Development of a Phishing Detection System Using Ensemble Machine Learning Method

Oguntunde Bosede O.
Redeemer's University,
Ede Osun State, Nigeria.
oguntunden@run.edu.ng

Iwuh Chizor S
Redeemer's University,
Ede Osun State,Nigeria.
iwuh55548905gb@run.edu.ng

Ojewumi Theresa
Redeemer's University,
Ede Osun State, Nigeria.
ojewumit@run.edu.ng

Abolarinwa Michael O
Osun State University,
Osogbo Osun State, Nigeria.
gbenga1abolarinwa@gmail.com

ABSTRACT

Over the years, phishing has been a major problem and has caused different people to lose sensitive information, hence leading to loss of financial assets. Different machine learning algorithms have been used in the assessment of phishing in different aspects: websites, emails, texts amongst others. However, phishing attacks continue to increase frequency and sophistication despite the numerous attempts to combat it, there is therefore a need for improved detection mechanisms. This study therefore assessed four machine learning algorithms (Random Forest (RF), Logistic Regression (LR), Naive (NB) and Support Vector Machine (SVM)), built an ensemble model with them and developed a system using this model to detect phishing websites. A dataset obtained from Kaggle machine learning repository containing 549,347 records of websites was split into two, 70% to train the ensemble model and 30% to test the model. Two categories of features were selected: Lexical based features and Domain based features of the URL. The performance of the four algorithms were evaluated using accuracy, precision, recall and f1-score. The model was implemented with Python programming language in Jupyter Notebook and 97.42% accuracy was recorded. The results obtained showed that proposed model is comparable to existing models with accuracies of 96%, 98%, 72% and 97% for LR, SVM, RF and NB respectively. The model was used to develop a user-friendly system where users can paste URLs in order to check the safety of the address. The system however is limited to HTTP protocols and might not be equipped to handle short URLs.

Key words: Phishing, Machine Learning, Machine Learning algorithms, Random Forest, Logistic Regression, Naïve Bayes, Support Vector Machine.

1. INTRODUCTION

The importance of data privacy, protection, and security against cyberattacks cannot be overstated, given the massive expansion in data caused by social networks, Internet usage, and IoT devices over time and the rapid improvement of technology. Phishing is one of the most common cyberattacks where malicious actor sends an email posing as from a respected firm to obtain classified and detailed information about its host to gain access to data, applications, and networks without authorization. Most phishing victims are unaware the email they received contains malicious software or could direct them to fraudulent websites in order to trick them into providing information as sensitive as credit card numbers, account IDs and login detail. Attackers use phishing to trick users into clicking on a malicious link that appears to be trustworthy. Jang-Jaccard and Nepal (2014) emphasized that cybersecurity concerns must be addressed as attackers employ more sophisticated phishing techniques. According Kuala (2020), 91% of all cyber-attacks initiated by sending a disguised email to an unsuspecting victim, and phishing techniques account for 32% of successful infiltrations. Despite multiple attacks throughout the years, people are continuously targeted by this most aged malware.

Cybercrimes including phishing continue to have a devastating impact on Nigeria's economy. According to Microsoft, phishing assaults have grown by 250 percent both nationally and internationally (Adepetun, 2019). Ogbonnaya (2020) opined that cybercrimes and electronic fraud cost Nigeria's commercial banks a combined $39 million in 2018, the bulk of which were committed through phishing and identity theft, which raised the loss from similar crimes from the previous year. During the COVID_19 pandemic, Deloitte Nigeria reports that fraudsters are exploiting the epidemic to deceive people into downloading ransomware that is masquerading as a legitimate COVID application. The cost of cyberattacks on Nigerian enterprises, including phishing and SQL injection assaults, is estimated at $649 million annually by Serianu's Nigeria Cyber security Report (Serianu, 2017). Nigeria has experienced an evolution in cyberattacks throughout time. Online fraudsters have developed new methods to collect personal information in response to numerous phishing attack warnings and awareness campaigns. Given the tremendous increase in IT equipment, one would wonder if businesses are prepared to manage breaches and phishing attempts. According to the report (FireEye, 2020), 51% of firms believe they are unprepared. The Anti-Phishing Working Group (APWG, 2020) discovered 182,465 phishing sites in Q2 of 2019 and 266,387 just in Q3, during the COVID-19 epidemic in the Q2 of 2020, 165,772 phishing websites were discovered while 112,163 attacks occurred in the Q2, 122,359 in Q3 and 132,553 in the Q4 of the same year. This goes to show how ineffective the current systems and measures against phishing have been. Over time, phishing assaults have improved and gotten more sophisticated, discovery and combating techniques also need to be enhanced for efficiency. Enterprises have built phishing detection systems to detect and report phishing attacks using machine learning techniques (Choudhary, et al, 2023). Machine learning is a branch of artificial intelligence that builds models from sample data. Machine learning uses a combination of methods to analyze the structure, feature and content of suspected URLs and compare with known phishing URLs.

