy, generalization capabilities of any ML technique depends on the training data set i.e. choice of the techniques used in the preprocessing stage have an overall effect on quality of the classifier (Biro *et al.,* 2008; Karima *et al.,*; 2012; Dey Sarkar et, al., 2014; Wang, Ma & Zhang, 2016; Chao & Sirmorya, 2016; Singh *et al.,* 2017)

E. Evaluating the Classifier

To test the performance of the web page classifier, some evaluation metrics are utilized to do this. A confusion matrix is one of the widely used metric for evaluating the performance, which is shown in table 1 below. A confusion matrix is a table that showcases the correct label of a category again the predicted label of a category. In the table, the total number of true positive classification is represented by “i”, while that of false positive classification is denoted by “j”. Also, the number of false negative classifications are denoted by “k”, while that of true negative classification are denoted by “l”. For a classifier to be of optimal performance, both j and k must be zero (Jindal *et al.,* 2015).

Table 1: Confusion Matrix

		Predicted Class	
		T_n	Not T_n
Actual Class	T_n	I	j
	Not T_n	K	L

The accuracy of the classifier can be calculate from the confusion matrix using the value obtained from this calculation: (i+j) / (i+j+k+l). Other metrics used for evaluating the performance of web page classification problems are known as Precision (Pr) and Recall (Re). The value of recall is calculated as i/(i + k) which is the total proportion of dataset in category Tn that are correctly predicted has been in that class. While the whole ratio of dataset which are correctly predicted to belong to that category Tn which belongs to that category. At every point of recall, a precision is associated with it. Most times, it is standard practice to combine both Pr and Re as a standalone metric called F1 score which is calculated from the computation of (Jindal *et al.,* 2015).

From the review above, it is has been highlighted that the web page classification process proceeds with the creation of corpus, pre-processing/document representation, obtaining the required features, building the WPC model and finally evaluating the classifier

3.2 RQ2: What are the corpuses used in web page classification?

To delve in to this question, we look in to reviewing the datasets utilized by several authors in building web page classification systems. Dataset utilized for web page classification include: Reuters datasets, which is a dataset created from Reuter's newswire and it contains 118 different category of news. Web Kb is another well-known dataset which has been utilized in several text classification problems. It is freely distributed online and it contains about 1065 web pages which is categorized in to two categories. Also the 20Newsgroups dataset is another popular dataset that is utilized for text categorization. It contains 20 categories of news items which is made up of 18846 different news in different categories. Another popular dataset is the yahoo news dataset which contains user's activities of yahoo websites and applications such as sports, finance and real estate (Wang *et al.,* 2016). Another corpus that is being utilized as training data set is the Imdb dataset which contains 1094 movie scripts downloaded from the Internet Movie Script Database (IMSDB) in HTML format. The movie scripts in this dataset are American Hollywood movies released from 1935 to 2015. The distribution of the genres of the movie in the corpus are drama, thriller, comedy, action, crime, romance, adventure, sci-fi, horror. Also, SOS dataset are frequently utilized which contains several categories of articles such as history, language studies, music, religion and so on, summing up to 4,625.

Figure 2 reveals the most important dataset utilized by several authors in web page classification systems. From the chart it is shown that Reuter's dataset is still the most utilized dataset by several authors.

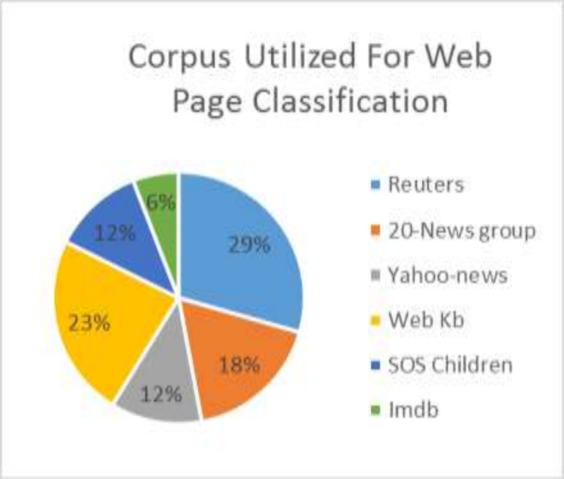


Figure 2. Corpuses utlilzed for Web Page Classification

RQ3: What are current document representation techniques utilized in web page classification?

According to Google index the amount of web resources available online is over 130 trillion pages and its growing at a rapid rate as due to fact that new users are added to the existing users every day. Retrieving information as soon as possible from this web documents is becoming necessary for many real life application (Azam & Yao, 2012). The accuracy and generalization capabilities of the classifier in assigning a web page to its correct category is heavily dependent on the document representation (Wang *et al.,* 2016). Several authors have applied various DR techniques to improve the quality of the input dataset which inherently will increase the general performance of the WPC system. Each technique is fraught by one challenge or the other. Some of them are highlighted below:

Term Frequency-Inverse Document Frequency (TFIDF)

This is the traditional and most popular method for document representation that is often used in information retrieval. It is a model that measures the importance of a word across a document. It weighs the important words increasingly based on how frequently they appear in the document but decreases the weight proportionally as it occurs in other documents. The TF-IDF weighting function is shown below:

$$W_{i,j} = tf_{i,j} * idf_j = = tf_{i,j} * \log_2 \left[\frac{N}{df}\right]$$

The Term Frequency, $tf_{i,j}$ measures the no of occurrences of a word in a document:

$$tf_{i,j} = \frac{Number\ of\ times\ word\ t\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document}$$

The Inverse Document Frequency, idf_j, measures the importance of a word by reducing the word's weighting score if it frequently occurs in other documents

$$idf_i = log_2 \left(\frac{Total\ number\ of\ document}{Number\ of\ documents\ with\ term\ t\ in\ it}\right)$$

TF-IDF can represent a document well by removing stop words from the documents. Some of the drawbacks of tf-idf are that it does not capture semantic similarity, does not respect word order and it is an unordered collection of words.

Latent Semantic Indexing (LSI)

Another popular method used in information retrieval method which utilizes linear algebra index technique to tackle the sparse matrix produced by TF-IDF methods is referred to as Latent Semantic Indexing (LSI) (Deerwester, *et al.,* 1990). LSI uses a vector model to build a matrix of word co-occurrences. It identifies the position on a vector space where each term and a document in a collection are. It works on the assumption that groups of words are semantically related will cluster together (Landauer *et al.,* 1997). To create a low dimensional representation of the document, it utilizes SVD algorithm on the sparse bag of words matrix, to create a denser matrix that approximately models the original document. It composes frequencies of terms as a term-document matrix. LSI was used to solve the synonym and polysemy problem of TF-IDF.

However, a major drawback of LSI are that, it does not capture multiple meanings of a word and it does not respect word order (Zhang *et al.,* 2011). Also, LSA models a document as a Gaussian distribution while in most situation a Poisson distribution is observed and the resulting dimensions might be difficult to interpret (Biro *et al.,* 2008).

Probabilistic Latent Semantic Indexing (PLSI)

To overcome some of the afore-mentioned problems with LSI, (Hofmann, 1999) proposed a more sound approach referred to as Probabilistic Latent Semantic Indexing (PLSA), which uses a generative approach for enhancing the capabilities of latent semantic indexing (LSI). The model obtained by PLSI is usually a probabilistic co-occurrence of words as a mixture of generative words. It uses EM Algorithm for its learning (Oneata, 1999). PLSI is usually viewed as a more sound method as it provides a probabilistic interpretation, whereas LSI achieves the factorization by using only mathematical foundations (more precisely, LSI uses the singular value decomposition method) (Batra & Bawa, 2010). Also, PLSA deals with synonyms and polysemy words by taking a deeper look at different forms of words and meanings. The core foundation of probabilistic latent semantic analysis are statistical models.

The introduction of PLSA shows promising results but it has two major drawbacks which are: the hyper-parameters are linear in nature while real life web documents are not, which impacts on predicting of new documents (Biro *et al.,* 2008).

N-Gram Model

Another popular method for document representation is the N-gram model. It is based on the assumption that any given word can be predicted based on the probability of its proceeding n-1 word, where n = 1, 2, 3.........x, x is a whole number. If n = 1, it is referred to as a unigram model, when n = 2, it is a bigram, 3 is a trigram. N-gram approach to feature representation converts a corpus of text in to the corresponding feature vector by taking record of the n-gram frequency counts which will serve as input vector to the classifier (Cianflone & Kosseim, 2017). The N-gram model can be of two forms. The first is referred to as character n-grams model, it rest on the assumption that sequence of unique occurring letters in a corpus while the second refers to as word n-gram model which relies on sequence of unique and occurring words in a corpus. Word N-gram model outperforms character n-gram model in many real word applications (Giannakopoulos *et al.,* 2012). According to (Elberrichi & Aljohar, 2007), some of the major strengths of N-GRAM are: No need to performing word segmentation. Capturing of root words automatically by the model. All languages are independent of each other. It has a low tolerance with distortion of words and mistakes usually made with spellings. In addition, no dictionary or language specific techniques are needed (Wei *et al.*, 2009).

N-GRAM suffers from data sparcity and high dimensionality (Mikolov *et al.,* 2013).

Latent Dirichlet Analysis (LDA)

Latent dirichlet analysis also known as LDA is a probabilistic topic model that generates what is referred to as latent topics based on the occurrence of a word in a text corpus or documents (Blei *et al.,* 2003). It assumes documents are a blend of several topics and that each word in the document can be grouped under the document's topics. LDA is typically handy is situations where there is need to find accurate mixture of topics within a given document. LDA is an unsupervised language model that transforms words from bag of words counts into continuous representative matrix. According to (Blei *et al.*, 2003), LDA works with the assumption that the generative process for each document in a collection of documents D is as follows:

1. $Choose\ N \sim Poisson\ (\varepsilon)$
2. $Choose\ \emptyset \sim Dir\ (\alpha)$
3. $For\ each\ of\ the\ N\ words\ W_n$
4. $(a) Choose\ Z_n\ Multinomial\ (\emptyset)$
5. $(b) Choose\ a\ word\ W_n\ from\ p(W_n|Z_n,\beta) a\ multinomial\ probability$
6. $Condition\ on\ the\ topic\ Z_n$

A major drawback of LDA is that improper calibration of these parameters could lead to sub-optimal results. Also, LDA uses an unsupervised learning function which depends on words in the corpus which will determine the matching degree and thus will suffer from vocabulary mismatching problem (Dit *et al.,* 2013).

Word2vec

This is a neural network language model that can learn word embedding's. The 2 main architectures are CBOW (Continuous-Bag-of-word) and Skip-gram (Continuous-Skipgram Model). The first architecture tries to predict words from the context of words while skip-gram tries to predict the context from the words. In the CBOW model each input vector u (i) is a column in the Matrix U. The CBOW model predicts a word u (i) utilizing the context u (i − n)... u (i − 1), u (i + 1)..., u (i + n), while the Skip-gram model predicts each word in the context utilizing the word u (i). The Word2Vec framework aims at predicting the context of word or word based on their context. The word embedding's are learned through maximizing the objective function. With these word embedding's it can capture distributed representations of text to capture similarities among concepts (Mikolov *et al.,* 2013) which is one of the major advantages of Word2Vec. However, a major drawback of word2Vec is that it does not model the global relationship between documents to topics (Wang *et al.,* 2016).

According to (Singh *et al.*, 2017), many new hybrid techniques have been formulated by several authors to harness the strength of each of the technique highlighted above for adequate preprocessing of the input data: LSI and TF-IDF (Chen & Hsieh, 2006; Zhang *et al.*, 2011), N-Gram and TF-IDF (Karima *et al.,* 2012), Word2Vec and TF-IDF (Lilleberg *et al.*, 2015), Word2Vec and LDA (Wang *et al.*, 2016), TF-IDF and firefly Algorithm (Raj *et al.*, 2016; Ma *et al.,* 2016), TF-IDF and K-means clustering (Milios *et al.,* 2006; Dey Sarkar *et al.*, 2014), LDA and TF-IDF (Sriurai *et al.,* 2010), Doc2Vec and Affinity propagation (Ma *et al.,* 2016).

Figure 3 below shows that the dominant document representation technique utilized by most researchers is the bags of words model TF-IDF followed by LDA then Word2Vec. The chart also shows that the hybrid techniques are gradually becoming popular.

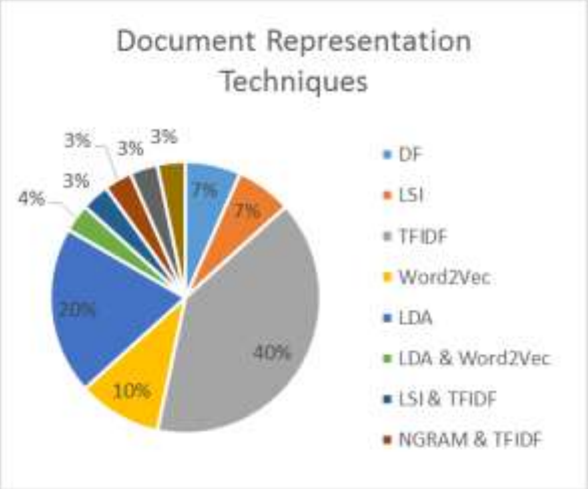


Figure 3: Document Representation Techniques

According to literature, there are many document representation techniques used for the preprocessing stage of the WPC and bag of words model (Bow) TF-IDF is still the most used DR technique

3.4 RQ4: What feature of the web page is used for web classification?

To answer this question, we focus on reviewing the feature utilized in creating the training dataset for the WPC system. This is because web pages are semi-structured with HTML tags which is made of several parts such as the page content, Meta tags, links & URL and HTML structure. According to (Shibu *et al.,* 2010), building WPC system using the page content involves utilizing the content of the web page to determine the category in which the web page belongs to. For Meta tags, the WPC system rely solely on attributes of Meta tags i.e. (<META name=”Keywords”> and <META name=”description”). For links and URL, this method is dependent on exploiting the contexts surrounding a link in an HTML document to extract useful information. HTML structures approach exploits both the content, html structure, images, placement of links contained in the page for building the WPC system. Figure 4 below shows that page content is the most used feature utilized for building the WPC system.

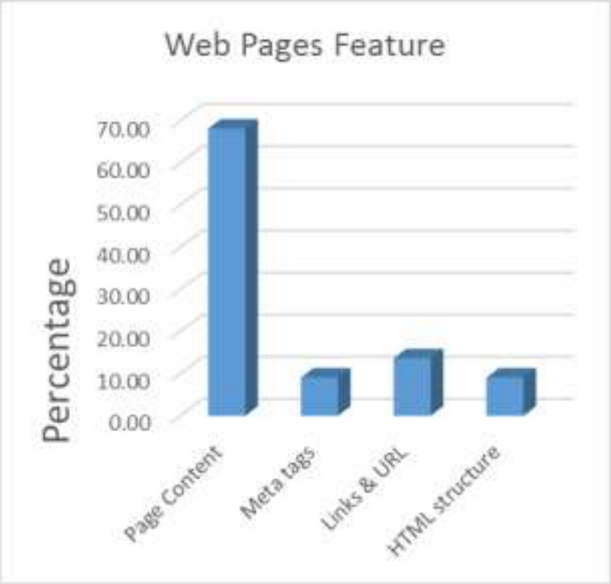


Figure 4: Web Page Features Utilized for WPC System

RQ5: What kind of methods are used for web page classification

According to (Xu *et al.,* 2011), a URL-based web page categorization using n-grams as the DR technique was proposed. Most updates users sends on social media like twitter, Facebook and so on contain links to email and webpages, there URL can be a means of categorization this information. Recently the influx of multimedia content on the web such as videos and images makes categorization by web pages a cumbersome process. In their work, they use n-Gram Language Model (LM) to classify textual data using the URL links of the web pages. The proposed method was applied to three datasets (webKb, DMOZ and GVO datasets). Results obtained showed an increase in the F1 measure for their method when compared with earlier methods

In the work of (Dey Sarkar *et al.,* 2014), they tried to solve the document representation and feature selection problem in web page classification. The methodology employed involved using chi-square metric to select the important words. The selected words are represented by their occurrence in various documents by simply taking the transpose of the term document matrix. K-means clustering is used to prune the feature space further to reduce the dimensionality of the term document matrix. Naive Bayes classifier is then fitted against the document to classify the document in to its appropriate class. Their proposed methods was applied to thirteen datasets and experimental results shows that their method outperforms other earlier feature selection techniques. A major gap identified in their work is that document representation technique utilized was Document Frequency (DF) which is bag of words model and it is an inherent problem of not capturing the semantic similarity and word order of the document (Singh *et al.,* 2017).

Deri *et al.,* (2015), formulated a classification tool for TLD (top level domain). The methodology preceded by creating a custom made crawler that is able to download web pages starting from the index page, removing non HTML web pages and automatically discarding irrelevant pages such as about us page and so on. Then using python NLTK processor to perform traditional text processing such as lemmatization, stop words removal and so on. They used a bag of words model (tf-idf) to construct the term document matrix. Then the terms were then trained using classification algorithm (Naive Bayes and SVM). Results obtained shows that Naive Bayes classifier performs better than the SVM algorithm using Precision, recall and F1 score. A major gap identified in their work is that the document representation technique used was TF-IDF which is a bag of words model which does not capture semantic similarity and word order of the document being transformed (Wang *et al.,* 2017).

According to (Lilleberg *et al.,* 2015), they applied neural network model (Word2Vec) with bags of words model (tf-idf) to solve the document representation problem of web classification. Accurate Representation of documents affect the correct classification or categorization of new documents. To solve the document representation problem, they created a hybrid of Word2Vec weighted with tf-idf with stop words to correctly represent the feature vectors of a document. The proposed method was applied to 20 newsgroup text dataset. Results obtained shows that there proposed method outperforms tf-idf with/without stops words and word2vec with/without stop words. A major drawback with their work is that, stops words increase the dimensionality of the feature vectors which impacts badly on the classification accuracy and computational burden (Wang *et al.,* 2016). Also the classification algorithm used was a linear SVM, other kernels such as string and RBF kernels could produce better results (Nayak *et al.,* 2015).

In the works of (Raj *et al.,* 2016), they proposed a method to automatically classify web pages into different categories via three stages, which are: FE, information learning and classification. In the methodology adopted, term document matrix is created using tf-idf, then the terms are used to extract object based features. Decision tree algorithm is then used to generate rules from the features set. The rules extracted are then used as input in to the hybrid of optimal firefly algorithm based Naive Bayes Classifier (FA-NBC). The proposed method was applied to WebKB datasets. Experimental results shows that their proposed method outperforms earlier methods such as KNN. Drawbacks identified in their work include: using tf-idf to construct the term document matrix does not capture any semantic similarity or form of grammatical analysis (Wang *et al.,* 2016).

Moiseev (2016), proposed a method to analyze and categorize e-commerce websites automatically. In their methodology, e-commerce website were crawled, text preprocessing and the terms of the document were derived using tf-idf. The proposed method was applied 1312 e-commerce and 1077 non e-commerce web site, preprocessing of the webpages, term weighted with tf-idf and classified using SVM. Experimental results shows that the produced method outperforms pure TF-IDF. Also the results shows a substantial increase in the accuracy of the classifier. A major gap identified in their word is that bag of words model like TF-IDF does not capture semantic similarity and respect word order of the document being represented (Singh, Devi & Mahanta, 2017).

In the works of (Wang *et al.,* 2016), they proposed the use of a hybrid strategy that consist of Latent Dirichlet Allocation (LDA) and Word2Vec for document representation. Word2Vec create a vector representation of the document which shows the semantic relationship between the words of the document. Euclidean distance was used to measure and interpret similarity between document and topic in sparse space. Their methods was applied to 20 News group data using SVM classifier. Results obtained shows that their proposed methods outperforms earlier methods such as TF-IDF+ SVM, Word2Vec + SVM, LDA + SVM. One of the major drawback of their method is that hyper-parameter tuning of LDA parameters i.e. # of topics, could produce unsatisfactory results as most of the parameters for the LDA are imported from natural language community (Dit *et al.,* 2013).

