An Open Access Journal Available Online

Covenant Journal of Informatics & Communication Technology (CJICT)

Vol. 8 No. 1, June, 2020

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.

http://journals.covenantuniversity.edu.ng/index.php/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

URL: *http://journals.covenantuniversity.edu.ng/index.php/cjict*

Articles

Covenant Journal of Informatics & Communication Technology. Vol. 8 No. 1, June, 2020
ISSN: *print* 2354 – 3566 *electronics* 2354 – 3507 DOI:

An Open Access Journal Available Online

Approach for Identifying Phishing Uniform Resource Locators (URLs)

Nureni Azeez, Oluwaseyi Awotunde, Florence Oladeji

Department of Computer Sciences, Faculty of Science, University of Lagos, Lagos, Nigeria.
nurayhn1@gmail.com, seyi.juliana@gmail.com, foladeji@unilag.edu.ng

Received: 06.01.2020 Accepted: 14.05.2020
Date of Publication: June, 2020

Abstract—Phishing attacks are still very rampant and do not show signs of ever stopping. According to Santander Bank Customer Service, reports of phishing attacks have doubled each year since 2001. This work is based on identifying phishing Uniform Resource Locators (URLs). It focuses on preventing the issue of phishing attacks and detecting phishing URLs by using a total of 8 distinctive features that are extracted from the URLs. The sample size of study is 96,018 URLs. A total of four supervised machine learning algorithms: Naive Bayes Classifier, Support Vector Machine, Decision Tree and Random Forest were used to train the model and evaluate which of the algorithms performs better. Based on the analysis and evaluation, Random Forest performs best with an accuracy of 84.57% on the validation data set. The uniqueness of this work is in the choice of the selected features considered for the implementation.

Keywords/Index Terms—Cyber-attacks, Decision Tree, Phishing, Random Forest, Support Vector Machine

1. Introduction

Phishing is a cyber-attack carry out by fraudulent people to defraud people of their confidential information, login credentials and also finances. They do this for either their personal gain and this attack is not just done on individuals but also on Organizations. Phishers use legitimate sites to steal internet users' private and confidential information (Nureni and Irwin, 2010).

The term "Phishing" can be backdated to the early 1990s where a group of scammers came together and created an algorithm that allows them to generate random credit card numbers which they used to create accounts on America Online (AOL) (Adebowale, et. al., 2019). This was stopped by AOL but the "phishers" did not stop there but started pretending to be AOL employees and were messaging customers for their information. As people started becoming inclined to these scams, the group of scammers moved on to emails which was harder to track. They sent multiple emails to different people and robbed them of their information. These threats started becoming rampant and these scammers moved on from emails to other platforms and started hitting other major businesses.

Phishing is a really big and serious threat which keeps increasing year by year. In 2017, phishing attacks increased by 65% and over 1 million phishing sites were created. 76% of businesses were affected by these attacks in 2018 (Azeez, et. al., 2020). These few statistics go to show how serious and dangerous phishing is to the regular users, to businesses and organizations.

There are various types of phishing attacks and some of the popular ones are:

1. Spear phishing (Gupta, et. al., 2018): with this phishing attack, attackers pose as authentic company owner by using some of the features of the authentic and target sites to trick customers into giving out their personal and confidential information
2. Pharming: attackers convert the domain name system (DNS) to numerical Internet Protocol (IP) address, so users will put in the correct website link of their choice but they get redirected to the phishing site without knowing
3. Vishing: this is where phishers call people in the pretence of family members or relatives and collect information or funds from them
4. Smishing (Gupta, et. al., 2018): phishers send SMS messages to people with fake link for them to put in their information

There was a 65% increase of phishing in 2016, with a total of 1,220,525 attacks for the year and half a billion dollars was reportedly lost to phishing in the United States every year (Azeez, et. al., 2020). Phishing attacks are still very rampant and do not show signs of ever stopping. Reports of phishing attacks have doubled each year since 2001 (Azeez and Ademolu, 2016). This goes to show that many people still fall victim to this cyber-attack. These attacks are done with precision on the part of the attackers.

Phishers study their victims to know the sites they visit regularly and ensure to contact these victims stating the need for them to change their passwords as their account could be blocked or disabled. The victims who want to preserve their accounts, will go ahead and change their password or login details, providing access for the attack. Due to this danger, a lot of individuals and companies have lost valuable information and a lot of money (Nureni and Irwin ,2010).

Because the victims do not notice the minute details that differentiate these sites from the legitimate ones, they fall prey to the attack. Through the adoption of whitelist, users will be notified when such changes occur, thereby saving them from impending danger and monumental loss.

Different techniques to combat phishing and prevent phishing have been implemented over the years. One of them is the Whitelist approach.

Whitelist, being the opposite of a blacklist is a list of sites that a user frequently visits that requires login details and are considered to be legitimate. This in turn blocks other sites that are not on the list from accessing the user's information.

This basically checks the sites that are safe and notifies the user if the site is not legitimate or if the site is not on the whitelist. Efforts were made to adopt Machine Learning (ML) approach: Naive Bayes Classifier, Support Vector Machines (SVM), Decision Tree and Random Forest for the implementation of this work.

The work aims at preventing phishing through the Whitelist approach. The objectives include:

1. To prevent phishing through the Whitelist approach
2. To identify illegal sites
3. To protect the interest and privacy of users while surfing the internet

To reduce the rapid increase in phishing attacks to a minimal level

2. Methodology

Machine learning algorithms were used in the implementation of this system. The steps taken to implement this are:

1. Data gathering: Data was gathered from PhishStorm (Azeez and Ademolu, 2016). 48,009 non-phishing sites were gathered, and 48,009 phishing sites were gathered from site. 10 features were extracted from the data.
2. Data Cleaning: Incorrect data entry was manually filtered out of the data gathered to allow the models to train using correct and authentic data
3. Model Training: Models were trained using some selected supervised machine learning algorithms (Support Vector Machines, Decision Tree, Naive Bayes Classifier and Random Forest).
4. Model Comparison: Trained models were compared based on their performances by using the following metrics: True Positive, True Negative, False Positive and False Negative.
5. Creation of Web Browser Extension: The best model based on its performance was then used to create a dataset which was used to detect phishing sites as an extension on the web browser

2.1 Data Gathering

This is the first step in implementation of the solution. It involves collecting several phishing and non-phishing sites.

Data was gathered from PhishStorm. A total number of 48,009 non-phishing sites and 48,009 phishing sites were gathered.

2.2 Data Cleaning

The data gathered contained some inaccurate entries which were inconsequential to the research. Data cleaning was done by manually going through the data and filtering out the incorrect entries in order to help the models to better understand what a phishing and a non-phishing URL looks like.

2.3 Feature Extraction

This is where the data is converted to dataset of lesser number of variables based on the features selected containing the right amount of information to work with. Some features were selected to check the URLs and how well the models perform. A total of 8 features were selected to check the legitimacy of the URLs.

The Features are:

1. Length of URL
2. HTTPS token
3. Number of dots
4. Number of sub-domains
5. Digit count in the URL
6. Suspicious characters like @ and %40
7. Multiple occurrence of https, http

The features were divided into numerical and categorical features.

2.3.1 Numerical Features

These are the features that have continuous numeric data. They are data that signify a measurement or a count of values.

1. Length of URL: Most phishing sites are very lengthy because they are trying to cover the illegitimacy of their sites such that users will not be able to see it due to the length. Because URLs are broken down into three major parts with various sub-parts, this feature will be broken down to best classify the site

$$f_1 = length\ of\ the\ host\ name$$

$$f_2 = length\ of\ path$$

2. Number of dots: Phishing sites tend to have a lot of dots in their host name unlike legitimate sites with less than two dots. URLs that have many numbers of dots are most times categorised as phishing sites.

$$f_3 = number\ of\ dots$$

3. Number of sub-domains: Phishing sites are known to want to duplicate original sites and they tend to use the same name but add extra words to it, making the user think he is on a safe site. These extras are most times added between domain of a legitimate site and they are most times more than one.

$$f_4 = number\ of\ sub-domains$$

4. Digit count in the URL: The occurrence of digits in a legitimate URL is very rare and if it exists, the digits are always very few. Phishing sites tend to have a lot of digits in their URL.

$$f_5 = number\ of\ digits$$

2.3.2: Categorical Features

These are the features that have discrete numeric data. They are data that signify uncountable data and data that can be described using intervals.

1. HTTPS token: Websites are said to be secure when they have an https token but illegitimate and not secure sites do not have that but instead have http

$$f_6 = \begin{cases} 1, & \mathit{HTTPS\ token} \\ 0, & \mathit{not\ HTTPS\ token} \end{cases}$$

2. Suspicious characters like @ and %40: Legitimate sites do not have the occurrence of '@', '_' and '% in their URLs. URLs that have any one of these suspicious characters can be categorized as phishing sites

$$f_7 = \begin{cases} 1, \mathit{has\ no\ suspicious\ characters} \\ 0. \ \mathit{has\ suspicious\ characters} \end{cases}$$

3. Multiple occurrence of https, http: Websites are required to have just one occurrence of https or http but when a URL has more than one of these tokens, it can be said to be a phishing site

$$f_8 = \begin{cases} 1, & \mathit{one\ occurrence\ of\ https,http} \\ 0, & \mathit{multiple\ occurrences\ of\ https,http} \end{cases}$$

2.4 Model Training and Algorithms Used

Decision was reached on the three algorithms because of their popularity along with observable contradictory results obtained on them from previous researches. What is more, they can also provide relatively good performance on the classification task in this work.

The data collected was separated into training and testing sets. Some part of the data was used to train the model using the features extracted based on the aforementioned supervised machine learning algorithms (Naive Bayes Classifier, Support Vector Machine, Decision Tree and Random Forest) and results were obtained. The testing data was then fed into the model to see how well it has trained.

2.4.1: Naive Bayes Classifier

Naive Bayes Classifier is a machine learning model or classifier that uses the Naive Bayes' theorem of probability. It is used to predict a class of unknown circumstances. The classifier assumes that the predictions on a class are not dependent on each other.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \text{..............} \ (1)$$

Where P (c | x) is the posterior probability of class given predictor
P (x | c) is the likelihood i.e. probability of attribute given class
P(c) is the prior probability of class
P(x) is the prior probability of predictor

This algorithm assumes that the features are independent of each other so it tests the data based on the features individually (Jain and Gupta, 2016). How it works is that:

1. It converts the data set into a frequency table
2. Creates a table of likelihood to derive the probabilities of each feature
3. The algorithm was implemented using Python

2.4.2: Support Vector Machine

Support Vector Machine (SVM), is a supervised learning model that is used to analyse data for classification and regression problems. It is a model that best splits data. It works as follows: each data item is plotted as a point in ndimensional space (n is the number of features) and the values of each of the features is the value for a specific coordinate.

A margin of best fit is plotted to show how best the data can be split and this margin is referred to as a Hyperplane (Azeez and Babatope, 2016). The points closest to the hyperplane on opposite sides are referred to as the Support Vectors. The distance between the support vectors and the hyperplane should be as far as possible.

The algorithm uses support vectors and hyperplanes, where support vectors are the vectors closest to the plane and the hyperplane is the line of best fit that passes through the points or vectors (Nivedha et. al., 2017). The steps taken to use this are:

1. Identify the right hyperplane
2. Classify the two classes in the data
3. Implement it using Scikit-learn libraries

2.4.3: Decision Tree

A Decision Tree is a prediction model used in machine learning to solve problems of classification and regression. It is designed in the form of a tree-like graph and the data set is split using different features or conditions. It represents decisions and decision making. It represents the if-else statement (Nivedha et. al., 2017).

How this works is:

1. Start with a training data set that has attributes and classification
2. Ascertain the best attribute in the dataset
3. Split this set into subsets with values of this best attribute
4. Generate decision tree nodes based on the best attribute
5. Keep generating nodes using the subset from (3) till you cannot classify further

2.4.4: Random Forest

This model makes use of many decision trees, hence the word “Forest”. It is used for classification and regression. To classify a new instance, each decision tree provides a classification for the input data. The classification from all the trees are taken and the prediction with the highest “vote” is selected (Chiew et. al., 2020).

How it works is:

1. When classifying a new object, different decision trees are used
2. Each decision tree classifies the input data
3. All the classifications made by the trees are taken and compared
4. Vote is taken for the classification
5. The classification with the highest vote is selected

2.5: Model Evaluation

The model results were accessed for each of the machine learning algorithms used. The models were accessed based on their performances (Wu et. al., 2018). The following metrics were used to evaluate the models:

1. **Confusion Matrix** (True Positive, False Positive, True Negative and False Negative): ***True positive*** is when the assumed class of a data is 1 (true) and the predicted result

is 1 (true) (Al-Janabi et. al., 2017). ***False Positive*** is when the assumed class is 0 (false) and the predicted is 1 (true). True Negative is when both the assumed and the predicted result are 0 (false) and ***False Negative*** is when the assumed data class is 1 (true) and the predicted result is 0 (false)

2. **Accuracy**: This refers to the amount of correct predictions made by the model

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total}$$
… (2)

3. Precision: This refers to how concise and exact the predictions are in that, the sites we predicted as phishing sites are actually phishing sites, same for legitimate sites

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Negative}$$
… (3)

4. Recall or Sensitivity: This refers to the correctness of the models in diagnosing the sites as phishing or non-phishing or legitimate. The sites which are phishing should be predicted as phishing, same for non-phishing

$$Sensitivity = \frac{True\ Positive}{Actual\ True} \ldots (4)$$

5. Specificity: This refers to the correctness also in that, the sites that are legitimate were predicted to be legitimate by the models

$$Specificity = \frac{True\ Negative}{Actual\ False} \ldots\ldots. (5)$$

2.6 Implementation of Web Browser Extension

Implementation was carried out using JavaScript, HTML, and CSS and it is categorized into five steps.

1. Create the project: This is where the file and the folder to house these files were created. A manifest file is created which tells the browser what it needs to know in order to open the extension. The HTML and CSS files are also created which contains the display of the extension. A separate file was created to hold any script file and it references the HTML file
2. Update the manifest file: Code was added to the manifest file which is in a JSON format
3. Create the UI: Writing of the code in the HTML page that allows you to click on the extension icon
4. Implement how the UI should work: Write the script such as event listeners
5. Test the Implementation: This is where the extension created was tested to know if it is working fine or needs any improvement

3.0: System Design

The application is in the form of a web browser extension where once there is a change in the URL, the Whitelist system scans the URL and compares it to the ones already on the whitelist.

If there are similarities between the new URL and one of the URLs on the list, the user will be notified that progress can be made. Whereas, if there is no similarity, the user is notified about the change and required to stop all transactions on that site.

4. Machine Learning Techniques

The model was evaluated using the four machine learning algorithms (Naïve Bayes, Support Vector Machine, Decision Tree, and Random Forest). The result gotten from the comparison of the evaluation was used to determine the algorithm that will then be used to create the Web Extension.

4.1 Naïve Bayes

The confusion matrix for Naive Bayes was able to correctly classify 8663 URLs as authentic (True negatives), wrongly classified 3600 URLs as authentic (False negatives), wrongly classified 1015 URLs as phishing (False positives) and correctly classified just 5904 URLs as phishing (True positives).

Table 1 shows a total Precision of 0.71 and 0.85 for both non Phishing and Phishing when using Naïve Bayes. The corresponding graphical interpretation is shown in Figure 1.

Table 1 Model Evaluation for Naïve Bayes

Category	**Precision**	**Recall**	**F1-Score**	**Support**
Non-Phishing	0.71	0.90	0.79	9678
Phishing	0.85	0.68	0.72	9504
Total/Average	0.78	0.76	0.75	19182

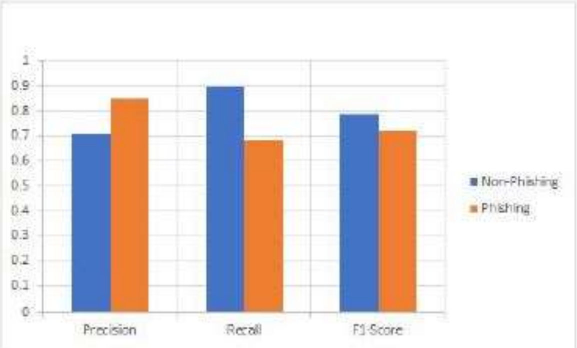

Figure 1. Graph of Model Evaluation for Naïve Bayes

4.2 Support Vector Machine (SVM)

The confusion matrix for SVM was able to correctly classify 8762 URLs as authentic (True negatives), wrongly classified 3663 URLs as authentic (False negatives), wrongly classified 916 URLs as phishing (False positives) and correctly classified just 5841 URLs as phishing (True positives)

Table 2 Model Evaluation for SVM

Category	**Precision**	**Recall**	**F1-Score**	**Support**
Non-Phishing	0.71	0.91	0.79	9678
Phishing	0.86	0.61	0.72	9504
Total/Average	0.78	0.76	0.76	19182

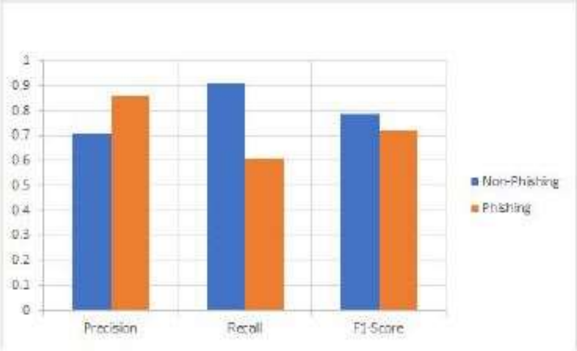

Figure 2. Graph of Model Evaluation for SVM

Table 2 shows 0.71 and 0.86 as values for Precision for both non-Phishing and Phishing with SVM. The graphical interpretation is shown in Figure 2.

4.3 Decision Tree

The confusion matrix for Decision Tree was able to correctly classify 8624 URLs as authentic (True negatives), wrongly classified 2178 URLs as authentic (False negatives), wrongly classified 1054 URLs as phishing (False positives) and correctly classified just 7326 URLs as phishing (True positives).

Table 3 Model Evaluation For Decision Tree

Category	**Precision**	**Recall**	**F1-Score**	**Support**
Non-Phishing	0.80	0.89	0.84	9678
Phishing	0.87	0.77	0.82	9504
Total/Average	0.84	0.83	0.83	19182

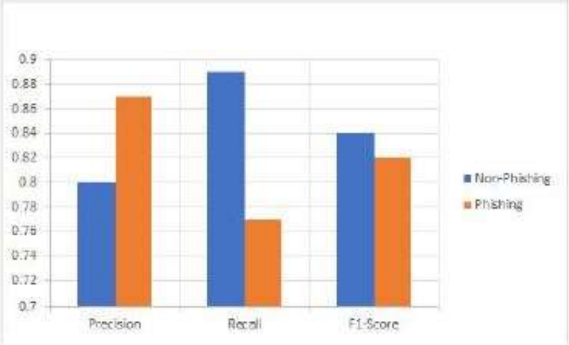

Figure 3. Graph of Model Evaluation for Decision Tree

Table 3 and Figure 3 provide the values obtained for both categories (Non Phishing and Phishing) when Decision Tree was considered.

4.4 Random Forest

The confusion matrix shows Random Forest was able to correctly classify 8545 URLs as authentic (True negatives), wrongly classified 1895 URLs as authentic (False negatives), wrongly classified 1133 URLs as phishing (False positives) and correctly classified just 7609 URLs as phishing (True positives).

Table 4 Model Evaluation for Random Forest

Category	Precision	Recall	F1-Score	Support
Non-Phishing	0.82	0.88	0.85	9678
Phishing	0.87	0.80	0.83	9504
Total/Average	0.84	0.84	0.84	19182

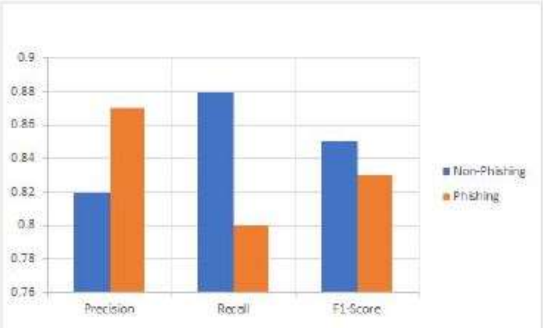

Figure 4. Graph of Model Evaluation for Random Forest

Table 4 and Figure 4 provide the values obtained for both categories (Non Phishing and Phishing) when Random Forest was considered.

Based on this comparison shown below, Random Forest has a higher model evaluation compared to the rest. It has the highest recall, that is, it is correctly classifying the non-phishing URLs as non-phishing, therefore, it is the best option to use to create the Web Extension.

Table 5. Comparison of the algorithms performances

Category	**Precision**	**Recall**	**F1-Score**
Naïve Bayes	0.78	0.76	0.75
SVM	0.78	0.76	0.76
Decision Tree	0.84	0.83	0.83
Random Forest	0.84	0.84	0.84

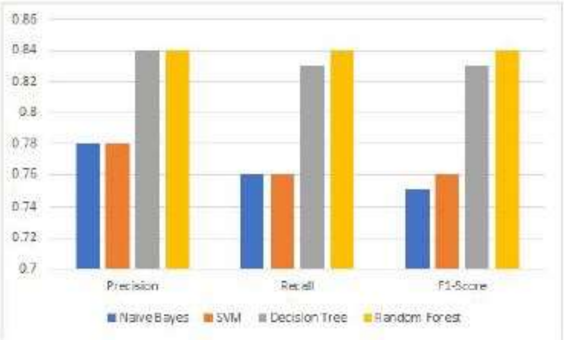

Figure 5. Graph of Comparison of the Algorithms Performances

Table 5 and Figure 5 provide the summary of the values obtained for both categories (Non Phishing and Phishing) when all the Machine Learning algorithms were evaluated.