2. RELATED WORKS

Wenyin et al.'s (2010) study suggests a novel approach for predicting harmful websites. The model Link network suggests a method for identifying phishing websites based on four convergent situations. The Intelligent Phishing Website Detection and Prevention System provided by (Naresh, 2013) has the drawback of indicating positive where not the case for any non-fraudulent website addressed by just the IP, not including a domain name which presented inaccurate mistakes and inaccurate model results (Madhuri et al., 2013). Baykara and Gürel (2018) suggests an anti-phishing simulator to be used to prevent phishing attacks. This model provides comprehensive details on the frequency of phishing attempts, including methods to pointing them out using Bayes method to determine spam content from its acquired dataset. This simulator worked by detecting potentially malicious keywords from a given sample text. Han et al. (2012) provides a strategy that combines visual similarity-based algorithms with white lists to give the system the ability to fight against phishing assaults. Blum et al. (2010) expressively declares that lexical characteristics perform well and are simpler to compare than other features. The URL length in addition to the number of special characters proved to be the statistical aspects of the website string that Srinivasan et al., (2021) found to be the most often used attributes. For the purpose of detecting URL phishing, Vinayakumar et al. (2019) emphasized the need to compare feature engineering techniques and character level embedding through Machine Learning and Deep learning. Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM), deep learning architectures outperformed conventional machine learning algorithms, as reported by Bahnsen et al. (2017) for URL identification using random forest classifier, lexical characteristics, and statistical URL analysis. Despite the excellent performance of both models, the LSTM outperformed the conventional machine learning models. Traditional detection classifiers with an emphasis on URL predictive filtering employed content analysis and blocklists. This study created a model to identify malicious URLs on the Splunk platform. Malicious and benign URLs were divided into categories using the Support Vector Machine (SVM) and Random Forest (RF). SVM has a 90% accuracy score and an 88% recall score compared to the RF algorithm's 85% precision and 87% recall (Christou et al, 2020).

DeepURLDetect (DUD), a service provided by Alazab & Fellow (2020) helped improve the accuracy of phishing website identification by offering a feature selection method. This method is paired with a majority-voting-related ensemble learning model. Afterward, it was contrasted with several classification models, such as Random Forest, Logistic Regression, and Prediction model. The experiment had a 95% accuracy rate, and the learning model used indicates that the suggested approach would be more useful for identifying URL phishing. Sahingoz et al.'s (2019) suggested anti-phishing model used distinct classification algorithms and NLP-based characteristics. The RF algorithm, when used exclusively with features centered around NLP, produced 97.98% accuracy for the detecting malicious links, according to the findings of testing. Alexa and PhishTank had datasets taken and were used by the authors to construct a methodology (Basit et al., 2020). DT, KNN, SVM, and NB were the four classifiers used. The PhishTank dataset with 11055 URLs was compared for model content and characteristics in Abdelhamid et al., (2017). According to the creators, the first ML and DL algorithm to be utilized by an application for anti-phishing is Dynamic rule induction (eDRI). Page structure property and 49 malicious websites obtained from the PhishTank.com dataset were utilized in another method by Mao et al. (2019). All ML classifiers listed — SVM, AdaBoost, DT, and RF — were evaluated on analyses of more than 20,000 message sample data points. Jain & Gupta (2018) shows that by using RF, SVM, LR, and NB on various datasets, detection accuracy could be boosted to 99.09%. The first dataset came from PhishTank and featured 1528 phishing sites. The next two datasets, from OpenPhish and Alexa, each comprised 613 phishing sites. Ubing et al., (2019) suggests another study based on ensemble learning, using the three approaches bagging, boosting, and stacking. The data set had 30 characteristics, and the result column contained 5126 entries. A remove replace feature selection strategy (RRFST) was presented by (Hota et al., 2018) and evaluated on a dataset of phishing emails from the khoonji's anti-phishing website, which comprised 47 features. Their proposed model condensed 30 characteristics to just 11. The Random Forest method, which is used to identify phishing attacks, is another well-liked technique. Many researchers have used RF in the past, with promising outcomes. Joshi and Pattanshetti (2019) presents the Relief Feature Selection (reliefF) method utilized as a forward selection strategy employing 48 features and the Random Forest algorithm as a binary classifier. Choudhary et al (2023) evaluated the performance of five machine learning models for detecting phishing websites; XGBoost, DT, LR. RF AND SVM, on PhishTank and UCL datasets. Random forest showed a superior performance with 98.8% and 97.87% accuracy respectively. In Dinesh (2023), XGBoost outperformed Random Forest and Logistic Regression in identifying phishing attacks with 94.2% accuracy. Jayaraji et al (2024) employs Hybrid Ensemble Feature Selection (HESF) method to develop a machine learning based phishing detection system. The system achieves 96.8% accuracy. Bahaghighat, Ghasemi & Ozen (2023) proposed an improved predictive machine learning model based on six different algorithms, LR, KNN, NB, RF, SVM and XGBoost. XgBoost shows a superior performance over others with 99.2% accuracy.