Based on the review conducted, several authors have proposed myriads of methods of improving the accuracy of the WPC system at the pre-processing or document representation stage, therefore showing this stage is still open to more research.

Observations

Representation of the input data (DR) is a crucial issue in web page classification and text classification systems at large. The performance of an algorithm is determined by the function of the input data available (Oyelade et al., 2010). Several feature selection techniques have been proposed to solve the issue of semantic matching of unstructured data, but are marred with one issue or the other. Recently, there has been an increase in the use of SVM and KNN for text classification (Khan et al., 2010; Jindal et. al, 2015). Also from extant literature, SVM, KNN and Naïve bayes are one of the most widely used ML algorithm for text classification (Joachims, 1998; Kwon & Lee, 2000; Asirvatham & Ravi, 2001; Sun et al., 2002; Khan et al., 2010; Sriurai et al., 2010; Krestel, 2012; Mangai et al., 2013; Lin & Wang, 2014; Lilleberg et al., 2015; Dixit & Gupta, 2015; Raj et al., 2016).

In the work of (Dey Sarkar et al., 2014), they decided to investigate this issue and compared SVM, KNN and Naïve Bayes on text classification tasks. Results obtained shows that SVM was not a clear winner, despite quite good overall performance. If a suitable pre-processing is applied to KNN and Naïve Bayes theory, these algorithms will achieve very good results and scales up to the performance of SVM. In light of this, there is need for an adequate document representation technique to retrieve the semantics of a web document. Optimized document representation techniques such as hybridizing neural network language models (Word2Vec) and topic model (LDA) or Word2Vec and TF-Idf with optimizing the parameters of LDA with search algorithms (such as GA) will provide better semantics of the document in WPC. This hybrid approaches has shown to perform better (obtain the semantic features) by harnessing the strength of the individual technique in the arrangement Word2Vec and LDA (Wang et al., 2016) or Word2Vec and TF-Idf (Lilleberg et al., 2015]. Also, proper calibration of the parameters of LDA with a search algorithm would produce better latent topics across words in a document (Dit et al., 2013).

Conclusion

In this paper, we gave an overview of web page classification system. Different application areas and an in-depth analysis of the web page classification process were looked into. Analysis of state of art techniques for feature selection techniques used in WPC was looked in to with a view to identify challenges fraught by each one. Also related works in the areas of WPC was reviewed to identify the latest works in this domain. It clearly shows that document representation phase is one of the areas that are receiving interest by researchers. Most currently used methods of document representation are Vector Space Model (VSM), Probabilistic Topic Model and Statistical Language Models and Neural network language models [47]. The chosen document representation technique have a direct impact on the classification results as it captures the semantics of document and also contribute to reduce the problem of high dimensionality. Combining different DR technique are new areas of research because each technique perform differently depending on the dataset. Future work in WPC should focus on improving the semantic relationship of web document by hybridizing difference DR technique which will inherently improve the classification result. Also, ontology based techniques can be used to capture the real semantics of unstructured text (Daramola et al., 2013).

References

Abdelbadie B., Abdellah I., & Mohammed B. (2013). Web Classification Approach Using Reduced Vector Representation Model Based On Html Tags. *Journal of Theoretical and Applied Information Technology,* 55(1).

AbdulHussien, A. A. (2017). Comparison of Machine Learning Algorithms to Classify

Web Pages. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 8(11), 205-209.

Alamelu Mangai, J., Santhosh Kumar, V., & Sugumaran, V. (2010). Recent Research in Web Page Classification–A Review. *International Journal of Computer Engineering & Technology (IJCET)*, 1(1), 112-122.

Asirvatham, A. P., & Ravi, K. K. (2001). Web page categorization based on document structure. *Centre for Visual Information Technology*.

Ayyasamy, R. K., Tahayna, B., Alhashmi, S., Eu-gene, S., & Egerton, S. (2010, May). Mining Wikipedia knowledge to improve document indexing and classification. In Information *Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on IEEE*, (pp. 806-809).

Azam, N., & Yao, J. (2012). Comparison of term frequency and document frequency based feature selection metrics in text categorization. *Expert Systems with Applications*, 39(5), 4760-4768.

Bang, S. L., Yang, J. D., & Yang, H. J. (2006). Hierarchical document categorization with k-NN and concept-based thesauri. Information processing & management, 42(2), 387-406.

Batra, S., & Bawa, S. (2010). Using lsi and its variants in text classification. Advanced Techniques in Computing Sciences and Software Engineering, 313-316.

Biro, I., Benczur, A., Szabo, J., & Maguitman, A. (2008, October). A comparative analysis of latent variable models for web page classification. In Latin American Web Conference, 2008. LA-WEB'08. (pp. 23-28). IEEE.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Chao, B., & Sirmorya, A. (2016). Automated Movie Genre Classification with LDA-based Topic Modeling. *International Journal of Computer Applications*, 145(13).

Chen, R. C., & Hsieh, C. H. (2006). Web page classification based on a support vector machine using a weighted vote schema. *Expert Systems with Applications*, 31(2), 427-435.

Cianflone, A., & Kosseim, L. (2017). N-gram and Neural Language Models for Discriminating Similar Languages. arXiv preprint arXiv:1708.03421.

Cui, H., Kan, M. Y., Chua, T. S., & Xiao, J. (2004, July). A comparative study on sentence retrieval for definitional question answering. *In SIGIR Workshop on Information Retrieval for Question Answering (IR4QA)* (pp. 383-390).

Daramola, O., Afolabi, I. T., Akinyemi, I., & Oladipupo, O. O. (2013). Using ontology-based information extraction for subject-based auto-grading.

ResearchGate, 373-378.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science,* 41(6), 391.

Deri, L., Martinelli, M., Sartiano, D., & Sideri, L. (2015, November). Large scale web-content classification. *In Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference* on (Vol. 1, pp. 545-554). IEEE.

Dey Sarkar, S., Goswami, S., Agarwal, A., & Aktar, J. (2014). A Novel Feature Selection Technique for Text Classification Using Naïve Bayes. *International Scholarly Research Notices*, 2014.

Dit, B., Panichella, A., Moritz, E., Oliveto, R., Di Penta, M., Poshyvanyk, D., & De Lucia, A. (2013, May). Configuring topic models for software engineering tasks in tracelab. *In Traceability in Emerging Forms of Software Engineering (TEFSE), 2013 International Workshop on* (pp. 105-109). IEEE.

Dixit, S., & Gupta, R. K. (2015). Layered Approach to Classify Web Pages using Firefly Feature Selection by Support Vector Machine (SVM). *International Journal of u-and e-Service, Science and Technology*, 8(5), 355-364.

Elberrichi, Z., & Aljohar, B. (2007). N-grams in Texts Categorization. *Scientific Journal of King Faisal University (Basic and Applied Sciences)*, 8(2), 1428H.

Fatima, S., & Srinivasu, B. (2017). Text Document categorization using support vector machine.

Giannakopoulos, G., Mavridi, P., Paliouras, G., Papadakis, G., & Tserpes, K. (2012, June). Representation Models for Text Classification: a comparative analysis over three Web document types. *In Proceedings of the 2nd international conference on web intelligence, mining and semantics* (p. 13). ACM.

Hammami, M., Chahir, Y., & Chen, L. (2003, October). WebGuard: Web based adult content detection and filtering system. *In Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on* (pp. 574-578). IEEE.

Haveliwala, T. H. (2003). Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering*, 15(4), 784-796.

Hofmann, T. (1999, August). Probabilistic latent semantic indexing. *In ACM SIGIR Forum* (Vol. 51, No. 2, pp. 211-218). ACM.

Huang, C. C., Chuang, S. L., & Chien, L. F. (2004). Using a web-based categorization approach to generate thematic metadata from texts. *ACM Transactions on Asian Language Information Processing (TALIP),* 3(3), 190-212.

Ishii, N., Murai, T., Yamada, T., & Bao,

Y. (2006, July). Text classification by combining grouping, LSA and kNN. *In Computer and Information Science, 2006 and 2006 1st IEEE/ACIS International Workshop on Component-Based Software Engineering, Software Architecture and Reuse. ICIS-COMSAR 2006. 5th IEEE/ACIS International Conference on* (pp. 148-154). IEEE.

Jindal, R., Malhotra, R., & Jain, A. (2015). Techniques for text classification: *Literature review and current trends. Webology,* 12(2), 1.

Joachims, T. (1998). Text categorization with support vector machines: *Learning with many relevant features. Machine learning: ECML*-98, 137-142.

Karima, A., Zakaria, E., Yamina, T. G., Mohammed, A. A. S., Selvam, R. P., & VENKATAKRISHNAN, V. (2012). Arabic text categorization: a comparative study of different representation modes. *Journal of Theoretical and Applied Information Technology*, 38(1), 1-5.

Kato, R., & Goto, H. (2016, March). Categorization of web news documents using word2vec and deep learning. *In Proceedings of the 2016 International Conference on Industrial Engineering and Operations Management Kuala Lumpur, Malaysia*.

Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology,* 1(1), 4-20.

Kim, J., Lee, B., Shaw, M., Chang, H., Nelson, W (2001). Application of Decision-Tree Induction Techniques to Personalized Advertisements on Internet Storefronts. *International Journal of Electronic Commerce,* 5(3), 45-62.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University,* 33(2004), 1-26.

Kojadinovic, I., & Wottka, T. (2000, September). Comparison between a filter and a wrapper approach to variable subset selection in regression problems. *In Proc. European Symposium on Intelligent Techniques (ESIT).*

Krestel, R. (2012). On the Use of Language Models and Topic Models in the Web: New Algorithms for Filtering, Classification, Ranking, and Recommendation (Doctoral dissertation).

Kwon, O. W., & Lee, J. H. (2000, November). Web page classification based on k-nearest neighbor approach. *In Proceedings of the fifth international workshop on on Information retrieval with Asian languages* (pp. 9-15). ACM.

Landauer, T. K., Dutnais, S. T., Anderson, R., Carroll, D., Fbltz, P., Pumas, G., Streeter, L. (1997). A Solution to Plato’s Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. Psychological Review, 1(2), 211–240. Retrieved from http://www.indiana.edu/~clcl/Q550_WWW/Papers/Landauer_Dumais_1997.pdf

Li, S., Xia, R., Zong, C., & Huang, C. R. (2009, August). A framework of feature selection methods for text categorization. *In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP:* Volume 2-Volume 2 (pp. 692-700). Association for Computational Linguistics.

Lilleberg, J., Zhu, Y., & Zhang, Y. (2015, July). Support vector machines and word2vec for text classification with semantic features. *In Cognitive Informatics & Cognitive Computing (ICCI* CC), 2015 IEEE 14th International Conference* on (pp. 136-140). IEEE.

Lin, Y., & Wang, J. (2014, June). Research on text classification based on SVM-KNN. *In Software Engineering and Service Science (ICSESS), 2014 5th IEEE International Conference on* (pp. 842-844). IEEE.

Ma, S., Zhang, C., & He, D. (2016). Document representation methods for clustering bilingual documents. *Proceedings of the Association for Information Science and Technology*, 53(1), 1-10.

Mangai, J. A., Kothari, D. D., & Kumar, V. S. (2012). A Novel Approach for Automatic Web Page Classification using Feature Intervals. *International Journal of Computer Science Issues* (IJCSI), 9(5).

Mangai, J. A., Kumar, V. S., & Alias Balamurugan, S. A. (2012). A novel feature selection framework for automatic web page classification. *International Journal of Automation and Computing*, 9(4), 442-448.

Masada, T., Kiyasu, S., & Miyahara, S. (2008, February). Comparing LDA with pLSI as a dimensionality reduction method in document clustering. *In LKR* (pp. 13-26).

Miao, D., Duan, Q., Zhang, H., & Jiao, N. (2009). Rough set based hybrid algorithm for text classification. *Expert Systems with Applications*, 36(5), 9168-9174.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *In Advances in neural information processing systems* (pp. 3111-3119).

Milios, E. E., Shafiei, M. M., Wang, S., Zhang, R., Tang, B., & Tougas, J. (2006). A systematic study on document representation and dimensionality reduction for text clustering. Technical report, Faculty of Computer Science, Dalhousie University.

Moiseev, G. (2016). Classification of E-commerce Websites by Product Categories. In *AIST* (Supplement) (pp. 237-247).

Nayak, J., Naik, B., & Behera, H. (2015). A comprehensive survey on support vector machine in data mining tasks: applications & challenges. *International Journal of Database Theory and Application*, 8(1), 169-186.

Oneata, D. (1999). Probabilistic Latent Semantic Analysis. *In Proceedings of the Fifteenth conference on Uncertainty* (pp. 1-7).

Oyelade, O. J., Oladipupo, O. O., & Obagbuwa, I. C. (2010). Application of k Means Clustering algorithm for prediction of Students Academic Performance. arXiv preprint arXiv:1002.2425.

Patil, A. S., & Pawar, B. V. (2012, March). Automated classification of web sites using Naive Bayesian algorithm. *In Proceedings of the international multiconference of engineers and computer scientists* (Vol. 1, pp. 519-523).

Qi, X., & Davison, B. D. (2009). Web page classification: Features and algorithms. *ACM computing surveys* (CSUR), 41(2), 12.

Raj, A. J., Francis, F. S., & Benadit, P. J. (2016). Optimal Web Page Classification Technique Based on Informative Content Extraction and FA-NBC. *Computer Science and Engineering*, 6(1), 7-13.

Ruiz, M. E., & Srinivasan, P. (1998, October). Automatic text categorization using neural networks. *In Proceedings of the 8th ASIS SIG/CR Workshop on Classification Research* (pp. 59-72).

Said, D. A. (2007). Dimensionality reduction techniques for enhancing automatic text categorization (Doctoral dissertation, Faculty of Engineering at Cairo University in Partial Fulfillment of the Requirements for the Degree of MASTER OF SCIENCE in COMPUTER ENGINEERING FACULTY OF ENGINEERING, CAIRO UNIVERSITY GIZA).

Shibu, S., Vishwakarma, A., & Bhargava, N. (2010). A combination approach for web page Classification using Page Rank and Feature Selection Technique. *International Journal of Computer Theory and Engineering*, 2(6), 897.

Singh, K. N., Devi, H. M., & Mahanta, A. K. (2017). Document representation techniques and their effect on the document Clustering and Classification: A Review. *International Journal of Advanced Research in Computer Science*, 8(5).

Sriurai, W., Meesad, P., & Haruechaiyasak, C. (2010, June). Improving Web Page Classification by Integrating Neighboring Pages via a Topic Model. *In IICS* (pp. 238-246).

Sun, A., Lim, E. P., & Ng, W. K. (2002, November). Web classification using support vector machine. *In Proceedings of the 4th international workshop on Web information and data* *management* (pp. 96-99). ACM.

Talavera, L. (2005). An evaluation of filter and wrapper methods for feature selection in categorical clustering. *Advances in Intelligent Data Analysis* VI, 742-742.

Tsukada, M., Washio, T., & Motoda, H. (2001). Automatic web-page classification by using machine learning methods. Web Intelligence: *Research and Development*, 303-313.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37, 141-188.

Wang, X., Chen, R., Jia, Y., & Zhou, B. (2013, November). Short Text Classification using Wikipedia Concept based Document Representation. *In Information Technology and Applications (ITA), 2013 International Conference on* (pp. 471-474). IEEE.

Wang, Y., Wang, X., Jiang, Y., Liang, Y., & Liu, Y. (2016). A code reviewer assignment model incorporating the competence differences and participant preferences. *Foundations of Computing and Decision Sciences*, 41(1), 77–91. https://doi.org/10.1515/fcds-2016-0004

Wang, Z., Ma, L., & Zhang, Y. (2016, June). A Hybrid Document Feature Extraction Method Using Latent Dirichlet Allocation and Word2Vec. *In Data Science in Cyberspace (DSC), IEEE International Conference on* (pp. 98-103). IEEE.

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS quarterly*, xiii-xxiii.

Wei, Z., Miao, D., Chauchat, J. H., Zhao, R., & Li, W. (2009). N-grams based feature selection and text representation for Chinese Text Classification. *International Journal of Computational Intelligence Systems*, 2(4), 365-374.

Xu, Z., Yan, F., Qin, J., & Zhu, H. (2011, October). A web page classification algorithm based on link information. *In Distributed Computing and Applications to Business, Engineering and Science (DCABES), 2011 Tenth International Symposium* on (pp. 82-86). IEEE.

Yin, D., Hu, Y., Tang, J., Daly, T., Zhou, M., Ouyang, H., ... & Langlois, J. M. (2016, August). Ranking relevance in yahoo search. In Proceedings of the 22nd ACM SIGKDD *International Conference on Knowledge Discovery and Data Mining* (pp. 323-332). ACM.

Yu, B., Xu, Z. B., & Li, C. H. (2008). Latent semantic analysis for text categorization using neural network. *Knowledge-Based Systems*, 21(8), 900-904.

Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF* IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, 38(3), 2758-2765.

Covenant Journal of Informatics & Communication Technology. Vol. 6 No. 2, Dec., 2018

An Open Access Journal Available Online

Optimized Model Simulation of a Capacitated Vehicle Routing problem based on Firefly Algorithm

Mayo Zion O[1], Muhammad Bashir Mu'azu[2],
Adewale Emmanuel Adedokun[2],
Salawudeen Ahmed Tijani[2] & Ibrahim Ahmad Bello[2]

[1]CodeLagos, Lagos State Government, Nigeria.
[1]mayozion@gmail.com
[2]Department of Computer Engineering,
Ahmadu Bello University, Zaria, Nigeria.
[2]{mbmuazu, wale, tasalawudeen, ibrahimbello}@abu.edu.ng

Abstract: This paper presents an optimized solution to a capacitated vehicle routing (CVRP) model using firefly algorithm (FFA). The main objective of a CVRP is to obtain the minimum possible total travelled distance across a search space. The conventional model is a formal description involving mathematical equations formulated to simplify a more complex structure of logistic problems. These logistic problems are generalized as the vehicle routing problem (VRP). When the capacity of the vehicle is considered, the resulting formulation is termed the capacitated vehicle routing problem (CVRP). In a practical scenario, the complexity of CVRP increases when the number of pickup or drop-off points increase making it difficult to solve using exact methods. Thus, this paper employed the intelligent behavior of FFA for solving the CVRP model. Two instances of solid waste management and supply chain problems is used to evaluate the performance of the FFA approach. In comparison with particle swarm optimization and few other ascribed metaheuristic techniques for CVRP, results showed that this approach is very efficient in solving a CVRP model.

Keywords: Optimization, CVRP, Firefly, Solid Waste Management, logistics.

1. Introduction

The rapid advancement in technologies have made logistics systems have become very important for revenue and budgetary considerations for government and its establishments, most importantly for companies in the private sector. The fact that anybody on the planet can be all around connected has prompted complex transport networks that are exceptionally requesting and are winding up progressively critical. Hence, an efficient logistic network will be beneficial to companies and relevant business operations. To highlight the importance of logistics in some sectors, like groceries delivery, online stores delivery of goods, waste management, intra-city public transportation, distribution costs can increase in the production price up to 70%. Thus, the need for vehicle routing become necessary.

Vehicle routing problem (VRP) define a class of optimization problems that involve optimizing itineraries of a fleet of vehicles. Researchers have over the years, developed a serious research interest in VRP due to its practical importance, as well as its complexity. The framework is employed in modelling an extremely broad range of logistic issues in various applications like, supply chain management, delivery services, public transportation, telecommunications and production planning.

However, because of the real-life applications and complexity of these problems, a class of optimization algorithms is used in obtaining optimal solutions. Although various VRP problems can be combined in the form of a multi-objective decision problem which consider providing convenient service distribution for demands between predefined points in the search space. The aim of this study is to maximize route optimization, minimizing the total route distance in a search space using a firefly based capacitated vehicle routing problem model (FFA-CVRP).