5. Related Work

This part shows the review of articles of journals, documents from the internet on what phishing is about and the methods or approaches used to detect and prevent phishing. These methods were reviewed based on their benefits and their weaknesses in solving phishing.

In the work of Dudhe and Ramteke, they discussed the use of various approaches to detect phishing. The use of known and new features was applied in preventing phishing. They made use of Blacklist-Whitelist based approach, Fizzy rule-based approaches, Machine learning approaches, heuristic approach, CANTINA based approaches and Image based approaches to prevent and detect phishing for users (Dudhe and Ramteke, 2015). These approaches were used to determine which of them is the best among the anti-phishing techniques listed and the heuristic approach was said to be the best or at least better than the other approaches. The weakness of this is in its inability to work on server-side security.

A desktop application called PhishShield that takes a URL as input and brings the status of the URL (either phishing or legitimate website) as the output was implemented and discussed in the work of Rao and Ali. It has an accuracy rate of 96.57% and it can detect phishing sites that trick users by changing the contents to images (Rao and Ali, 2015). This implementation made use of the heuristic approach and was said to detect phishing attacks that blacklists cannot detect. It is considered to be faster than visual based assessment techniques that have been used in phishing detection (Strinzel, 2019). However, the result can still be improved upon in terms of its performance and the cost of computation using techniques like generic algorithms, neural network.

In 2015, Sedgewick et. al., developed Application Whitelisting which uses whitelists to determine the applications that are allowed to execute on a host, thereby preventing malware and other unapproved software. They wanted to educate organizations on the use and implementation of application whitelisting (Sedgewick et al., 2015). They discussed how highly recommended these solutions are when it comes to security. Organizations who want to make use of these solutions should be risk conscious when it comes to deploying the solutions. It requires diligence among staff to maintain and manage the solutions.

A very promising method to avoid phishing, Zero Knowledge Authentication (ZeKo), was developed by Shar et al., in 2015. The solution protects users from phishing attacks. The reasons phishing ((Matumba et. al., 2019). is still a rampant and growing attack is due to the ignorance of the users when it comes to computer and its usage. Users fail to see the slightest change in the URL; they fail to notice security warnings when they are on a website. They studied human behaviour in relation to phishing and realised that the attackers go for users that are gullible and extract their classified information directly from them. The attackers do this either via SMS, known as SMSishing and Voice conversation, known as Vishing. With this solution in place, phisher can easily be checked and prevented from carrying out his nefarious activities (Shar et al., 2015).

A content-based approach to detecting phishing using CANTINA as a good phishing site detector was implemented (Dudhe and Ramteke, 2015). The implementation made use of PHP and MYSQL, also making use of web crawlers. It basically crawls the original website URL, the location of the server and 'whois' information. When a user gets an email attached with a phishing link, the system takes the URL, that is, the link, and compares it with the original URL. It also does that for the location of the server and the 'whois' information. It analyses these for similarities, then conveys the result to the user. This implementation is said to be effective as it has a 6% false positive performance, then coupled with the heuristic approach, has a 1% false positive performance but it still needs to be improved on as because it is not user friendly (Gupta et al., 2015).

Fraud Website Detection application which discovers fraud websites through the use of RIPPER algorithm to categorize the websites was implemented by Prajapati et al., in 2016. This application takes corrective measures against fraudulent websites by reporting the prospective sites to the concerned authority. They went on to discuss different approaches used to detect fraud websites and how Heuristic approach is the better approach as it can detect fraud websites before they are blacklisted (Rao and Ali, 2015). The application still needs to be improved upon as it can be a plug-in to the browser, thereby, notifying the users when they are surfing the internet.

A novel approach for phishing protection that makes use of auto-updated whitelist of all authentic sites that a user access was implemented. A whitelist has a list of all the legitimate sites a user can visit while blacklist contains all the sites that a user should not visit as it is a phishing site. This approach has the likelihood of detecting attacks very well and very fast. It is sufficient for a real-time environment and it can be improved upon by using other features to detect phishing and legitimate sites even if these new features will increase running time complexity of the system (Gupta and Jain, 2016).

Rao and Ali made use of an enhanced heuristic approach to combat phishing where blacklist and whitelist were made use of. Websites that are not legitimate and are not already on the blacklist are discovered and the blacklist is updated, same for the whitelist where it is updated on the legitimate sites that are not already on it (Rao and Ali, 2015). The solution was implemented using PHP programming and Database and has a high accuracy level (Okunoye et al., 2016). It is said to be highly effective and user-friendly but it still needs to be further worked on as it does not use visual similarities approach which makes it time consuming.

6. Conclusion

Having fully known the danger of phishing in the global community, it is an understatement to say that it has caused financial damages in most financial institutions. The essence of carrying out this research is, therefore, in the right direction. The machine approach adopted has clearly revealed how the adopted approach can be fully utilized in identifying phishing URLs and curtailing phishers. The summary of the results obtained as shown in Table 5 revealed that Random Forest performed has the best performance with the metrics considered. Phishing URLs can easily be detected if users are conscious of the change in the URLs and also when web extensions can notify the user if the URL is a phishing or non-phishing one. In order to achieve maximum accuracy, we propose that neural networks should be used for future research instead of traditional ML approach adopted in this case. Consequently, the proposed application can identify phishing URLs with an accuracy of 84.57%.

References

Adebowale MA, Lwin KT, Sánchez E, Hossain MA (2019) Intelligent web-phishing detection and protection scheme using integrated features of Images, frames and text. Expert Systems with Applications 115 (2019) 300–313.

Azeez NA, Salaudeen BB, Misra S, Damaševičius R, Maskeliūnas R et al (2020). Identifying phishing attacks in communication networks using URL consistency features. Int. J. Electronic Security and Digital Forensics, Vol. 12, No. 2, pp 200-213

Gupta BB, Arachchilage NAG, Psannis KE (2018) Defending against Phishing Attacks: Taxonomy of Methods, Current Issues and Future Directions. Telecommunication Systems volume 67, pp 247–267, https://doi.org/10.1007/s11235-017-0334-z

Azeez NA, Ademolu O (2016) CyberProtector: Identifying Compromised URLs in Electronic Mails with Bayesian Classification. 2016 International Conference Computational Science and Computational Intelligence (CSCI) Las Vegas, NV, USA: IEEE, 959-965.

Jain AK, Gupta BB (2016) A novel approach to protect against phishing attacks at client side using auto-updated whitelist. EURASIP Journal on Information Security. doi:10.1186/s13635-016-0034-3

Azeez NA, Babatope AB (2016) AANtID: an alternative approach to network intrusion detection. The Journal of Computer Science and its Applications. An International Journal of the Nigeria Computer Society, 129-143.

Nivedha S, Gokulan S, Karthik C, Gopinath R et al (2017). Improving Phishing URL Detection Using Fuzzy Association Mining. International Journal of Engineering and Science (IJES), 21-31.

Chiew KL, Tan CL, Wong K, Yong KSC. , Tiong WK (2020) A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. Information Sciences 484 (2019) 153–166

Wu C, Shi J, Yang Y, Li W (2018) Enhancing Machine Learning Based Malware Detection Model by Reinforcement Learning. ICCNS 2018, November 2–4, 2018, Qingdao, China.

Al-Janabi M, Quincey E, Andras P (2017) Using supervised machine learning algorithms to detect suspicious URLs in online social networks. 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Nureni AA, Irwin B (2010). Cyber security: Challenges and the way forward. Computer Science & Telecommunications, 29, 56-69.

Nivedha S, Gokulan S, Karthik C, Gopinath R et al (2017). Improving Phishing URL Detection Using Fuzzy Association Mining. International Journal of Engineering and Science (IJES), 21-31.

Dudhe P.D, R. P. (2015). A Review on

Phishing Detection Approaches. International Journal of Computer Science and Mobile Computing, 166-170.

Rao R.S, A. S. (2015). PhishShield: A Desktop Application to Detect Phishing Webpages through Heuristic Approach. Eleventh International Multi-Conference on Information Processing-2015, 147-156.

Sedgewick A, S. M. (2015). *Guide to Application Whitelisting.* National Institute of Standards and Technology Special Publication 800-167.

Shar K, S. T. (2015). Phishing: An Evolving Threat. *International Journal of Students Research in Technology & Management*, 216-222.

Gupta B.B, A. A. (2015). Defending against Phishing Attacks: Taxonomy of Methods, Current Issue and Future Directions.

Prajapati U, S. N. (2016). Fraud Website Detection using Data Mining. *International Journal of Computer Applications*, 0975-8887.

Gupta B.B., J. A. (2016). A novel approach to protect against phishing attacks at client side using auto-updated white-list. *EURASIP Journal on Information Security*.

Okunoye B, A. N. (2016). PHISHDETECT: A Web Enabled Anti-Phishing Technique using Enhanced Heuristic Approach. *Department of Computer Sciences, University of Lagos, Nigeria*.

Matumba L, M. F. (2019). Blacklisting or Whitelisting? Deterring Faculty in Developing Countries from Publishing in Substandard Journals. *University of Toronto Press*, 83-95.

Strinzel M, S. M. (2019). Blacklists and Whitelists To Tackle Predatory Publishing: a Cross-Sectional Comparison and Thematic Analysis. *Swiss National Science Foundation, Bern, Switzerland*.

Covenant Journal of Informatics & Communication Technology. Vol. 8 No. 1, June, 2020
ISSN: *print* 2354 – 3566 *electronics* 2354 – 3507 DOI:

An Open Access Journal Available Online

Exploring the Readability of Terms of Service of Social Networking Sites

William Kodom Gyasi & Manasseh Jonah Bangmarigu

Department of Communication Studies, University of Cape Coast, Cape Coast, Ghana

William.gyasi@ucc.edu.gh/wkodomgyasi@gmail.com
manasbangmarigu95@gmail.com

Abstract—Terms of service of social networking sites provides relevant information for end-users to make inform decision regarding their use of social networking site. The purpose of this paper was to evaluate the readability of the top five social networking sites namely Facebook, WhatsApp, Messenger, YouTube, and WeChat. Using Flesch reading ease and grade level indexes, the authors measured the readability of the selected terms of service. Measures of central tendencies, and one-way analysis of variance, with bootstrapping, were used to analyse the data. The results showed that the ToS were difficult to read when compared to the standard score of public documents. Moreover, the readability levels of the ToS required end users to have attained at least 13 years of formal education, which is equivalent to first year college student, in order for the users to find the ToS readable. In addition, no statistical difference existed among ToS of the five social networking sites meaning all the ToS of the five networks were unilaterally difficult for the average reader (end user). It is recommended that the ToS of these networks be revised to make them more readable since substantial number of their audience are young with low number of years of education.

Keywords/Index Terms— Readability, Readability Indexes, Terms of Service, Social Networking, Web Application

1. Introduction

Social networks are part of the main drivers for the growing global connections of businesses and individuals (Bahera, Rath, Damasevicius & Maskeliunas, 2019). As global networks, social networking sites connect users of diverse cultural backgrounds to interact and share information. The developers of social networking sites as part of the legal and ethical requirements add terms of service to the software packages of the networks that users are supposed to consent to before they can install and use the social networking site package. Terms of Service (ToS) refers to rules, conditions, and (or) stipulations that a user of any service must abide before services are rendered. Concerning social networking sites, terms of use defines the obligations and privileges of the end user as well as the service provider. It is a very important document that legally binds users to agree with developers specifications.

Due to its vital importance, ToS is found on the platforms of every online service provider. It is mostly a necessity for end users to agree to such ToS during account creation on the online platforms such as Social Networking Websites (SNW). Users who agree to ToS imply that they have consented to the terms and are legally bound to the agreement (Scott, 2007). In essence, agreeing to terms of use of social networking site is equivalent to signing a contract where the parties are legally bound to relate according to the articles contained in the contract, in this case terms of use.

Despite the contractual function of terms of use in social networking sites, there is a growing observation that many people consent to many ToSs without proper scrutiny when it comes to their digital activities (Gyasi & Bangmarigu, 2019). This is the case because consenting to ToS is simply by ticking a check box that says ‘I agree’ (Luger, Moran, & Rodden, 2013). So bad is the situation such that earlier researchers have quantified that less than 1% of end users actually pause to read these ToS (Bakos, Marotta-Wurgler, & Trossen, 2009), and even those who should know better (educated ones) tend not to bother (Hillman, 2005).

A critical reflection on the importance of ToS as an agreement document that spells out the responsibilities of both developers and users with regards to the social networking sites connotes the need for users to do a thorough reading to comprehend the ToS before signing to it. Understanding Terms of Service is vital to ensure responsible use of the social networking sites. For instance, terms of use of social networking sites like Facebook prohibit pornographic content for legal and ethical reasons. Moreover, other social networking sites (Twitter and WhatsApp) states the number of characters that are allowed in composing a content in the site or the volume of data that is allowed to be sent to other users through the social networking site. When this kind of information is known to users, it is possible it could reduce irresponsible use of the social network sites to promote violence and other content that are abusive to other users of the network. In one study, Azeta, Omoregbe, Ayo, Raymond, Oroge and Misra (2014) discovered that university students resorted to social media as a platform to perpetrate cultism and violence acts in their institutions. Moreover, cases of people leaking nude videos on social media have been a potential threat to privacy of other users especially celebrities. The terms of service therefore, becomes a tool to educate and spell the legal implications of actions that are prohibited in the social networking sites. But if users do not read the ToS, how will they engage in responsible use of the social networking site?

The lack of readership of ToS has been discussed on many platforms, and has gained attention in academic circles. Studies have identified three major factors that account for the lack of reading of ToS. First, many users do not consider ToS as important document when they are signing to a social networking site. As Abascal et al. (2015) discovered with regards to ToS that many users read or skim through ToS when large sums of money are involved (e.g. buying a house), medical treatment is in question (e.g. before operation), and in other similar circumstances. In such cases, users consider ToS services as least important and without any repercussion to their decisions. Another reason that accounts for users' lack of readership of ToS is the lengthy nature of almost all ToS. Terms of Service are known to be very lengthy. It has been estimated that on the average, ToS contains about 2500 words (about 6 – 7 pages long; using 1 – 1.5 spacing) (McDonald & Cranor, 2008). Due to its length, many ToS are buried deep inside the website to improve the website's attractiveness such that users do not even get to see it before ticking "I agree" (Meiselwitz, 2013).

A third reason that accounts for the lack of readership of terms of service of Social Networking Websites (SNWs) is the readability of the ToS. Terms of Service of many websites have been indicated to be very poor in terms of readability. For example, a number of authors have warned that the complexity of privacy policies, and terms and conditions hinders their readability and creates one of the key usability problems of website design (Fiesler & Brudcman, 2014; Luger et al., 2013; McDonald & Cranor, 2008). Luger et al., (2014) explicitly indicated that ToS are very unreadable, and feared that if not checked, it will result to a world of "digital under class without the capabilities to make meaningful consent choices to control their digital identities" (p. 2688). Such a situation will have a much profound effect on the masses if the ToS of a platform with large number of users, such as Social Networking Websites (SNWs), is not written at a readable grade.

Social Network Websites have become very useful and powerful means of disseminating information in this era. Humans are now able to 'connect' to others just by clicking on few buttons on websites. What makes social networking so powerful is the speed in which information can spread. With the click of a button, one can send a 'tweet' not only to one's followers (say 10,000) on Twitter, but potentially to the individuals in the network. The speed and exponential growth of information sharing is what makes social networking websites so powerful, and for which reason the use of SNWs has become deeply embedded in human society in this era. Access to social networking websites has increased due to the proliferation of internet devices such as mobile smart phones, iPads, tablets among others. As more and more users join these websites, there is a growing need to provide readable terms of service content so that almost every user will understand what they are signing in for or to. The need to produce readable terms of use is even demanding when one consider the fact that users of social network websites have different educational, cultural, age and demographics backgrounds that affect how readable a text will be to them.

However by using readability formulas, developers can obtain reading ease scores that will help them predict the readability of their ToS as public document. For instance, the flesch reading ease formula recommended a score of 60-70 in its scale of 0-100 as standard score for public document. This implies that if Terms of service are measured by these readability formulas, readability scores will help them predict the reading ease of the content before attaching it for the end-users.

In recent study published by Gyasi and Bangmarigu (2019) on the readability of terms of service of software packages, the authors discovered that the terms of service of the software packages were difficult to read, and, therefore, require revision to make them readable to users. Though the readability of terms of service of social networking websites is relevant to improve readers' readership and comprehensibility of the ToS before signing to the ToS, there is limited empirical studies on the readability of terms of service of social networking sites. This study therefore seeks to fill the lacuna by exploring the readability of terms of service of selected social networking sites namely, Facebook, Twitter, Messenger, WhatsApp, YouTube and Wechat.

The specific objectives of this paper are to determine:

1.The readability of ToS of top five SNWs by using readability formulas.

2.The difference(s) in readability of ToS among these selected SNWs.

Terms of Service (ToS) of social networking sites is regarded as the contractual document in social networking websites; therefore, it is important for users of social networks to be able to read and understand the ToS of these websites. Moreover, it is the firm belief of the authors that users who read and understand terms of service will better understand the repercussion that comes with flouting the agreement signed in the terms of use. Moreover, certain warnings on the use of the sites are embedded in the terms of service and hence reading and understanding the terms of service will help readers to use the service responsibly related consequences

2. Review of Related Works

Bahera et al. (2019) asserted that social network analysis is essential in understanding the interactions and relationships among users as well as the patterns of online users' behavior. The social network analysis however, has been focused on understanding the nodes and edges that are involved in networks. The nodes refers to the individuals, groups and organisation that are involved in a social network. Edges, on the other hand, refers to the relationships, friendship, kinship and financial transaction among social network users (Olajde, Adeshakon, Misra & Ayo, 2016). The nodes and edges make up networks, and networks can be categorized into offline and online networks. Offline networks involve interaction among nodes while online networks are in the form of online social networking websites like Facebook, Instagram and Twitter that allow users to interact and share information (Olajde et al. 2016). The online networks which are basically social networking websites involve more than examining the nodes and edges but also the terms of service that enjoin the developers and end users.

In the context of social networking sites, terms of service refers to the agreement document between social network website developer and the end-user of the website. Almost all social networking websites have a terms of service package added to the software packages that is usually downloaded and used by end-users to access the service of the social network website. End-user predominantly cannot access the service of SNWs without first consenting to the terms of service of the SNWs. The terms of service usually covers important areas such as description of the social network website, licensing for users, restrictions of use, termination of use, limitation of liability, disclaimers of warranties, copyright infringement and contact information (Gyasi & Bangmarigu, 2019). The content is therefore lengthy in providing details as to what is permitted and what is prohibited as far as the developer/manufacturer and end-user are concerned.

Despite the overarching importance of the content of ToS to end-user, the evidence from previous studies showed that terms of use are less read by end users. Meiselwitz (2013) who studied the policies and procedures of social networking websites, thus Facebook and Twitter, discovered that readers do not read terms of service because they found them too lengthy, inaccessible and their file format unappealing. The author also discovered that policies and procedures of the social networking websites are difficult to read and users do not usually notice mostly the changes to the policies because the files are placed at inaccessible location.

The consequences of end users not reading terms of service include users breaching the terms unknowingly, which can lead to their termination, or the users may try to use the website for unspecified function that the website does not support. It is also a legal flout to claim ownership of parts or blame damages of the site on developers when those claims and damages were specified as part of the terms of service. Therefore, if developers can compose readable and comprehensible terms of service for users, there is a great chance of promoting mutual beneficial use of the websites among all users and the developers. Moreover, users could avoid legal implications that could be because of their lack of knowledge of the terms of service. Finally, when users read and understand terms of service, they are better equipped to make informed decisions on which social networking websites to install on their devices. This is because as part of the terms of service, some social networking websites add promotional features such as sending automatic ads to the end user, which are sometimes having offensive content for some users. Noticing such added features, for example, through reading the terms of use could save users from making wrong installation that will affect their devices negatively. Terms of service is therefore, a vital communication tool that provides relevant information from developers to end-users so that the end-users could make informed decision on whether to use or not to use the social networking websites.

Not only is readability of terms of service a gateway to improving the relationship and communication between developers and end-users of social networking websites, but generally, the quality of websites can be measured by several factors such a accessibility, usability, functionality, efficiency and readability (Niazi & Kamram, 2012; Ismail et al. 2016). According to Pawn et al (2018), the web has become a popular and important medium for transmitting information from one place to another. A web content can be considered accessible if it follows the web 2.0 guidelines. It is useful and functional if it provides the service at an optimum efficiency it promises to offer. While it is undoubtedly true that the social networking websites are accessible, usable and efficient communication tools for diverse users, the readability of the content of social networking websites has not received much attention. Social networking website's content is considered readable and understandable to a reader if it uses simple language. Polysyllabic words and lengthy sentence affect the readability of any text (DuBay, 2004).

Studies on the text variables of social networking websites have discovered that their content contain variables that pose difficulty to measuring their readability. For instance, Davenport and DeLine (2014) argued that hashtags, acronyms, multilingual tweets and abbreviations render social networking websites content unreadable. In fact, same authors found that such variables make it difficult to use traditional readability formulas to measure the readability of social media user generated content. To compensate the limitation of traditional readability formulas efficiency to measure social media content, Pawn et al. (2015) identified some computational formulas that can help measure web content such as the GUI evaluator, age rank algorithm and many others that can help to measure web content readability by considering the non-traditional variables.