3. METHODOLOGY

The architecture of the proposed system is depicted by figure 1 showing the stages of building the system. Data in form of URLs were collected from a public repository, consisting of authentic and known malicious URLs, the data were preprocessed and split into two for training and testing on the selected machine learning algorithms. The details of the processes are discussed in this section.

3.1 Data Collection

A total of 549347 records was collected from Kaggle data repository, containing 393424 legitimate site and 156422 phishing sites. The dataset was split into two in ratio 70:30 for training and testing the model respectively. The dataset of URLs contains two distinct categories; phishing URLs and legal URLs.

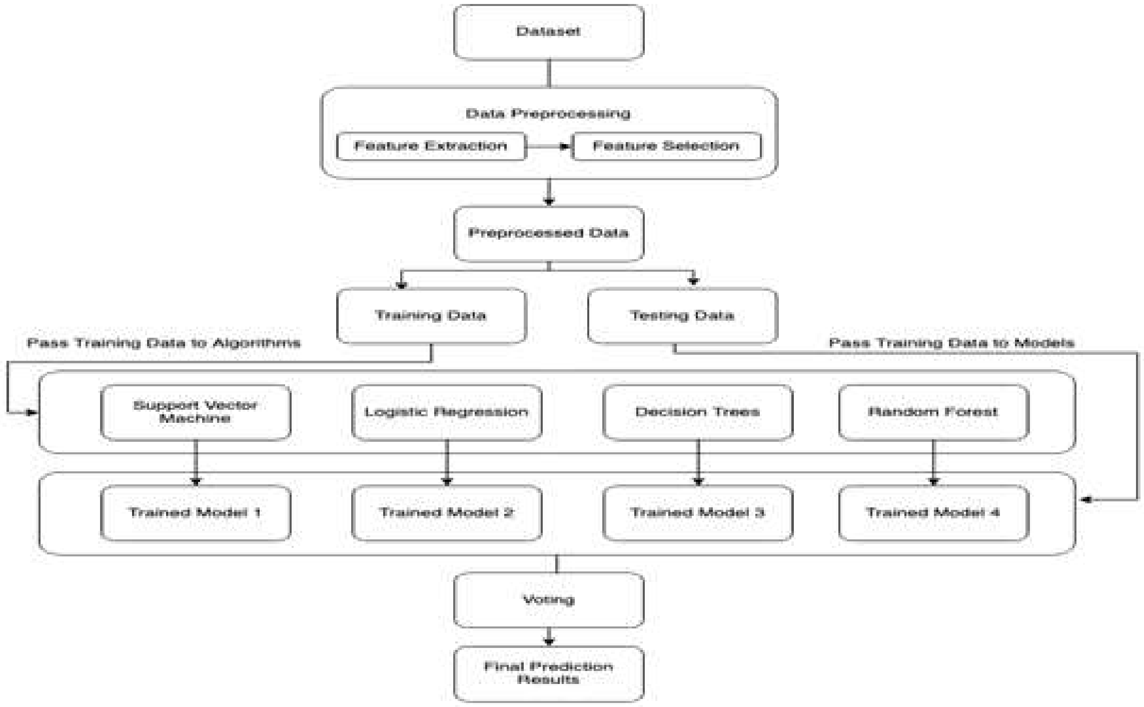
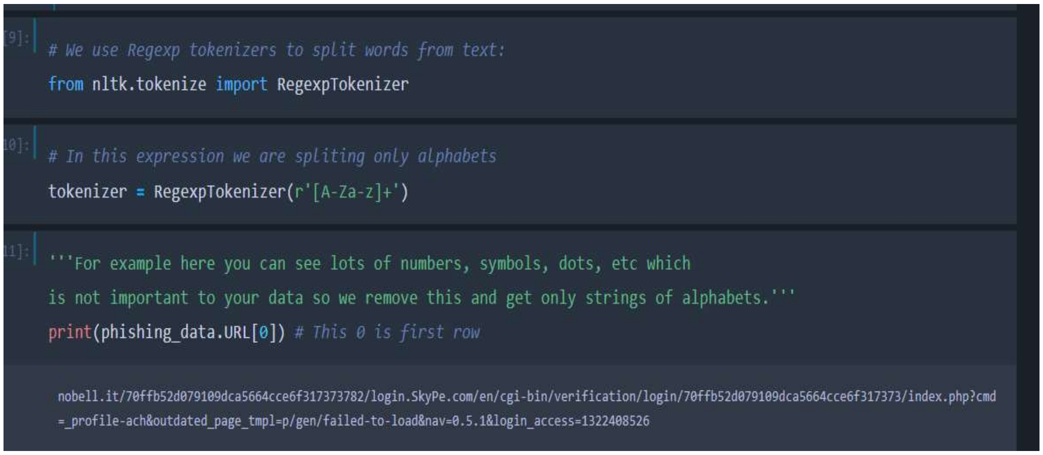