There are several variants to name a few are the

i. Capacitated VRP: the capacity of the vehicles is considered for the modelling of the objective function.
ii. VRP performing pickup and delivery simultaneously: models a payload being dropped off and collected at the same node point for all nodes.
iii. VRP with Mixed pickup and delivery: a payload is dropped and picked up but not necessarily from the same node.
iv. Multi depot VRP: more than one depot is considered in simulating this VRP also it can be combined with any other variants
v. VRP with access time windows: time limitations are being implemented in modelling this type of VRP hence, the deliveries are performed in pre-defined periods.

2. Methods

Instances are the arrangement or scenarios formulated by some attributes like number of customers, number of routes in some cases the route duration, the route distance all these will be discussed. Literature demonstrated that the number of point (including the depot) and the number of routes should reflect the naming and formulation of instances. An example of a naming nomenclature $E-n101-k8$ is an instance that has 1 depot, 100 customers and 8 routes. The series are usually named randomly by the authors. In the E series by (Nicos Christofides & Eilon, 1969) where locations are generated at random from a uniform distribution, some of the instances actually come from (Dantzig & Ramser, 1959) and (Gaskell, 1967) while some are modifications on the capacity suggested by (Gillett & Miller, 1974). For the M series, customers are grouped into clusters as an attempt to represent practical cases and some instances are modifications of the E series by considering increment in customers and capacity. For example, instances $M-n200-k17$ and $M-n200-k16$ differs only by the number of routes. These new instances were formulated because $M-n200-k16$ had tightness very close to 1 (0.995625) that finding any feasible solutions maybe difficult. However, the optimal solution of M-n200-k16 instance may costs less than the optimal solution of $M-n200-k17$ (Christofides *et al.*, 1979). The F series presents instances with data set from real-world applications, from grocery deliveries and delivery of goods to a gasoline service station (Fisher, 1994) etc. The A, B and P series by (Augerat *et al.*, 1998) proposed a situation where the customers and depots are randomly positioned in the A series and clustered in the B series while the demands are picked from a uniform distribution in both series. The P series are just modifications in the capacity and the routes of some instances in A, B and E. (N Christofides et al., 1979) defined a CMT benchmark set, which consists of modifications of some E and M series whereby the number of routes are not fixed. This set also has an addition of maximum route duration and service time values while the vehicles are assumed to travel at unitary speed.

Various algorithms have been applied to CVRP to name a few are: an ant colony algorithm building parallel routes other than sequential routes for its route optimization (Mazzeo *et al.*, 2004). A string model based simulated annealing algorithm is used in optimizing fuel consumption (Xiao *et al.*, 2011). A hybrid genetic algorithm and particle swarm optimization for solving a capacitated vehicle routing problem with fuzzy demand, the study used GA to modify the PSO with the hope of improving its performance and used fuzzy variables to deal with the uncertain parameters in developing the CVRP model. However, the concept of smart bin data was not implemented for the collection, yielding a limited experiment (Kuo *et al.*, 2012). A hybrid algorithm consisting of an iterated local search and a set partitioning formulation which could solve small size instances (Subramanian *et al.*, 2013). An integration of lagrangian spilt and variable neighborhood search (VNS) although its resolution is impractical for relatively large instances (Bauzid *et al.*, 2015). An architecture and intelligent sensing algorithm to detect solid waste at real time in a bin monitoring system which will contribute to solid waste collection, however the sensor sometimes produces inaccurate output data, due to the irregularities of the solid waste pattern (Al Mamun et al., 2016). A new set of Benchmark Instances proposed by (Uchoa et al., 2017) presents a more detailed and balanced experimental scenarios using iterated local search set partitioning (ILS-SP) and unified hybrid genetic search (UHGS) but the UHGS had poor quality solutions for instances of small sizes while the ILS-SP had slow convergence towards the solution for large instances. Furthermore, (Hannan et al., 2018) proposed modified PSO for a CVRP model for waste collection was initiated, the Instances were generated from the A, B and P series, a threshold waste level and scheduling concepts were implemented and however, the optimization technique used could not attain an optimal value for some instances.

2.1 Firefly Algorithm (FFA)

This algorithm is used to improve the route within the search space. It is modelled after the behavior of the flashing characteristics and movement of the Firefly. (XS Yang 2009). The Firefly algorithm (FFA) like the glow-worm swarm optimization algorithm (GISO) and the bioluminescent swarm optimization algorithm (BiSO) is in the classification of the luminous inspired insect algorithms which all belong to the Biological Inspired Algorithms. (Bo Xing and Wen-Jing Gao 2013). In this study, extracting the rules of the FFA, the ideology of the algorithm in relationship to CVRP are as follows: The nodes have high mobility due to the versatility in attractiveness variations, hence, the search space is explored more efficiently i.e. The best route will be more efficiently identified and exploited for vehicles to deliver to customers. The brightness is proportional to the attractiveness. i.e. a less bright firefly will move towards a brighter one. Thus, considering the fitness at each stage of motion, for each iteration, the nodes move to get a better result dropping the previous result to be replaced and continues until the maximum iteration is reach where there are no brighter fireflies, it searches randomly. The nodes represent each firefly. Finally, all fireflies are considered as unisex, one firefly will be attracted to other fireflies regardless of their gender which means the nodes can be heterogeneous relating to vehicles and the customers and still function on the model.

The distinction of light intensity and creation of the attractiveness are two critical issues in the FFA.

The attractiveness of a firefly is determined by its brightness which is a function of the objective function. Usually, the brightness I at a location **x** can be chosen as $I(x)\alpha f(x)$. In a scenario where the light absorption coefficient γ is fixed, the light intensity I vary with the distance r, where I_0 is the original light intensity. To eliminate the singularity problem at $r = 0$ in the expression I_s/r^2 where, I_s is source light intensity, the combined effect of both the absorption and inverse square law can be approximated using the Gaussian form (Arora & Singh, 2013).

$$I(r) = I_0 e^{-\gamma r^2} \qquad (1)$$

The attractiveness β of a given firefly is relative, since its proportional to light intensity of a pre-established firefly (Yang, 2010). Thus, leads to a variation with the distance r_{ij} between firefly i and firefly j. Hence, with an increase in the distance from its source, there is a measurable decrease in the light intensity, and light, is absorbed in the transmission so the attractiveness will vary with the degree of absorption, where, β_0 connotes the attractiveness at $r = 0$.

$$\beta(r) = \beta_0 e^{-\gamma r^m} \qquad (2)$$

The distance between two fireflies i and j at x_i and x_j, is represented as the cartesian distance where x_{ik} is the kth element of the spatial coordinate xi of ith firefly (Yang & Deb, 2010).

$$r_{ij} = \left\| x_i - x_j \right\| = \sqrt{\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2} \qquad (3)$$

The movement of a firefly i which is attracted to a firefly j with higher attractiveness (brightness) is determined by equation (4).

$$x_i = x_i + \beta_0 e^{-\gamma r_{ij}^{2}}(x_j - x_i) + \alpha(rand - \frac{1}{2}) \quad (4)$$

The second segment of equation (4) is due to the attraction while the third segment is randomization with *α* being the randomization parameter (Sayadi *et al.*, 2010)

2.2 CVRP

Capacitated vehicle routing problem defines the optimal set of routes for a fleet of vehicles to navigate from a depot to a specified set of customers ensuring the vehicle capacity is not exceeded. Figure 1 shows an instance of a capacitated vehicle routing problem. The figure contains 77 nodes (bins), with 1 depot located at the center of the grid across 10 routes (they are segmented in different color codes). Where a vehicle takes off from a depot, moves from one node to another and back to the depot, over a certain distance to form a route.

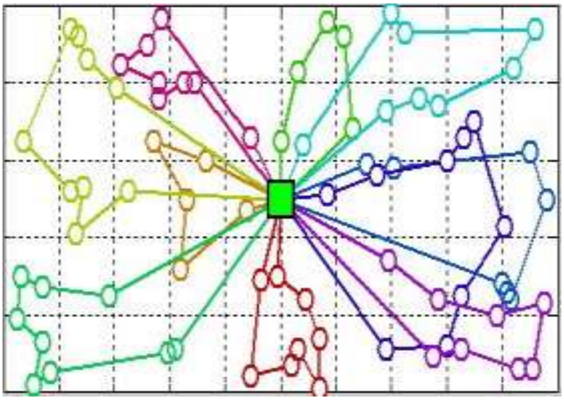


Figure 1. A Scenario of CVRP

The basic concept of VRP is to serve a set of customers to find the least travelled distance but when the capacity of the vehicle is factored, it becomes CVRP. The objective of this model is to develop an optimized routing scheme in other to determine a viable route that minimizes the total distance travelled by the vehicles which invariably reduces the total cost. There are some constraints accredited to the modelling of a CVRP explained in this study. Where *N* is the number of customers, a nonnegative distance cost d_{ij} represents distance from bins i to j, where $i \neq j$. A set of homogenous vehicles $k = \{1,2,...,K\}$ is available at the depot to either collect or deliver demand as the case maybe.

A route is established by the summation of multiple links. A link is formed with the notation P_{ij}^{k} which moves from customer i through to customer j, by a vehicle *k*, where the decision variables are dependent of the vehicle capacity and the customer demand which are modelled as follows:

$$P_{ij}^{k} = \begin{cases} 1, \ if\ vehicle\, travels\ from\ customer\ i\ to\ j \\ 0, \ if\ otherwise \end{cases} \quad (5)$$

The variables take only the integer (s) 0, 1 because the number of customers, vehicles and route cannot be a fraction,

$$P_{ij}^{k} \in \{0,1\},\ j = 0,1,2,...,N;k = 1,2,...,K \quad (6)$$

All vehicles begin and end at the depot i.e. each vehicle isn't used more than once,

$$\sum_{\substack{i=1\\j=1}}^{N} P_{\substack{ai\\ja}}^{k} \leq 1,\ \ k = 1,2,...,K \quad (7)$$

The vehicle must not be re-used, the inequality considers when a vehicle is also not being used at all, out of the pool of vehicles at the depot. When all vehicles are used, the expression will be an equal sign. Where *a* represents the depot.

A customer is visited once, by only one vehicle each time,

$$\sum_{k=1}^{K}\sum_{\substack{i=0\\j=0}}^{N} P_{ij}^{k} = 1, \quad j = 1,2,...,N;\ i = 1,2,...,N \qquad (8)$$

There must be route continuity,

$$\sum_{i=0}^{N} P_{it}^{k} - \sum_{i=0}^{N} P_{tj}^{k} = 0, \quad k = 1,2,...,N \qquad (9)$$

A route distance has a limit (not exceeding the total travel distance)

$$\sum_{i=0}^{N}\sum_{j=0}^{N} d_{ij}^{k} P_{ij}^{k} \le D_{k} \quad k = 1,2,...,K \qquad (10)$$

The number of routes and vehicles must be above 1, else the model becomes a TSP and not a VRP, where the former deals with a vehicle and a single route.

$$\sum_{i=2}^{N}\sum_{j=2}^{N} P_{ij}^{k} > 1 \quad i,j = 2,3,...,N; k = 2,3,...,K \qquad (11)$$

The capacity of the vehicle must not exceed its maximum, there must be no overloading,

$$Q_{k} \le Q_{k_{\max}} \quad k = 1,2,...,K \qquad (12)$$

The total demand q_T on each route must not exceed the vehicle capacity,

$$\sum_{j=0}^{N} q_{j}\left(\sum_{i=0}^{N} P_{ij}^{k}\right) \le Q_{k}, \ k = 1,2,...,k \qquad (13)$$

All the demand must be accomplished,

$$q_{T} = \begin{cases} \sum_{j=0}^{N} q_{j}\left(\sum_{i=0}^{N} P_{ij}^{k}\right) & if \ \ q_{ij} \ne 0 \\ 0 & if \ \ Otherwise \end{cases} \qquad (14)$$

The total cost and travel distance are minimized,

$$S = \min \sum_{i=1}^{N}\sum_{j=0}^{N}\sum_{k=1}^{K} d_{ij} P_{ijk} \qquad (15)$$

In the implementation of these constraints there are some parameters to consider

Vehicle capacity: this is the ability of the vehicle to accommodate a certain amount of payload without an overload.

Number of vehicles: One major difference between the TSP (travelling salesman problem) and VRP (vehicle routing problem) is that in the latter, more than one vehicle is used to visit the customers in the search space. The number of vehicles to be used for a VRP determines the speed at which customers can be served and also contributes in achieving a shorter service time.

Demand: This is the amount of payload that is required by the customer(s), which inevitably determines the number of vehicles to be used in a specified space to oblige with the constraints where, the total demand for every route, must not exceed the capacity of the vehicle.

Number of customers: the number of customers that are involved in the logistics is a prime factor as it can be used to guide a model in determining the other parameters and variables dependent on the design. It is assumed that the number of customers equals the number of nodes.

Customer positioning: the positions and locations of customers are paramount in the result of an optimum solution because factors like distance and distribution plays part in the architecture and modelling of the solution method. Customers can be positioned randomly, in clusters or both cases. In this study, customers will be randomly positioned.

Route size: this is the number of routes that the distribution can be sectioned into.

Route distance: this is the length of the course taken from the depot to the serve a set of customers and back to the depot. It is the dimension of travel which will determine the total time taken and also the optimum solution for that given set of instances. Although some methods are best used for shorter distances while some for long distances, but in this work will create a common ground for such uprising.

3. CVRP Optimization Using Firefly Algrithm

The firefly based technique simply solves the CVRP model by identifying the nodes (customer points) as the stationary fireflies and a vehicle as the moving fireflies. Evaluating all the points and the given parameters. Then, the vehicles are evaluated knowing which one is to be assigned to which route, after which it is attracted to the nearest customer location guided by the set constraints. This process continues until the CVRP is solved. Illustrations in this research shows two scenarios. First is a total of thirty-six cases of waste management problem and ten cases of supply chain problem was used to validate the model. This information was used along with the parameters for the optimization of the CVRP model as described in subsection 2.2. The total cost and travel distance of the CVRP described in equation (15) was then optimized using the firefly optimization algorithm.

The simulation parameters showing the range of values used to achieve the results for both the Solid Waste Management and Retail Supply Chain are quantified in the given Table below.

Table 1: Simulation Parameters

SN	Parameters	Values	Units
1	Number of customers, *N*	2 - 10	--
2	Number of Vehicles, *V*	11 - 100	--
3	Capacity of vehicle, *Q*	100 - 400	kg
4	Capacity / Quantity of demand, *q*	10	kg
5	Travelled distance, *d*	20 - 1500	km
6	Iteration (SWM & Supply Chain)	120 & 500	--

In developing the Optimized routing scheme for the CVRP model, the parameters vehicle capacity (*Q*), number of customers (*N*) which correspond to the number of fireflies, number of vehicles (*V*) which correspond to the search dimensions and the quantity of load (*q*) were initialized. The parameters of the FFA algorithm which are the initial customer points (*i*), the next customer point (*j*), number of iterations, and population were also initialized.

The fitness of these initial positions was evaluated, and each firefly are ranked according to their fitness. The vehicle moves from firefly *i* to firefly *j* and progresses in that order from the initial customer points (*i*), the next customer point (*j*), to the next point (*i+1),* then to (*j+1*) until the maximum number of fireflies is reached.

The FFA solution search process was then performed in an enclosed loop and the fitness of the new positions were evaluated. The entire process was then evaluated over a number of iterations continuously until the maximum number of iteration is reached and the firefly with the overall best position is taken as the optimum solution as structured in Fig. 2.

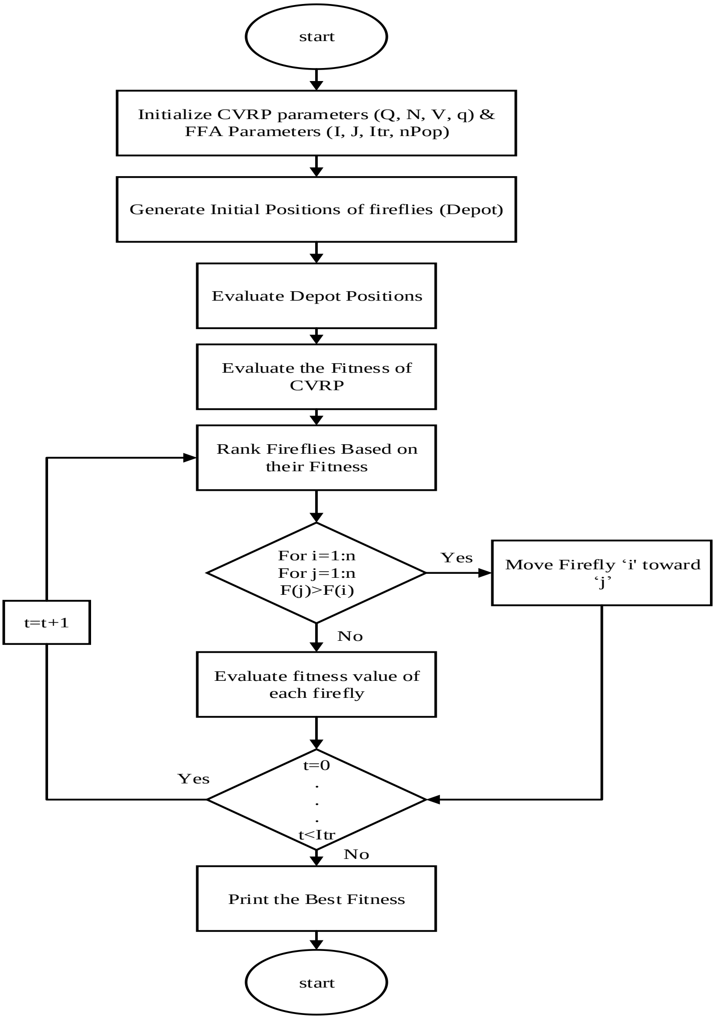

start
Initialize CVRP parameters (Q, N, V, q) & FFA Parameters (I, J, Itr, nPop)
Generate Initial Positions of fireflies (Depot)
Evaluate Depot Positions
Evaluate the Fitness of CVRP
Rank Fireflies Based on their Fitness
For i=1:n
For j=1:n
F(j)>F(i)
Yes
Move Firefly 'i' toward 'j'
t=t+1
No
Evaluate fitness value of each firefly
Yes
t=0
.
.
.
t<Itr
No
Print the Best Fitness
start

Fig 2. Flowchart of the FFA-CVRP model

Fig.2 shows the flow of the processes involved in the FFA implementation featured in a Flowchart.

4. Results and Discussions

The simulation was conducted in MATLAB R2015b environment, on a computer with Intel Core i3 @ 2.00GHz Processor with 4GB RAM. The main objective of this study is to minimize the total route distance applying all the constraints and using the parameters as earlier explained. It is assumed that a reduction in the total route distance, connotes a reduction in cost and time.

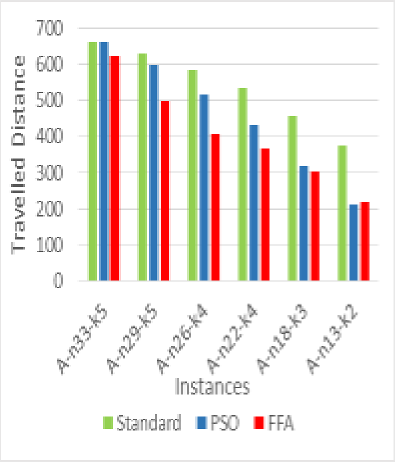
Table 2 below shows the actual values used in formulating the thirty-six instances featured in (Hannan et al., 2018).