Moreover, social media content sometimes contain multilingual content. In terms of measuring the readability of multilingual content, Pawn et al. (2015) suggested language specific formulas such as fernandezhuetainde for Spanish text, Djoko formula for Indonesians text and lix for French text as alternative readability formulas that can be used to measure the readability of multilingual content of social networking websites. The advantage however, with terms of use of social networking websites is that there are predominantly written in the English language, and if not, there can be translated in the site. Also, the terms of service of social networking websites does not include the variables such as hashtags, pictures and others that affect effectiveness of readability formulas prediction. It is imperatively easier to determine the readability of terms of service of social networking websites than other content on social networking websites.

Exploring the terms of services' readability is paramount since they are communication tool between developer and user of any social networking site. Besides, the terms of service of social networking websites specify the developer and user rights, their responsibilities and their claims. Also, as indicated earlier, the terms of service is one of the content that are written in specific language, especially English Language, with less web centered variables such as hashtags, acronyms among others. It therefore implies, terms of service readability can be predicted by traditional readability metrics because the content of terms of use has no special characters that will pose difficulty to measuring their readability using traditional readability formulas.

2.1 Empirical Studies on Social Networking Websites

Social networking analysis has been a crucial aspect of studies on technology and its impact on society. Olajde, Adeshakon, Misra and Ayo (2016) studied social network analysis impact on e-commerce in south-west Nigeria. Based on their analysis of centrality of the social network sites, the authors discovered that social network analysis especially on the recommender system of e-commerce, has led to good customer satisfaction, effective product ranking and sales, high recommendation and purchasing and better understanding of customer preferences. The study of Olajde et al. (2016) provided evidence that social network interconnect people from diverse backgrounds and as such the importance of individuals of the community understanding the terms of service that regulate the producer and end user is vital. Moreover, the benefits discovered by Olajde et al. (2016) on social networks provide a better reason to argue that enhancing such benefits will be by providing readable terms of service so that users of the network will use the networks more responsible to avoid conflict of interest.

Azeta et al. (2014) investigated how to reduce or curb the perpetration of cultism and terrorism on social networking sites among university students in Nigeria. The authors developed a software package that had the capacity to filter abusive language and channel it to the university administration for further scrutiny and investigation. The authors discovered that cultism behaviors on social media websites have led to vices such as violence, rapes and recalcitrant behaviors that are uncalled for in an intellectual and moral based university setting. While the application developed by the authors provided features to deal with the cultism on social networking websites, it is ultimately important to understand the place of terms of service of social networking sites as a yardstick to measuring the agreement between programmers and users. Improving terms of service of networking websites by adding terms that prohibit uncouth behaviors will also be a way of reducing foul language among users of social media sites. However, even with the inclusion of termination of user as part of the terms of service, end users do not usually know ToS because most of them do not read the terms of use.

Also, few previous studies on the terms of service of websites have produced insightful evidence of their difficulty for readers. First, Luger et al., (2013) evaluated the readability of ToS of six UK Energy services using SMOG readability formula. The researchers observed that ToSs were written at a level that was far beyond what a functionally literate adult could be expected to understand.

Similarly, Prichard and Hayden (2008) compared the readability of freeware end-user licensing agreements of 91 freeware programs frequently downloaded in 2003 and 100 freeware programs frequently downloaded in 2008. Seven different formulas (Dale – Chall, Coleman – Liau, ARI, Fog, SMOG, Flesch reading ease and Flesch – Kincaid Grade level) were used to evaluate the readability. They found that "anywhere from 55% to 97% of the 2003 agreements studied were either difficult or very difficult to read" and that "anywhere from 61% to 97% of the 2008 agreements were categorized as either difficult or very difficult" to read. On the average, the authors found that the readability grade level score for both 2003 and 2008 agreements were between 11.76 years (i.e., senior high school) and 14.5 years (2nd year college). None of these earlier works focused on the readability of online communication services platforms. Thus, this work seeks to fill the information gap.

In a recent study, Gyasi and Bangmarigu (2019) studied the readability of thirty-eight (38) terms of service of software packages. The researchers used readability formulas to objectively measure the comprehensibility of the terms of service documents in relations to how they achieve their intended purpose of communicating developer-user agreement. The authors argued that determining the readability of terms of use of software packages is beneficial for both programmers and users. It helps programmers to communicate their intended rights and restrictions for users and it helps users to understand their rights and limitations in using the software package. Using Flesch reading ease and Flesch Kinkaid grade level to measure the reading ease and grade level of the terms of service respectively, Gyasi and Bangamarigu (2019) discovered that majority of the terms of use were difficult to read and understand. The authors therefore recommended a revision of the terms of service to make them readable for users.

From these studies, it is evident that the terms of service of most websites have been found to be difficult to read. Meanwhile, terms of service of social networking sites contain vital information such as authorization for use of the social networking website as well as restrictions of users in terms of publishing certain information or using certain content. As part of terms of service, users can be terminated if they use the social networking websites wrongly. Copyright infringement, disclaimers of warranties and contact information of developers are all included in the terms of service (Gyasi & Bangmarigu, 1995). Based on the content of terms of service, it is obvious that the content is vital for programmers and users as far as ethical and legal issues may be concerned. The recent growing interest to control publication of obnoxious content on social media such as the use of foul language and use of false identities to defraud others has prompted the need to explore how readability of terms of use can help defend programmers and promote proper use of social networking websites. It is therefore, vital to examine the readability of social networking websites' terms of use in order to ascertain whether there are readable or difficult to read.

2.2 The Concept of Readability

Readability is a crucial concept in language as far as reading and understanding of a text is concerned. Basically, readability refers to ease of reading a text (Ahmed, Zeeshan, Shaukat, & Islam, 2013). The study of readability is the study of those features of written text that aid or hinder the effective communication of ideas and information to a reader (Bailin & Grafstein, 2016). Readability of a text is largely determine by the writing style such as the choice of words and sentence length. Readability has been used as a proxy for predicting comprehension of a material (Richards & Van Staden, 2015), since direct measuring of comprehension is difficult.

Readability is usually measured by using readability formulas. A readability formula (index) is a mathematical tool used to quantify the ease of reading a text (Abascal et al., 2015). "Readability formulas are derived from regression analysis, and are based on semantic characteristics of written text (including sentence length, syllables, characters etc.) (Bailin & Grafstein, 2016). Readability formulas are based on the assumption that how difficult a text is to read is related to whether or not the words in the text are understood, and whether or not these words are put together in an easy-to-follow manner (Bailin & Grafstein, 2001)."Therefore, readability formulas hang on vocabulary difficulty and syntactic complexity. Vocabulary difficulty refers to the degree to which a text contains words that are unfamiliar and/or difficult to understand. Similarly, syntactic complexity refers to the degree to which the sentences in a text have complicated grammatical structures. That is, the longer a sentence, the more difficult it is to comprehend.

Currently, there are hundreds of readability formulas in use (Benjamin, 2012). These include the Fry formula, SMOG, and Flesch tests (Flesch-Kincaid and Flesch Reading Ease). Though the formulas vary, they estimate difficulty based on what is easy to count at the level of individual words and sentences, such as the length of words and sentences (US Dept. of Health and Human Services, 2012). "The increasing number of indexes used to evaluate readability is a function of the inadequacies or flaws underlying classical readability indexes. For example, classical readability indexes do not take into account the meaningfulness of a text during evaluation. Relying on a grade level score can thus mislead an individual into thinking that the materials are clear and effective when they are not. Because of these inadequacies of existing readability indexes, newer indexes are designed frequently and are aimed at 'fixing' these flaws. Nevertheless, classical readability indexes are adjudged as good enough for assessing readability, and they are used not in a conclusive manner (DuBay, 2004). It helps to give a writer a fair idea of the readability of his written text.

Although there are many readability indexes, not all are fit for certain texts.

For example, SMOG and Fry readability indexes have been identified to work well for almost all written text. Yet, it has been indicated that the SMOG index does not work well with low literacy texts. Therefore, in this work, the Flesch tests (Reading Ease and Grade Level) have been used since they are approved and noted in readability studies to be the most reliable metrics (DuBay, 2004; Benjamin, 2012)."

The Flesch Reading Ease (FRE) is one of the oldest and it is considered to be the most accurate of all the formulas. This formula is mostly used for academic text. It is largely used to assess the difficulty of a reading text written in English language. According to Owu- Ewie (2014) instead of using grade levels, this formula uses a scale from 0 to 100; whereby 0 corresponds to the 12th grade (Senior High School 3) and 100 is also equivalent to 4th grade (Primary 4). This simply means that the higher the score the easier the passage is to be read and the lower the score the more difficult the passage.

Flesch-Kincaid Grade Level (FKGL) Test is a related test, which translates the Flesch Reading Ease Test scores to grade level. Peter J. Kincaid and his team propounded the formula in 1975. It is mostly used in pedagogy. This formula is used to determine the readability level of a variety of educational materials especially books. This formula makes it easier for parents, teachers, and librarians to select suitable reading texts for their children/learners (Owu-Ewie, 2014). The combination of Flesch reading ease and Flesch-Kincaid grade level helps to predict the reading difficulty level of the text and the grade level (formal education) readers require to find the terms of service (use) readable and understandable.

2.3 Theoretical Framework

A study of this sort requires a theory to guide the analysis. In this study, Ehri's theory of stages of reading development and fluency is adopted. According Ehri (1995), there are four stages of reading development and fluency which are pre-alphabetic stage, partial alphabetic stage, fully alphabetic stage and skilled reading level.

The pre-alphabetic stage is where readers lack understanding of alphabetic principle, which is letters and their sounds and hence have difficulty pronouncing words, except by doing association of letters based on their visual components. In addition, the partial alphabetic stage is a stage where readers learn the letters and their sounds but their knowledge of sounds are limited hence they can find it difficult to pronounce unfamiliar words. Fully alphabetic stage refers to readers having the ability to use pronunciation and hence can pronounce unfamiliar words based on the sounds combinations. They however, may not be fluent readers as in reading fast.

The skilled level is where readers develop the skill of knowing words by sight. At this stage, readers can read fast. According to Ehri (1998), the building blocks of fluency are graphophonic letter familiarity, phonemic awareness and knowledge of graphemes. To Ehri, the ability of readers to decode a text is dependent on reading fluency.

In using Ehris's theory, Gyasi and Bangmarigu (2019) asserted that Ehri's

(1998) theory provides an understanding of the foundation in language skills, which are syntax and lexis. The lexis (words) and syntax (sentence) are crucial for readability studies. This is because readability studies consider sentence and word length as the two most important predictors of text difficulty. In this current research, the researchers posit that Ehri's theory of stages of reading development is vital in understanding how different readers of diverse reading stages will face in reading the terms of use. Users of social networking sites range from young adults to the aged. It therefore means that the reading abilities of these readers may be at different levels further requiring low readability scores. Therefore, composers of terms of use of social networking sites should consider adopting a writing style, word choice and sentence structures, that appeal to the reading abilities of the mass users of social networking websites.

3.0 Research Design

Since the study sought to describe the readability of ToS, descriptive research approach was employed (Blessing & Chakrabarti, 2009). According to Streubert and Carpenter (1999: 49), descriptive research involves direct exploration, analysis and description of the particular phenomena, as free as possible from unexplained presuppositions, aiming at maximum intuitive presentation. To Reinard (1994), 'descriptive empirical research is invited when a research problem's questions ask about current description of things and explore explanations that characterize things as they are now.' In this study, the researchers seeks to ascertain the readability levels of terms of service of social networking websites in order to ascertain whether there is the need to recommend revision of the content of terms of use.

3.1 Sample and Sampling Technique

Readability of Terms of Service of all social networks were the target population. Out of these, the ToS of the topmost five of the world's largest social networks ranked in terms of reported or estimated global Monthly Active Users or MAUs were purposively selected. That is, the sample size of ToS was five. A list of these networks is given in the Table 1.

Table 1- Top 5 Social Networks in the World (Ranking as of April 2017)

Name of Network	**Estimated MAU**	**Rank**
Facebook	1.9 Billion	1
WhatsApp	1.2 Billion	2
Messenger	1.2 Billion	3
YouTube	1 Billion	4
WeChat	889 Million	5

Source: Sparks, 2017

The ToSs of these SNWs were selected since they have a very large number of users. As such, users failure to understand the ToS of these networks due to poor readability will affect a substantial number of individuals.

After selecting these five ToS, readability scores for the entire ToS of each SNW were calculated. The selection of texts in each ToS was done such that each subheading constituted one sample of text. This approach was used since different subheadings treated different legal requirements. Hence, different readability levels were expected across texts of different subheadings.

3.2 Data Collection

Selected texts for which readability scores were to be calculated were first prepared by removing things (e.g. punctuation marks) that would have otherwise confused and mislead the computer in the calculation. This was done in line with the recommendations of earlier researchers (US Dept. of Health and Human Services, 2012). Since the computer interprets any period (full stop) as the end of a sentence, embedded punctuation such as periods that were used for abbreviations were first removed. In addition, texts that were not in full sentences, such as titles, headings, and bulleted points that are not full sentences were excluded.

This preparatory stage was important in order to achieve accurate readability scores since MS Word readability program tells the computer to sense the end of a sentence by looking for the type of punctuation that normally marks the end of a sentence, such as a period, question mark, or exclamation point. Because sometimes this punctuation falls within a sentence, rather than at the end, the computer cannot distinguish this, and thus result in errors in its computations. Therefore, each selected text was prepared before measurement with the readability formulas.

Prepared texts were copied into Microsoft Word. Readability scores of the terms of service were subsequently calculated using the inbuilt Flesch readability calculator in MS Word.

Two readability indexes were used to calculate the readability of the ToS, namely, Flesch Readability Ease and Flesch – Kincaid Grade Level. The Flesch Reading Ease and Flesch – Kincaid Grade Level were employed because they are among the oldest readability indexes, are considered to be the most accurate of all the readability formulas.

3.3 Data Analysis

With the help of IBM Statistical Products and Services Solutions (SPSS) version 24.0, frequencies, percentages, means, and standard deviations were used to describe the readability of the terms of use of the selected networks.

A one-way analysis of variance, using bootstrapping technique, was performed for samples of 1000 to ensure robust estimates of significance or p-value, standard errors and the confident intervals (IBM, 2013). To ensure robust confidence intervals, Bias corrected and accelerated (BCa) intervals were used since it ensures adjusted intervals that are more accurate (IBM, 2013). Mersenne Twister Random Number Generator was set to replicate a sequence of random numbers. This helped to preserve the original state of the random number generator and restore that state after the analysis was completed (Arbuckle, 2010). The Simple method was used since it helps to resample with replacement from the original dataset. No post-hoc analysis was conducted because the results of the ANOVA was statistically insignificant.

4.0 Results and Discussion

The main purpose of this study was to explore the readability of terms of service of social networking websites of top five social media websites. The readability of the terms of service of the selected social networking websites were determined by using Flesch reading ease and Flesch Kincaid grade level readability formulas.

Table 2 presents descriptive statistics of the readability of ToS of the top five social networking websites. The table 2 also included the means and standard deviations of the readability scores of the terms of service for better interpretation the readability levels.

Table 1 -Descriptive Statistics showing readability scores of ToS of top 5 SNWs

Social Network		**Min**	**Max**	**Mean**	**SD**
Facebook	FRE	3.60	61.40	36.85	14.82
	FKGL	10.10	21.80	14.57	3.56
WhatsApp	FRE	10.40	62.60	39.19	14.10
	FKGL	8.60	19.60	13.83	3.20
Messenger	FRE	.00	55.20	30.52	25.33
	FKGL	9.00	36.20	18.67	12.37
YouTube	FRE	25.80	66.70	36.75	11.00
	FKGL	9.90	18.20	15.05	2.56
WeChat	FRE	16.10	48.10	35.24	10.04
	FKGL	11.70	18.60	15.16	2.04

According to Owu- Ewie (2014), Flesch reading ease predicts the level of reading difficulty of a text based on the higher the score the easier the text is to read and understand, and the lower the score the difficult the text is to read and understand. From that perspective, it is evident that the mean readability scores of the selected terms of service were Facebook (36.85), WhatsApp (39.19), Messenger (30.52), YouTube (36.75) and WeChat (35.24). From these scores, the researchers ascertain that the readability of the terms of service were generally difficult. To better understand the implication of these scores on readers, the researchers included the Flesch Kincaid grade scores, which translated the Flesch scores to grade level scores. The grade level scores predicted the number years of formal education a reader should have attained before he/she could find the text readable and understandable. From the means scores of Flesch Kincaid grade level, a end-user requires at least 13 years of formal education, thus equivalent to college level, before there can find any of the terms of service readable.

It is succinct that the ToS of the top five SNWs were generally ‘difficult to read’ when measured in terms of the Flesch reading ease. The highest mean Flesch reading ease score was recorded for the ToS of WhatsApp social network (FRE = 39.19, SD = 14.10). This implies that on the average, the easiest to read ToS was that of WhatsApp social network. Although WhatsApp's ToS emerged as the easiest to read on the average, the readability level was still 'difficult' to read when considered on the Flesch Kincaid grade level. One required over 13 years of education to be able to comprehend this ToS (FKGL = 13.83, SD = 3.20). In contrast, the most difficult to read ToS was that of Messenger (= 30.52, SD = 25.33), requiring readers to attain nearly 19 years of formal education in order to be able to comprehend the terms of service (FKGL = 18.67, SD = 12.37). From the mean score of the ToS of Messenger social network (app), the readability of the ToS was 'very difficult' to read when measured in terms of the Flesch reading ease.

Generally, one major reason that accounts for end-user of networking websites avoiding reading of the terms of service of the sites is the readability of the terms of service. Meiselwitz (2013) discovered that placement of files of social media policies and procedures affected readers' access to the policies but then, the readability of the policies and procedures of the social networking websites were difficult to read. In the same lens, Prichard and Hayden (2008) concluded from their study that majority of end user agreements were difficult to read. Prichard and Hayden (2008) found grade level scores of end user agreement to be between 11.76 to 14.5 years. This implies the authors grade level scores were slightly lower than the current study which recorded grade level scores between 13.83 to 18.67 years. Similar findings were also reported by Luger et al., (2013) with respect to ToS of Energy companies in the UK. In this study, the readability scores were high because the terms of service included complex syntactical and lexical items.

As Ehri indicated in the reading development and fluency theory that the graphophonic letter familiarity, phonemic awareness and knowledge of graphemes contributes to reading difficulty as well as fluency. As far as the readability scores are concerned, it is clear that the word choice and sentence structures of the selected terms of use were not written in a readable style, which therefore negatively affected the readability scores. From Ehri's theory, the authors glean that the unfamiliarity of the words, the phonemic difficulties and lack of knowledge of the graphemes will woefully affect readers' understanding of the terms. Worst of all is that as DuBay (2004) argued that when a text is difficult to read, readers are likely to get bored and therefore, avoid reading the text entirely. DuBay (2004) added that even when the readers persist to read, the reading difficulty of the text will divert their concentration as they grapple to find meaning of polysyllabic words or complex grammatical structures. From this point of view, it logical to agree with Gyasi and Bangmarigu (2019) that the readability of terms of use greatly affect the readership and comprehensibility of terms of use.

Meanwhile, considering that ToS outlines the legal requirements and privileges of end users of this medium of communication (social media), it is reasonable to assume that such information should be, at a minimum, written in such a way as to be understandable to a functionally literate individual. This assumption is met if texts are written at a plain language level (FRE score of 60 – 70; FKGL score of 7-8) since the reading level of majority of the masses is at best that of plain language level. Therefore, any ToS falling below plain language readability level will be practically difficult to read and understand by majority of users of these SNWs. Considering this study, none of the ToS scored a readability of a plain language level or below and this is suggesting that all the ToS are far difficult to comprehend by target audience.

If SNWs genuinely wish to communicate effectively with a wider proportion of their customer base, then it is expedient that the ToS are revised to grade level of 7 – 8. This is particularly important since people of varying educational backgrounds use SNWs. As has been suggested by Luger et al., (2013), individuals with poor literacy skills are (a) more literal in their interpretation of words, (b) ignore words unfamiliar to them, (c) place over-emphasis upon the insignificant details within a text. Moreover, such readers: (d) read more slowly than those with higher literacy levels, (e) find difficulty in identifying the key concepts within a text, and (f) often do not consider the context of the narrative (Luger et al., 2013). Therefore, writing complex ToS will make it extra difficult for audience to interpret it as meant by the authors. Thus, SNW users would not be able to make informed choice about which network to use because the ToS are generally written above the recommended grade level for public documents.

Objective 2- Are there statistically significant differences in the readability of ToS of the five SNWs?

While it is important to know the individual scores of each network terms of service, the authors examined the statistical difference among the network sites in terms of their readability scores. To achieve the objective two, the authors evaluated the relative difference of the readability scores of the top five social media sites using one-way analysis of variance (with bootstrapping). Preliminary analysis was conducted to ensure that the assumption of homogeneity of variance was not violated. The results from the ANOVA are presented in Table 3.

Table 3- Results of One-Way Analysis of Variance of FRE scores among the ToS of the top five SNWs.

FRE	**Sum of Squares**	**df**	**Mean Square**	**F**	**Sig.**
Between Groups	268.521	4	67.130	.367	.831
Within Groups	11336.321	62	182.844		
Total	11604.842	66			

From Table 3, the results indicated no significant difference existed among the readability scores of ToS of the topmost five SNWs (F $_{(62, 66)}$ = 0.367; p = 0.831). The results means that the readability scores of the terms of service of the top five social networking websites were apparently similar with no relative difference. Since there was no statistical difference among the readability scores, no posthoc analysis was conducted. The implication of this finding is that, the topmost five SNWs all have ToS of equal readability. Therefore, one cannot decide on which SNWs to choose based on this decision-making tool, the ToS, of any of these networks since they are all of equal readability.