Figure 1 Architecture of the Model

3.2 Data Preprocessing

The data were preprocessed, recurring items were erased from the dataset, omitted properties were filled in. Row of all missing labels was removed during data cleaning. A class that lacks values would have the mean of the properties under such field take its place. The URLs were vectorized, using Count Vectorizer to aggregate words, Regexp tokenizer was used, it uses a regular expression to split a string by matching either the tokens or the separators between tokens as presented by figure 2. Snowball stemmer was used to get the root words out of tokenized text.



```
# We use Regexp tokenizers to split words from text:
from nltk.tokenize import RegexpTokenizer

# In this expression we are spliting only alphabets
tokenizer = RegexpTokenizer(r'[A-Za-z]+')

'''For example here you can see lots of numbers, symbols, dots, etc which
is not important to your data so we remove this and get only strings of alphabets.'''
print(phishing_data.URL[0]) # This 0 is first row

nobell.it/70ffb52d079109dca5664cce6f317373782/login.SkyPe.com/en/cgi-bin/verification/login/70ffb52d079109dca5664cce6f317373/index.php?cmd
=_profile-ach&outdated_page_tmpl=p/gen/failed-to-load&nav=0.5.1&login_access=1322408526
```

Figure 2: Tokenizer

3.3 Feature extraction and selection

The data was represented in the features form to train the machine learning algorithms.: Lexical based features and Domain based features, the sum of thirteen (13) features were selected from two categories of features, nine lexical based features and four domain-based features. Value 0 is given to URLs that do not contain features which may make up a phishing URL and 1 otherwise.

Lexical Features

i. The presence of the @ symbol in a URL: The presence of "@" symbol in 20% of phishing website makes URLs which contain them hold a value of 1 other wise 0.
ii. HTTPS in the middle of the URL: "https" string appended to the domain of the URL will assume a value of 1 else a value of 0.
iii. Length of URL: Any URL which includes more than 54 characters will be flagged with a value of 1.
iv. Redirect Request: Double slash "//" should only be included after the protocol. Any position beyond that would make that URL suspicious and hence flagged with a value of 1.
v. Prefix or Suffix "-" in Domain: The "-" character could be the only differentiator between the string of a legitimate site and a malicious hence it could, and has been used to create exploits which would earn URLs containing the "-" character a value of 1.
vi. Using URL Shortening Services "TinyURL": Shortened URLs could lead to blind redirect. This is an easy exploit for a bad actor. URLs created using shortening services would be flagged with a value of 1.
vii. Delimiter Characters: Suspicious URLs have been known to contain delimiters such as #, & over the years earning URLs containing such characters a value of 1.
viii. Dot count**:** When a domain goes beyond the third level, it makes the URL suspicious earning it a value 1.
ix. Having IP Address: URLs containing an IP address are suspicious and earn a 1.