Table 2. Result of FFA on CVRP Model for Instances of Solid Waste Management

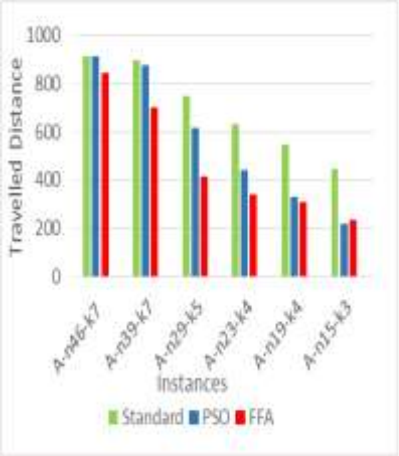
5	Datasets	Q (unit)	q (unit)	TWL (%)	N	V	Distance				
							FFA	Standard	Improvement (%)	PSO	Improvement (%)
1	A-n33-k5	100	10	0	32	5	622	661	5.87	661	5.87
2				60	28	5	499	629	20.6	599	16.63
3				70	25	4	407	585	30.51	518	21.52
4				75	21	4	367	533	31.12	430	14.62
5				80	17	3	304	457	33.48	316	3.8
6				90	12	2	219	374	41.57	212	-3.07
7	A-n46-k7	100	10	0	45	7	842	914	7.82	914	7.82
8				60	38	7	699	895	21.91	876	20.22
9				70	28	5	413	750	44.94	615	32.86
10				75	22	4	339	634	46.53	440	22.96
11				80	18	4	310	548	43.51	329	5.91
12				90	14	3	235	449	47.59	221	-6.47
13	A-n60-k9	100	10	0	59	9	1121	1371	18.21	1371	18.21
14				60	41	8	909	1258	27.75	1154	21.24
15				70	38	8	834	1223	31.81	1091	23.56
16				75	31	6	663	1048	36.77	801	17.27
17				80	29	6	528	979	46.04	699	24.43
18				90	19	4	317	693	54.19	350	9.29
19	P-n40-k5	140	10	0	39	5	359	458	21.62	458	21.62
20				60	34	4	345	417	17.27	380	9.21
21				70	32	4	334	388	13.92	329	-1.52
22				75	25	4	333	352	5.4	271	-22.88
23				80	18	3	266	294	9.52	189	-40.74
24				90	12	2	192	232	17.24	118	-62.71
25	B-n78-k10	100	10	0	77	10	1091	1263	13.6	1263	13.6
26				60	54	9	828	1124	26.33	1000	17.19
27				70	43	8	732	1069	31.49	912	19.69
28				75	27	6	409	732	44.16	424	3.6
29				80	21	4	304	613	50.41	298	-2.01
30				90	11	2	111	346	68.01	95	-16.52

31	P-n101-k4	400	10	0	100	4	489	705	30.64	705	30.64
32				60	81	4	442	616	28.25	538	17.84
33				70	70	4	436	564	22.7	451	3.33
34				75	62	3	424	545	22.2	421	-0.71
35				80	55	3	411	494	16.8	346	-18.79
36				90	33	2	193	351	45.01	175	-10.29

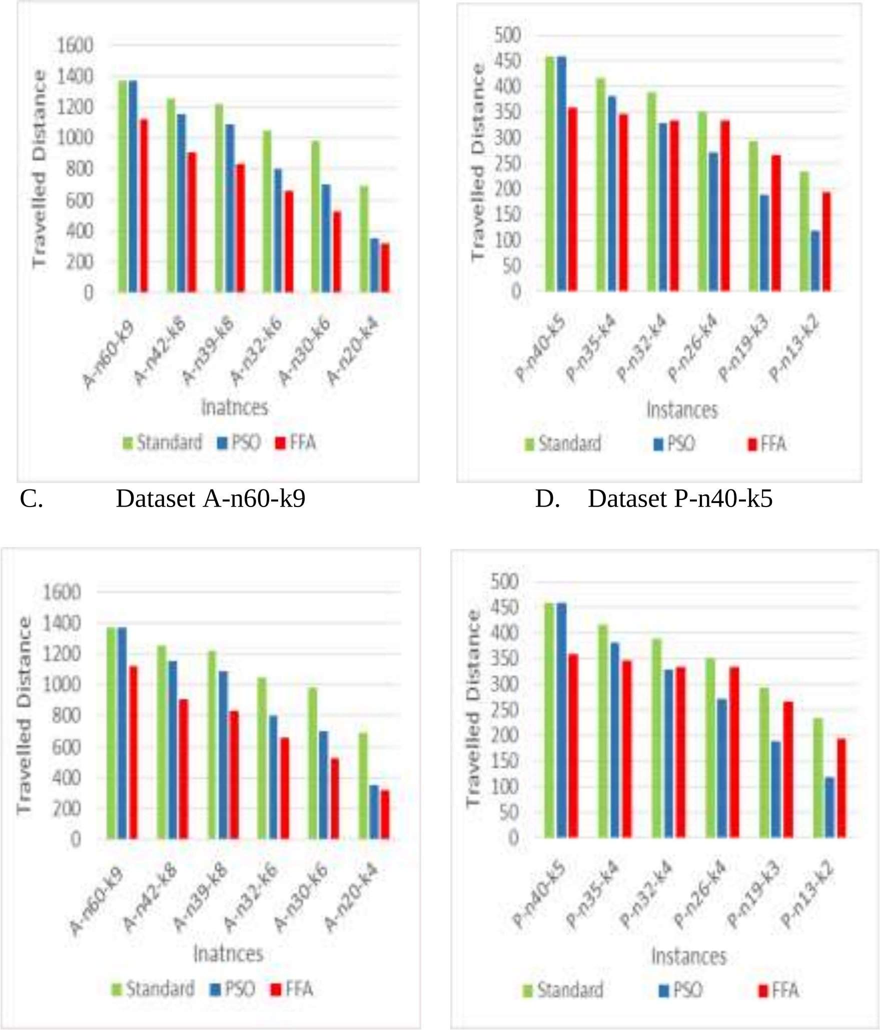
The result obtained using the FFA on the CVRP model shows improvement on the distance across all instances. Each set of instances has same capacity of all vehicles while the number of service points and TWL (quantity of demand) varies. The TWL which is the threshold waste level, provides the information on the actual percentage filled capacity of the bin. As the number of nodes (bins) decreases, the route length logically decreases, it is expected that the distance decreases, thus fewer vehicles are needed. Although, the customer positions are randomly located. Each improvement is realized by the percentage difference between the FFA acquired distance and the standard from literature. When comparing with the standard result from (Hannan et al., 2018), the first set of instances A-n33-k5, gives a collective improvement of 27.19%, A-n46-k7 gives a collective improvement of 35.39%, A-n60-k9 gives a collective improvement of 35.80%, P-n40-k5 gives a collective improvement of 14.16%, B-n78-k10 gives a collective improvement of 39.00% and P-n101-k4 gives a collective improvement of 27.60%.



A. Dataset A-n33-k5



B. Dataset A-n46-k7



C. Dataset A-n60-k9

D. Dataset P-n40-k5

E. Dataset B-n78-k10

F. Dataset P-n101-k4

Fig. 3 Plot of Travelled distance against the Instances

The collective improvement is the average of the individual improvement in each set of instances. From the table above, using the FFA metaheuristic approach, it is observed that there is total improvement on all instances, this interprets a reduced total route distance. When comparing with the result from the PSO technique from (Hannan et al., 2018), the percentage difference between the FFA acquired distance and the PSO approach is the % improvement of the FFA based model. For the first set of instances A-n33-k5 has an improvement of 9.89%, A-n46-k7 has an improvement of 13.88%, A-n60-k9 has an improvement of 19.00%, P-n40-k5 gives a decline of -16.17%, B-n78-k10 has an improvement of 5.93% and P-n101-k4 has an improvement of 3.67%. From the table above, using the FFA metaheuristic approach, it is observed that five out of the six set of instances have substantial improvements on the total route distance.

From Table 2, as the number of vehicles and customer points decrease, even with an increasing threshold waste level (TWL) from 0 – 90%, the total travel distance reduces. This is because technically, with a smaller number of vehicles and customers interprets a smaller number of routes which invariably gives a reduced travelled distance. Of the 36 instances where the FFA-CVRP model is tested on, the FFA has a 72% better results over the PSO. The graphical representation of these result can be seen in Fig 3.

The developed model in this research was validated using the iterated local search with set partitioning (ILS-SP), unified hybrid genetic search (UHGS) and branch and cut price (BCP) methods presented in the work of (Uchoa et al., 2017). The data from the result is analyzed in the Table below.

Table 3. Result of FFA on CVRP Model for Instances of Supply Chain

		Instance Characteristics				Travelled distance achieved through Metaheuristic					Improvement	
#	Name	n	Dep	Cust	Q	ILS-SP	UHGS	BCP	BKS	FFA	Distance	(%)
1	X-n101-k25	100	R	RC (7)	206	27591	27591	27591	27591	**22572**	5019	18.19
2	X-n153-k22	152	C	C (3)	144	21340	21220	21140	21140	**20538**	602	2.85
3	X-n200-k36	199	R	C (8)	402	58626	58578	58455	58455	**52052**	6403	10.95
4	X-n303-k21	302	C	C (8)	794	21812	21748	21546	21546	**19784**	1762	8.18
5	X-n401-k29	400	E	C (6)	745	66453	66243	65971	65971	**60194**	5777	8.76
6	X-n502-k39	501	E	C (3)	13	69284	69253	69120	69120	**65785**	3335	4.82
7	X-n613-k62	612	C	R	523	60229	59778	59323	59323	**55361**	3962	6.68
8	X-n701-k44	700	E	RC (7)	87	82888	82293	81694	81694	**78617**	3077	3.77
9	X-n801-k40	800	E	R	20	73830	73587	73124	73124	**70175**	2949	4.03
10	X-n1001-k43	1000	R	R	131	73776	72742	71812	71812	**67927**	3885	5.41

Table 3 shows the outcome the FFA-CVRP model on the supply chain instances. These set of instances is used to validate the FFA approach on the CVRP model. The result obtained from the FFA-CVRP simulation is compared to the best-known solution amongst iterated local search-set partitioning (ILS-SP), the unified hybrid genetic search (UHGS), the branch and cut price (BCP) methods which were used on the Instances (Uchoa et al., 2017). In this scenario, demand is dropped-off at each customer site, unlike the solid waste management where demand is picked. In Table 3, it is observed that in all cases there were improvements in the result. The Table depicts the BKS that was obtained considering the previously used three algorithms (ILS-SP, UHGS and BCP). The BKS was then used to compare the results obtained by the FFA. It is seen that applying the FFA on the CVRP model minimized the total travelled distance for X-n101-k25 by 5019m, for X-n153-k22 by 602m, for X-n200-k36 by 6403m, for X-n303-k21 by 1762m, for X-n401-k29 by 5777m, for X-n502-k39 by 3335m, for X-n613-k62 by 3962m, for X-n701-k44 3077m, for X-n801-k40 2949m and for X-n1001-k43 by 3885m. This result was then implemented to calculate the percentage improvement for each of the Instances considered. Although, a slight percentage is observed in the improvement, this is because the distance covered is large, hence, the percentage difference compared to the largely covered distance will not have a high magnitude.

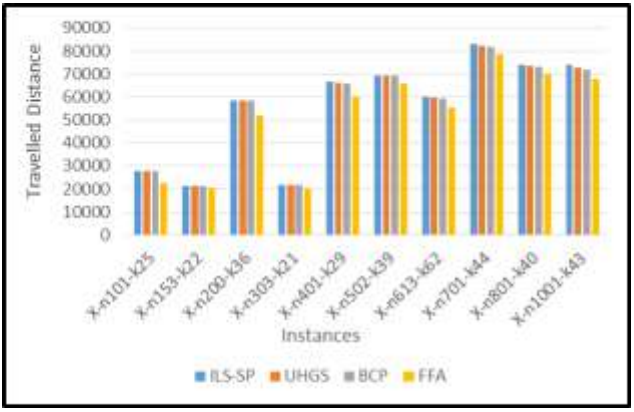


Fig 4 Best Known Solution for supply chain Instances

Fig 4 shows the plot of the travelled distance against the Instances for the supply chain. The result of FFA outperforms the Best Known Solution among the algorithms used in (Uchoa et al., 2017). For all the 10 instances in serving 100 to 1000 customers, it is certified that the FFA now provides the new best known solution (BKS) amongst the four techniques tested on the Instances. In order to further evaluate the performance of the developed method, the performance test given in Fig. 5 was generated.

Results for Performance of the FFA-CVRP Model

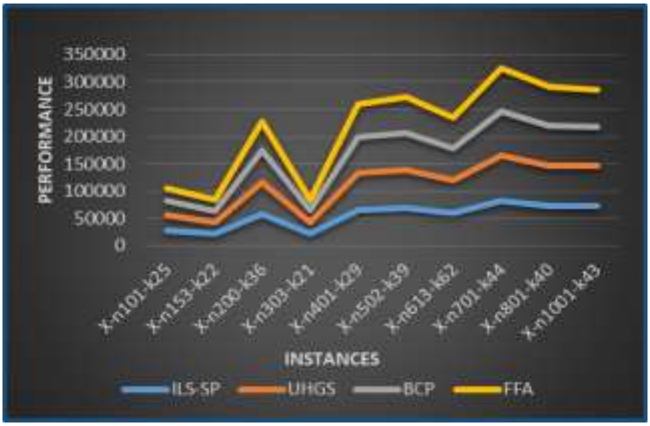


Fig 5 Perfromance of the FFA-CVRP Model

Fig 5 shows a graphical representation of the performance of the FFA-CVRP Model which has shown to provide better results over the other methods used in solving both large and small scale instances for supply chain across all instances.

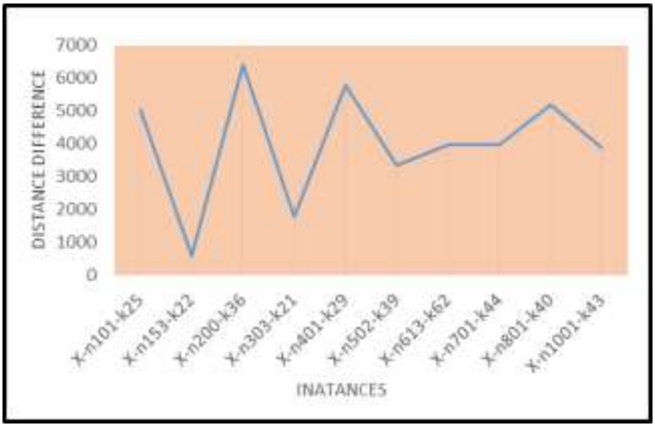


Fig 6 Performance of the Instances

Fig 6 shows the performance if each of the instances, deduced from the difference in the BKS of (Uchoa et al., 2017) and the FFA-CVRP Model. The instance X-n200-k36 has the highest value, which means the FFA-CVRP model is able to navigate channels faster and better with efficient productivity to obtain a more improved solution. This is due to the low ratio of the number of routes and the vehicle capacity to the number of customers and their demand distribution. The dip in the X-n153-k22 instance, shows it possess the lowest difference between the BKS of the earlier techniques used and the FFA based model.

The developed capacitated vehicle routing model using firefly algorithm significantly improved the total route distance on both large and small sized instances. For the solid waste management instances, the FFA-CVRP model contributed an overall improvement of 29.86% to the standard method and a 6.03% over PSO. The model outperformed the best known solution of the ILS, UHGS and BCP approach used on the set of instances for supply chain with an average improvement of 7.36%. All the observations were made assuming same conditions as other techniques used. The developed model achieved a distinct travel path and search in actualizing the best route and position to locate a depot.

5. Conclusion

This paper has presented an optimization of a capacitated vehicle routing model using firefly algorithm. The paper employed two instances comprising of waste management problem and supply chain problem to evaluate the performance of the developed approach. Several simulations were performed using MATLAB R2015b simulation environment. Results when compared with particle swarm optimization, iterated local search set partitioning, unified hybrid genetic search and branch and cut price approaches, showed that this approach is very effective in solving CVRP of different cases. For future research, modelling the time windows to the customer availability, considering the effect of variable positions of depot and hybridizing FFA with other algorithms such as smell agent optimization (SAO) for improved performance can be considered.

Acknowledgement

The authors would like to acknowledge the inputs and support provided by the Computer and Control Engineering Research Group of Ahmadu Bello University, Zaria. Their technical help brought about the successful modelling and system design.

References

Agatz, N., Campbell, A., Fleischmann, M., & Savelsbergh, M. (2011). Time slot management in attended home delivery. Transportation Science, 45(3), 435-449.

Al Mamun, M. A., Hannan, M., Hussain, A., & Basri, H. (2016). Theoretical model and implementation of a real time intelligent bin status monitoring system using rule based decision algorithms. Expert Systems with Applications, 48, 76-88.

Archetti, C., Fernández, E., & Huerta-Muñoz, D. L. (2017). The flexible periodic vehicle routing problem. Computers & Operations Research, 85, 58-70.

Arora, S., & Singh, S. (2013). The firefly optimization algorithm: convergence analysis and parameter selection. International Journal of Computer Applications, 69(3).

Augerat, P., Belenguer, J. M., Benavent, E., Corberán, A., Naddef, D., & Rinaldi, G. (1998). Computational results with a branch-and-cut code for the capacitated vehicle routing problem.

Bianchessi, N., Drexl, M., & Irnich, S. (2017). The Split Delivery Vehicle Routing Problem with Time Windows and Customer Inconvenience Constraints.

Bocewicz, G., Banaszak, Z., & Nielsen, I. (2017). Delivery-flow routing and scheduling subject to constraints imposed by vehicle flows in fractal-like networks. Archives of Control Sciences, 27(2), 135-150.

BoussaïD, I., Lepagnot, J., & Siarry, P. (2013). A survey on optimization metaheuristics. Information Sciences, 237, 82-117.

Bouzid, M. C., Haddadene, H. A., & Salhi, S. (2017). An integration of Lagrangian split and VNS: The case of the capacitated vehicle routing problem. Computers & Operations Research, 78, 513-525.

Budzianowski, W. M. (2016). A review of potential innovations for production, conditioning and utilization of biogas with multiple-criteria assessment. Renewable and sustainable energy reviews, 54, 1148-1171.

Cattaruzza, D., Absi, N., Feillet, D., & González-Feliu, J. (2017). Vehicle routing problems for city logistics. EURO Journal on Transportation and Logistics, 6(1), 51-79.

Chen, A.-l., Yang, G.-k., & Wu, Z.-m.

(2006). Hybrid discrete particle swarm optimization algorithm for capacitated vehicle routing problem. Journal of Zhejiang University-Science A, 7(4), 607-614.

Christofides, N., & Eilon, S. (1969). An algorithm for the vehicle-dispatching problem. Journal of the Operational Research Society, 20(3), 309-318.

Christofides, N., Mingozzi, A., & Toth, P. (1979). Loading problems. N. Christofides and al., editors, Combinatorial Optimization, 339-369.

Dantzig, G. B., & Ramser, J. H. (1959). The truck dispatching problem. Management science, 6(1), 80-91.

Du, K.-L., & Swamy, M. (2016). Search and optimization by metaheuristics: Springer.

Eberhart, R., & Kennedy, J. (1995). A new optimizer using particle swarm theory. Paper presented at the Micro Machine and Human Science, 1995. MHS'95., Proceedings of the Sixth International Symposium on.

Erdoğan, S., & Miller-Hooks, E. (2012). A green vehicle routing problem. Transportation Research Part E: Logistics and Transportation Review, 48(1), 100-114.

Fisher, M. L. (1994). Optimal solution of vehicle routing problems using minimum k-trees. Operations research, 42(4), 626-642.

Gaskell, T. (1967). Bases for vehicle fleet scheduling. Journal of the Operational Research Society, 18(3), 281-295.

Gillett, B. E., & Miller, L. R. (1974). A heuristic algorithm for the vehicle-dispatch problem. Operations research, 22(2), 340-349.

Grangier, P., Gendreau, M., Lehuédé, F., & Rousseau, L.-M. (2017). A matheuristic based on large neighborhood search for the vehicle routing problem with cross-docking. Computers & Operations Research, 84, 116-126.