From this finding therefore, any attempt to encourage SNWs to revise their ToS to make them readable should target all these five SNWs with the same amount of seriousness and attention. With respect to Ehri's theory of fluency, the authors of this study opine that the text will be difficult for most readers because of the unfamiliar words, complex sentence structures among others that are affecting the readability of the terms of service of social networking applications. It is therefore, the firm position of the researchers that the terms of service of social networking sites be revised to meet the standard reading levels of most users of social networking sites.

5.0 Conclusion

In this study, the researchers examined the readability of ToS of the world's top five social networks. Using the readability formulas and descriptive statistics, the researchers measured the readability of the terms of use of the social networking sites and analysed the data respectively. The analysis revealed that the ToS of these social networks were generally difficult to comprehend. Amongst the five networks, the ToS of WhatsApp was found to be the easiest to read, requiring 13 years of education in order to comprehend. In contrast, ToS of Messenger social network platform was the most difficult to read, requiring users to have attained nearly 19 years of formal education to comprehend the text. In addition, the study showed that the readability level of ToS of all SNWs were similar with no significant statistical difference among them. This finding was not surprising since all ToS are typically written in legalistic format. The researchers firmly believe that revision of the terms of service of the social networking sites will improve their communicative effectiveness as well as their usefulness to end-users who require the information in terms of use to make informed decision. Moreover, revising the content of the terms of service to plain language will entice readers to read the terms rather than ignore as it is the usual practice of most end-users.

References

Abascal, J., Barbosa, S., Fetter, M., Gross, T., Palanque, P., Winckler, M., & Hutchison, D. (2015). "I Agree": The Elects of Embedding Terms of Service Key Points in Online User Registration Form. In J. Abascal, S. Barbosa, M. Fetter, T. Gross, P. Palanque, & M. Winckler (Eds.), Human-Computer Interaction INTERACT 2015 (pp. 420–428). Bamberg: Springer.

Anais, G., Davis, S. N., Havath, C., & Nina, B., (2017). How tweet readability and brand hedonism affect consumer engagement: In Na_Advance In

Consumer Research 45, Association of Consumer Research, 628-633

Arbuckle, J., L. (2010). IBM SPSS Amos 19 user's guide. Crawfordville, FL: Amos

Azeta, A. A., Omoregbe, N. A., Ayo, C. K., Raymond, A., Oroge, A., & Misra, S. (2014, June). An anti-cultism social education media system. In 2014 Global Summit on Computer & Information Technology (GSCIT) (pp. 1-5). IEEE.

Bakos, Y. Marotta-Wurgler, F. Trossen, D. R. (2009). Does Anyone Read the Fine Print? Testing a Law and Economics Approach to Standard Form Contracts, New York: University Law and Economics Working Paper.

Behera, R. K., Rath, S. K., Misra, S., Damaševičius, R., & Maskeliūnas,R. (2017). Large scale community detection using a small world model. Applied Sciences, 7(11), 1173.

Gyasi, W. K. & Bangmarigu, J. M. (2019). Readability, communication and terms of use of software packages, International Journal of Research Studies Management, 07, 10.5861/ijrsm.2019.4401, 75-88.

Blessing, L. T. M., & Chakrabarti, A. (2009). Descriptive Study I: Understanding Design. DRM, a Design Research Methodology. In Field & Andy (2013). Discovering statistics using IBM SPSS statistics. Sage.

Davenport, J. R, & DeLine, R., (2014). Readability of tweets and their geographic Correlations with Education:arXivpreprent arXiv: n 1401.6058

DuBay, W. H. (2004). The principles of readability. Costa Messa, California: Impact Information.

Fiesler, C., & Bruckman, A. (2014). Copyright terms in online creative communities. In: CHIEA 2551–2556. ACM, New York.

Hillman, R. A., (2005). On-Line Consumer Standard-Form Contracting Practices: A Survey and Discussion of Legal Implications, Cornell Law Faculty Publications, Paper No. 29.

Ismail, A., Kuppusamy, K. S., Kumar, A., & Ojha, P. K. (2016). Connect the dots: Accessibility, readability and site ranking - An investigation with reference to top ranked websites of Government of India. Journal of King Saud University - Computer and Information Sciences. https://doiorg10.1016/j.jksuci.2017.05. 007.

Kumar Behera, R., Kumar Rath, S., Misra, S., Damaševičius, R., & Maskeliūnas, R. (2019). Distributed Centrality Analysis of Social Network Data Using MapReduce. Algorithms, 12(8), 161.

Luger, E., et al. (2013). Consent for all: revealing the hidden complexity of terms and conditions. In: CHI 2013, p. 2687. ACM, New York.

McDonald, A. M., & Cranor, L. F. (2008). The cost of reading privacy policies. A J. Law Policy Inf. Soc. 543(4), 1–22.

Meiselwitz, G. (2013). Readability assessment of policies and procedures of social networking sites. In A. Ozok & P. Zaphiris (Eds.), Online communities and social computing (pp. 67-75). Berlin: Springer. https://doi.org/10.1007/978-3-642-39371-6_8

Niazi, M. G. & Kamran, M. K. A. (2016). Evaluating Iranian state university websites using WebQEM", The Electronic Library, 34(6), 1031-1050.

Olajide, F., Adeshakin, K., Misra, S., & Ayo, C. K. (2016). On the Investigation of Social Network Analysis for E-Commerce Transaction in South-West Region of Nigeria. International Journal of Pharmacy & Technology, 8(4), 23108- 23114.

Scott, M. D. (2007). Scott on Information Technology Law (3rd. ed.). New York, NY: Aspen Publishers Online.

Temnikova, I., Vieweg, S., & Castillo, C., (2015). The Case for Readability of Crises

Communication in Social Media: International World Wide Web Conference Companion.

Covenant Journal of Informatics & Communication Technology. Vol. 8 No. 1, June, 2020
ISSN: *print* 2354 – 3566 *electronics* 2354 – 3507 DOI:

An Open Access Journal Available Online

A Hybrid Fuzzy Time Series Technique for Forecasting Univariate Data

Alhassan Mohammed B., Muhammad Bashir Mu'azu, Yusuf Abubakar Sha'aban, Shehu Mohammed Yusuf, Salawudeen Ahmed Tijani & Suleiman Garba

[2]Department of Computer Engineering, Ahmadu Bello University, Zaria, Nigeria.
Corresponding Author: changedmookie@yahoo.com

Received: 20.08.2019 Accepted: 29.11.2019
Date of Publication: June, 2020

Abstract: In this paper a hybrid forecasting technique that integrates Cat Swarm optimization Clustering (CSO-C) and Particle Swarm Optimization (PSO) with Fuzzy Time Series (FTS) forecasting is presented. In the three stages of FTS, CSO-C found application at the fuzzification module where its efficient capability in terms of data classification was utilized to neutrally divide the universe of discourse into unequal parts. Then, disambiguated fuzzy relationships were obtained using Fuzzy Set Group (FSG). In the final stage, PSO was adopted for optimization; by tuning weights assigned to fuzzy sets in a rule. This rule is a fuzzy logical relationship induced from FSG. The forecasting results showed that the proposed method outperformed other existing methods; using RMSE and MAPE as performance metrics.

Keywords: Forecasting, Fuzzy Time Series, Cat Swarm Optimization based Clustering, Particle Swarm Optimization.

1. Introduction

The estimation of what is likely to happen in the future especially in business and relative financial investment or practice is an unavoidable task (Singh, 2016). It is a key function that aids decision, planning and development in science, technology and engineering (Songh & Chissom, 1993). As the application of information technology is growing very rapidly, the proper utilization of data becomes undoubtedly necessary. The most applicable form of data is mostly available in time series. Fuzzy Time Series (FTS) forecasting technique is capable of handling both numeric and linguistic time series data. In terms of development, forecasting plays a role everywhere in our lives ranging from economy to technology and every aspects in between. Economically, accurate forecasting results can lead to improvement of a country's Gross National Product and Gross Domestic Product, increase strength of currency and expand industries. Technically, forecasts predict new practical developments that could change the operations of an organization. For example, in computing; the introduction of transistors removed vacuum tubes from business.

In contrast with the traditional methods, Fuzzy time series forecasting is the application of linguistic mathematical reasoning to model and predict the future value of a variable from a time series historically presented or available in either numeric or imprecise form. It is a reliable method that deals with uncertainties in observations recorded over successive period. Not only that but also, assumptions and too much back ground knowledge of the variable under forecast is not required.

Song and Chissom were the first to introduce FTS in 1993. It comprise of three stages namely; fuzzification, determination of fuzzy relationships (fuzzy inference) and defuzzification. In the fuzzification stage crisp observations are converted into linguistic values by identifying variations in the crisp data. In this first stage, decision on interval length is highly significant for dividing the universe of discourse. In this work interval lengths were objectively obtained using Cat Swarm Optimization based Clustering (CSO-C), define memberships that best explains the unknown structure of the observations; having taken these steps, defining universe of discourse becomes unnecessary, in fuzzy time series forecasting. Defuzzification is the final stage in fuzzy time series forecasting. This is the process of deriving future crisp forecasts from fuzzy forecasting rules. In a bid to remove recurrent fuzzy relationship; Fuzzy Set Group (FSG) was introduced. Finally, particle swarm optimization was proposed to assign weights to elements of forecasting rules and obtain defuzzified forecasts.

The rest of the paper is presented as follows: a brief discussion of Cat Swarm Optimization based Clustering (CSO-C) and Particle Swarm Optimization (PSO) were explained in section 2. Section 3 discusses Fuzzy Time Series (FTS). Section 4 discusses the results obtained from the application of the proposed forecasting model to two data sets. Finally, the conclusion was presented in Section 6.

2. Methods

Over decades, different methods have been used to improve FTS forecasting. The beauty in FTS forecasting is that regardless of the stage a researcher chooses to work upon, an appreciative reduction of error or increase in accuracy can be achieved in the FTS forecasting technique (Egrioglu *et al*, 2016). The complicated maximum minimum composition operation was replaced by a simplified arithmetic operation (Chen 1996), in order to achieve effective interval length, it is advisable to set the heuristic in a way that at least half the fluctuation in the time series will be reflected by the chosen lengths of intervals (Huarng 2001), Differential Evaluation Algorithm (DEA) was utilized to avoid subjective judgments for determining the interval lengths while discrete weights assigned to fuzzy relation that occurred in the defuzzification process. Consequently, improved result was presented (Bas *et al.,* 2013), temporal information was utilized to partition the universe of discourse into intervals with unequal lengths through Gath-Geva clustering (Wang *et al.*, 2013), a hybrid FTS forecasting model with empirical mode decomposition to partition universe of discourse, three layer back propagation artificial neural network for the determination of fuzzy relation and particle swarm optimization to optimize the weights and threshold of bpANN (Huang and Wu, 2017), an adaptive FTS model for multivariate forecasting of Shanghai Stock Exchange; Cuckoo search was utilized to partition a training data set into unequal intervals. Then relationships were generated using Fuzzy Logic Relationship Group. The results showed an improvement in forecasting accuracy. However, computational complexities involved might make the model not work effectively for higher variable sets (FLRG) (Zhang *et al.,* 2017).

2.1 Cat Swarm Optimization based Clustering (CSO-C)

This algorithm was used to code the fuzzification module of the hybrid FTS forecasting model, so as to objectively determine interval length among other steps in the first stage. Cat Swarm Optimization for Clustering was first proposed by Santosa and Ningrum in 2009 (Bahrami *et al*., 2018). According to Bahrami *et al.,* 2018; CSO-C is made up of two parts namely:

i) Clustering of data and
ii) Searching for the best cluster center.

The following are inputs for clustering CSO:

i) Population of data
ii) Number of clusters
iii) Number of copy

The phases of CSO-C are described as follows:

Phase 1: Define initial cluster center: In this phase, k point is chosen arbitrarily from the collected data in order to form the initial cluster center.

Phase 2: To Group data into clusters: Data is imputed into cluster with the closest cluster center. Distance between data and cluster data can be obtained by (Bahrami *et al*., 2018):

$$d(x,y)=\|x-y\| =\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \quad (1)$$

Phase 3: To Calculating the Sum of Squared-Error (SSE): The fitness function of the algorithm can be obtained by:

$$SSE = \sum_{i=1}^{k} \sum_{x \in Di} (\|x - m_i\|^2) \qquad (2)$$

Where:

$x = data,\ member\ of\ cluster\ D$

$m_i = cluster\ center\ i$

$k = number\ of\ cluster$

Phase 4: Clustering optimization with CSO: With regard to this algorithm, the cat is represented by a cluster center, while the new cluster center will be the solution set and is expected to come up with a smaller SSE value than before. A few adjustments are necessary in order to gain efficiency in the application of CSO to CSO-C. The adjustments are:

i. In order to allow every cat pass through the seeking and tracing mode; it becomes necessary to remove the Mixture Ratio (MR). Consequently, the time needed to find the best cluster centre will be reduced.
ii. If the value of CDC were always assumed to be 100% in the seeking mode, it will allow a change for every dimension of cat copy.

Phase 4.1: Seeking mode: The function of seeking mode is to enable the possession of the ability to search for best points around the cluster centres that have possibilities of attaining optimal fitness value. This is the reason that necessitated the need to define three parameters as outlined below:

i) Seeking Memory Pool (SMP): this will represent the number of copy a cluster have.
ii) Seeking Range of the Selected Dimension (SRD): this declares the mutative ratio, with a value between [0, 1].
iii) Self Position Considering (SPC): it is a Boolean random value (Amjad *et al.*, 2012).

The algorithm for seeking mode in CSO-C is given as follows (Santosa & Ningrum, 2009):

1. Evaluation of the parameter of seeking mode which include; SMP, SRD, SPC
2. For i = 1 to k (number of cluster center), do Copy cluster center (i) position as many as SMP.

Determine j value

Compute the shifting value (SRD*cluster center (i))

3. For m = 1 to SMP, do

Addition or subtraction of cluster centres with shifting value is performed randomly.

The output will be (SMP x k) cluster center candidates

4. Compute the distance, sub classify data into clusters, and compute SSE
5. Choose a candidate to be the new cluster centre roulette wheel selection

Phase 4.2: Updating SSE and cluster centre

The numerical quantity of SSE obtained from seeking mode is compared with the previous result of SSE. If the SSE numerical quantity obtained from seeking mode is less than the earlier SSE, then the result obtained from the seeking mode becomes the new cluster center. Conversely, if the numerical quantity obtained from seeking SSE is greater than or equal to the value of earlier SSE, we use the previous cluster centre.

Phase 4.3: Tracing Mode: The aim of the tracing mode is to shift point of concentration to a better position for obtaining optimal fitness value.

The Tracing Mode algorithm for CSO Clustering is as follows (Bahrami *et al.*, 2018):

1. For i = 1 to k, do
   Update velocity (i)
   Update position (i),
   get the new cluster center (i)
2. Calculate the distance, grouping data into clusters, and calculate SSE

Phase 4.4: Repeat step 4.2 for tracing SSE and cluster centre: With regard to SSE, the numerical quantity obtained from tracing mode is compared with the previous result of SSE. If the numerical quantity happens to be less than the previous, it will be used or considered as the cluster center. Conversely, if the result of tracing SSE is greater than or equal to earlier SSE, the previous cluster centre is used.

Phase 5: Repeat phase 4 until it reaches the stopping criteria.

Table 3.1: CSO-C Parameters and Specifications

Parameters	**Specifications**
SMP	5
CDC	100%
SRD	0.2
Const1	2
r1	[0,1]
Velmax	0.9
Number of clusters	7
Maximum number of iterations	100

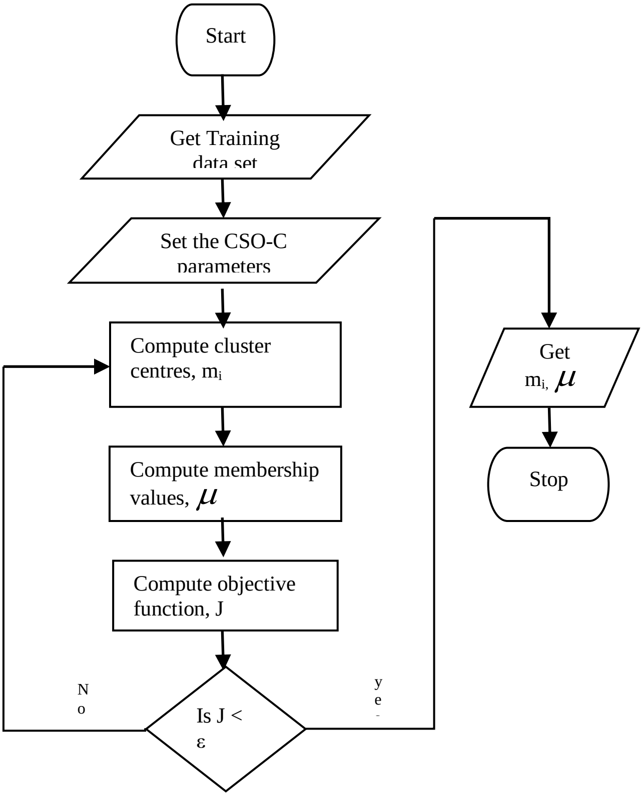
Start
Get Training data set
Set the CSO-C parameters
Compute cluster centres, m_i
Compute membership values, μ
Compute objective function, J
Is J < ε
No
yes
Get m_i, μ
Stop

Fig 1. Flowchart of the CSO-C modification Module

2.2. Particle Swarm Optimization

It was first introduced in 1995 by Kennedy and Eberhart (Kennedy and Eberhart 1995). PSO is a population-based evolutional algorithm that mimics the behaviour of birds flocking or fish grouping in search for the location of food (Eleruja *et al.*, 2012). Particle Swarm Optimization (PSO) is in the class of evolutionary computation (EC) and it is related to genetic algorithm and evolutionary programming, (Kennedy & Eberhart, 1999). The only thing it needs is traditional mathematical operators and in terms of speed and memory requirement it is computationally economical, (Amjad *et al.*, 2012). Empirically, it has been proven to be effective with different kinds of problems not only in the forecasting domain.

Problem optimization in PSO is achieved by having a population of candidate solution, in this process dubbed particles are moved around within a search space in accordance with simple mathematical formula over the particles position. The particles are also guided towards the best known position in the search space and are updated as better positions are found by other particles (Amjad *et al.*, 2012).

3. Fuzzy Time Series

The concept of fuzzy time series was first introduced by Song and Chissom (1993a).

The most important advantage of the fuzzy time series approach is to be able to work with a very small set of data.

Definition 1: Let U be the universe of discourse, where $U=\{u_1,u_2,...,u_n\}$. then a fuzzy set A_i of U can be defined as(Bas *et al.*, 2013):

$$A_i=\mu_{A_i}(\mathrm{u}_1)/\mathrm{u}_1+\mu_{A_i}(\mathrm{u}_2)/\mathrm{u}_2+,...,+\mu_{A_i}(\mathrm{u}_n)/\mathrm{u}_n \quad (3)$$

Where μ_{A_i} is the membership function of the fuzzy set A_i and $\mu_{A_i};U\rightarrow[0,1]$. In addition to $\mu_{A_i}(\mathrm{u}_j)$, j=1,2,...,n denote the generic elements of fuzzy set $A_i;\mu_{A_i}(\mathrm{u}_j)$ is the degree of belongingness of u_j to $A_i;\mu_{A_i}(\mathrm{u}_j)\in[0,1]$.

Definition 2: Fuzzy Time Series; let $Y(\mathrm{t})\,(\mathrm{t}=...,0,1,2,...)$, a subset of real numbers, be the universe of discourse by which fuzzy sets $f_i(\mathrm{t})(\mathrm{i}=1,2,3,...)$ are defined. If $\mathrm{F(t)}$ is a collection of $f_i(\mathrm{t})(\mathrm{i}=1,2,3,...)$, then $\mathrm{F(t)}$ is called a fuzzy time series defined on $Y(\mathrm{t})(\mathrm{t}=1,2,3,...)$ (Yusuf *et al.*, 2015).

Definition 3: Fuzzy Logic Relation (FLR); if there exist a fuzzy logic relationship $R(\mathrm{t}-1,t)$, such that $\mathrm{F(t)}=\mathrm{F(t-1)}\circ\mathrm{R(t-1,t)}$, where $\circ$ represents an operator, then $\mathrm{F(t)}$ is said to be caused by $\mathrm{F(t-1)}$. The relationship between $\mathrm{F(t)}$ and $\mathrm{F(t-1)}$ is denoted by;

$$\mathrm{F(t-1)}\rightarrow\mathrm{F(t)} \quad (4)$$

If $\mathrm{F(t-1)}=\mathrm{A}_i$ and $F(\mathrm{t})=\mathrm{A}_j$ then $A_i\rightarrow A_j$

Definition 4: Fuzzy Logic Relationship Group (FLRG).

Relationships with the same fuzzy set on the left hand side can further be grouped into a relationship group. Relationship groups are also referred to as Fuzzy Logic Relationship Groups (FLRG). Suppose that: $A_i\rightarrow A_{j1}, A_i\rightarrow A_{j2},...A_i\rightarrow A_{jn}$, then, they can be grouped into a relationship group as follows: $A_i\rightarrow A_{j1},A_{j2},...,A_{jn}$ (Yusuf *et al.*, 2015).The simulation parameters used to achieve the results are quantified in given in Table 1.

4. Results and Discussions

The FTS forecasting modules were coded in MATLAB 2016a, on a laptop computer with Intel core (TM) i3-3250M micro processor, frequency speed rate (2.30 GHz) and Random-Access Memory (RAM) of 4.00 GB.

The main objective of this study is to increase forecasting accuracy by using CSO-C in the fuzzification stage of FTS to objectively and unequally determine interval lengths. PSO was integrated into the defuzzification stage to optimize the process by tuning the "if-then" rules as discussed earlier.