Domain-Based Features

i. Domain Age: Sites that have lived no longer than a year would be suspicious and its URL would earn a value of 1.
ii. Website Validity: If the website is no longer active then it would be flagged with a value of 1.
iii. Name Server Record: If there is no public record of this URL that makes it makes a suspicious one, earning it a value of 1.
iv. Web Traffic**:** These metrics should how much users interact with the site. Malicious site should not have as much activity on them and so any website not ranking about 100,000 will be flagged with a value of 1.

4. MODEL BUILDING

Four machine learning classification algorithms were used to build classification models, they are Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB) and Support Vector Machine (SVM) algorithms. The algorithms were trained, tested and their individual performances evaluated based on accuracy, precision and recall. The proposed ensemble model was built with these four algorithms and its base learners and soft voting by the base learners determines the classification. Figure 3 shows the process flow involved in the phishing detection system.

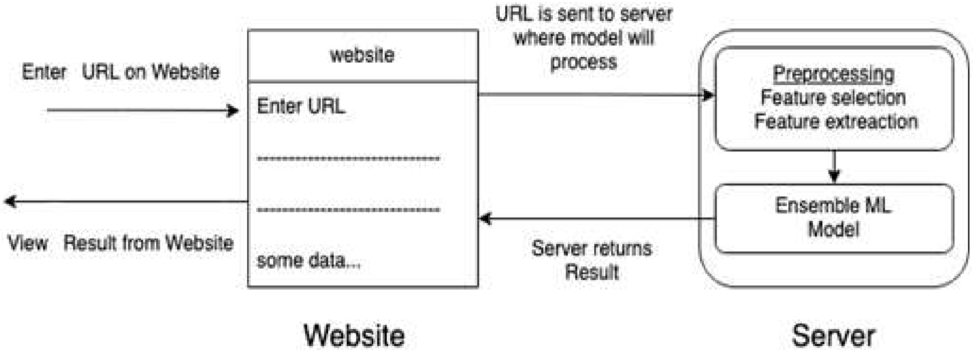


Figure 3: **Process Flow of the Proposed System**

The link (URL) to check is entered into the system using the web interface built to interact with our developed model. The link is then uploaded the web application where it is sent to the model. The machine learning model is a pretrained model which classifies whether the link is safe to use or not.

5. RESULTS AND EVALUATION

The parameters used to tweak the model and for the re-evaluation process to achieve the optimal results are described. Recall, F1-score and Accuracy (ACC) were used to measure each classified effectiveness. Then, as inputs, feed the various URLs into the trained model. It predicts whether the URL is good or bad and delivers a positive or negative result. To begin, import Logistics Regression, Support Vector Machine, and Random Forest and |Naïve Bayes. Then, using fit (), fit the model to the train set and predict on the test set.

5.1 Logistic Regression

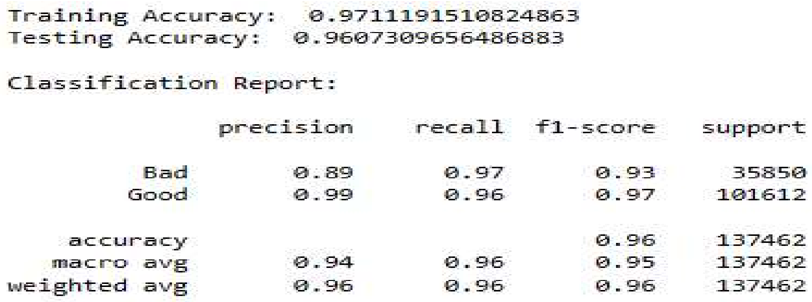
Logistic regression is a statistical model (sometimes known as a logit model) that is frequently used in classification and predictive analytics. Based on a given dataset of factors, logistic regression assesses the chances of an event occurring, such as voting or not voting. Because the outcome is a probability, the dependent variable has a range of 0 to 1. A logit transformation is used to the odds (that is, the likelihood of success divided by the probability of failure) in logistic regression. This is often referred to as the log odds or the natural logarithm of odds, which is represented by the formular given in equation 1.

$$\text{Logit}(pi) = 1/(1+ \exp(-pi))$$

$$\ln(pi/(1-pi)) = Beta\_0 + Beta\_1*X\_1 + \ldots + B\_k*K\_k$$

Eq.1
Logistic regression model gave an accuracy of 97% as shown in table 1, while its confusion matrix is presented in figure 4.