Hannan, M., Akhtar, M., Begum, R., Basri, H., Hussain, A., & Scavino, E. (2018). Capacitated vehicle-routing problem model for scheduled solid waste collection and route optimization using PSO algorithm. Waste Management, 71, 31-41.

Haruna, Z., Mu'azu, M. B., Abubilal, K. A., & Tijani, S. A. (2017, November). Development of a modified bat algorithm using elite opposition—Based learning. In Electro-Technology for National Development (NIGERCON), 2017 IEEE 3rd International Conference on (pp. 144-151). IEEE.

Hasle, G., & Kloster, O. (2007). Industrial vehicle routing Geometric modelling, numerical simulation, and optimization (pp. 397-435): Springer.

Hernandez, F., Gendreau, M., & Potvin, J. Y. (2017). Heuristics for tactical time slot management: a periodic vehicle routing problem view. International Transactions in Operational Research.

Innocente, M., & Sienz, J. Particle Swarm Optimization:

Development of a General-Purpose Optimizer. Paper presented at the Proceedings of the 6th ASMO UK Conference on Engineering Design Optimization.

Kennedy, J. (2011). Particle swarm optimization Encyclopedia of machine learning (pp. 760-766): Springer.

Kirci, P. (2016). An optimization algorithm for a capacitated vehicle routing problem with time windows. Sādhanā, 41(5), 519-529.

Kumar, S. N., & Panneerselvam, R. (2012). A survey on the vehicle routing problem and its variants. Intelligent Information Management, 4(03), 66.

Kuo, R., Zulvia, F. E., & Suryadi, K. (2012). Hybrid particle swarm optimization with genetic algorithm for solving capacitated vehicle routing problem with fuzzy demand–A case study on garbage collection system. Applied Mathematics and Computation, 219(5), 2574-2588.

Lin, C., Choy, K. L., Ho, G. T., Chung, S. H., & Lam, H. (2014). Survey of green vehicle routing problem: past and future trends. Expert Systems with Applications, 41(4), 1118-1138.

Mahmoudi, M., & Zhou, X. (2016). Finding optimal solutions for vehicle routing problem with pickup and delivery services with time windows: A dynamic programming approach based on state–space–time network representations. Transportation Research Part B: Methodological, 89, 19-42.

Moh, Y. C., & Manaf, L. A. (2014). Overview of household solid waste recycling policy status and challenges in Malaysia. Resources, Conservation and Recycling, 82, 50-61.

Osaba, E., Yang, X.-S., Diaz, F., Onieva, E., Masegosa, A. D., & Perallos, A. (2017). A discrete firefly algorithm to solve a rich vehicle routing problem modelling a newspaper distribution system with recycling policy. Soft Computing, 21(18), 5295-5308.

Pecin, D., Pessoa, A., Poggi, M., & Uchoa, E. (2017). Improved branch-cut-and-price for capacitated vehicle routing. Mathematical Programming Computation, 9(1), 61-100.

Penna, P. H. V., Afsar, H. M., Prins, C., & Prodhon, C. (2016). A Hybrid Iterative Local Search Algorithm for The Electric Fleet Size and Mix Vehicle Routing Problem with Time Windows and Recharging Stations. IFAC-PapersOnLine, 49(12), 955-960.

Ralphs, T. K., Kopman, L., Pulleyblank, W. R., & Trotter, L. E. (2003). On the capacitated vehicle routing problem. Mathematical programming, 94(2-3), 343-359.

Sayadi, M., Ramezanian, R., & Ghaffari-Nasab, N. (2010). A discrete firefly meta-heuristic with local search for makespan minimization in permutation flow shop scheduling problems. International Journal of Industrial Engineering Computations, 1(1), 1-10.

Subramanian, A. (2012). Heuristic, exact and hybrid approaches for vehicle routing problems. Universidad Federal Fluminense. Tesis Doctoral. Niteroi, 13.

Subramanian, A., Uchoa, E., & Ochi, L. S. (2013). A hybrid algorithm for a class of vehicle routing problems. Computers & Operations Research, 40(10), 2519-2531.

Uchoa, E., Pecin, D., Pessoa, A., Poggi, M., Vidal, T., & Subramanian, A. (2017). New benchmark instances for the capacitated vehicle routing problem. European Journal of Operational Research, 257(3), 845-858.

Veenstra, M., Roodbergen, K. J., Coelho, L. C., & Zhu, S. X. (2016). A simultaneous facility location and vehicle routing problem arising in health care logistics in the Netherlands: CIRRELT.

Xiao, Y., Zhao, Q., Kaku, I., & Xu, Y. (2012). Development of a fuel consumption optimization model for the capacitated vehicle routing problem. Computers & Operations Research, 39(7), 1419-1431.

Yang, X.-S. (2009). Firefly algorithms for multimodal optimization. Paper presented at the International symposium on stochastic algorithms.

Yang, X.-S. (2010). Firefly algorithm, stochastic test functions and design optimisation. International Journal of Bio-Inspired Computation, 2(2), 78-84.

Yang, X.-S., & Deb, S. (2010). Eagle strategy using Lévy walk and firefly algorithms for stochastic optimization Nature Inspired Cooperative Strategies for Optimization (NICSO 2010) (pp. 101-111):

Covenant Journal of Informatics & Communication Technology. Vol. 6 No. 2, Dec., 2018

An Open Access Journal Available Online

Security Algorithm for Preventing Malicious Attacks in Software Defined Network (SDN)

Oluwasogo Adekunle Okunade [1], Oluwaseyi Osunade [2] & Emmanuel GbengaDada [3]

[1] Department of Computer Science, Faculty of Sciences,
National Open University of Nigeria, Abuja, Nigeria.
[2]Department of Computer Science, University of Ibadan,
Ibadan, Nigeria.
[3]Department of Computer Engineering, Faculty of Engineering,
University of Maiduguri, Nigeria.

aokunade@noun.edu.ng1, seyiosunade@gmail.com[2],
gbengadada@unimaid.edu.ng[3]

Abstract—This paper explores the success record of the Internet as well as its shortcoming in the area of network configuration, response to fault(s), load and change(s) that led to the concept of Software Defined Network (SDN).These are the factors that separated combined network's control from forwarding planes for easier optimization, programming of network and centralization of control logic capabilities. These had also led to new different challenges, that open doors for new threats that were not existing or harder to exploit. SDN prototype embraces third-party improvementas a result of hard work, that later makes the SDN vulnerable to potential trust issue on its applications (apps).This makes it possible for an intruder toinsert malicious content/programs into the network packets and then forward into the network.Codes were written to implement the designed algorithm using white/blacklist source identification combined with Hash Bayes' Theorem (W/B+HBT) content filter as a security measure to prevent the malicious attack(s). It was shown that new transaction(s) from known attack source(s) are classified as Blacklist and dropped, while those known as whitelist are forwarded to their respective destination as a legitimate packet(s) (W/B). Those from unknown sources were treated using Hash Bayes' Theorem (HBT) content filter. The result of the implementation is able to record 10% false positive (FP) and false negative (FN) and 90% true positive (TP) and true negative (TN) (accurate classification of packets) for the presented algorithm.

Keywords/Index Terms— OpenFlow, Flow table, Control plane, Hash Bayes' Theorem, Security Algorithm

1. Introduction

Software Defined Network (SDN)is anemerging innovative technology for enabling open programmable network environment to realize network with efficient and dynamic nature. it isdynamic, manageable, inexpensive network components and high-speed network emerging services according (Yutaka, Hung-Hsuan&Kyoji, 2013 and Raphael, Dietmar&Mark, 2015).Before the advent of dynamic nature SDN, the complexities of traditional computer networks were being managed with theadding of more protocols suites to meet up with the required expectation despite its complexity according to (Muhammad *et al.*,2014).Open Networking Foundation (ONF) is a profitless organization dedicated to the development,standardization, and commercialization of SDN according to (Wenfeng*et al.*, 2015). However, the openness of the SDN has resulted in security challenges that could jeopardize its purpose of existence if left unaddressed. This had made security a major concerned for SDN, as a result of its distinguishing features, conventional network security approaches cannot be directly applied. For the fact thatSDN improves network performance, yet it creates some peculiar challenges due to its centralized control and programmability features. It introduces security control challenges(Diego *et al.*,2013; Phillip*et al.*, 2012; Ali*et al.*, 2015) in Matthew, Mahamadou*et al.*,(2016).SDN can be seen as an eye-catchinghoneypot for intruders and a source of challenges for less equipped network operators such asamplifiedprospective for denial-of-service (DoS) attacks.OpenFlow is exposed to man-in-the-middle attacks when Transport Layer Security (TLS) is not used and network breaches may result when network controllersare shared by multiple users or applications(Ali *et al.*, 2015) in (Matthew, Mahamadou&Sarhan. 2016). Rapid changes in position and strength of flows requires flexible move toward successful network resource(s) management, various number of devices such as smartphones, tablets, and notebooks had increased much fold to put pressure on enterprise resources to bring about rapid changes to network resources and as such security challengesto the management of Quality of Service (QoS) (Muhammad *et al.*, 2014).

Internet with the use of traditional IP based protocol has exploited it functionality and there is a need for a network paradigm that will take the network to a new level,suitable for today'sdemand of internet and its functionality. Software Defined Network (SDN) promised potential basic change in network configuration and real-time traffic management performed (Taimur, 2017). It separates between the network control plane and the data plane, which provides user applications with a centralized view of the distributed network states (Ian *et al.*, 2016). It moves the control plane outside the switches and enables an external centralized control of data through a logical software entity known as the SDN controller., it decouples software from hardware and centralizes network state in the control layer(Ian*et al.*, 2016).This makes the network administration, provisioning, arrangement, resource optimization, and network protection flexible using robotic SDN programs (Vandana, 2016).This enables researchers and practitioners to design much easier, flexible and powerful innovative network functions and protocolscalled SDN (Seungwon*et al.*, 2013). It enables direct programming of network operation(s) using an ordinary computer, programmer, operating system and programming languages.SDNs are logically segmented on three general regions: Application layerthis is the management plane responsible for the network programming section. Control layer hosting the network intelligent and Datalayer(Bruce & Rossi, 2016).

The remainder of this paper is organized as follows: Section 2 is the background of the work, Section 3 introduces the framework for preventing Software Defined Networks (SDN) from Malicious Attacks, Section 4 describes the result derived from the given framework in Section 3. Finally, an important conclusion is discussed in Section 5.

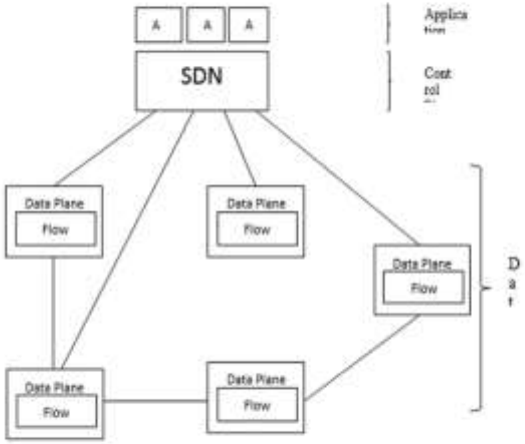


Figure 1:Overview of Software Defined Network (Sdn) (Okunade & Osunade, 2014)

2. Background of the Work

2.1. Northbound Application Programming Interfaces (APIs)

This is an open source-based application interface representing the software interface between the software modules of the controller platform and the SDN applications. The northbound interface facilitates the operation by providing the abstract view of the underlying network and empower the direct expression of network behavior and requirements.

2.2. Application Plane/Layer

Application plane is the topmost SDN planethat process request of incoming traffic and request services from the lower layers on behalf of the received traffic for further processing

(Hrishikesh, 2015)is composed of network service applications, business services, security services, and others that benefit from abstracted global view of the network according to their own purposes (Cabaj*et al.,*2014).this is an example of Northbound Application Programming Interfaces.

2.3. Control Plane

Control Plane handle the network intelligence control and states, it implement the network policies to globally regulate the network states and activities of the SDN. The logical centralization of controller enabled better decision making and maintaining of a global view of the entire network (Chienhung, Kuochen, and Guocin, 2017). It is the brain behind the successful execution of any SDN activities. According to Daojing, Sammy and Mohsen (2016), control plane manages the configuration of networking devices (such as switches and routers) and their forwarding functions. The data plane consists of protocols to execute the forwarding functions according to the rules configured by the control plane protocols. SDN controller is the central point of the network that enables the administrator to apply custom policies/protocols across the network hardware; control plane directs the data plane on flow forwarding and modifications processes. The controller is accountable for the conversion of applications' orders to the lower level communication protocol used by the data plane devices(Cabaj*et al.,*2014). The most widely deployed controller is a network operating system (NOX), controller. Naga*et al.* (2015) made to understand that controller can exercise it dynamic nature to modified the switchesthrough commands to adjust to traffic requests and equipment failures that may be observed through an event.

2.4. Southbound Application Programming Interfaces (APIs)

This is an interface through which the controllers are able to communicate with the network devices such as switches and data plane.It empowers the direct expression of network behavior and requirements. A controller can implement its responsibilities on data plane by communicating its command to the data plane through the southbound such as changing of forwarding behavior of a switch through altering offlow rule. Southbound Application Interface (APIs) are communication protocols between the controllers and the data planes examples are;OpenFlow (SDN most widely used communication protocol), OVSDB, OpenDaylight, Onix and HP VAN, and so on.

2.5. OpenFlow(OF)

OpenFlow communicate between the SDN controller via southbound open interfaces (such as OF protocol)and the data plane.OFwas created and hosted at the University of Stanford in 2008 for evangelizing and supporting the OpenFlow Community. OpenFlow is the most widely used SDN protocol; it is an open standard based communication protocol that enables the control plane to communicate with the data plane according to (Mateus, Bruno and Katia, 2013).Wolfgang and Michael (2014) stated that OpenFlow mainly focuses its consideration on switches whereas other SDN approaches focused on other network elements such as routers. According to Jad, David, Covington, Guido and Nick (2008) OpenFlow pushes difficulty to controller software so that the controller administrator has full control over it. This is done by pushing forwarding decisions to a "logically" centralized controller and allowing the controller to add and remove forwarding entries in OpenFlow switches.

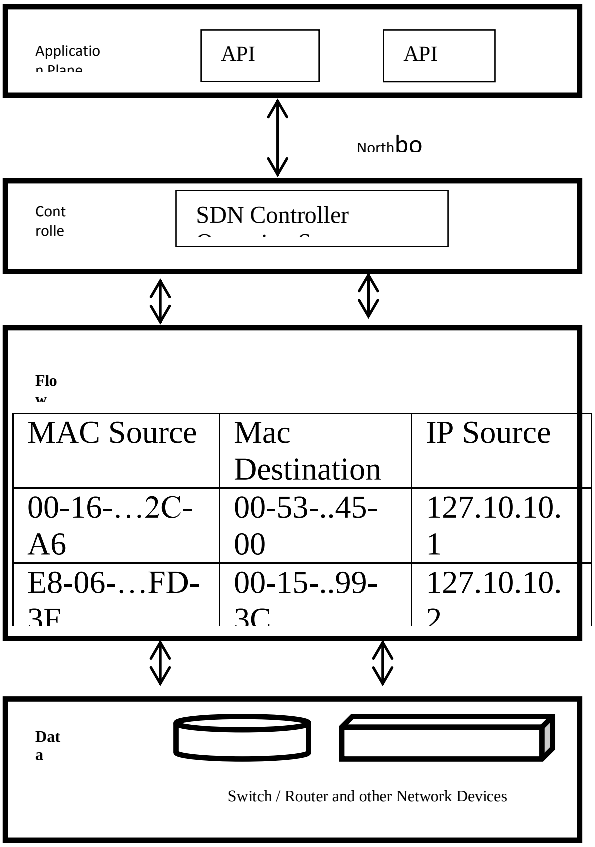
3. Algorithm and Implementation

3.1. Methodology

To address the aforementioned problem, a code was written to implement the Security Algorithm presented with embedded security extension of SDN OFtable rule (figure 2). This introduced security controlextending the SDN flow table with black/white list, which helped to secureSDN paradigm, where control plane will check for the authentication of users' application through the API foruser's confirmation usingwhite / blacklist for legitimacy confirmation of users' request who is requesting to make use of control plane by sending signals.

3.2. White / Black List plus Hash Bayes theorem (W/B+HBT) Algorithm Model

Figure 3 is the W/B+HBT Security Algorithm for preventing malicious attacks in Software Defined Network and process model that shows the incoming packet/request from the network, parsing the header field and match against the flow table to check if flow rule is already presented for the source address. If checked result is (NO) it means no existing flow rule for the packet source address, implying that packet/request source is communicating with that particular destination for the first time. The algorithm then requests from the controller for the creation of new flow rule for the newly arrived requestpacket transmitting from an unknown source. If the test checked result is (YES) it means there is an existing flow rule between the source and destination of the newly arrived packetrequesting from the network. The algorithm further its test to check if the identified flow rule between the basis and target of requesting packet is enlisted within the black or white list security extension of the SDN OF Architecture. If the flow rule is within the white list, the transaction is successfully executed by adding an entry for it in each of the switches along the path. Otherwise (if the flow rule falls within the blacklist), the algorithm generates an alarm that is sent to the controller and it also replies by sending a drop action to block or discard the transaction



MAC Source	Mac Destination	IP Source
00-16-…2C-A6	00-53-..45-00	127.10.10.1
E8-06-…FD-3E	00-15-..99-3C	127.10.10.2

Figure 2:Extended SDN Openflow (Of) Table with White/Blacklist Security Features

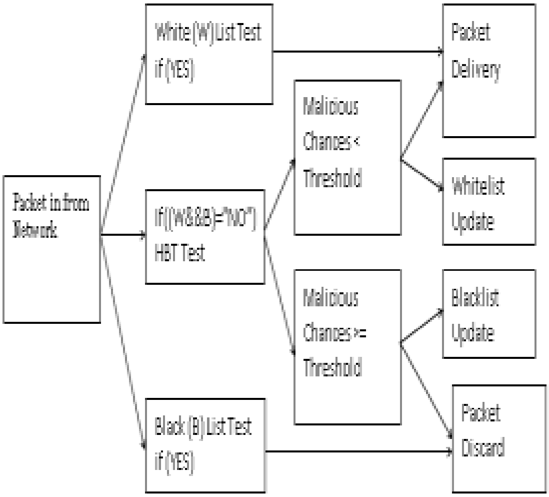
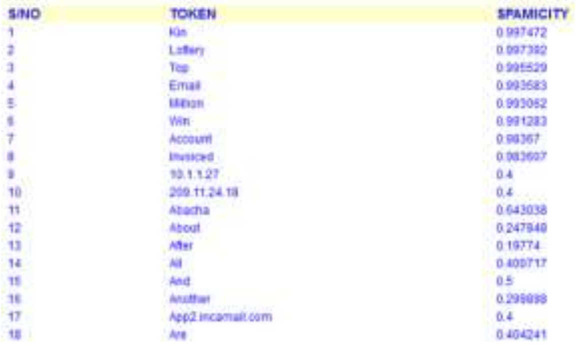

Packet in from Network
White (W) List Test if (YES)
If((W&&B)="NO") HBT Test
Black (B) List Test if (YES)
Malicious Chances < Threshold
Malicious Chances >= Threshold
Packet Delivery
Whitelist Update
Blacklist Update
Packet Discard

Figure 3: W/B+HBT Algorithm Model

If otherwise (newly arrived Packet source) not in either Black/White list, then the algorithm applies HBT content-based filter to calculate the Malicious chancesof the incoming packet using statistical Bayes' theorem (Okunade &Osunade, 2014).If (Malicious value) less than (<) the set Threshold of 0.5 the packet is forwarded/delivered to the appropriate quarters. If the packet calculated Malicious chance/value is greater than (>) the set Threshold of 0.5 the packet is discarded. Whatever the case may be the result is used to update the white/blacklist for a subsequent transaction(s).