Table 4.1 below shows the shows the set of values for the PSO parameters used (Yusuf *et al*., 2015).

Table 4.1: The PSO parameters used

Parameters	Specifications
Swarm Size	5
Maximum Number of Iterations	500
Target Fitness Value as MSE	1
Min and Max Particles Position Limited to	[0,1]
Min. and Max. Vel. Range	[-0.01,0.01]
Learning Factors C_1 and C_2	2
Inertial Coefficient, w	1.4
Maximum number of iterations	100

In the course of application, so as to verify the performance of the proposed hybrid forecasting model, it was applied to two different time series data sets: yearly deaths in car road accidents in Belgium data set and enrollments at the University of Alabama data set. The obtained results were compared with the results obtained from other fuzzy time series models in the literature using Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) as performance metrics which are mathematically represented as shown:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{T}\left(x_t - \hat{x}_t\right)^2} \quad (5)$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{\left|x_t - \hat{x}_t\right|}{x_t}\times 100 \quad (6)$$

For each time series observations, the decided CSO-C parameters are as shown in table 3.1. The definitions of the PSO parameters are also shown in table 4.1.

4.1 Forecasting Car Road Accident in Belgium

In a bid to verify efficiency of the developed model, the first implementation was carried out on a time series data set of the observation in the occurrence of car road accident in Belgium. Table 4.2 shows the cluster centers and their linguistic values which were obtained in ascending order. The FSG and optimal weights assigned to the forecasting rule contents as observed in the enrollment sets is shown in table 4.4. In addition a tabular presentation was made to compare between enrollment forecasts for both proposed and previous techniques. In terms of; RMSE and MAPE.

Table 4.2: Cluster Centers and Defined Fuzzy Sets for Yearly Deaths from Road Accidents in Belgium

Cluster	Center	Fuzzy Set
m1	1172.10	A_4
m2	1380.00	A_5
m3	1432.00	A_6
m4	1478.10	A_6
m5	1574.06	A_6
m6	1616.00	A_7
m7	1644.00	A_6

4.2 Forecasting Enrollments at University of Alabama

The implementation of the developed model was also carried out on University of Alabama student enrollment time series standard data set. Table 4.3 comprises of the cluster centers and their linguistic values, they were obtained in ascending order. Table 4.5 shows the fuzzy set groups and optimal weights assigned to the forecasting rule contents as observed in the enrollment sets. A comparative presentation of enrollments forecasts and the RMSE and MAPE values for the proposed methods and some other methods are given.

Table 4.3: Cluster Centers and Defined Fuzzy Sets for Students Enrollment in University of Alabama.

Cluster	Center	Fuzzy Set
m1	13055.11	A_1
m2	13565.35	A_2
m3	15164.65	A_2
m4	15862.01	A_3
m5	16917.99	A_3
m6	18149.95	A_3
m7	19333.69	A_4

Table 4.4: Fuzzy Set Groups and Optimal Weights for Accidents in Belgium.

Data points	**Maps**	**Optimal weight(s)**
1	#, #→ A_4	#,#
2	#,A_5→ A_5	#,#
3	A_5, A_6→ A_6	0.022442, 0.977558
4	A_5, A_6, A_7→ A_6	0.23847,0.0023311, 0.81188
5	A_7, A_7→ A_6	0.98302, 0
6	$A_{7,}$ A_7 , A_6→A_7	0.47619, 0.28692, 0.25285
7	A_7, A_6, A_7→ A_6	0.45287, 0.08873, 0.43831
8	A_6, $A_{7,}$ A_6→ A_4	0.93558, 0, 0
9	A_7, A_6, A_5→ A_3	0.081525,0.041537, 0.89499
10	A_6,$A_{5,}$ A_5→ A_3	0.47014, 0.19109, 0.25047

11	A_5, A_5, $A_4 \rightarrow A_2$	0.34784, 0.14902, 0.44225
12	A_5,A_5, A_4, $A_4 \rightarrow A_4$	0.983841, 0, 0, 0.016159
13	A_5,A_5,A_4, A_4, $A_5 \rightarrow A_2$	0.42501,0, 0.5797, 0,0
14	A_4, A_5, $A_4 \rightarrow A_3$	0.21641, 0.7953, 0
15	A_5, A_4, $A_5 \rightarrow A_4$	0, 0.076319, 0.96584
16	A_4, A_5, $A_5 \rightarrow A_5$	0.46637, 0.40469, 0.25499
17	A_5, A_5 ,$A_6 \rightarrow A_4$	0.54514, 0.24358, 0.21717
18	A_5, A_6, $A_5 \rightarrow A_2$	0.69881, 0, 0.26337
19	A_6,A_5,$A_4 \rightarrow A_2$	0 , 0
20	A_6,A_5, A_4,$A_4 \rightarrow A_3$	0.32258,0.42598,0.13994,0.082344
21	A_6,A_5,A_4,A_4,$A_5 \rightarrow A_1$	0.0073059,0, 0.034287, 0.86937
22	A_5, $A_3 \rightarrow A_1$	0.15219, 0.73786
23	A_3, $A_1 \rightarrow A_1$	0.72378, 0.23485
24	A_1, $A_2 \rightarrow A_1$	0.044945, 0.955055
25	A_1, A_2, $A_3 \rightarrow A_1$	0.33401, 0.68015, 0
26	A_3, $A_2 \rightarrow A_1$	0.99895, 0.025133
27	A_3, A_2, $A_3 \rightarrow A_2$	0, 0.13372, 0.92348
28	A_3, $A_4 \rightarrow A_1$	0.93486, 0
29	A_4, $A_1 \rightarrow A_1$	0.34751, 0.50543
30	A_1,$A_1 \rightarrow A_1$	0.023115, 0.82616

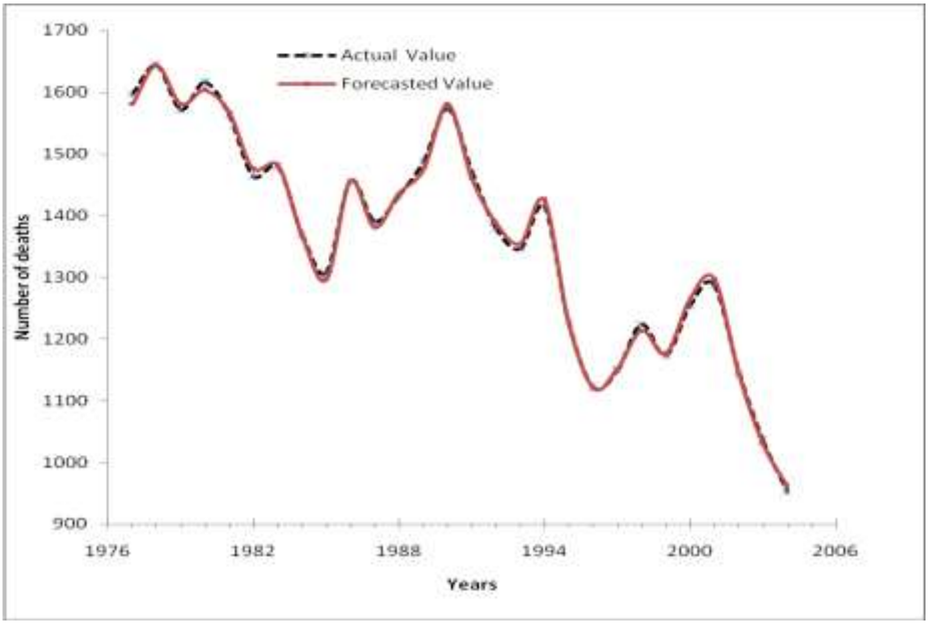

Figure 2: Graph of Forecasts of the Proposed Method and Actual Observations of Yearly Deaths from Accidents in Belgium

Table 4.5: Generated Fuzzy Set Groups and Optimal Weights for Enrollments

Data Points	**Maps**	**Optimal Weight(S)**
1	#, #→ A_1	#, #
2	#, $A_1 \rightarrow A_2$	#, #
3	A_1 , $A_1 \rightarrow A_2$	0 , 0.955055
4	$A_1, A_1, A_1 \rightarrow A_3$	0.85229, 0, 0.20749
5	$A_1, A_2 \rightarrow A_3$	0, 0.56155
6	$A_1, A_2, A_2 \rightarrow A_3$	0, 0.42887, 0.56148
7	$A_1, A_2, A_2, A_2 \rightarrow A_4$	0.9149, 0, 0, 0.18862
8	A_2 , $A_3 \rightarrow A_4$	0.9743, 0.0257
9	$A_3, A_4 \rightarrow A_5$	0.052746, 0.947254
10	$A_3, A_4, A_6 \rightarrow A_5$	0, 0, 0.9743
11	$A_6, A_6 \rightarrow A_5$	0.21978, 0.74883
12	$A_6, A_5 \rightarrow A_3$	0.14587, 0.79113
13	$A_5, A_2 \rightarrow A_3$	0, 0.18867
14	$A_5, A_2, A_2 \rightarrow A_3$	0, 0.37964, 0.59998
15	$A_5, A_2, A_2, A_2 \rightarrow A_3$	0.80891, 0, 0
16	$A_2, A_2, A_2, A_2 \rightarrow A_4$	0, 0.033894, 0
17	$A_2, A_4 \rightarrow A_5$	0.95174, 0.13425
18	$A_2, A_4, A_6 \rightarrow A_6$	0.090562,0.28332,0.72234
19	$A_6, A_7 \rightarrow A_7$	0.090562,0.28332,0.72234
20	$A_6, A_7, A_7 \rightarrow A_7$	0.02116, 0, 0.97884
21	$A_6, A_7, A_7, A_7 \rightarrow A_7$	1,0, 0.085242, 0.042223
22	$A_7, A_7, A_7, A_7 \rightarrow A_7$	0, 0, 0.41191, 0.58313

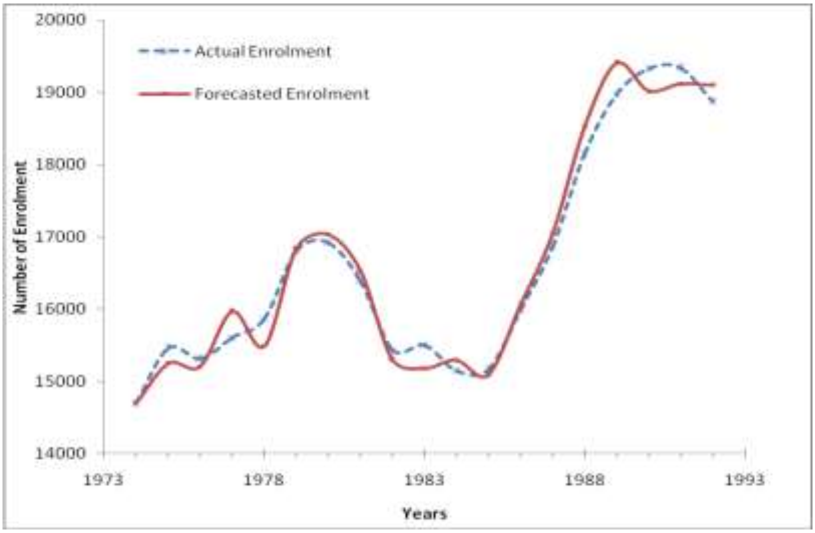

Figure 3: Graph of Forecasts of the Proposed Method and Actual Observations of Students Enrollment

Table 4.7: A Comparative Presentation of Yearly Deaths from Car Road Accidents Forecasts

Year	Actual	Egrioglu *et al* 2010	Jilani *et al* 2007	Uslu *et al* 2014	Yusuf *et al* 2015	Proposed Model
1974	1574		1497			
1975	1460		1497	1506		
1976	1536		1497	1453		1542
1977	1597	1500	1497	1598	1597	1588
1978	1644	1500	1497	1584	1643	1650
1979	1572	1500	1497	1584	1573	1560
1980	1616	1500	1497	1506	1633	1607
1981	1564	1500	1497	1584	1566	1572
1982	1464	1500	1497	1506	1464	1463
1983	1479	1500	1497	1453	1479	1487
1984	1369	1500	1497	1375	1369	1371
1985	1308	1400	1396	1383	1308	1315
1986	1456	1300	1296	1454	1457	1447
1987	1390	1500	1497	1453	1389	1390
1988	1432	1400	1396	1383	1432	1434
1989	1488	1400	1396	1509	1489	1484
1990	1574	1500	1497	1598	1574	1580
1991	1471	1500	1497	1506	1470	1462

1992	1380	1500	1497	1375	1380	1382
1993	1346	1400	1396	1383	1346	1338
1994	1415	1300	1296	1383	1414	1417
1995	1228	1400	1396	1231	1228	1229
1996	1122	1100	1095	1135	1065	1123
1997	1150	1200	1196	1180	1113	1148
1998	1224	1200	1196	1245	1223	1223
1999	1173	1200	1196	1135	1112	1177
2000	1253	1300	1296	1245	1212	1252
2001	1288	1300	1296	1284	1287	1288
2002	11445	1100	1095	1143	1146	1152
2003	1035	1000	995	970	1036	1041
2004	953	1000	995	970	954	945
	RMSE	85.35	83.12	41.61	19.2	5.931
	MAPE	5.25%	5.06%	2.29%	0.67%	0.34%

Table 4.8: Comparative Presentation of Enrollments Forecasts.

Year	**Actual Enroll-ment**	**S&C 1993**	**Chen 1996**	**Huarng 2001**	**Huarng *et al* 2006**	**Uslu *et al* 2014**	**Yusuf *et al* 2015**	**Proposed Model**
1971	13055							
1972	13563	14000	14000	14000	14242	13650		
1973	13867	14000	14000	14000	14242	13650	13873.3	13874
1974	14696	14000	14000	14000	14242	14836	14685	14701
1975	15460	15500	15500	15500	15474.3	15332	15465.6	15453
1976	15311	16000	16000	15500	15474.3	15447	15312.1	15307
1977	15603	16000	16000	16000	15474.3	15447	15600.7	15611
1978	15861	16000	16000	16000	15474.3	15447	15860	15860
1979	16807	16000	16000	16000	16146.5	16746	16813.5	16809
1980	16919	16813	16813	17500	16988.3	17075	16913.7	16921
1981	16388	16813	16813	16000	16988.3	16380	16389.9	16393
1982	15433	16789	16789	16000	16146.5	15457	15435.3	15430
1983	15497	16000	16000	16000	15474.3	15457	15508	15493
1984	15145	16000	16000	15500	15474.3	15457	15136.7	15150
1985	15163	16000	16000	16000	15474.3	15332	15174.3	15152
1986	15984	16000	16000	16000	15474.3	16027	15988.5	15985

1987	16859	16000	16000	16000	16146.5	16746	16860.9	16858
1988	18150	16813	16833	17500	16988.3	18211	18146.5	18162
1989	18970	19000	19000	19000	19144	19059	18979.3	18961
1990	19328	19000	19000	19000	19144	19059	19330.7	19340
1991	19337	19000	19000	19500	19144	19059	19348.4	19349
1992	18876		19000	19000	19144	19059	18887.4	18882
	RMSE	650	638	476	478	178	7.02	6.669
	MAPE	3.22%	3.11%	2.45%	2.20%	0.90%	0.04%	0.03%

5. Significance of Forecast Results

A smaller value for both performance metrics (RMSE and MAPE) in comparison to results obtained by previously used models is an indication of an improved forecasting. Meanwhile, forecast results obtained for car road accidents are (RMSE of 5.931 and MAPE of 0.34%). The results obtained for Alabama University student enrolment are (RMSE of 6.669 and MAPE of 0.03%).

The graphical representation of car road accidents is shown in Fig. 2. It compares actual value and forecasted value of proposed technique. A record of yearly deaths from car accident was presented. From visual inspection one can see the accuracy in pattern followed for forecasted values in relation to the actual values.

The graphical illustration in Fig. 3 shows forecasted results for student enrollment. The proposed technique followed actual pattern with few points of mismatch. An error as a result of method of collection of data, chosen parameters for tuning in the problem solving technique used and the like. On a lengthy note, the proposed technique followed trend of actual forecast.

Table 4.7 shows a comparative presentation between the proposed technique and previous techniques for car road accident. Similarly, table 4.8 compares the performance of the proposed model in relation to other previous techniques for student enrollment.

6. Conclusion

Researchers' observation has revealed that, objective partitioning of universe of discourse and the use of optimization technique to improve both fuzzification and defuzzification stages of the FTS forecasting process brings about accuracy in obtained results. This study presented an improved hybrid FTS forecasting technique used to handle any form of univariate dataset. Cat Swarm Optimization based Clustering (CSO-C) algorithm was utilized to objectively partition the universe of discourse and learn membership in datasets, while Particle Swarm Optimization algorithm was utilized in assigning optimal weights to elements of a fuzzy rule at the defuzzification stage. The results obtained demonstrate that the proposed forecasting technique provides more accurate forecasts

In a future development of forecasting model, it is necessary to consider other soft computing techniques that will be incorporated with FTS in order to form hybrid FTS techniques capable of handling errors caused by recurrence number of fuzzy relations and make objective choice of interval lengths.

Acknowledgement

The authors wish to utilize this medium to acknowledge the inputs and support provided by the Computer and Control Engineering Research Group of Ahmadu Bello University, Zaria. Their intellectual and technical help brought about the success of this work.

References

Amjad, U., Jilani, T. A., & Yasmeen, F. (2012). A two phase algorithm for fuzzy time series forecasting using genetic algorithm and particle swarm optimization techniques. International Journal of Computer Applications, 55(16).

Anderson, D. R., Sweeney, D. J., Williams, T. A., Camm, J. D., & Cochran, J. J. (2015). An introduction to management science: quantitative approaches to decision making: Cengage learning.

Bahrami, M., Bozorg-Haddad, O., & Chu, X. (2018). Cat Swarm Optimization (CSO) Algorithm. In O. Bozorg-Haddad (Ed.), Advanced Optimization by Nature-Inspired Algorithms (pp. 9-18). Fargo, North Dakota: Springer Nature Singapore Pte Ltd.

Bas, E., Egrioglu, E., Aladag, C. H., & Yolcu, U. (2015). Fuzzy-time-series network used to forecast linear and nonlinear time series. Applied Intelligence, 43(2), 343-355.

Bas, E., Uslu, V. R., Yolcu, U., & Egrioglu, E. (2013). A Fuzzy Time Series Analysis Approach by Using Differential Evolution Algorithm Based on the Number of Recurrences of Fuzzy Relations. American Journal of Intelligent Systems, 3(2), 75-82.

Chen, M.-Y., & Chen, B.-T. (2015). A hybrid fuzzy time series model based on granular computing for stock price forecasting. Information Sciences, 294, 227-241.

Chen, S.-M. (1996). Forecasting enrollments based on fuzzy time series. Fuzzy sets and systems, 81(3), 311-319.

Cheng, C.-H., Cheng, G.-W., & Wang, J.-W. (2008). Multi-attribute fuzzy time series method based on fuzzy clustering. Expert systems with applications, 34(2), 1235-1242.

Chu, S.-C., & Tsai, P.-W. (2007). Computational intelligence based on the behavior of cats. International Journal of Innovative Computing, Information and Control, 3(1), 163-173.

Eğrioglu, E., Aladag, C. H., Yolcu, U., & Dalar, A. Z. (2016). A Hybrid High Order Fuzzy Time Series Forecasting Approach Based on PSO and ANNs Methods. American Journal of Intelligent Systems, 6(1), 22-29.

Eleruja, S. e. A., Mu'azu, M. B., & Dajab, D. D. (2012). Application of trapezoidal fuzzification approach (TFA) and particle swarm optimization (PSO) in fuzzy time series (FTS) forecasting. Paper presented at the Proceedings on the International Conference on Artificial Intelligence (ICAI).

Huang, D., & Wu, Z. (2017). Forecasting outpatient visits using empirical mode decomposition coupled with back-propagation artificial neural networks optimized by particle swarm optimization. PloS one, 12(2), e0172539.

Huarng, K. (2001). Effective lengths of intervals to improve forecasting in fuzzy time series. Fuzzy sets and systems, 123(3), 387-394.

Kennedy, J., & Eberhart, R. C. (1999). The particle swarm: social adaptation in information-processing systems. Paper presented at the New ideas in optimization.

Lu, W., Chen, X., Pedrycz, W., Liu, X., & Yang, J. (2015). Using interval information granules to improve forecasting in fuzzy time series. International Journal of Approximate Reasoning, 57, 1-18.

Pei, A. (2015). Load forecasting based on fuzzy time series. Paper presented at the Proceedings of the 3rd international conference mater. mech. manuf. eng.(IC3ME 2015), Guangzhou, China.

Qiu, W., Zhang, P., & Wang, Y. (2015). Fuzzy Time Series Forecasting Model Based on Automatic Clustering Techniques and Generalized Fuzzy Logical Relationship. Mathematical Problems in Engineering, 2015. doi:http://dx.doi.org/10.1155/2015/962597

Santosa, B., & Ningrum, M. K. (2009). Cat Swarm Optimization for Clustering. International Conference of Soft Computing and Pattern Recognition, 23, 54-59. doi:10.1109/SoCPaR.2009.23

Sets, F., & Zadeh, L. (1965). Inform. Control, 8, 338-353.

Singh, P. (2016). Fuzzy Time Series Modeling Approaches: A Review Applications of Soft Computing in Time Series Forecasting (pp. 11-39): Springer.

Song, Q., & Chissom, B. S. (1993). Forecasting enrollments with fuzzy time series—part I. Fuzzy sets and systems, 54(1), 1-9.

Wang, L., Liu, X., & Pedrycz, W. (2013). Effective intervals determined by information granules to improve forecasting in fuzzy time series. Expert Systems with Applications, 40(14), 5673-5679.