Table 1: Logistic regression



```
Training Accuracy:  0.9711191510824863
Testing Accuracy:  0.9607309656486883

Classification Report:

              precision    recall  f1-score   support

         Bad       0.89      0.97      0.93     35850
        Good       0.99      0.96      0.97    101612

    accuracy                           0.96    137462
   macro avg       0.94      0.96      0.95    137462
weighted avg       0.96      0.96      0.96    137462
```

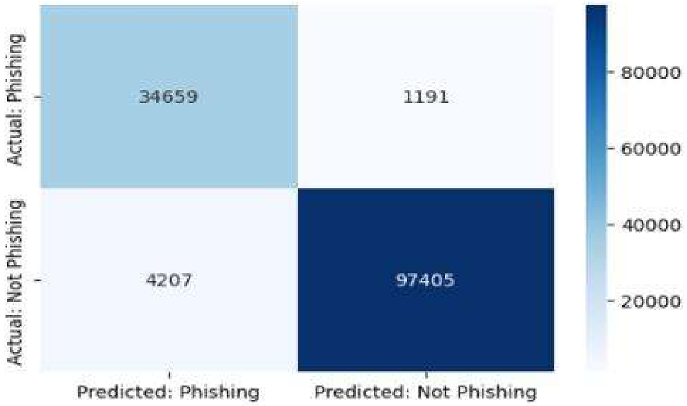
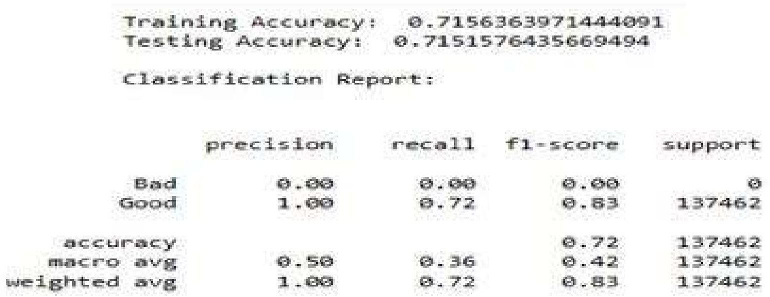


Figure 4: Confusion matric for Logistic Regression score result

5.2 The Random Forest,

Random decision forest is a machine-learning ensemble method that can be used for regression and/or classification. First, a number of decision trees are built on a randomly selected subset of training sets in this sort of categorization. Following that, they aggregate the decisions of the trees and obtain an average of them, not only to improve forecast accuracy but also to control over-fitting; additionally, the instabilities of the individual decision trees can be eliminated by its forest structure. Random forest model obtained an accuracy of 71.56% as presented in table 2 with its confusion matrix in figure 5.

Table 2: Random Forest



```
Training Accuracy:  0.7156363971444091
Testing Accuracy:  0.7151576435669494

Classification Report:


              precision    recall  f1-score   support

         Bad       0.00      0.00      0.00         0
        Good       1.00      0.72      0.83    137462

    accuracy                           0.72    137462
   macro avg       0.50      0.36      0.42    137462
weighted avg       1.00      0.72      0.83    137462
```