S/NO	TOKEN	SPAMICITY
1	Kin	0.997472
2	Lottery	0.997392
3	Top	0.995529
4	Email	0.993583
5	Million	0.993062
6	Win	0.991283
7	Account	0.98367
8	Invoiced	0.983607
9	10.1.1.27	0.4
10	209.11.24.18	0.4
11	Abacha	0.643038
12	About	0.247840
13	After	0.19774
14	All	0.400717
15	And	0.5
16	Another	0.298898
17	App2.incamail.com	0.4
18	Are	0.404241

Figure 4: Some Suspicious Tokens and Associated Spamicity /Malicious Values

3.3. Word Hashing Operation

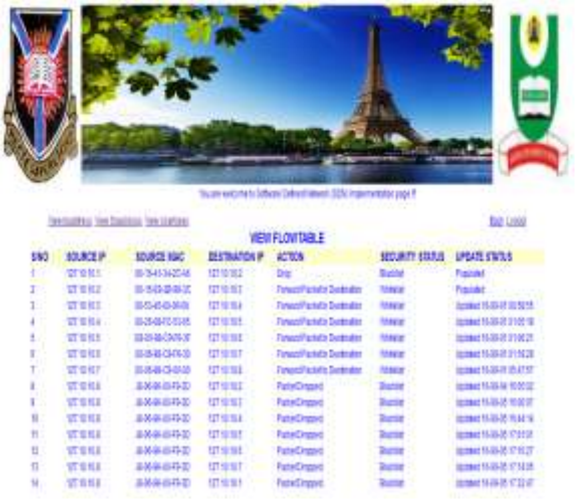
Word hashing operation is foremost executed on the newly arrived packet content, this is the removal of all unwanted prefixes, affixes and suffixes in the word(s) in order to deal with actual root/real word. The security algorithm contains inbuilt word hashing filtering technique that removed all unwanted prefixes, affixes and suffixes special characters used around the word(s) (especially around the suspicious terms) by intruders to misspelled/manipulate/ modified/mismanage tokens (such as $, /, \, |, =, !, @, #, %, ^, &, , (,), <, >, ?, :, ", ', {, [, },] and so on) used to foil the filters. This is done on the words in order to deal with actual root/real word, needed to calculate the malicious chances value using the Bayes' theorem.Then, algorithm will match packet token one after the other against suspicious table's token (Figure 4) in the database one after the other till the end of the suspicious table's token and then take the next token/word from the packet and do the same thing till the end of the tokens/words in the packet and match it against the list of tokens in the suspicious table.Then if there is matched the spamicity value of that particular matched token in the suspicious table (Figure 4) will be retrieved and assign against "a" been the first matched suspicious token follow by next matched identify suspicious term/token and assigned "b" been the second matched suspicious token, up to the last matched suspicious token and assigned it "z" is the last matched suspicious token. This assigned alphabet a to z are the alphabet finds in the Bayes formula, and spamicity values in (Figure 4) assigned to each of this alphabet (a-z) will be substituted into the Bayes formula as showed:

$$p(a,b,c...z)=\frac{a*b*c*......*z}{a*b*c*.....*z+[(1-a)*(1-b)*(1-c)*........*(1-z)]}$$

Then use the Bayes formula to calculate malicious chances, result gotten out of values substituted into the formula will then check against the threshold value that could set to any of: minimum with threshold value of 0.2, medium with threshold value of 0.3 and maximum with threshold value of 0.5 to give if (maliciousChances<= threshold) the entire newly arrived packet is forwarded to the appropriate port and then populate packet table of SDN database whitelist. But if otherwise (maliciousChances> threshold) the entire newly arrived packet is discarded and then populate a malicious table of SDN database blacklist.

4. Results and Discussion

This report the results of basic evaluation of a prototype implementation of Software Defined Network (SDN) Security Access control Algorithm using PhP/HTML code, Running/loading the Algorithm is depicted in figure 5 below.It shows that the contents of flow table consist of previous transactions status between nodes that could be used to predict further transaction, it contain source and destination of transactions nodes IP and MAC addresses, action(s) performed on such transaction which could either be “drop” or “forward to the appropriate quarters”, security status that could be grouped into “blacklist” or “whitelist” and update status that signified if the flow table was initially populated at the starting point of implementation or updated by the application based on encountered during the execution of the application and also stated the date and time updated.

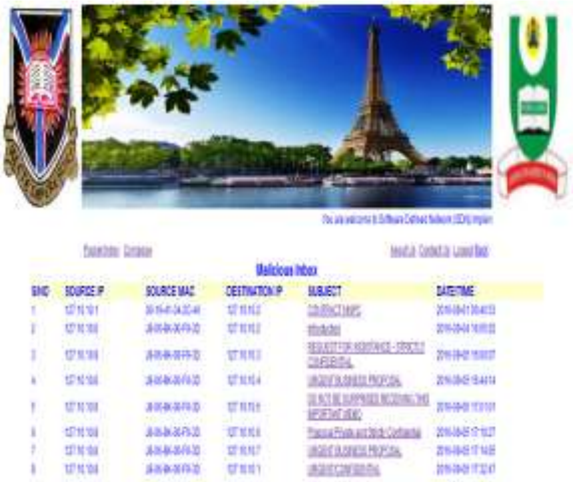


VIEW FLOWTABLE

S/NO	SOURCE IP	SOURCE MAC	DESTINATION IP	ACTION	SECURITY STATUS	UPDATE STATUS
1	[illegible]	[illegible]	[illegible]	Drop	Blacklist	Populated
2	[illegible]	[illegible]	[illegible]	Forward Packet to Destination	Whitelist	Populated
3	[illegible]	[illegible]	[illegible]	Forward Packet to Destination	Whitelist	Updated [illegible]
4	[illegible]	[illegible]	[illegible]	Forward Packet to Destination	Whitelist	Updated [illegible]
5	[illegible]	[illegible]	[illegible]	Forward Packet to Destination	Whitelist	Updated [illegible]
6	[illegible]	[illegible]	[illegible]	Forward Packet to Destination	Whitelist	Updated [illegible]
7	[illegible]	[illegible]	[illegible]	Forward Packet to Destination	Whitelist	Updated [illegible]
8	[illegible]	[illegible]	[illegible]	Packet Dropped	Blacklist	Updated [illegible]
9	[illegible]	[illegible]	[illegible]	Packet Dropped	Blacklist	Updated [illegible]
10	[illegible]	[illegible]	[illegible]	Packet Dropped	Blacklist	Updated [illegible]
11	[illegible]	[illegible]	[illegible]	Packet Dropped	Blacklist	Updated [illegible]
12	[illegible]	[illegible]	[illegible]	Packet Dropped	Blacklist	Updated [illegible]
13	[illegible]	[illegible]	[illegible]	Packet Dropped	Blacklist	Updated [illegible]
14	[illegible]	[illegible]	[illegible]	Packet Dropped	Blacklist	Updated [illegible]

Figure 5:View Flowtable

Malicious Inbox (Figure 6) is the list of received malicious packets, these are the list of incoming packets that are classified to be malicious rather than been packet. They are an unwanted packet and identified to be dangerous.The experiment was able to successfully group the entire algorithm tested malicious packets as such (malicious) therefore recorded 10% false positive and false negative.



You are welcome to Software Defined Network (SDN) implem

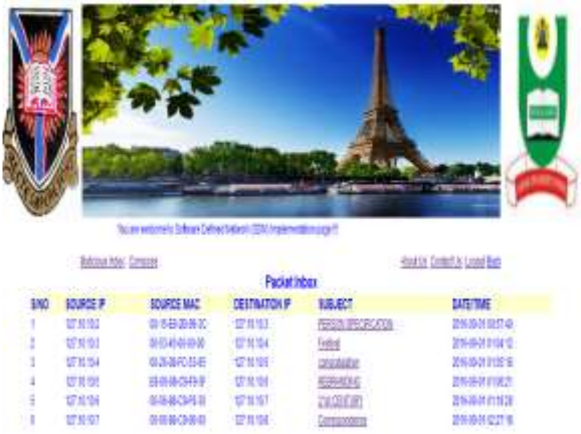
Packet Inbox Compose About Us Contact Us Logout

Malicious Inbox

S/NO	SOURCE IP	SOURCE MAC	DESTINATION IP	SUBJECT	DATE/TIME
1	[illegible]	[illegible]	[illegible]	CONTRACT INFO	[illegible]
2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
3	[illegible]	[illegible]	[illegible]	REQUEST FOR ASSISTANCE - STRICTLY CONFIDENTIAL	[illegible]
4	[illegible]	[illegible]	[illegible]	URGENT BUSINESS PROPOSAL	[illegible]
5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
7	[illegible]	[illegible]	[illegible]	URGENT BUSINESS PROPOSAL	[illegible]
8	[illegible]	[illegible]	[illegible]	URGENT CONFIDENTIAL	[illegible]

Figure 6: Malicious Inbox

Packet Inbox in Figure 7 is the list of received legitimate packets, these are the list of incoming packets that are classified to be legitimate rather than been malicious. The experiment was able to successfully group the entire algorithm tested legitimate packets as such (legitimate) therefore recorded 90% true positive and negative.



You are welcome to Software Defined Network (SDN) Implementation page !!!

Malicious Inbox Compose About Us Contact Us Logout

Packet Inbox

S/NO	SOURCE IP	SOURCE MAC	DESTINATION IP	SUBJECT	DATE/TIME
1	[illegible]	[illegible]	[illegible]	PERSON SPECIFICATION	[illegible]
2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Figure 7:Packet Inbox

4.1. Evaluation of Algorithm with the Existing TopoGuard Security Method

In an existing TopoGuard Security Method in Figure 7, once a packet send to an host could be hijacked, subsequent packets supplied to that particular host would be completely hijacked and redirected to the hijackers. The chart shown in figure 8 represents an evaluation of the White/Blacklist plus Hash Bayes Theorem (W/B + HBT) Algorithm with the Existing TopoGuard

Security where the two security methods were tested with the same data. The implemented TopoGuard Security algorithm indicates that 80% legitimate and malicious packets were classified as True positive (+ve) and true negative (-ve) where 20% legitimate and malicious packets were classified as False positive (+ve) and false negative (-ve). Whereas the White/Blacklist plus Hash Bayes Theorem (W/B + HBT) gives success record of 90% True positive (+ve) and true negative (-ve) and 10% record of False positive (+ve) and false negative (-ve) of legitimate and malicious packets classification.

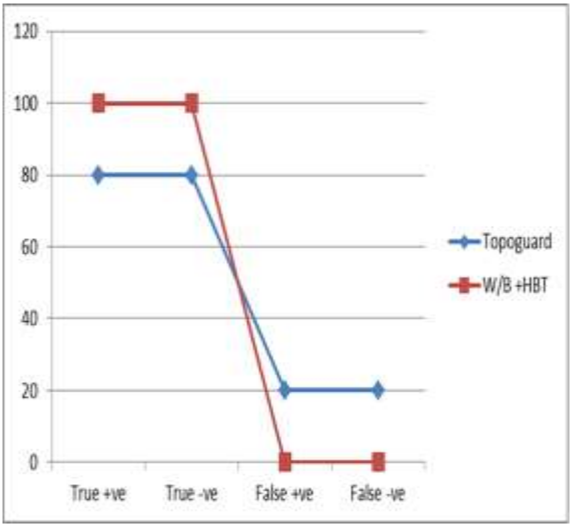


Figure 8: Evaluation of Algorithm with the Existing Topoguard Security Method

4.2. Discussion of Result

The presented algorithm having combined three examination levels. White/Blacklist plus Hash Bayes Theorem (W/B + HBT) Algorithm implementation prevented false positive or negative packets from being present. Unlike the existing Topoguard security method that discovers and prevents packet(s) from being sent to a changed or modified host address/location only, but does not prevent the host address/location from being changed or modified. The algorithm (W/B + HBT) prevents the insecure source from sending a packet to the targeted host and also prevents insecure (malicious packet(s)) from been sent. Whereas the existing Topoguard security method only considered already hijacked host (using source modification) from receiving the packet.

The result of evaluation of presented White/Blacklist plus Hash Bayes Theorem (W/B + HBT) security Algorithm compared against the existing Topoguard security Algorithm, recorded that the existing Topoguard security Algorithm has in its records 20% false positive and false negative and 80% true positive and true negative. Whereas the W/B + HBT Algorithmhave the result of 10% false positive and false negative and 90% true positive and true negative, which is accurate packets classification and far better, compared with the existing Topoguard security Algorithm.

5. Conclusion and Recommendation

5.1. Conclusion

This paper discussed in details the developed algorithm that prevents Software Defined Network (SDN) from malicious attack. As a proof of concept, it has been demonstrated and concluded from findings that algorithm combined source identification/authentication (using white/blacklist) and content filtering (using word hashing and Bayes' theorem) (W/B + HBT) method of malicious identification/authentication and packet grouping, provides effective solution to legitimate/malicious mail identification/authentication and as such prevents malicious attack from accessing their targeted host in Software Defined Network. The experiment was a successful one recorded 10% false positive and false negative, and 90% true positive and true negative.

5.2. Recommendation

This paper recommends the use of combined methods of source identification using whitelist/blacklist combined with word hashing and Bayes' theorem for content filtering mechanisms/algorithm (W/B + HBT) as a preventive measure for intrusion prevention in Software Defined Network (SDN).

References

Ali, S. T.,Sivaraman, V., Radford, A., &Jha, S. J. (2015). A survey of securing networksusingsoftwaredefined networking. *IEEE Transactions on Reliability*, vol. 64, no. 3, pp. 1086– 1097.

Bruce, H. & Rossi, R. (2016). Software Defined Networking for Systems and NetworkAdministration Programs. *The USENIX Journal of Education in System Administration*. Volume 2, Number 1. *www.usenix.org/jesa/0201*

Cabaj, K., Wytrębowicz, J., Kukliński, S., Radziszewski, P. & Truong D. K. (2014). SDN.Architecture Impact on Network Security. Position papers of the Federated *Conference on Computer Science and Information Systems*, pp. 143–148. ACSIS, Vol.3

Chien-hung, L., Kuochen, W. &Guocin, D. (2017). A QoS-aware routing in SDN hybrid networks. *The12th International Conference on Future Networks and Communications*. Published by Elsevier B.V. Peer-review under the responsibility ofthe Conference Program Chairs. Procedia Computer, Science 110 pp. 242–249. Available online at www.sciencedirect.com

Daojing, H., Sammy, C.,& Mohsen, G. (2016).Securing Software Defined Wireless Networks. 0163-6804/16/. *IEEECommunications Magazine*.

Diego, K., Fernando, M. V. R. & Paulo, V. (2013). Towards Secure and Dependable Software-Defined Networks.HotSDN'13, Hong Kong, China. ACM 978-1-4503- 2178- 5/13/08

Hrishikesh, A. D. (2015). Software Defined Networks: Challenges, Opportunitiesand Trends.*International Journal of*

Science and Research (IJSR) Vol. 4 (9). www.ijsr.net Licensed Under Creative Commons Attribution CC BY

Ian, F. A., Ahyoung, L., Pu, W., Min, L. & Wu, C.(2016). Research Challenges for Traffic Engineering in SoftwareDefined Networks.*IEEE Network*. pp. 52-58

Jad, N., David, E., Covington, G. A., Guido, A. & Nick, M. (2008). Implementing an OpenFlow Switch on the NetFPGA Platform. *ANCS '08, San Jose, CA,* USA. ACM 978-1-60558-346-4/08/0011.

Mateus A. S. S., Bruno, A. A. N. & Katia, O. (2013). Software-Defined Networking Based Capacity Sharing in Hybrid Networks. *IEEE*. http://www.projectfloodlight.org/floo dlight/

Matthew, N. O. S., Mahamadou, T. &Sarhan, M. M. (2016).Software-Defined Networking Concepts.*Journal of Scientific and EngineeringResearch*Article 3(5): 92-94.

Muhammad, H. R., Shyamala, C. S., Ali, N. &Bill, R. (2014). A Comparison of Software Defined Network (SDN) Implementation. *2nd International Workshop on Survivable and V. Robust Optical Networks* (IWSRON). Procedia Computer Science 32, pp. 1050 – 1055.

Naga, K., Haoyu, Z., Michael, F. & Jennifer, R. (2015). Ravana: Controller Fault-Tolerance in Software-Defined Networking. SOSR. Santa Clara, CA, USA. *ACM* 978-1-4503-3451-8/15/06. *http://dx.doi.org/10.1145/277499 3.2 774996*

Okunade, O. A. &Osunade, O. (2014). A Security Architecture for Software Defined Networks(SDN). *International Journal of Computer Science and Information Security,* Vol. 12 (12).

Raphael, H., Dietmar, N. & Mark, S. (2015). A Literature Review on Challenges and Effects of SoftwareDefined Networking. *Conference on ENTERpriseInformation Systems / International Conference on Project Management / Conference on Healthand Social Care Information Systems and Technologies,* CENTERIS / ProjMAN / HCist. ProcediaComputer Science 64 pp. 552 – 561. Available online at www.sciencedirect.com

Seungwon, S., Vinod, Y., Phillip, P. &Guofei, G. (2013). AVANT- GUARD: Scalable and Vigilant Switch Flow Management in Software-Defined Networks. CCS'13, Berlin, *Germany. ACM 978- 1-4503-2477 9/13/11.http://dx.doi.org/10.1145 /25 08859.2516684. Systems* -KES2013.Published by Elsevier B.V. Selection and peer-review under responsibility of KES International. Science Direct.22, pp. 810 – *819* *www.sciencedirect.com*

Taimur, B. (2017). State of the Art and Recent Research Advances in Software Defined Networking. *Wireless Communications and Mobile Computing Hindawi.*Article ID 7191647, 35 pages https://doi.org/10.1155/2017/7191647

Vandana C.P. (2016). Security improvement in IoT based on Software Defined Networking (SDN). *International Journalof Science, Engineering and Technology Research (IJSETR*), Vol. 5(1). 292 ISSN: 2278 – 7798

Wenfeng, X., Yonggang, W., Chuan, H.F., Dusit, N. &Haiyong X. (2015).A Survey onSoftware-DefinedNetworking. *IEEE Communication Surveysand Tutorials*, Vol.17(1). pp. 27-51.http://www.ieee.org/publications_standards/publications/

Wolfgang, B. and Michael, M. (2014). Software-Defined Networking Using OpenFlow:Protocols, Applications and Architectural Design Choices. *Future Internet* 6, pp. 302-336; doi:10.3390/fi6020302 ISSN 1999- 5903. *www.mdpi.com/journal /future internet*

Yutaka, J., Hung-Hsuan, H. and Kyoji, K. (2013). Dynamic Isolation of Network Devices UsingOpenFlow for Keeping LAN Secure from Intra-LAN Attack. *17th International Conference in Knowledge-Based and Intelligent Information and Engineering Systems* - KES2013. Published by Elsevier B.V. pp. 810 – 819. ScienceDirect Available online at www.sciencedirect.com

Covenant Journal of Informatics & Communication Technology. Vol. 6 No. 2, Dec., 2018

An Open Access Journal Available Online

Unsupervised Retinal Blood Vessel Segmentation Technique using *pdAPSO* and Difference Image Methods for Detection of Diabetic Retinopathy

Emmanuel Gbenga Dada & Stephen Bassi Joseph

Department of Computer Engineering, University of Maiduguri,
Maiduguri, Nigeria
gbengadada@unimaid.edu.ng,
sjbaassi74@unimaid.edu.ng

Abstract—Retinal vessel segmentation is a practice that has the potential of enhancing accuracy in the diagnosis and timely prevention of illnesses that are related to blood vessels. Acute damage to the retinal vessel has been identified to be the main cause of blindness and impaired vision. A timely detection and control of these illnesses can greatly decrease the number of loss of sight cases. However, the manual protocol for such detection is laborious and although autonomous methods have been recommended, the accuracy of these methods is often unreliable. We propose the utilization of the Primal-Dual Asynchronous Particle Swarm Optimisation (*pdAPSO*) and differential image methods in addressing the drawbacks associated with segmentation of retinal vessels in this study. The fusion of *pdAPSO* and differential image (which focuses on the median filter) produced a significant enhancement in the segmentation of huge and miniscule retinal vessels. In addition, the method also decreased erroneous detection near the edge of the retinal (that is not sensitive to light). The results are favourable for the median filter when compared to mean filter and Gaussian filter. The accuracy rate of 0.9559 (with a specificity of sensitivity rate of 0.9855), and a sensitivity rate of 0.7218 were obtained when tested using the Digital Retinal Images for Vessel Extraction database. The above result is a pointer that our approach will help in detecting and diagnosing the damage done to the retinal and thereby preventing loss of sight.