Wang, W., Pedrycz, W., & Liu, X. (2015). Time series long-term forecasting model based on information granules and fuzzy clustering. Engineering Applications of Artificial Intelligence, 41, 17-24.

Yusuf, S., Mohammad, A., & Hamisu, A. (2017). A novel two–factor high order fuzzy time series with applications to temperature and futures exchange forecasting. Nigerian Journal of Technology, 36(4), 1124-1134.

Yusuf, S., Mu'azu, M. B., & Akinsanmi, O. (2015). A Novel Hybrid Fuzzy

Time Series Approach with Applications to Enrollments and Car Road Accidents. International Journal of Computer Applications, 129(2), 37-44.

Zhang, W., Zhang, S., Zhang, S., Yu, D., & Huang, N. (2017). A multi-factor and high-order stock forecast model based on Type-2 FTS using cuckoo search and self-adaptive harmony search. Neurocomputing, 240, 13-24.

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

Covenant Journal of Informatics & Communication Technology. Vol. 8 No. 1, June, 2020
ISSN: *print* 2354 – 3566 *electronics* 2354 – 3507 DOI:

An Open Access Journal Available Online

Fuzzy-PID Controller for Azimuth Position Control of Deep Space Antenna

Halima S. Yakubu, Suleiman U. Hussein, Gokhan Koyunlu,
Essien Ewang & Sadiq U. Abubakar

Nile University of Nigeria, Abuja, Nigeria
Centre for Satellite Technology Development, Obasanjo Space Centre, Abuja, Nigeria
halima.syakubu@gmail.com, elsuligh@gmail.com

Abstract: The Deep Space Antennas are essential in achieving communication over very large distances. However, the pointing accuracy of this antenna needs to be as precise as possible to enable effective communication with the satellite. Therefore, this work addressed the pointing accuracy for a Deep Space Antenna using Fuzzy-PID control technique by improving the performance objectives (settling time, percentage overshoot rise time and mainly steady-state error) of the system. In this work, the PID controller for the system was first of all designed and simulated after which, a fuzzy controller was also designed and simulated using MATLAB and Simulink respectively for the sake of comparison with the fuzzy-PID controller. Then, the fuzzy-PID controller for the system was also designed and simulated using MATLAB and Simulink and it gives a better performance objective (rise time of 1.0057s, settling time of 1.6019s, percentage overshoot of 1.8013, and steady-state error of 2.195e-6) over the PID and fuzzy controllers respectively. Therefore, the steady state error shows improved pointing accuracy of ± 2.195e-6.

Keywords/Index Terms— azimuth position control, deep space antenna, fuzzy logic control, fuzzy-PID, PID controller, pointing accuracy.

1. Introduction

The significance of pointing accuracy cannot be overemphasised as the development of radar and satellite systems progresses, and thus, the need to produce better control results with improved control techniques have become of great importance especially in communication industries. Communication over very large distances (e.g., deep space communication) is achieved by means of satellite communication. This can be established and maintained if the constellation of the communication satellites ensures that it is always possible to make contact with satellite, irrespective of the actual position on Earth.

Position control systems have, in recent years, been used extensively in applications such as in robotics, antennas, automation and many others . Amongst the most common and traditional techniques for position control is the Proportional-Integral-Derivative (PID) controller. Its straightforward configuration makes it easy to comprehend and its satisfactory performance causes it to maintain its status as the most widely used controller in industrial control system. However, the major challenges with the use of conventional PID are the tuning of the parameters and effect of non-linearity in the plant.

Therefore, Fuzzy Logic Control (FLC) which has the capacity of overcoming the issue of non-linearity in a plant can be considered. Furthermore, the exact mathematical model of a plant is not necessary when FLC is applied for the control of the system. However, the accuracy of the controller is subject to the expertise of the designer, which ultimately might impede the performance of the control system.

A technique which incorporates the concepts of both Fuzzy Logic and PID control, the Fuzzy-PID control, is explored. Fuzzy-PID is considered an extension of the conventional PID as it preserves the linear structure of the controller.

Several control methods have been proposed in literature for the position control of deep space antenna. For example, in the work of Okumuş *et al.* (2012), antenna azimuth position was controlled using two different controllers; classical PID and FLC that was tested with various fuzzy rules and membership functions. Results from both controllers were compared and the FLC was seen to give better results however, it requires high computational power to function. Also, Sahoo and Roy (2014), proposed a robust Quantitative Feedback Theory (QFT) controller which was designed for a 2-Degree of Freedom (DOF) azimuth position control of antenna with parametric uncertain. The QFT controller produced good results in terms of performance and stability specification but did not take into consideration system disturbances such as noise. In Zaber *et al.* (2015), a position control scheme of a Radio Telescope Antenna with wind disturbance using PID was presented. Although the controller succeeded in attenuating wind disturbances acting upon the radio telescope model however, better results could have been achieved using a more robust control technique. In addition, Fandaklı and Okumuş (2016) designed three different controllers (PID, Fuzzy Logic and Sliding Mode Control) for the azimuth position control of deep space antenna and compared their results in terms of performance. The results shows that the Sliding Mode Controller (SMC) outperformed the other controllers in terms of settling time and low sensitivity to noise disturbance, however modifications were not made to reduce chattering which is inherent in SMCs. However, in the work of E. G. Kumar (2018), the position control of the antenna azimuth was investigated using Proportional Integral (PI) and Fractional Order Lead Compensator controllers. Though the proposed lead compensator outperforms the PI controller when considering closed loop performances like response speed and settling time, it however had a high frequency gain, which amplifies the high frequency noise.

The robustness and efficiency of fuzzy-PID controller have been established in literatures for the DC motor control speed and Permanent Magnet (PM) synchronous motor. This controller also finds application in Automatic Generation Control (AGC) for multi-area interconnected power system (Mohanty *et al.* 2016). It can be used for the control of wind turbine pitch angle in (Civelek *et al.* 2016). Furthermore, it can be applied for the control of autonomous underwater vehicle (AUV) in heading and depth altitude and many others.

Therefore, in this work, the fuzzy-PID controller for azimuth position control of deep space antenna has been proposed.

The outline of the paper is as follows: system description, fuzzy-PID controller design, results and discussion, and the conclusion.

2. System Description

In this section, concepts such as antenna position system modelling of DC motor are discussed.

2.1 Antenna Position System

In position control systems, position input signals are converted to position output responses. For deep space antenna control, the aim is to make the antenna azimuth $\theta_0(t)$ track the reference $\theta_i(t)$ as much as possible by minimizing the tracking error.

A typical antenna should be able to rotate around azimuth (vertical) and elevation (horizontal) axes. These movements are independent, and their control systems are independent as well. The movement and rotation of the antenna are controlled by elevation and azimuth controllers respectively.

Figure 1 shows the control diagram of the antenna azimuth which represented servo-controlled mechanism with gears and feedback potentiometers.

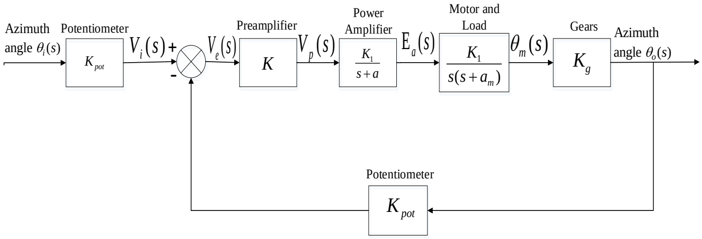

Figure 2: A Closed-Loop Antenna Azimuth Position Control for Deep Space Antenna

The closed-loop control diagram of the azimuth position control is shown in Figure 1. The input is an angular displacement which is converted into a voltage signal by a potentiometer. Similarly, the output angular displacement is also converted to a voltage signal for the feedback by the potentiometer. An error signal is generated at the comparator as a result of the difference between the input and output signals. Next, a differential amplifier checks the magnitude of the error as a result of the difference and passes it to the signal and power amplifiers which amplify the signal accordingly in order to drive the system.

The aim of the controller for the system is to drive the error to zero or as close as possible. When this is achieved the motor will not turn. The greater the error signal is, the higher the input voltage of the motor, which in turn makes the motor rotate faster. DC servo motor which is armature controlled is used for this system.

2.2 Modelling of DC Motor

DC motor that is armature controlled was chosen due to its high starting torque and relatively cheaper cost. The equivalent circuit of a DC motor with an armature controlled is shown in Figure 2.

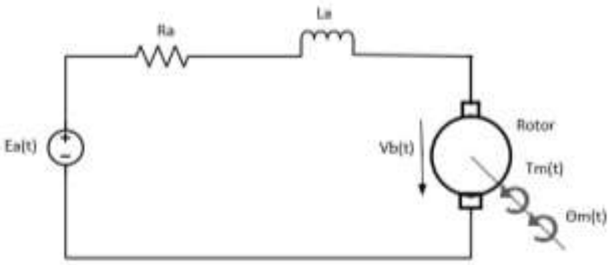

Figure 3: Equivalent Circuit Diagram of the Armature Controlled DC Motor

The dynamics of the electrical and mechanical subsystems of the armature controlled DC motor are given in Equations (1) to (6).

$$V_b = K_b \frac{d\theta_m}{dt} \quad (1)$$

$$R_a i_a(t) + L \frac{di_a(t)}{dt} + V_b(t) = E_a(t) \quad (2)$$

$$T_m = K_t i_a(t) \quad (3)$$

$$T_m = J_m \frac{d^2\theta_m}{dt^2} + D_m \frac{d\theta_m}{dt} \quad (4)$$

$$J_m = J_a + J_L \left(\frac{N_1}{N_2}\right)^2 \quad (5)$$

$$D_m = D_a + D_L \left(\frac{N_1}{N_2}\right)^2 \quad (6)$$

The parameters and units used in Equations (1) – (6) are given in Table I.

Table II: Parameters of the Antenna Dynamics

Parameter	Definitions	Values
V	Potentiometer voltage (V)	10
L	Motor inductance (H)	0.01
n	Potentiometer turns	10
K_1	Amplifier Gain Power	100
a	Pole of Power amplifier	100
R_a	Motor Resistance (Ω)	8
J_a	Motor inertial constant (kg-m2)	0.02
D_a	Motor Damping Constant (N-m s/rad)	0.01
K_b	Back EMF (V-s/rad)	0.5
K_t	Motor Torque Constant (N-m/A)	0.5
N_1, N_2, N_3	Gear teeth	25, 250, 250
J_L	Inertial constant of the load (kg-m2)	1

D_L	Load inertial constant (N-m s/rad)	1
K_{pot}	Gain of the Potentiometer	0.318
K_m	Load gain with motor	2.083
a_m	Pole of motor and load	1.71
K_g	Gear ratio	0.1

E_a – voltage across the motor (V)

θ_m – angular displacement of the motor (degree)

i – circuit current (A)

R – motor resistance (Ω)

T_m – motor torque (Nm)

V_b – voltage across the rotor (back emf) (V)

J – inertia of the motor rotor and load (Nms^2/rad)

D – damping of the motor rotor and load (Nms/rad)

L – armature inductance (H)

K_t – torque constant (Nm/A)

N – gear teeth

Through a series of substitutions using Equations (1) to (6), a mathematical expression of the armature controlled DC motor with respect to the output, θ_m to the input, E_a is derived and is given in Equation (7) as

$$K_b \frac{d\theta_m}{dt} = E_a - \frac{R_a}{K_t}\left(J_m \frac{d^2\theta_m}{dt^2} + D_m \frac{d\theta_m}{dt}\right) - \frac{L}{K_t}\left(J_m \frac{d^3\theta_m}{dt^3} + D_m \frac{d^2\theta_m}{dt^2}\right) \quad (7)$$

Equation (7) is then modelled using MATLAB and Simulink in Figure 3.

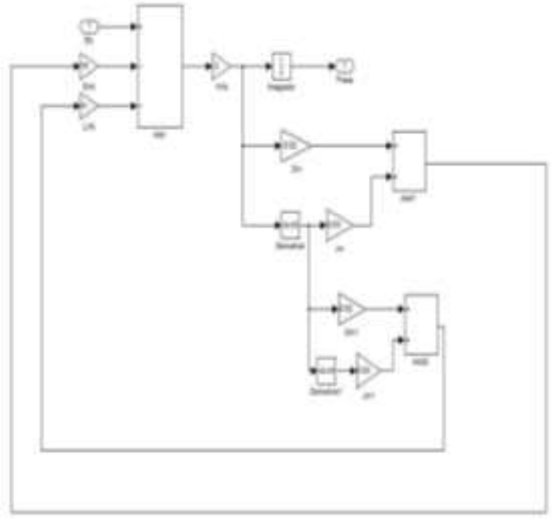

Figure 4: Simulink Model of Armature Controlled DC Motor

3. Fuzzy-PID Control Design

Firstly, we begin with the design of a fuzzy logic controller (FLC).

Lotfi A. Zadeh was the first to introduce fuzzy logic but was only later implemented by E. H. Mamdani almost ten years after its introduction. FLC have a wide range of applications in areas like industrial manufacturing and automation, automobile production, hospitals, banks, libraries and academic education, etc.

The basic structure of FLC system is shown in Figure 4. It comprises of four basic elements, which are: fuzzy knowledge base, fuzzification interface, inference engine (decision-making logic), and defuzzification interface .

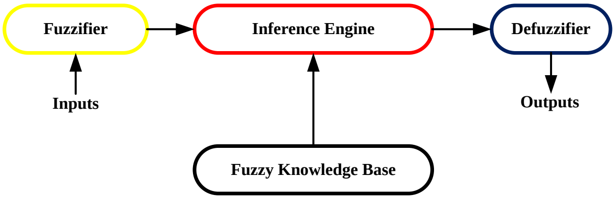

Figure 5: Basic Structure of a Mamdani-Type Fuzzy Logic System

3.1. *Fuzzification*

Here, the crisp inputs, ‘error (E)’ and ‘change in error (CE)’, are fuzzified, i.e. converted into fuzzy variables. In this research work, triangular membership function was selected for the inputs and output variables with its crossing μ = 0.5. The leftmost and rightmost fuzzy sets (with respect to inputs and output) are represented as shouldered ramps. The inputs and output are defined on a universe of discourse which was divided into 5 overlapping fuzzy sets: sets Negative Small (NS), Negative Large (NL), Zero (Z), Positive Large (PL), and Positive Small (PS). Figure 5 and Figure 6 show the two input variables for the fuzzy controller. The single output of the fuzzy controller is defined similarly to the inputs and is shown in Figure 7.

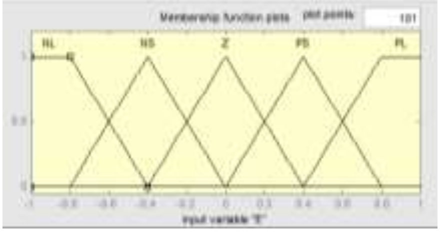

Figure 6: ‘Error (E)’ Input Variable

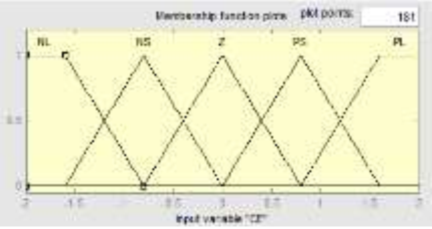

Figure 7: ‘Change In Error (Ce)’ Input Variable

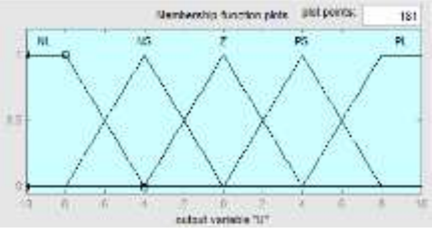

Figure 8: ‘U’ Output Variable

3.2. *Inference Engine*

It is knowledge base where rules are defined as if-then statement that guides the relationship between the input and output variables in terms of membership functions. At this level, the inference engine processes E and CE which executes 25 rules (5x5) as shown in Table II, where max-product inference method is used . The weight of all the rules is given as 1 (which actually has no effect on the implication process).

Table III: Fuzzy Rule Base for Controller Design

CE/E	**NL**	**NS**	**Z**	**PS**	**PL**
NL	NL	NL	NL	NS	Z
NS	NL	NS	NS	Z	PS
Z	NL	NS	Z	PS	PL
PS	NS	Z	PS	PS	PL
PL	Z	PS	PL	PL	PL

3.3. *Defuzzification*

This stage entails the generation of a usable output for the control of the system. Here, the internal fuzzy output variables are converted by the FLC into crisp values that can actually be used by the control system. Bisector method is used for defuzzification. The outputs are singletons, whose positions were derived by the cumulative of peak positions of the input sets.

Next, the PID controller is designed and tuned. Figure 8 shows the Simulink model of the fuzzy-PID configuration.

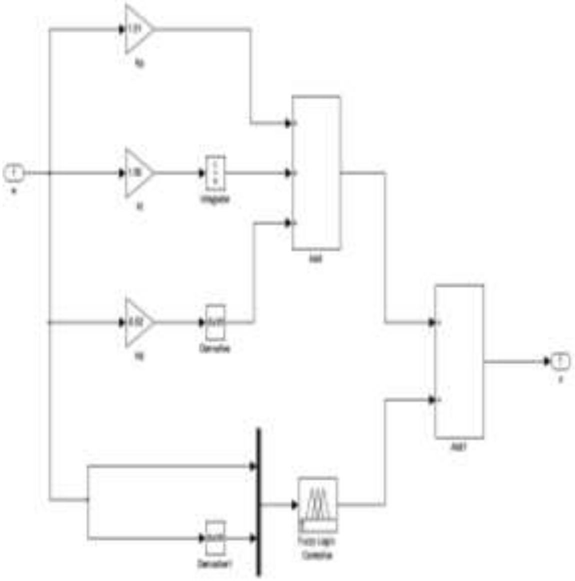

Figure 9: Simulink Model of the Fuzzy-PID Configuration

Figure 9 shows the Simulink model of the antenna with FLC for the azimuth position control of the system.

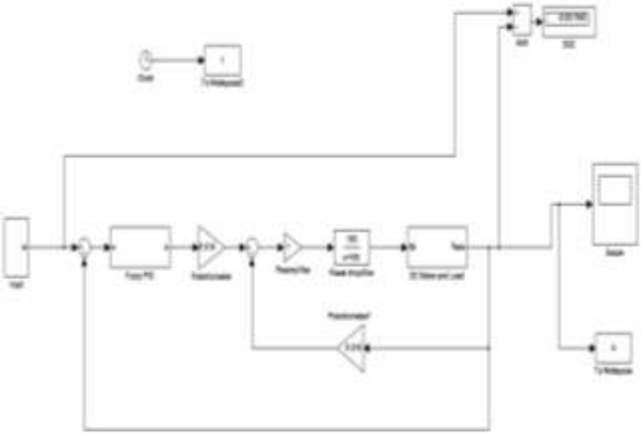

Figure 10. Simulink Model of the Antenna with Fuzzy-Pid Controller

The response of the fuzzy-PID controller was compared with that of the fuzzy logic and Proportion-Integral-Derivative (PID) controllers to determine its performance. Therefore, PID controller was first of all designed in MATLAB and Simulink to control the system after which the FLC was designed, and finally the fuzzy-PID controller was then designed for the antenna system.

4. Results and Discussion

In this section, the responses of the deep space antenna with respect to the three different controllers are presented here.

4.1 *Response of Deep Space Antenna with PID Controller*

Figure 10 shows the step response of the antenna position system with PID controller.

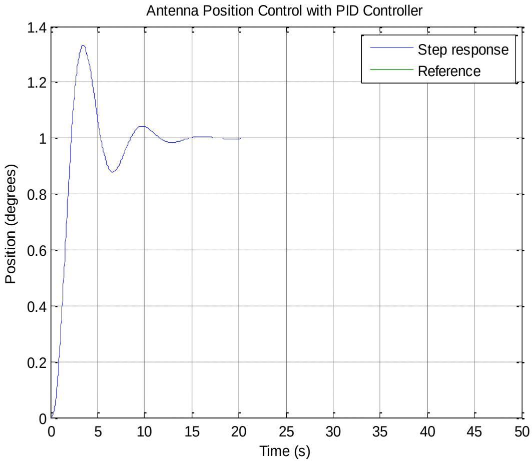

Figure 11: Unit Step Response of the Deep Space Antenna with PID Controller

URL: ***http://journals.covenantuniversity.edu.ng/index.php/cjict***

Table IV: Parameters of the Pid Controller

Parameter	Value
Rise Time	1.3726s
Settling Time	10.9478s
Overshoot	33.1750%
Peak Time	3.4675s
Steady-state Error	1.368e-007

From Table III above, it evident that the PID controller has a rise time of 1.3726s, and a lower steady-state error which is indicative of a high pointing accuracy. However, the PID controller has an overshoot of 33.1750% which is much higher than the accepted value of between 0 and 10% and a large settling time which makes the PID an undesirable controller. The large overshoot could lead to actuator (motor) damage during the transient state of the deep antenna operation.

4.2 Response of Deep Space Antenna with Fuzzy Logic Controller

Figure 11 shows the step response of the antenna with respect to the azimuth position control with FLC.