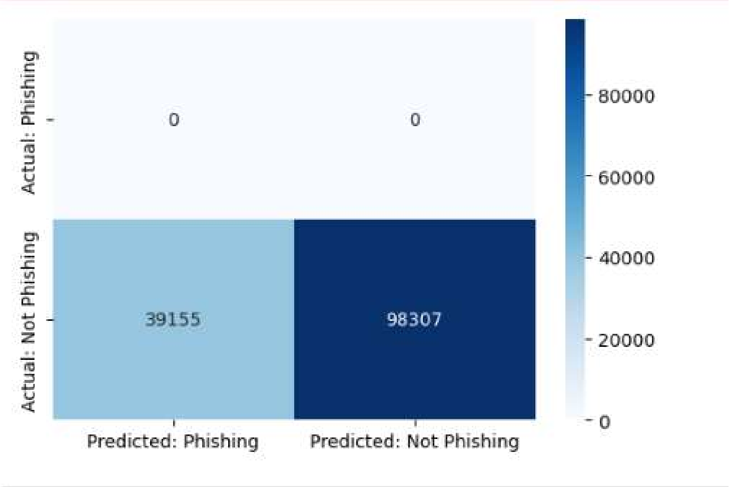


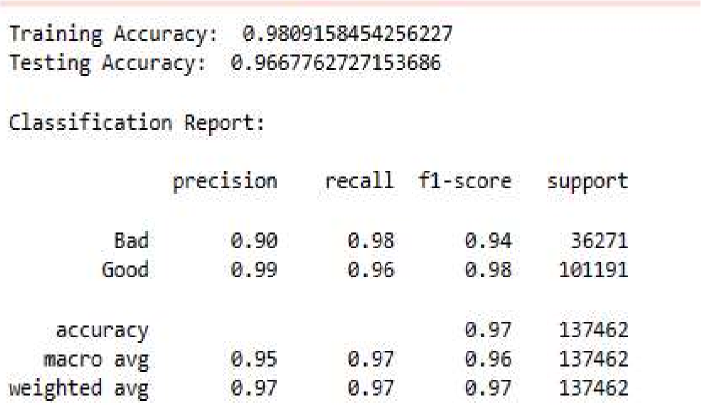
Figure 5: Confusion matrix for Random Forest

5.3 Naïve Bayes

The Nave Bayes classification is a probabilistic machine learning algorithm that is both simple and powerful. It is based on the Bayes theorem, which describes the relationship between statistical quantities' conditional probabilities. It is predicated on the concept of attribute value independence. Equation (2) is used to calculate the conditional probability.

$$P(H|E) = \frac{P(E|H)*P(H)}{P(E)} \qquad (2)$$

Table 3: Naïve Bayes



```
Training Accuracy:  0.9809158454256227
Testing Accuracy:  0.9667762727153686

Classification Report:

              precision    recall  f1-score   support

         Bad       0.90      0.98      0.94     36271
        Good       0.99      0.96      0.98    101191

    accuracy                           0.97    137462
   macro avg       0.95      0.97      0.96    137462
weighted avg       0.97      0.97      0.97    137462
```

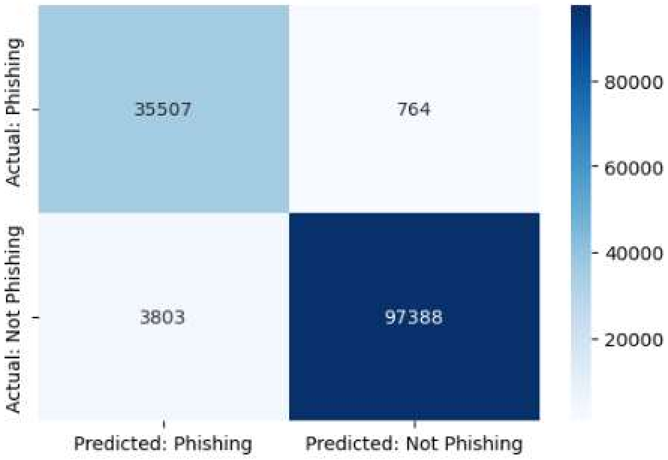
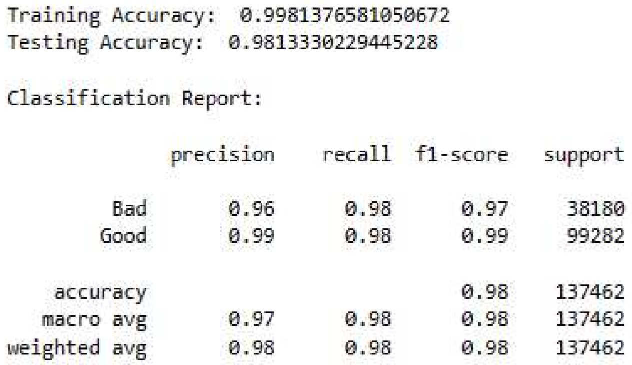


Figure 6: Confusion matrix for Naïve Bayes

5.4 Support vector machine (SVM)

Support Vector Machine is an important classifier in the machine learning concept, which is an important classifier in the machine learning concept, which searches non-linear decision boundaries using kernel tricks in the trained data

Table 4: Support vector machine



```
Training Accuracy:  0.9981376581050672
Testing Accuracy:  0.9813330229445228

Classification Report:

              precision    recall  f1-score   support

         Bad       0.96      0.98      0.97     38180
        Good       0.99      0.98      0.99     99282

    accuracy                           0.98    137462
   macro avg       0.97      0.98      0.98    137462
weighted avg       0.98      0.98      0.98    137462
```