Keywords/Index Terms—Retinal Vessel, Segmentation, Asynchronous Particle Swarm Optimisation, Primal-Dual, Diabetic Retinopathy

1. Introduction

Retinopathy is the subdivision of medicine that makes it possible to determine the cause of infections and ailments of the eye and treat them immediately. Digital photography and image analysis of retinal vessel are already gaining ground. The studies of (Kanski, 2007) and (Klonoff & Schwartz, 2000) observed that these techniques have been recently recognised as beneficial techniques in the identification and treatment of some diseases like diabetic retinopathy (DR), retinopathy of prematurity (ROP) which is a pathological disarrange of the retina and cardiovascular diseases (Mapayi, Tapamo & Viriri, 2014).

According to World Health Organisation (2016), DR and ROP are the chief reasons of eye defect and impaired vision universally, timely identification of the cause and control of these ailments will assist in a notable decrease of instances of impaired vision (Gergely & Gerinec, 2009). Ophthalmologists find vessel network very useful as they concentrate on retinal vessel quality assessment at some stage of diseases diagnosis. Detection by physical examination and testing of the retinal vessels in the hollow part of the images is quite cumbersome and laborious job which needs competent and skillful people who are not readily available (Varughese, *et al.,* 2008). Nevertheless, according to Marrugo *et al.,* (2012), it is possible for the ophthalmologist to diagnose and effectively control the diseases with the help of automated segmentation and systematic inspection of the arrangement of blood vessels in the retinal. Retinal vessel segmentation is the partitioning of a retinal image into sections that have related attributes such as grey level, colour, texture, brightness, and contrast (Zhang, Zhou & Bao, 2015). The image part is removed from the initial image during image analysis and image segmentation algorithms are used to divide the original image into different segments. The main purpose of the retinal vessel segmentation is to divide the retinal image into equally exclusive sections so that each section in relation to the pixel concentration is harmonised to a predetermined benchmark (Zhang, Zhou & Bao, 2015). This study presents a new technique for segmenting network vessels in retinal images through Difference image and *pdAPSO* techniques.

The remainder of this paper is structured as follows: Section 2 discusses the related works. Section 3 discusses our proposed methodology, filtering techniques, difference image, primal-dual particle swarm optimisation and post-processing. The results generated by our proposed approach on DRIVE datasets and discussions are specified in Section 4, and lastly the study is concluded in Section 5.

2. Related Works

Many research has been conducted in the area of retinal segmentation. Akram & Khan (2013) used a multi-layered thresholding-based blood vessel segmentation method for investigating cases of retinal diabetic that can result in blindness. Jiang & Mojon (2003) used an adaptive local thresholding prototype employing an authentication based multi-threshold analytical system for detecting the retinal image. The weakness of this approach is its lackluster performance in detecting the reedier vessels and some isolated vessels in the retinal. Mapayi, Viriri & Tapamo (2015) proposed a novel adaptive thresholding method for retinal vessel segmentation using local information that is uniform in nature. Qin *et al.,* (2006) employed a multiscale method that uses Gabor filters and scale multiplication for the segmentation of retinal vessels. Marin *et al.,* (2011) developed a novel supervised approach for blood vessel segmentation in retinal images through gray-level and moment invariants-based features for pixel depiction, whereas the vessel segmentation was done by neural network algorithm. Szpak & Tapamo (2008) used the gradient based technique and level set method for retinal vessels segmentation. Their approach was unsuccessful in detecting the tinnier retinal vessels. Wang *et al.,* (2013) used the combination of multi-wavelet kernels and multiscale hierarchical decomposition to segment retinal vessels. Xiao *et al.,* (2013) developed a retinal segmentation technique based on the Bayesian method and spatial constraint. Yin *et al.,* (2013) proposed an unsupervised segmentation technique using probabilistic formulation.

Lupascu & Tegolo (2011) implemented an unsupervised segmentation of retinal vessels using self-organizing maps (SOM) and k-means clustering. The SOM is trained on retinal images and the map was again partitioned by k-means clustering method into two groups. The complete image is again passed to SOM and the group with the most ideal identical section on SOM is allocated to each pixel. A hill climbing scheme on associated components is employed to detect the vessel network during the post-processing operation. Ramaswamy *et al.,* (2011) proposed the combination of k-means and fuzzy c-means clustering methods for the categorisation of discharges and non-discharges in retinal images. Saffarzadeh *et al.,* (2014) developed a technique that used k-mean for pre-processing after which the multi-scale line operator is utilised for the detection of retinal vessel network. The K-means assists the vessels to be more conspicuous and there is a significant decrease in the effect of bright gashes. The line detection operator in three scales is used in detecting the retinal vessels. Wen *et al.,* (2007) assessed the performance of the k-means algorithm in enhancing the detection of retinal vessels by decreasing the colour space. However, the output of this approach was not good.

Sreejini & Govindan (2015) applied PSO to find the best filter parameters of the multiscale Gaussian matched filter to attain increased accuracy of retinal vessel segmentation. Their technique have better performance than many of the existing retinal vessel segmentation methods. There is still need to improve on the performance of the proposed system as the mutiscale matched filter does not completely overcome the problem of undesirable performance figures of matched filters. Son, Park & Jung (2017) used generative adversarial neural network to produce the exact map of retinal vessels on DRIVE and STARE datasets. The drawback of their approach is that it was unsuccessful in detecting extremely tinny vessels. Li *et al.,* (2017) applied reinforcement local descriptions and SVM to segment retinal blood vessel. The system achieved a very high performance but the segmentation process can be time consuming. Mohsen *et al.,* (2018) proposed neural network hardware implementation and FPGA for retinal vessel segmentation. The major drawback of their technique is the complexity of the system. Memari et al., (2017) used the hybrid of matched filter and AdaBoost classifier for enhanced retinal vessel segmentation. Sumathi, Vivekanandan & Ravikanth (2018) proposed neural network for efficiently segmenting retinal vessel.

Particle swarm optimisation (PSO) algorithm is an unsupervised segmentation approach. PSO has found application in the field of image segmentation. Saatchi & Cheng-Hung (2007) did a survey of image segmentation application of swarm intelligence algorithms (PSO and ACO) and their hybrid with k-means and simple competitive learning algorithms. Gopi & Nageswara (2013) used PSO methods for image segmentation to detfect breast cancer. Their experimental result shows that the fusion of Fuzzy C-means (FCM) and Fractional Order Darwinian PSO (FODPSO) algorithm performs better than the PSO alone, and Darwinian PSO (DPSO). There is, however, need to develop the more efficient approach that can remove background noise from the images. Mahalakshmi & Velmurugan (2015) used PSO to segment a brain tumour medical images. The effectiveness of the approach is not certain as there is no clear parameter for measuring the performance of the technique.

While much progress has been made in developing efficient methods by the earlier research works, the results indicates that there is need for more work to be done in tackling the problem of high false detection close to the edge of the point where the optic nerve enters the retinal and the segmentation of big and reedier retinal vessels. In this study, we propose the combination of the Primal-Dual Asynchronous particle swarm optimisation and difference image approach for the segmentation of retinal blood vessel segmentation for easy detecting of thinner layers in the vessel and effective diagnosis of retinopathy diabetics. The major shortcoming of many of the vessel segmentation techniques include: low contrast of vessels, low quality images and large disparities and unpredictability in the size of the vessels. Achieving high accuracy in segmentation is a big problem because minute vessels are subjugated by various image noises such as Gaussian (Sreejini & Govindan, 2015).

3. Methodology

Normally, the presence of noise as a result of fluctuating lighting and contrast in retinal fundus images pose a challenge to retinal image clustering unless there is pre-processing operation. Realising the fact that efficient detection of the network of the vessel is a crucial phase required in the detection of diseases of the eye for dependable retinal vessel classification, there is need to develop an effective method that can handle the segmentation of sizable and small vessels in a well-timed and successful style. The green section of the pigmented retinal image is utilised for segmentation because it offers the most excellent brightness setting for the vessel (Kande, Subbaiah & Savithri, 2010). An in-depth explanation of our proposed technique is outlined below:

(1) Removal of the green channels of the coloured retina image.

(2) Filtering of the retinal image using median, mean, and Gaussian filtering methods.

(3) Producing the difference image.

(4) Partitioning of the retinal vessels from the difference image produced by the *pdAPSO* algorithm.

(5) Execution of a post-processing stage using median filter for the elimination of miscategorisation.

Figure 1 below depicts the framework of the proposed system.

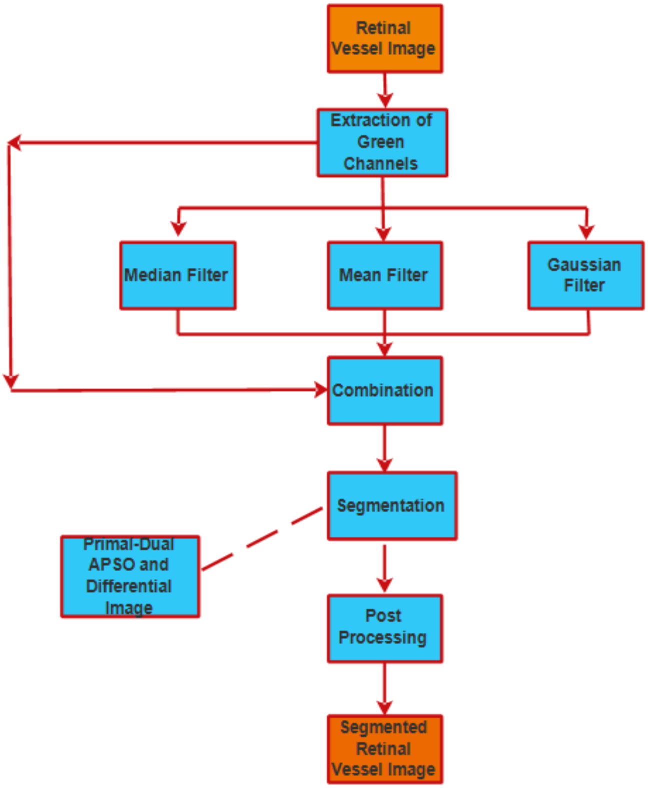

Retinal Vessel Image
Extraction of Green Channels
Median Filter
Mean Filter
Gaussian Filter
Combination
Segmentation
Primal-Dual APSO and Differential Image
Post Processing
Segmented Retinal Vessel Image

Figure 1: Framework of the proposed technique (the technique comprises of basically three steps using different filters and combinations of filters and a segmentation technique).

3.1 Filtering Techniques

The improvement of the green channel of the fundus retinal image is by achieved applying diverse filtering methods. The mean filter and Gaussian filter which are linear filters are employed to free the images from roughness and unevenness of the surface. These filters help to decrease image noise. However, they cannot effectively protect the boundaries of an image. The median filter which is a non-linear filter is very effective in eliminating image noise in addition to safeguarding information at the borderline of images. It is very essential for us to state that the dimensions of the frame of the chosen image should not be too big to be able to proficiently handle the noise resulting from brightness deviation that usually characterises the retinal image. Being meticulous in the choice of window size that has adequate data points is very crucial to achieving for good enhancement. Choosing window sizes with adequate data points is crucial to attaining superior enhancement in the segmentation process. In this study, we employ the mean, Gaussian and median filters because of the complexity of the retinal image. The complexity of the retinal image is explained below:

$$R = F \otimes G \qquad (1)$$

$$R_{(i,\, j)} = \sum_{(x,\, y)\ \varepsilon\ F\ (i-x,\, j-y)\ \varepsilon\ G} F\ (x, y)\ G\ (i\text{-}x, j\text{-}y) \qquad (2)$$

Given that R is the twisted retinal image, G represents the green channel of the retinal image and the twisted cover F is used to denote the filtering method used in this research work

3.2 Difference Image

The subtraction of the green channel of the coloured retinal image from the twisted retinal image produced the difference image. The formula for the difference image D(i, j) is as follows:

$$D(i,j) = R(i,j) - G\ [i,j] \qquad (3)$$

where $D(i,j)= \{D_q(i,j), D_G(i,j), D_\alpha(i,j)\}$, here $D_q(i,j)$ is the difference image created by median filter (DIMDF), $D_G(i,j)$ is the difference image created by mean filter (DIMNF) and $D_\alpha(i,j)$ is the difference image created by Gaussian filter (DIGF). We also experimented with using the combination of median and mean filters (DIMDMNF), median and Gaussian filters (DIMDGF), and mean and Gaussian filters (DIMNGF). The possible permutations are

$$D_q^G = D_G(i,j) + D_q(i,j)$$

$$D_G^q = D_q(i,j) + D_G(i,j) \qquad (5)$$

$$D_\alpha^G = D_\alpha(i,j) + D_G(i,j) \qquad (6)$$

where D_q^G represents the DIMDMNF, and D_G^q is used to denote the DIMDGF, while D_α^G represents the DIMNGF. The outputs of the equations (4) to (6) are regularised to the interval [0, 255].

3.3 Primal-Dual Asynchronous Particle Swarm Optimisation (pdAPSO) Algorithm

The Asynchronous PSO (APSO) is a variant of PSO proposed by Dada and Effirul (2015). The current personal best (*pbest*$_{i,\mathrm{m}}$) and the global best (*gbest*$_{i,\mathrm{m}}$) of a particle, its velocities, and positions of particles are instantly modified to be up to date after computing their fitness. As a result of this, the parameters are updated using partial or deficient information about the neighbourhood. This leads to diversities in the swarm of particles as the current swarm is a mixture of certain information from the earlier iteration and the ones from the current iteration. For detail information on the Primal-Dual PSO, the reader is referred to our previous work (Dada & Effirul, 2015). This present study adapted the *pdAPSO* algorithm for efficient segmentation of retinal images. In this approach, a single particle x_i denotes N cluster such that $x_i=(y_{i1},...,y_{ij},\ldots,y_{iN})$ where y_{ij} correspond to the jth cluster centre of inertia trajectory of the i^{th} particle. Consequently, a swarm typifies a number of contending cluster centers. The fitness of each group of cluster is computed based on the formula below:

$f(x_i,m_i) = w_1 dist_{max}(m_i,x_i) + w_2(m_{max} - dist_{min}(x_i))$ (7)

where z_{max} = 2n-1 for an n-bit image; M is a matrix denoting the allocation of pixels to clusters of particle i. Each element m_{ijp} specifies if the pixel m_p is a member of cluster C_{ij} of particle i. The coefficients $w_1$and w_2 are constants that are specified by the user. Similarly, the highest value of the mean Euclidean distance of particles to their related clusters is

$dist_{max}(m_i, x_i) = max_{j=1,\dots,N_e}\{\sum_{\forall m_p \epsilon c_i} dist(m_p, y_{ij}) / |c_{ij}|\}$ (8)

and the distance between any set of clusters with the lowest Euclidean distance value is $dist_{min}(x_i)$ =

$\underset{\forall j_1, j_2, j_1 \neq j_2}{min}\{d(y_{ij1}, y_{ij2})\}$ (9)

The flowchart of the *pdAPSO* algorithm is shown in figure 1 below.

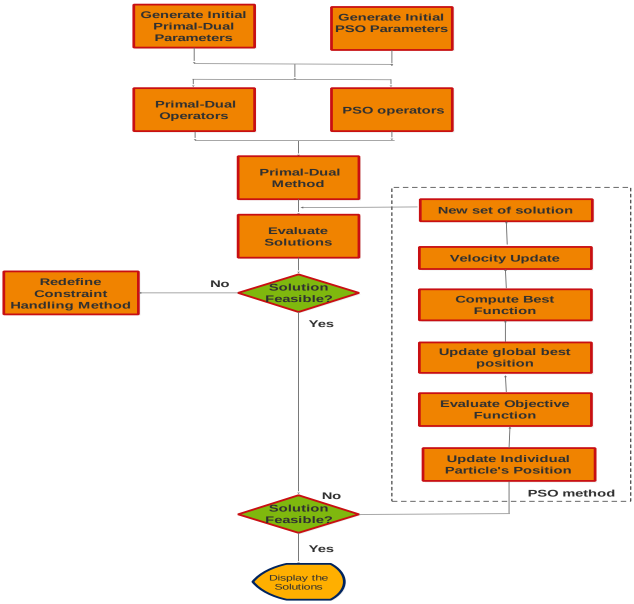


Figure 2: Primal- Dual-APSO (*Pdapso*) Flowchart

The proposed algorithm for retinal vessel segmentation is outlined below:

Algorithm 1 *pdAPSO* Segmentation algorithm

Step 1: Haphazardly assign an initial value to cluster centers for each particle.

Step 2: On behalf of each particle, designate each pixel to a cluster with the shortest distance to its cluster center.

Step 3: Compute the fitness function for each particle and obtain the global best solution.

Step 4: Save the best solution found so far as the *pbest* or personal best solution.

Step 5: Use the equations below to update the cluster centers

$$v_{i,m}^{(t+1)} = w * v_{i,m}^{(t)} + c_1 * rand1() * (pbest_{i,m} - x_{i,m}^{(t)} + c_2 * rand2() * (gbest_m - x_{i,m}^{(t)}) \quad (10)$$

$$x_{i,m}^{(t+1)} = x_{i,m}^{(t)} + v_{i,m}^{(t+1)} \quad (11)$$

Step 6: If the stopping condition is fulfilled move to the subsequent step. Else, move to Step 3.

Step 7: Display the best solution.

The Primal-Dual Asynchronous Particle Swarm Optimisation method explained in Algorithm 1 is utilised in segmenting the retinal vessel network from the contextual tissue in the retinal images by utilising the outputs produced of equations (10) and (11).

3.4 Post-Processing

Median filter is utilised for the post-processing stage to ensure the reduction to the barest minimum of erroneously detected pixels in the vessel. We adopted a 2×2 median filter to remove the noisy pixels in the image to so that we can get the segmented vessel network. The median filter works by taking into consideration each pixel in the vessel image in turn and considers at its close neighbours to determine if it is a true exemplification of its surroundings. Rather than just merely changing the pixel values, it substitutes it with the median of these values. A 2x2 neighbourhood is considered here as larger neighbourhoods will produce better smoothing in the retinal vessel image. The median filter is very robust and it does not generate fresh impracticable pixel values when the filter overlaps a border. This makes it to be much better at maintaining sharp edges than some other filters.

3.5 Performance Measures

In this section, the experimental setup used to evaluate three algorithms proposed for retinal vessel segmentation. The sensitivity, specificity, and accuracy are the main performance metrics used in this work. The algorithms were experimented on the DRIVE database explained in section 3.3.

Sensitivity = TP/ (TP + FN) (12)

Specificity = TN/ (TN + FP) (13)

Accuracy = (TP + TN)/ (TP + TN + FP + FN) (14)

where TP = True Positive, TN = True Negative, FP = False Positive and FN = False Negative.

The amount of pixels that are rightly categorised as vessel pixels is the TP. The amount of pixels that are rightly categorised as non-vessel pixels is the TN, while FN is the number of pixels wrongly categorised as non-vessel pixels. The FP is the number of pixels wrongly categorised as vessel pixels. The sensitivity (SE) is the division of TP by the sum of vessel pixels in the ground truth segmentation, while the specificity (SP) is the division of TN by the sum of non-vessel pixels in the ground truth.