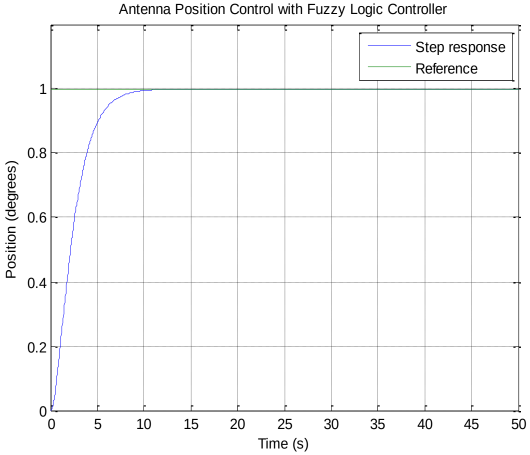

Figure 12: Figure XI. Step Response of System with Fuzzy Logic Controller

Table V: Fuzzy Controller Parameters

Parameter	Value
Rise Time	4.3381s
Settling Time	7.4146s
Overshoot	0%
Peak Time	10s
Steady-state Error	0.004358

Table IV shows the fuzzy controller performance with an overshoot of 0% and a settling time of 7.4146s which is highly favourable to the actuator (motor) for driving the gears of the deep space antenna. But 0.004358 steady state error is present which is also favourable. This implies that the pointing accuracy of the deep space antenna to a satellite would be $\pm 0.004358m$ which is very good.

4.3 Response of Deep Space Antenna with Fuzzy-PID Control

Figure 12 shows step response of the antenna azimuth position control with fuzzy-PID controller.

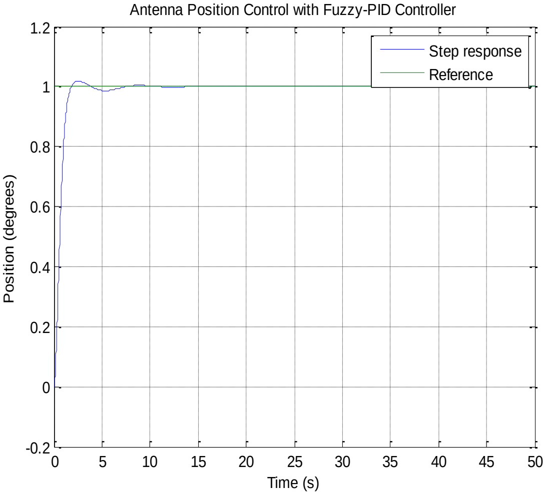

Figure 13: Unit Step Response of the Antenna with Fuzzy-PID Controller

Table VI: Fuzzy-Pid Controller Parameters

Parameter	**Value**
Rise Time	1.0057s
Settling Time	1.6019s
Overshoot	1.8013%
Peak Time	2.6091s
Steady-state Error	2.195e-006

From Table V, it is evident that the fuzzy-PID controller has a fast rise time of 1.0057s and settling time of 1.6019s. It has an overshoot of 1.8013% which is acceptable for a control system. Also, the steady-state error implies a high pointing accuracy $\pm 2.195e-006m$ between the deep space antenna and the satellite which is very good.

4.4 *Comparing the Response of the Fuzzy Logic and PID Controllers for the Deep Space Antenna*

Figure 13 shows the comparison of the step response of the antenna with respect to the control of the azimuth position with PID, fuzzy and fuzzy-PID controllers.

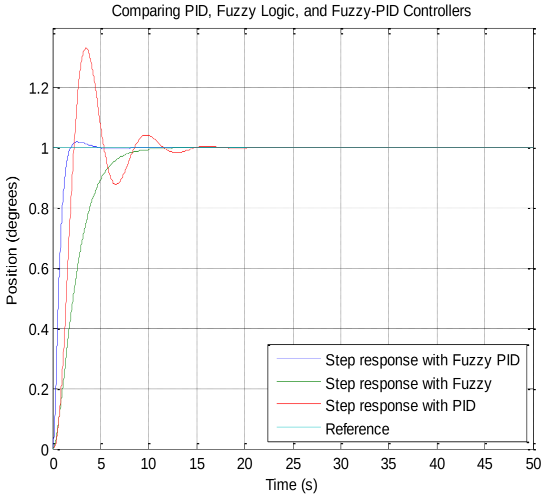

Figure 14: Step Responses of Azimuth Position Control with Fuzzy-PID, Fuzzy and PID Controllers

Table VI gives the comparison of PID, fuzzy and fuzzy-PID controller with respect to performance.

Table VII. Performances Comparison of PID, Fuzzy and Fuzzy-PID Controllers

Parameter	**PID**	**Fuzzy**	**Fuzzy-PID**
Rise Time	1.3726s	4.3381s	1.0057s
Settling Time	10.9478s	7.4146s	1.6019s
Overshoot	33.1750%	0%	1.8013%
Peak Time	3.4675s	10s	2.6091s
Steady-state Error	1.368e-007	0.004358	2.195e-006

From Table VI, it shows that the PID controller has the best performance with respect to steady state error but a large overshoot and slow settling time undermines its overall performance. To put it further, an overshoot of 33.1750% is well above the prescribed value for a control system and is likely to cause a fault to the antenna system which will render the steady-state performance irrelevant.

The fuzzy controller performs better in terms of overshoot. However, the values of the rise time and settling time are large which implies a sluggish response of the system to the controller.

The fuzzy-PID controller has the best performance as regard settling time, rise time and peak time. It also has an acceptable overshoot of 1.8013% and a small steady-state error. These imply that the system exhibits a fast response and a high pointing accuracy.

It is clear that the fuzzy-PID controller outperforms the PID and fuzzy controllers if one considers the parameters of each controller relative to the other.

5. Conclusion

This work aimed at improving the Azimuth Position Control of a Deep Space Antenna by increasing the pointing accuracy through a small steady-state error and very low overshoot. From the work done, it shows that the fuzzy-PID controller (out of all the controllers used) has the best performance to achieve this aim.

A PID controller for the system was first of all designed and simulated in this research work, after which a fuzzy controller was also designed and simulated using MATLAB and Simulink respectively for the sake of comparison with the fuzzy-PID controller. Then, the fuzzy-PID controller for the system was also designed and simulated using MATLAB and Simulink and it gives the best performance objectives (rise time of 1.0057s, settling time of 1.6019s, percentage overshoot of 1.8013, and steady-state error of 2.195e-6) over the PID and fuzzy controllers.

The contribution to knowledge of this work is the improved pointing accuracy of ±2.195e-6 using fuzzy-PID control that will enable the deep space antenna track the satellite.

5.1 Further Work

A fuzzy-PID controller tuned using heuristic methods such as Genetic Algorithm and neural networks for the control of deep space antenna can be looked into for improved performance.

References

Abed, W. (2015). Design of Armature and Field Control Systems based Bacterial Foraging Optimization Technique for Speed Control of DC Motor. International Journal of u-and e-Service, Science and Technology, 8(2), 385-394.

Bansal, U. K., & Narvey, R. (2013). Speed control of DC motor using fuzzy PID controller. Advance in Electronic and Electric Engineering, 3(9), 1209-1220.

Burns, R. (2001). Advanced control engineering: Elsevier.

Choi, H. H., Yun, H. M., & Kim, Y. (2015). Implementation of evolutionary fuzzy PID speed controller for PM synchronous motor. IEEE Transactions on Industrial Informatics, 11(2), 540-547.

Civelek, Z., Lüy, M., Çam, E., & Barışçı, N. (2016). Control of pitch angle of wind turbine by fuzzy PID controller. Intelligent Automation & Soft Computing, 22(3), 463-471.

E. G. Kumar, R. P., S. Rishivanth, S. A. Anburaja, A. G. Krishna. (2018). Control of Antenna Azimuth Position using Fractional order Lead Compensator. International Journal of Engineering & Technology, 7(2), 166-171.

Fandaklı, S. A., & Okumuş, H. İ. (2016). Antenna azimuth position control with PID, fuzzy logic and sliding mode controllers. Paper presented at the Innovations in Intelligent Systems and Applications (INISTA), 2016 International Symposium on.

Gawronski, W. (2004). Control and pointing challenges of antennas and (Radio) telescopes. IPN Progress Report, 42-159.

Gil, P., Lucena, C., Cardoso, A., & Palma, L. B. (2015). Gain tuning of fuzzy PID controllers for MIMO systems: a performance-driven approach. IEEE Transactions on Fuzzy Systems, 23(4), 757-768.

Gulley, N. (1996). Fuzzy logic toolbox for use with MATLAB.

Jantzen, J. (1998). Design of fuzzy controllers. Technical University of Denmark, Department of Automation, Bldg, 326, 362-367.

Khodayari, M. H., & Balochian, S. (2015). Modeling and control of autonomous underwater vehicle (AUV) in heading and depth attitude via self-adaptive fuzzy PID controller. Journal of Marine Science and Technology, 20(3), 559-578.

Mamdani, E. H. (1974). Application of fuzzy algorithms for the control of a simple dynamic plant. Proc IEEE P, 121-159.

Mohanty, P. K., Sahu, B. K., Pati, T. K., Panda, S., & Kar, S. K. (2016). Design and analysis of fuzzy PID controller with derivative filter for AGC in multi-area interconnected power system. IET Generation, Transmission & Distribution, 10(15), 3764-3776.

Nise, N. S. (2007). CONTROL SYSTEMS ENGINEERING: John Wiley & Sons.

Okumus, H. I., Sahin, E., & Akyazi, O. (2012). Antenna azimuth position control with classical PID and fuzzy logic controllers. Paper presented at the Innovations in Intelligent Systems and Applications (INISTA), 2012 International Symposium on.

Omar, A., Aous, A., & Balasem, S. (2015). Comparison between the Effects of Different Types of Membership Functions on Fuzzy Logic Controller Performance. International Journal of Emerging Engineering Research and Technology, 3(3), 76-83.

P., H. (1998). Mathematics of Fuzzy Logic. Kluwer Academic Publishers(Dordrecht, The Netherlands).

Sahoo, S. K., & Roy, B. (2014). Antenna azimuth position control using Quantitative feedback theory (QFT). Paper presented at the Information Communication and Embedded Systems (ICICES), 2014 International Conference on.

Temelkovskia, B., & Achkoskia, J. (2014). Modeling and Simulation of Antenna Azimuth Position Control System. International Journal of Multidisciplinary and Current Research, 4.

Vassallo, E., Martin, R., Madde, R., Lanucara, M., Besso, P., Droll, P., . . . De Vicente, J. (2007). The european space agency's deep-space antennas. Proceedings of the IEEE, 95(11), 2111-2131.

Yusof, A. M. (2013). Comparative study of conventional PID and fuzzy-PID for DC motor speed control. Universiti Tun Hussein Onn Malaysia.

Zaber, N. M., Ishak, A., Soh, A. C., Hassan, M., & Abidin, Z. Z.

(2015). Designing PID controller for position control with disturbance. Paper presented at the Computer, Communications, and Control Technology (I4CT), 2015 International Conference on.

Zadeh, L. A. (1965). Fuzzy Sets. Information and Control(8), 338-353.

Covenant Journal of Informatics & Communication Technology. Vol. 8 No. 1, June, 2020
ISSN: *print* 2354 – 3566 *electronics* 2354 – 3507 DOI:

An Open Access Journal Available Online

Sequential Feature Selection Using Hybridized Differential Evolution Algorithm and Haar Cascade for Object Detection Framework

Salefu Ngbede Odaudu, Emmanuel Adewale Adedokun,
Ahmed Tijani Salaudeen, Francis Franklin Marshall,
Yusuf Ibrahim & Donald Etim Ikpe

Ahmadu Bello University, Zaria, Kaduna, Nigeria.
ousalefu@abu.edu.ng, wale@abu.edu.ng, tasalawudeen@abu.edu.ng,
frankdidam14@gmail.com, yibrahim@abu.edu.ng, deikpe@abu.edu.ng

Received: 15.06.2019 Accepted: 06.12.2019
Date of Publication: June, 2020

Abstract: Intelligent systems an aspect of artificial intelligence have been developed to improve satellite image interpretation with several foci on object-based machine learning methods but lack an optimal feature selection technique. Existing techniques applied to satellite images for feature selection and object detection have been reported to be ineffective in detecting objects. In this paper, differential Evolution (DE) algorithm has been introduced as a technique for selecting and mapping features to Haarcascade machine learning classifier for optimal detection of satellite image was acquired, pre-processed and features engineering was carried out and mapped using adopted DE algorithm. The selected feature was trained using Haarcascade machine learning algorithm. The result shows that the proposed technique has performance Accuracy of 86.2%, sensitivity 89.7%, and Specificity 82.2% respectively.

Keywords/Index Terms: Differential Evolution, Haar-cascade, Machine learning, Satellite image

1. Introduction

Object detection is a computer vision approach of identifying objects in an image. It's an approach that remains a fundamental problem because, real-time images exhibit variation in resolution and when applied to a dynamic world, information about the object can be superseded or corrupted before it is ready for use if the algorithm used is a slow one (Ramisa *et al.*, 2008). Factors such as system or sensor noise, varying brightness, perspective changes, cluttered background, and others, contribute to why humans still have the capacity to recognize objects in images with lesser effort (Kurian, 2011). These factors necessitate the need for a robust model capable of detecting objects within the shortest possible time. The adoption of machine learning algorithm for the use of object detection has enabled process automation thereby making the process less dependent on human subjective procedures. The process takes large object samples as training dataset and compares further inputs with the existing training models to output a result that should look similar to the training set of objects. Common examples of object detection machine learning models include deep learning, Haar-cascades and etc (Kranthi, & Surekha, 2019). The algorithm here take image as an input and output it in the form labels (Kurian, 2011). The classification algorithm is an unsupervised method of learning that take a given data sample and classify them into a group base on the training rules (Sathya & Abraham, 2013). Localization, similar to classification is the training of an object detection algorithm to identify an object in a single image (Cinbis *et al.*, 2017).

Object detection in satellite images is a subset of object detection in optical sensing images. This detection entails the determination of an aerial or object contained in an image belonging to a localized area of interest and predictively locating an object in large set image dataset (Cheng & Han, 2016). The choice of a good object detection algorithm should be on the bases of (Kurian, 2011): Reliability, speed, and automation. As the algorithm is required to be robust in handling and image variation so that it will not degrade the image in the process. Speed is essential because the algorithm might be deployed to work online and it should work without human intervention (Leibe *et al.*, 2008).

Differential Evolution (DE) is a meta-heuristic based algorithm (Beheshti & Shamsuddin, 2013; Feoktistov, 2006) that is efficient with an aim to resolve non-linear, non-differential, non-continuous and real-parameters problem (Ecaterina *et al.*, 2011; Nunes *et al.*, 2017). From a randomized general population with a solution, Differential Evolution main objective becomes the selection of the best solution iteratively through some set of instructions. DE has been reported to have parameter that can be adjusted. Mutation factor $f \in [0,2]$, combination factor $c \in [0,1]$ and number of individual population N_{pop}

The challenge with existing machine learning techniques in identifying features or objects in satellite images effectively and accurately in an efficient manner with little or no delay in the processing is the lack of methods that can perform feature selection optimally. Hence, this work core contribution is the hybridization of differential Evolution algorithm for feature selection and classifier (Haar-cascade) for object detection and identification in satellite images.

2. Related Work

Several object detection algorithms exist in the literature (Cheng & Han, 2016). Most work on object detection in aerial image, in the past satellite images were object-oriented (Merchant *et al*., 2019). Methods such as template matching-based object detection, machine learning-based object detection, and object-based image analysis, knowledge-based object detection. (Kim *et al.*, 2004; Leninisha &Vani, 2015; Mayer *et al.*, 2006; Wang *et al.*, 2015) presented methods of detecting road networks and other objects in satellite images. (Zhang *et al.*, 2011) introduced a semi-automatic template matching technique to track roads. The work adopted spoke wheel algorithm to get direction of road width and starting point. Also (Kim *et al.*, 2004) used rectangular template against profile adopted in the work of Zhang to track ribbon road the use of least square correlation template matching. (Zhou *et al.*, 2006) proposed some road tracking techniques using profiles that are orthogonal and parallel to the road direction. (Baltsavias, 2004) present a review of knowledge-based object detection in RSI. Object-based image analysis currently has become the most widely used method for classifying and mapping VHR imagery into a well-defined object. It is a two steps image segmentation and classification (Blaschke *et al.*, 2014)

2.1.1 Haar cascade

Prior to the invention of machine Learning techniques for object detection such as Haar cascade for application in divert field, several other template and object matching algorithm had been actively use. Such as the Scale invariant feature transform (Dalai, 2019), Speed up Robust Feature (Sharma, 2019), oriented fast and rotated binary robust independent elementary features (Gollapudi, 2019). Though, these object detection algorithm have high Accuracy but require longer processing time. On the other hand, Haar-like – feature or Haar cascade is a machine learning object detection method developed by Viola and Jones (Ren *et al.*, 2017) for the purpose of detecting images with speed and accuracy in detection rate. The approach introduces a method of representing an image called Integral Image (Viola & Jones, 2001). This method of representing images allows features trained in another classifier to be computed very fast. When these classifiers are combined in the form of ensemble learning, the approach is called Haarcascade (Phuc, 2019). That is the combination of two or more classifiers trained with haar-like features to produce the best result (Leibe *et al.*, 2008). Haar cascade is an algorithm that operates on the fact that all human face has certain features and these features can be used when trained in a machine to detect objects in images. The feature in relation to human face is (Viola &

Jones, 2001): the eyes region is darker than the nose and upper cheek and the nose bridge is brighter than eyes.

$$f\left(x,y\right)=\sum_{a=0}^{x}\sum_{b}^{y}I\left(a,b\right) \quad (1)$$

From (1), the algorithm takes into account the sum difference between pixel value taken from the dark region and compared with the summed integral value f at localized area $\left(x,y\right)$ in a rectangle with range of $\left[0,0\right]$ to $\left[x,y\right]$.The generalized expression for detection and false positive in Haarcascade is given in (2) and (3).

$$w=\prod_{i=1}^{k}w_i \quad (2)$$

Equation (2) and (3), represents the learning process and the detecting process is shown in equation (3).

$$z=\prod_{i=1}^{k}z_i \quad (3)$$

Where w is the minimum accepted false positive rate, z is the minimum accepted detection rate, P = set of positive, N = set of negative, Feature Engineering

Features are extracted to give more insight into dataset. The process entails understanding the component and features that are contained in data. For classification related problems (unsupervised learning) classification algorithm is entrusted with interpretation to the dataset as it is expected to classify or cluster the data based on the instructed rules. Haarcasde is a feature rather than pixel-based classifier. Owning to its fastness when compared to pixel extracted features. Extracting features from satellite image will require good knowledge in satellite image processing tools such as ArcGIS, R language, MatLab, etc through the use of Raster library and other in-built libraries specifically dedicated to image processing.

2.1.2 Creating Haar Cascade for Object Dectection: A Theoretical Background

Haar cascade performance can be improved given the efficiency of Adaboost that allow the algorithm to contain a significantly large number of training example that in turn contributes to generalized performance of stronger classifier's error. Consequently, this makes small training image samples containing the need to find feature to be misclassified (Fan, 2019). Adaboost simultaneously associates learning procedures (Wang, 2019). The essence of associating the learning procedure by large was to construct classifiers for object recognition. The choice of the states in the learning process in Adaboost, is designer dependent but the first choice for each state will be created by the system on positive images and tested on negative images which will be used for subsequent use in bulding a second classifiers that mature into better detection rate. The process continues with the next classifer that is then used for the next state. the iterative process ends when the last state is completed (Kyrkou 2010). The cascaded stages discussed, are achieved by l training each classifier by means of Adaboost and minimizing thr error rate with the use of other compiling threshold algorithm.

3. Methodology

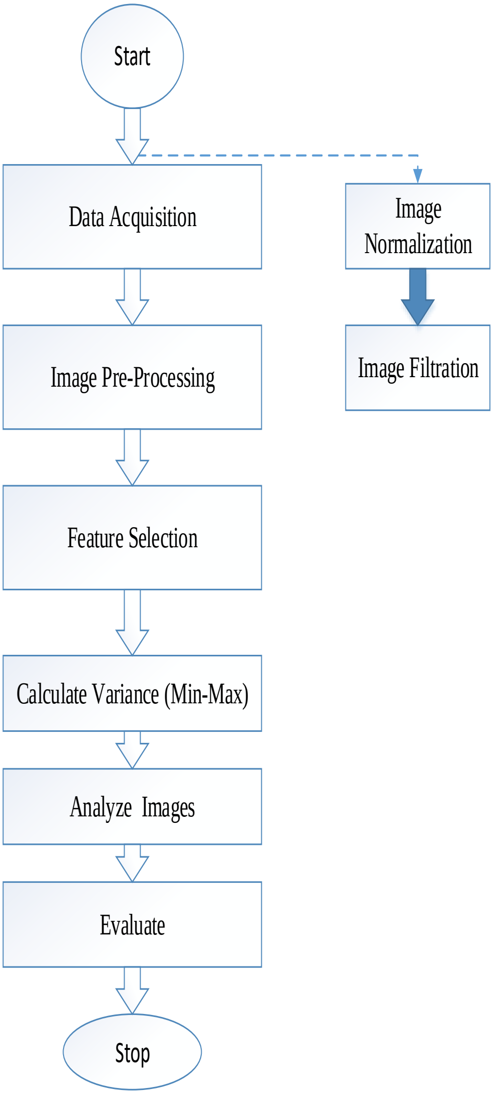

Figure 1: Work Flow chat

3.1.1 Data Collection and Pre-Processing

This work used landsat8 imagery extracted from https://usgs.gov/fm/data/ between the periods of January through June 2018 for Kaduna state Nigeria. Kaduna state capital is a commercial city while other regions of the state predominantly for farming and mining. The state is located in the Northwestern region of Nigeria with a population density of about 6,113,503 according to 2006 population census. The dataset contains satellite images of high resolution and has features such as water bodies, land, vegetation and others such as buildings. The aim of this work is the application of Haar features like machine learning algorithm in identifying objects in the dataset.