4. Results and Discussion

Some tests were carried out to ascertain the efficiency of our new technique. Our technique was tested using DRIVE database on http://www.isi.uu.nl/Research/Databases/DRIVE. Experiments were conducted using MATLAB 2015a on an AMD A 10-7300 Radeon R6, 10 Compute Cores 4C+6G, 1.90 GHz, 8.00GB of RAM. The optimal parameters for the filtering operation were discovered by inputting one of the images into the optimised median filter. The *pdAPSO* optimisation technique comprises of diverse arbitrary processes. It therefore means that using identical parameters for the filter does not guarantee getting the same result even when the program is restarted and executed.

4.1 Results

The results of our experiments indicate the performance of the diverse difference images hybridised with the pdAPSO algorithm. pdAPSO and DIMDF produced the most ideal segmentation result when compared to pdAPSO and DIMNF then pdAPSO and DIGF. pdAPSO and DIMDF generated the most superior accuracy rate 0.9559 and a specificity rate of 0.9855. In Figure 3, we provide a pictorial explanation that relates the results gotten from vessels segmented by pdAPSO and difference image using each of the filters.

4.2 Discussion

The fusion of DIMDF and *pdAPSO* detected many of the big and tinny vessels, whereas some few gauzier vessels go on without detection. There are few cases of erroneous detection coupled with previous leftover close to the edge of the point where the optic nerve enters the retinal. The erroneous detection near the edge of the optic disc is nevertheless greatly reduced on the segmented vessels by DIMDF and *pdAPSO* but much in number on segmented vessels generated by the combination of *pdAPSO* with DIMNF, and *pdAPSO* with DIGF. The reason for this is because the median filter maintains boundary information of the vessels in the improved retinal image. Furthermore, the DIMDMNF and DIMDGF while hybridised with *pdAPSO* produced good sensitivity of 0.7217 and 0.7205 respectively. The addition of DIMDF caused the increase in sensitivity of the retinal image. However, the accuracy of DIMDF is better than the ones generated by DIMDMNF, DIMDGF, and DIMNGF while fused with the *pdAPSO* method. The statistical results generated by our approach are presented in table 1.

In Fig. 4 there is a visual description of the result comparison of retina vessels segmented by the fusion of *pdAPSO* and DIMDF DIMDMNF, DIMDGF and DIMNGF. As illustrated in Fig. 3, a good number of the large and tinny vessels are detected, while the ones that remain undetected are very few. The erroneous detection near the edge of the optic nerve is greater on the outputs generated by DIMDMNF, DIMDGF and DIMNGF while fused with *pdAPSO* but smaller on the outputs produced through *pdAPSO* fused with DIMDF. This explains the reason why a better accuracy rate was attained.

4.3 Performance comparison of pdAPSO with other segmentation techniques on DRIVE dataset

The performance comparison of some of the segmentation techniques on DRIVE database is presented in Table 2. The combinations of the proposed approach with the best performance yield higher and specificity rates, when compared to the earlier approaches. It is worth noting that the work of Ricci & Perfetti (2007) generated the highest accuracy rate among all the techniques compared, the value for the sensitivity and specificity rates are not specified.

Table 1: Performances of Selected Segmentation Techniques on Drive

Technique	**Accuracy**	**Sensitivity**	**Specificity**
pdAPSO and DIMDF	0.9559	0.7218	0.9855
pdAPSO and DIMNF	0.9551	0.7092	0.9824
pdAPSO and DIGF	0.9552	0.7056	0.9869
pdAPSO and IMDMNF	0.9545	0.7217	0.9703
pdAPSO and DIMDGF	0.9542	0.7205	0.9726
pdAPSO and DIMNGF	0.9505	0.7108	0.9759

Table 2: Performances of Selected Segmentation Techniques on Drive

Technique	**Accuracy**	**Sensitivity**	**Specificity**
Human observer	0.9473	0.7761	0.9725
Mapayi *et al.,* (2015)	0.9469	0.7477	0.9680
Saffarzadeh *et al.,* (2014)	0.9387	N/A	N/A
Xiao *et al.,* (2013)	0.9529	0.7513	0.9792
Oliveira *et al.* (2016)	0.9464	N/A	N/A
Meng *et al.* (2016)	0.9630	0.7680	0.9827
Mohsen *et al.* (2017)	0.9469	0.5147	0.9950
Memari *et al.* (2017)	0.9321	0.8124	0.9505
Sumathi, Vivekanandan & Ravikanth (2018)	0.9671	0.8139	0.9822
Proposed Technique	0.9559	0.7218	0.9855

The result below shows the images produced by the fusion of difference image and pdAPSO algorithm.

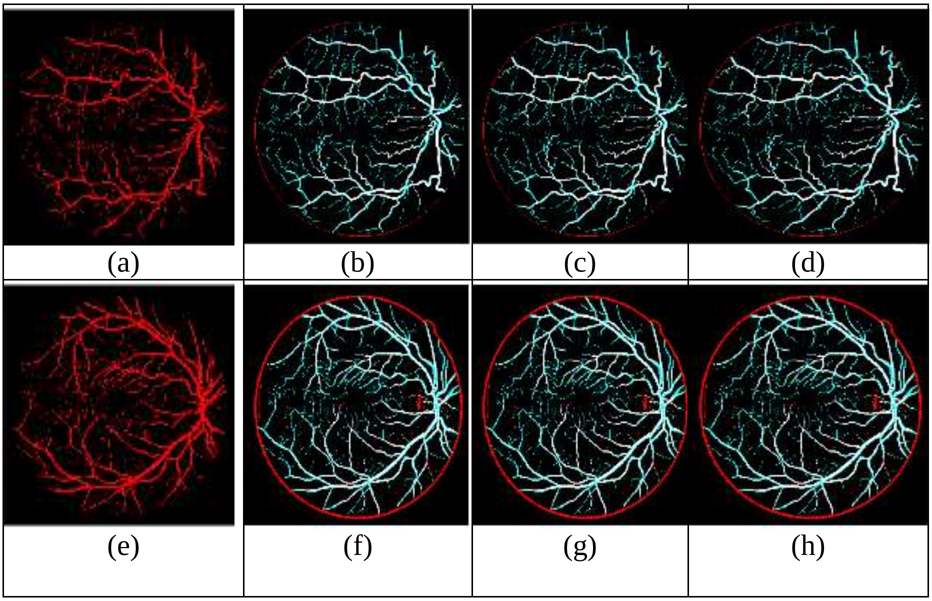


Fig. 3: (a) & (e) DRIVE Database Gold Standard. (b) & (f) Segmented Vessels Using *pdAPSO* and DIMDF. (c) & (g) Segmented Vessels Using *pdAPSO* and DIMNF. (d) & (h) Segmented Vessels Using *pdAPSO* and DIGF.

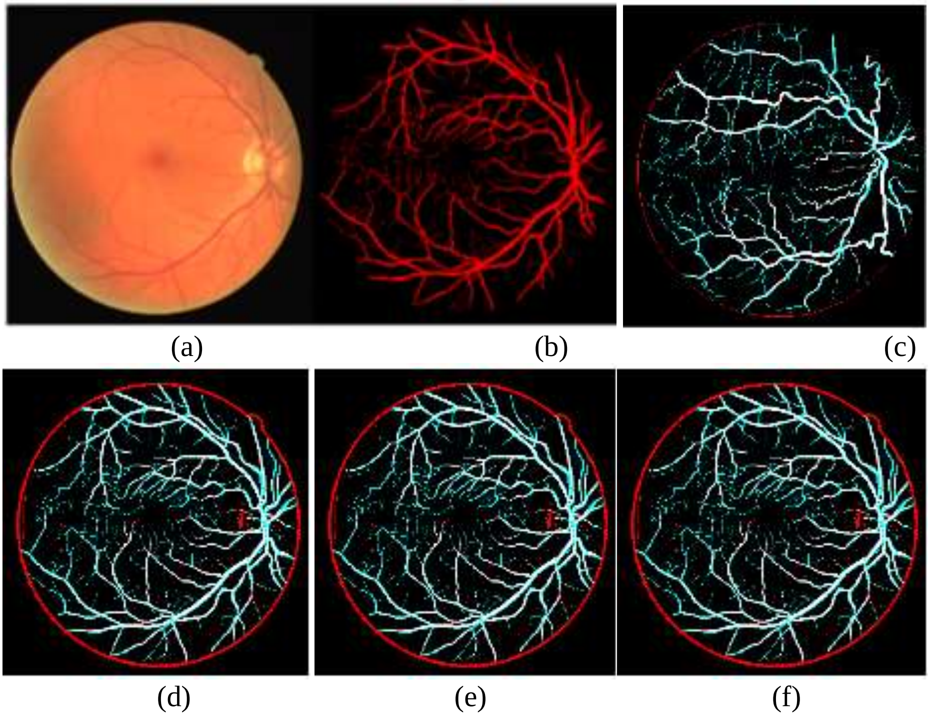


Fig. 4: (a) DRIVE Database Coloured Fundus Image (b) DRIVE Database Gold Standard (c) Segmented Vessels Using *pdAPSO* and DIMDF (d) Segmented Vessels Using *pdAPSO* and DIMNGF (e) Segmented Vessels Using *pdAPSO* and DIMDGF (f) Segmented Vessels Using *pdAPSO* and DIMDMNF

5. Conclusion

In this study, the hybrid of difference image and *pdAPSO* was used for the segmenting the retinal vessels. It was demonstrated that the proposed vessel segmentation technique is time efficient and yields a high accuracy, specificity rates, and sensitivity when compared to some segmentation methods on DRIVE database. The fusion of *pdAPSO* and difference image centered on median filter effectively segmented both large and tinny retinal vessels and also reduced erroneous detection near the edges of the optic disc. The results also proved the superiority of the hybrid of *pdAPSO* and difference image centered on median filter compared to difference image centered on mean filter and difference image centered on Gaussian filter fused with *pdAPSO* for the segmentation of retinal vessels. Moreover, the capacity of the median filter to preserve the boundaries of the retinal image is the reason for the outstanding performance. This work also proved that the hybridisation of *pdAPSO* with difference images centered on linear filtering technique and difference image centered on the median filter generated an excellent vessel segmentation result. Finally, we deduced from our experiments that our proposed method that integrates difference image with *pdAPSO* produced accuracy, specificity and sensitivity that are as good as that of other earlier methods on DRIVE database. Future work, will concentrate on developing more effective segmentation methods using Softcomputing algorithms such as Moth flame Optimisation and Grey Wolf Optimisation algorithms, and also exploit filters such as Gabor Wavelet filters, matched filter, and Frangi's filters.

References

Akram, M.U. & Khan, S.A. (2013). Multilayered thresholding-based blood vessel segmentation for screening of diabetic retinopathy, Eng Comput, vol. 29, pp. 165-173.

Dada, E. G. & Ramlan, E. I. (2015). Primal-Dual Interior-Point Method Particle Swarm Optimization (pdipmPSO) Algorithm. In: 3rd Int'l Conference on Advances in Engineering Sciences & Applied Mathematics (ICAESAM'2015), London (UK), pp. 117-124.

Gergely, K. & Gerinec, A. (2009). Retinopathy of prematurity epidemics, incidence, prevalence, blindness, Bratislavske lekarske listy, vol. 111 no. 9, pp. 514 - 517.

Gopi, R.N. & Nageswara, R. P. A. (2013). Particle Swarm Optimization Methods for Image Segmentation Applied In Mammography. Int. Journal of Engineering Research and Applications, vol. 3, Issue 6 pp. 1572-1579.

Jaemin, S., Sang, J. P. & Kyu-Hwan, J. (2017). Retinal Vessel Segmentation in Fundoscopic Images with Generative Adversarial Networks. arXiv:1706.09318v1 [cs.CV] 28 Jun 2017.

Jiang, X. & D. Mojon, D. (2003). Adaptive local thresholding by verification based multi-threshold probing with application to vessel detection in retinal images, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25 no. 1, pp. 131-137.

K.S. Sreejini, V.K. Govindan (2015). Improved multiscale matched filter for retina vessel segmentation using PSO algorithm, Egyptian Informatics Journal (2015) 16, 253–260.

Kande, G. B. Subbaiah, P. V. & Savithri, T. S. (2010). Unsupervised fuzzy based vessel segmentation in pathological digital fundus images, Journal of Medical Systems, vol. 34, no. 5, pp. 849–858.

Kanski, J. J. (2007). Clinical Ophthalmology: A Systematic Approach. 6th Edition, Edinburgh: Butterworth-Heinemann/Elsevier, 491 pages.

Klonoff, D. C. & Schwartz, D. M. (2000). An economic analysis of interventions for diabetes”, Diabetes Care, vol. 23, no. 3, pp. 390-404.

Lupascu, C. A. & Tegolo, D. (2011). Automatic unsupervised segmentation of retinal vessels using self-organizing maps and k-means clustering, In Computational Intelligence Methods for Bioinformatics and Biostatistics. Springer Berlin Heidelberg. pp. 263-274.

Mahalakshmi, S. & Velmurugan, T. (2015). Detection of Brain Tumor by Particle Swarm Optimization, Image Segmentation. Indian Journal of Science and Technology, vol 8 issue 22, pp. 13-19. DOI: 10.17485/ijst/2015/v8i22/79092.

Mapayi, T., Tapamo, J. R & Viriri, S. (2015). Retinal Vessel Segmentation: A Comparative Study of Fuzzy C-means and Sum Entropy Information on Phase Congruency International Journal of Advanced Robotic Systems, vol. 12, no. 133, pp 1-11, doi: 10.5772/60581.

Mapayi, T., Viriri, S. & Tapamo, J. R. (2014). A New Adaptive Thresholding Technique for Retinal Vessel Segmentation Based on Local Homogeneity Information, In Image and Signal Processing. Springer International Publishing, Ser. Lecture Notes in Computer Science, pp. 558-567

Mapayi, T., Viriri, S. & Tapamo, J.R. (2015). Comparative study of retinal vessel segmentation based on global thresholding techniques, Computational and Mathematical Methods in Medicine, vol. 2015 Article ID 895267.

Marin, D., Aquino, A., Gegundez-Arias, M. E. & Bravo, J. M. (January 2011). A New Supervised Method for Blood Vessel Segmentation in Retinal Images by Using Gray-Level and Moment Invariants-Based Features, IEEE transaction on medical imaging, vol.30 no. 1, pp. 146-158.

Marrugo, A. G., Milln, M. S., Cristbal, G., Gabarda, S., Sorel, M. & Sroubek, F. (June, 2012). Image analysis in modern ophthalmology: from acquisition to computer-assisted diagnosis and telemedicine, In SPIE Photonics Europe, International Society for Optics and Photonics, pp. 84360C-84360C.

Mendonca, M., & Campilho A.J. (2006). Segmentation of Retinal

Blood Vessels by Combining the Detection of Centerlines and Morphological Reconstruction, IEEE Trans Med Imag., vol 25, pp. 1200-1213.

Meng, L., Zhenshen, M., Chao, L., Guang, Z., & Zhe, H. (2017). Robust Retinal Blood Vessel Segmentation Based on Reinforcement Local Descriptions. Hindawi BioMed Research International, Volume 2017, Article ID 2028946, 9 pages https://doi.org/10.1155/2017/2028946

Mohsen, H., Nader, K.S.M., Reza, S., Shadrokh, S. & Kayvan, N. (2017). Retinal blood vessel segmentation for macula detachment surgery monitoring instruments, Int J Circ Theor Appl. 2018;1–15. DOI: 10.1002/cta.2462.

Niemeijer, M. Staal, J. Van Ginneken, B. Loog, M. & Abramoff, M. D. (2004). Comparative study of retinal vessel segmentation methods on a new publicly available database, Proc SPIE Med Imaging, vol. 5370, pp. 648-656.

Oliveira, W. S., Teixeira, J. V., Ren, T. I., Cavalcanti, G. D. C., Sijbers, J. (2016). Unsupervised Retinal Vessel Segmentation Using Combined Filters. PLoSONE, vol. 11, issue 2, e0149943. DOI:10.1371/journal.pone.0149943.

Qin, L. You, J., Zhang, D. & Bhattacharya, P. (2006). A Multiscale Approach to Retinal Vessel Segmentation Using Gabor Filters and Scale Multiplication, IEEE International Conference on Systems, Man and Cybernetics (SMC '06). vol.4, pp. 3521-3527.

Ramaswamy, M., Anitha, D., Priya Kuppamal, S., Sudha, R., Fepslin, S. A. M. (2011). A Study and Comparison of Automated Techniques for Exudate Detection Using Digital Fundus Images of Human Eye: A Review for Early Identification of Diabetic Retinopathy, Int. J. Comp. Tech. Appl., vol. 2 no. 5, pp. 15031516.

Research Section, Digital Retinal Image for Vessel Extraction (DRIVE) Database (2017). Utrecht, The Netherlands, Univ. Med. Center Utrecht, Image Sci. Inst. [Online]. Available: http://www.isi.uu.nl/Research/Databases/DRIVE.

Ricci, E. & Perfetti, R. (2007). Retinal blood vessel segmentation using line operators and support vector classification", IEEE Transactions on Medical Imaging, vol. 26 pp. 1357-1365.

Saatchi, S. & Cheng Hung, C. (December 2007). Swarm Intelligence and Image Segmentation. Swarm Intelligence: Focus on Ant and Particle Swarm Optimization, Book edited by Felix T. S. Chan and Manoj Kumar Tiwari, pp. 532, Itech Education and Publishing, Vienna, Austria ISBN 978-3-902613-09-7.

Saffarzadeh, V. M., Osareh, A., & Shadgar, B. (2014). Vessel Segmentation in Retinal Images Using Multi-Scale Line Operator and K-Means Clustering", Journal of medical signals and sensors, vol. 4 no. 2, pp. 1-22.

Sumathi, T., Vivekanandan, P., Ravikanth, B. (2018). Retinal vessel segmentation using neural network. IET Image Process., 2018, Vol. 12 Iss. 5, pp. 669-678, doi: 10.1049/iet-ipr.2017.0284.

Szpak, Z. L. & Tapamo, J. R. (2008). Automatic and Interactive Retinal Vessel Segmentation, South African Computer Journal, vol. 40, pp. 23-30.

Varughese, S., Gilbert, C., Pieper, C. & Cook, C. (2008). Retinopathy of prematurity in South Africa: an assessment of needs, resources, and requirements for screening programmes, British Journal of Ophthalmology, vol. 92 no. 7, pp. 879- 882.

Wang, Y., Ji, G., Lin, P. & Trucco, E. (2013). Retinal vessel segmentation using multiwavelet kernels and multiscale hierarchical decomposition, Pattern Recognition vol. 46, pp. 2117-2133.

Wen, Y. H., Bainbridge-Smith, A., Morris, A.B. (2007). Automated Assessment of Diabetic Retinal Image Quality Based on Blood Vessel Detection, Proceedings of Image and Vision Computing, Hamilton, New Zealand, pp. 132-136.

World Health Organization Prevention of Blindness and Visual Impairment (2016). http://www.who.int/blindness/causes/priority/en/index8.html.

Xiao, Z., Adel, M. & Bourennane, S. (2013). Bayesian Method with Spatial Constraint for Retinal Vessel Segmentation, Computational and mathematical methods in medicine vol. 2013, Article ID 260410.

Yin, Y., Adel, M. & Bourennane, S. (2013). Automatic Segmentation and Measurement of Vasculature in Retinal Fundus Images Using Probabilistic Formulation, Computational and mathematical methods in medicine 2013 Article ID260410.

Zhang, W., Zhou, C. & Bao, X. (2015). Investigation on digital media image processing algorithm based on asynchronous and inertia adaptive particle swarm optimization. International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 8, no. 2, pp. 65–76.