After the dataset was collected, it was observed that the images were of different sizes and colour intensity as such, image color channel switch was done and Guassian blur. The reason for guassian blur was to obtain a 2-D distribution fuction which can equally be achieved with convolution. To produce the desired convolution, discrete approximation to Guassian is done to output a weight average of each pixels. The choice of Guassian is because, Guassian output a more smoothing and preserves edge in an improve manner than similar methods. In order to reduce noise in a simpler and effective way, we used binary threshold method (Sezgin 2004; Senthilkumaran 2016).

The pre-processing of data is an essential part of the data mining process. It involves steps like data filtering, replacement of missing information (data cleaning), normalization and feature extraction. In this work, acquired dataset was ensured to be in uniform size, extends and formats. Hence, feature of interest was cropped out. In this work three features were cropped out and thence presents as $C^t = \left(C^t_{ux}, C^t_{uy}, C^t_{uz}\right)$ and C^t_{uc} as the coordinate of each feature in an image $\mathrm{H} \in = \{x, y, z\}$ of features $b \in \{1, ..., k\}$. From our available dataset, a set of features was extracted for water bodies and vegetation.

In order to attenuate noise, introduce into the images, each set of extracted features was first normalized for the purpose of uniformity in training and evaluation.

3.1.2 Feature Normalization

This operation on a dataset is a recurrent operation in machine learning domain (*Forman. et al* 2009). Data normalization for this work is done using the model in equation (4).

$$v_{ij} = \frac{V'_{ij} - \min\left(V'_{m.j}\right)}{\max\left(V'_{m.j}\right) - \min\left(V'_{m.j}\right)} \quad (4)$$

Where V'_{ij} is the feature being normalized, V_{ij} represents values of normalized features respectively and $V'_{m.j}$ is the column of j in the matrix V_m which represents the constructive arrangement of the same feature in the dataset, this operation is carried out for both training set and test set thereby resulting in V_m and V_{me} matrices. Hence, the set of features will be *V*. From the previous steps applied, V can be said to be a set-in matrix containing all features in the dataset.

$$V=\left[\left(V^{1}\right)^{T}\left(V^{2}\right)^{T}.......\left(V^{x}\right)^{T}.....\right], \qquad (5)$$

where V^{x} is a related sub-matrix of feature x.

This work assumes that there are reoccurring features in the dataset as it's possible to assume that each feature can be determined by just physical observation. Based on this assumption, this work performed other statistical analyses to determine the variation between the features in terms of variance in the composition of features.

$$\delta^{t}_{xy}=\sqrt{\sum_{d'}\left(g^{t}_{xd'}-g^{t}_{yd'}\right)^{2}} \qquad (6)$$

where g^{t}_{xy} and g^{t}_{yd} represents max and min respectively, the variance values of two features that look alike are therefore combined using equation (7).

$$G^{t}=\frac{1}{2}\sum_{j=1}^{m}\sum_{d}\left(h^{t}_{jd}\right)^{2}, \qquad (7)$$

Where $d\in\left\{x,y,z\right\}$ and the combination rule must satisfy equation (8)

$$G^{t}<G_{\min}, \qquad (8)$$

Where $G_{\min}$ is the threshold, the process is able to minimize the noise in the selected feature of dataset. The process, however, will produce a system that will have low computational cost and fast computing since it only requires features that are within the threshold.

For the case of feature selection for training and testing, selected features need to be combined and used to form a uniform set of similar sets of datasets. From equation (5), each set of set instance vector V^{a}_{w} can be constructed as

$$v^{a}_{wi}=\left[v^{a}_{wij}v^{b}_{wi}v^{a}_{wil}\right], \qquad (9)$$

Where v^{a}_{wif} and v^{a}_{wil} represent an instance of the selected features that belong to the same grouping.

3.1.3 Training Cascade Classifiers

Here in this work, it's assumed that the model is made up of several independent classifiers, the final detection rate and false-positive rate are given in equation and (1) and (2) respectively (Mutsuddy, 2019). Where k are the steps in the cascade, for selected, featured, the probability that a set of instance sample X will be trained in a classifier at a given stage of training and mapping task to independent classifiers is given as

$$P\left(S\cap H\,|\,X\right)=P\left(S\right)=P\left(S\,|\,X\right).1=P\left(S\,|\,X\right)P\left(H\,|\,X\right) \quad 10)$$

Where $P\left(S\,|\,X\right)$ represent the posterior probability of our output classifier. When S and X are independent $P\left(S\cap X\right)=P\left(S\right)P\left(X\right)$, likewise $P\left(S\,|\,X\right)=P\left(S\right)$. This implies the output of the classifier must not rely upon the input samples of instance. This happens when the instance is on the classifier boundary, where the output of the classifier corresponds to random prediction. This implies condition (a) to the left hand of equation can be forced by methods for the choice of the nearest instance to the boundary of classifier *S*. Then again, condition (b) is forced by the training procedure itself because of the way that the classifier train is fed with the selected training instance.

3.2 Evaluation Matrices

To evaluate the performance of classifiers in this work,five categories of positive image set were used. The categories were represented in percentages (100, 80, 60 ,40 and 20), three categories of negative images were also used. The image sets are of different sizes. Positive image set comprise of image with Water Body and other feature such as Rocks, Vegitation; negative image set has vegetations but no water patches in the image set. At the end of the process, the object correctly detected were saved in a separate location. The results however, showed that, the true positively detected objects represents the positive images with positive features. And the false positive represents images within the positive image set but does not have the object it was trained to detect. In this work, the performance of the classifier was measured using accuracy. For binary classification, accuracy is measured using the expression as follows (Liu, 2019):

$$Accuracy = \frac{T_P + T_n}{T_P + T_n + F_P + F_n} \qquad (11)$$

Where T_n represents true negative, which is denoted for correctly classified of negative instances, T_p (true positive) correctly classify positive instance, F_p, F_n represents false positive and false negative respectively. False-positive incorrectly classifies into negative while the false negative classify instances into positive classes and negative sample instances. The accuracy measurement does not consider unbalance dataset. So, therefore, accuracy measurement has a biased tendency towards the majority classes. Other evaluation parameter considered in this work includes precision and Recall as shown in equation (12 and 13) Bharadwaj (2019).

$$precision = \frac{T_p}{T_p + F_p} \qquad (12)$$

$$\mathrm{Re}\,call = \frac{T_p}{T_p + F_n} \qquad (13)$$

4.0 Result and Discussion

According to Table 1 and 2, Accuracy rate for object detection is shown in Figure 1 and 2. Results of Haarcascade's implemented on set of satellite images contains water body.The result showed boundary area box drawn around water bodies. The algorithm therefore, has high accuracy in detecting presence of water. Figure 2 represents the accuracy detection result of the algorithm when trained in other sets of datasets containing Vegetation. In the training rule, Green vegetation was denoted by cropping Region of Interest (ROI) that has been selected by equation (6)

Table 1 and 2 illustrate the performance accuracy, *TN, TP, FN, FP*, specificity and the sensitivity of the proposed method. The method achieves an accuracy performance of 85.89% when the training set is 100%. Sensitivity attains a performance of 84.7% and 81.3% for specificity.

Figure 3 illustrates the steady increase in accuracy as the size of the dataset increase from 20% to 100% while sensitivity experience a fall to 60% of the dataset. This, therefore, justify equation (13)

Similarly, Figure 4 illustrates the rise in the Specificity of the proposed methods. The specificity increased from 20% to 100%.

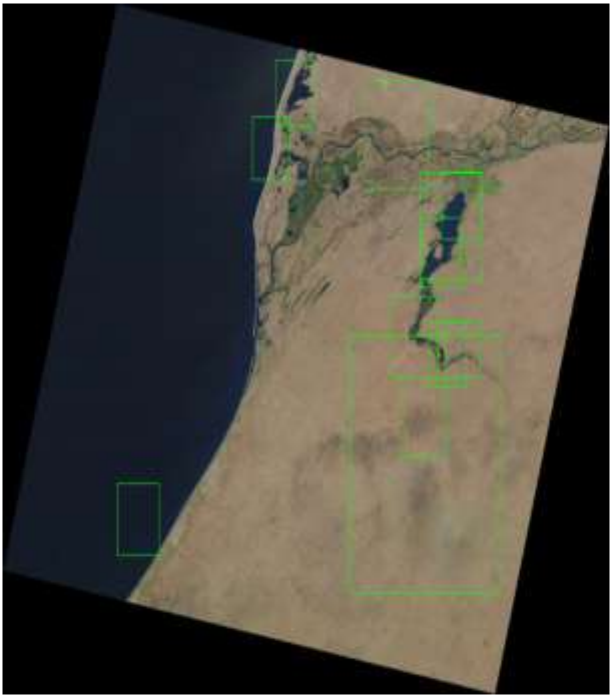

Figure 1: Detection of Water in Satellite Image

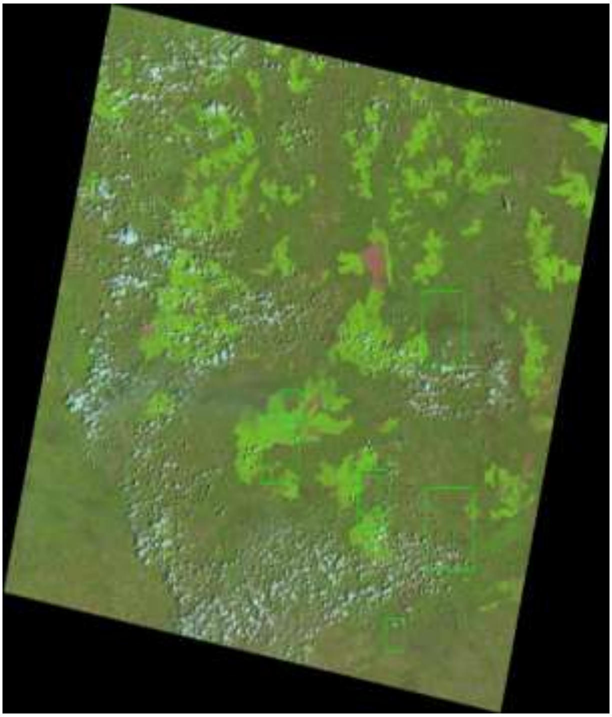

Figure 2: detection of Vegetation in Satellite Image

Table 1: Table of Confusion Matrices

Data size (%)	**TP**	**FP**	**TN**	**FN**
100	149	4	35	39
80	120	8	38	46
60	52	17	35	31
40	60	17	35	13
20	60	38	18	13

Table 2: Table showing specificity and Sensitivity

Data size (%)	**Accuracy**	**Sensitivity**	**Specificity**
100	86.2	89.7	82.2
80	82.55	82.6	72.2
60	65.1	75.3	63.7
40	63.81	77.9	82.2
20	63.76	77.9	82.2

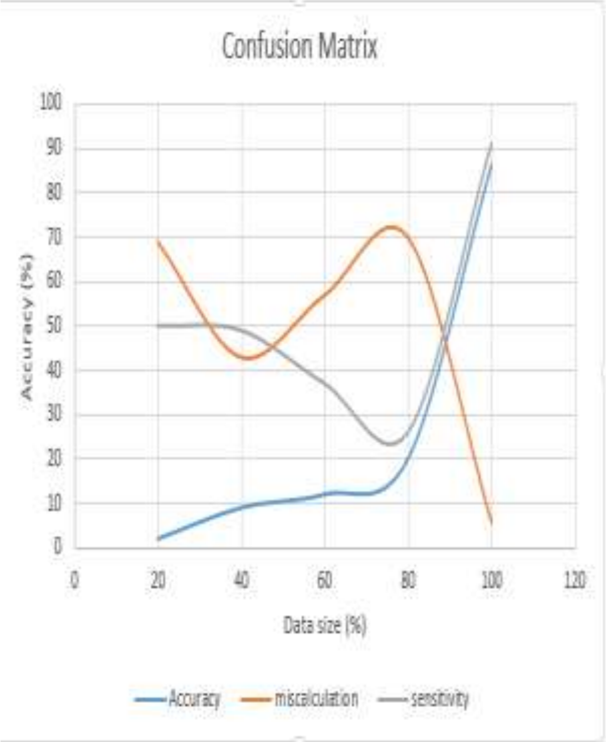

Figure 3: Confusion matrix chart for Haar cascade

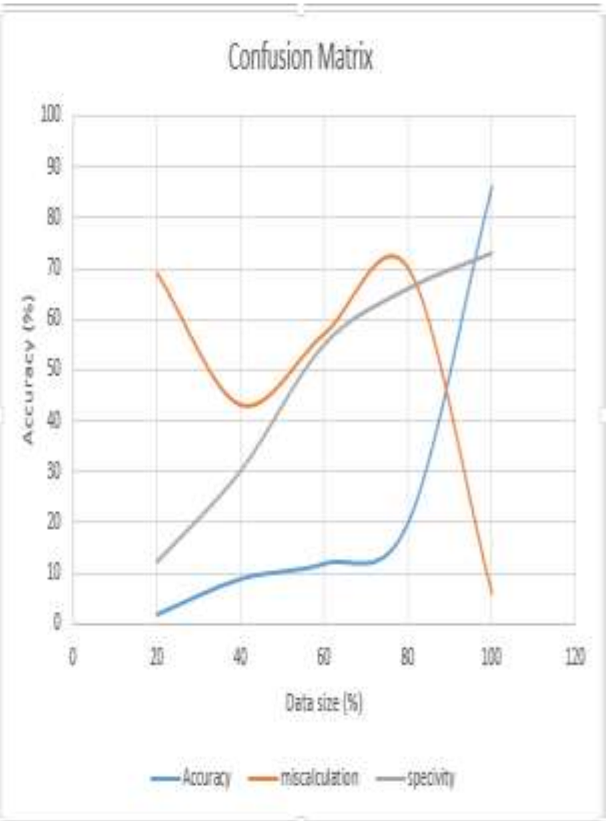

Figure 4: plot of Accuracy and Specificity

4. Conclusion

In this paper, we proposed the use of Differential Evolution algorithm for feature selection and mapping for the purpose of detecting objects and features in satellite images the selected features were trained using HaarCascade machine learning algorithm for detection. The proposed techniques hybridized DE a meta-heuristic algorithm and machine learning to achieve an improvement in reducing computational time and improving the accuracy of the Haar algorithm in detecting objects for satellite image. The result obtained shows that improvement in Accuracy, smaller number of False positive and increased true positive in Table 1 is an indication that the algorithm performed with high efficiency thereby leading Accuracy rate of 86.2,82.5,65.1,63.8 and 63.7 respectively. While Sensitivity and Specificity increases as the size of the training dataset increase which implies our proposed algorithm learn better with large set of data. Comparing the result obtained from this set of Satellite images with other satellite images of the same resolution and but from different location with similar features. This we will consider for future work as this will evaluate the performance of our model giving different location and source of dataset. Secondly, we had recommend the use of Deep Learning Techniques for multiple feature detection from satellite images to further reduce system overhead cost.

References

Baltsavias, E. (2004). Object extraction and revision by image analysis using existing geodata and knowledge: current status and steps towards operational systems. *ISPRS Journal of Photogrammetry and Remote Sensing, 58*(3-4), 129-151.

Beheshti, Z., & Shamsuddin, S. M. H. (2013). A review of population-based meta-heuristic algorithms. *Int. J. Adv. Soft Comput. Appl, 5*(1), 1-35.

Blaschke, T., Hay, G. J., Kelly, M., Lang, S., Hofmann, P., Addink, E., . . . van Coillie, F. (2014). Geographic object-based image analysis–towards a new paradigm. *ISPRS Journal of Photogrammetry and Remote Sensing, 87*, 180-191.

Bharadwaj, P., CN, A., Patel, T.S. and BR, K., 2019.

Drowsiness Detection and Accident Avoidance System in Vehicles.

Cheng, G., & Han, J. (2016). A survey on object detection in optical remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing, 117*, 11-28.

Cinbis, R. G., Verbeek, J., & Schmid, C. (2017). Weakly supervised object localization with multi-fold multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence, 39*(1), 189-203.

Dalai, R., & Senapati, K. K. (2019). A MASK-RCNN Based Approach Using Scale Invariant Feature Transform Key points for Object Detection from Uniform Background Scene. *Advances in Image and Video Processing, 7*(5), 01-08.

Ecaterina, V., Elisa, M., Mihaela, N., & Ovidiu, N. (2011). Differential Evolution in Parameters Estimation. *Journal of Electrical & Electronics Engineering, 4*(1).

Fan, K., Wang, P., & Zhuang, S. (2019). Human fall detection using slow feature analysis. *Multimedia Tools and Applications*, *78*(7), 9101-9128.

Faria, D. R., Vieira, M., Premebida, C., & Nunes, U. (2015, August). Probabilistic human daily activity recognition towards robot-assisted living. In *Robot and Human Interactive Communication (RO-MAN), 2015 24th IEEE International Symposium on* (pp. 582-587). IEEE.

Feoktistov, V. (2006). *Differential evolution*: Springer.

Gollapudi, S. (2019). Object Detection

and Recognition. In *Learn Computer Vision Using OpenCV* (pp. 97-117). Apress, Berkeley, CA.

https://deltas.usgs.gov/fm/data/data_ndwi.aspx

Kim, T., Park, S.-R., Kim, M.-G., Jeong, S., & Kim, K.-O. (2004). Tracking road centerlines from high-resolution remote sensing images by least squares correlation matching. *Photogrammetric Engineering & Remote Sensing, 70*(12), 1417-1422.

Kranthi, B. V., & Surekha, B. (2019). Real-Time Facial Recognition Using Deep Learning and Local Binary Patterns. In *Proceedings of International Ethical Hacking Conference 2018* (pp. 331-347). Springer, Singapore.

Kurian, M. (2011). Various Object Recognition Techniques for Computer Vision. *Journal of Analysis and Computation, 7*(1), 39-47.

Kyrkou, C., & Theocharides, T. (2010). A flexible parallel hardware architecture for AdaBoost-based real-time object detection. *IEEE Transactions on very large scale integration (VLSI) systems, 19*(6), 1034-1047.

Leibe, B., Leonardis, A., & Schiele, B. (2008). Robust object detection with interleaved categorization and segmentation. *International journal of computer vision, 77*(1-3), 259-289.

Leninisha, S., & Vani, K. (2015). Water flow-based geometric active deformable model for road network. *ISPRS Journal of Photogrammetry and Remote Sensing*, *102*, 140-147.

Liu, J., Yan, J., Chen, J., Sun, G. and Luo, W., 2019, ` July. Classification of Vitiligo Based on

Convolutional Neural Network. In *International Conference on Artificial Intelligence and Security* (pp. 214-223). Springer, Cham.

Mayer, H., Hinz, S., Bacher, U., & Baltsavias, E. (2006). A test of automatic road extraction approaches. *International Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences, 36*(3), 209-214.

Merchant, M. A., Warren, R. K., Edwards, R., & Kenyon, J. K. (2019). An Object-Based Assessment of Multi-Wavelength SAR, Optical Imagery and Topographical Datasets for Operational Wetland Mapping in Boreal Yukon, Canada. *Canadian Journal of Remote Sensing*, 1-25.

Mutsuddy, A., Deb, K., Khanam, T. and Jo, K.H., 2019,

August. Illegally Parked Vehicle Detection Based on Haar-Cascade Classifier. In *International Conference on Intelligent Computing* (pp. 602-614). Springer, Cham.

Nunes, U. M., Faria, D. R., & Peixoto, P. (2017). A human activity recognition framework using max-min features and key poses with differential evolution random forests classifier. *Pattern Recognition Letters, 99*, 21-31.

Phuc, L. T. H., Jeon, H., Truong, N. T. N., & Hak, J. J. (2019). Applying the Haar-cascade Algorithm for Detecting Safety Equipment in Safety Management Systems for Multiple Working Environments. *Electronics*, *8*(10), 1079.

Ramisa, A., Vasudevan, S., Scaramuzza, D., De Mántaras, R. L., & Siegwart, R. (2008). *A tale of two object recognition methods for mobile robots.* Paper presented at the International Conference on Computer Vision Systems.

Ren, S., He, K., Girshick, R., Zhang, X., & Sun, J. (2017). Object detection networks on convolutional feature maps. *IEEE transactions on pattern analysis and machine intelligence, 39*(7), 1476-1481.

Sathya, R., & Abraham, A. (2013). Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence, 2*(2), 34-38.

Sezgin, M., & Sankur, B. (2004). Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic imaging*, *13*(1), 146-166.

Senthilkumaran, N., & Vaithegi, S. (2016). Image segmentation by using thresholding techniques for medical images. *Computer Science & Engineering: An International Journal*, *6*(1), 1-13.

T Sharma, T., Rajurkar, S. D., Molangur, N., Verma, N. K., & Salour, A. (2019). Multi-faced object recognition in an image for inventory counting. In *Computational Intelligence: Theories, Applications and Future Directions-Volume II* (pp. 333-346). Springer, Singapore.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *CVPR (1)*, *1*(511-518), 3

.Wang, F., Jiang, D., Wen, H., & Song, H. (2019). Adaboost-based security level classification of mobile intelligent terminals. *The Journal of Supercomputing*, *75*(11), 7460-7478.

Wang, J., Song, J., Chen, M., & Yang, Z. (2015). Road network extraction: a neural-dynamic framework based on deep learning and a finite state machine. *International Journal of Remote Sensing*, *36*(12), 3144-3169.

Zhang, J., Lin, X., Liu, Z., & Shen, J.

(2011). Semi-automatic road tracking by template matching and distance transformation in urban areas. *International Journal of Remote Sensing, 32*(23), 8331-8347.

Zhou, J., Bischof, W. F., & Caelli, T. . (2006). Road tracking in aerial images based on human-computer interaction and Bayesian filtering. *ISPRS Journal of Photogrammetry and Remote Sensing, 61*(2), 108-124