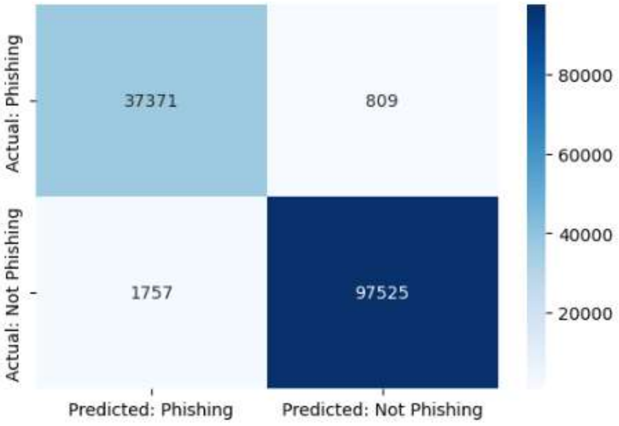
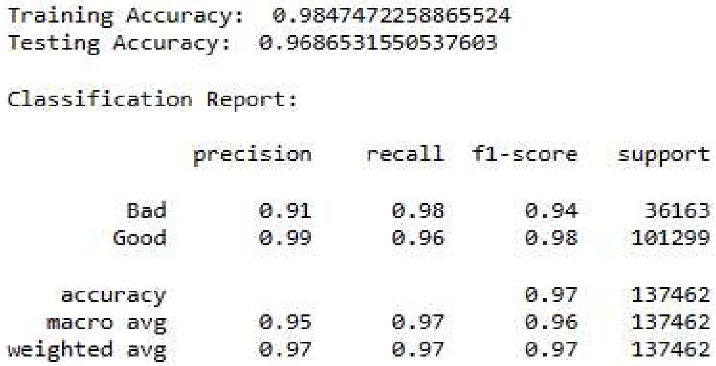


Figure 7: Confusion matrix for SVM

5.5 Proposed Voting Classifier

The proposed classifier is an ensemble method that combines four base learners and makes its prediction based on the combined outcome of the base learners. The proposed classifier in this work allows each of the four algorithms to make their individual predictions and the outcome is combined, soft voting technique by which the final prediction is made based on the mean probability of the base learners. The proposed voting classifier is applied to the same dataset with the same split ration and an accuracy of 98.47% was obtained as presented in table 5 and the confusion matrix is as shown in Figure 8.

Table 5: Voting Classifier



```
Training Accuracy:  0.9847472258865524
Testing Accuracy:  0.9686531550537603

Classification Report:

              precision    recall  f1-score   support

         Bad       0.91      0.98      0.94     36163
        Good       0.99      0.96      0.98    101299

    accuracy                           0.97    137462
   macro avg       0.95      0.97      0.96    137462
weighted avg       0.97      0.97      0.97    137462
```

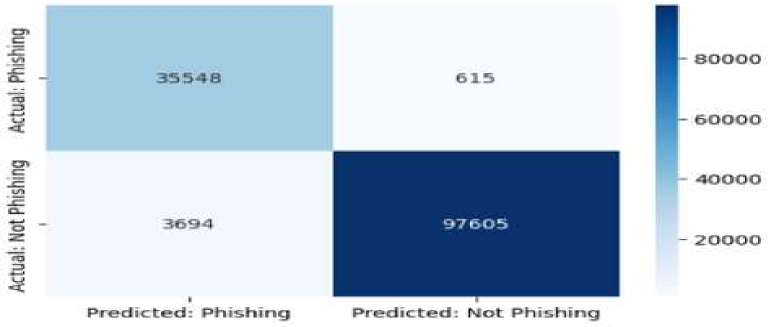


Figure 8: Confusion Matrix of Voting Classifier

The confusion matrix of the proposed model depicted by Figure 8 shows that the model achieved a true positive of 35548 and a true negative of 97605 for its predictions, 615 of its predictions were incorrect (false positive), predicting that a sample was a good site when it was not. It also has a false negative rate of 3694, which means that 3694 of its forecasts for good sites were incorrect.

Table 7: Comparison of Model Result

Technique	Precision	Recall	Fi-score	Accuracy
Logistic Regression	0.89	0.97	0.93	0.96
Random Forest	1.00	0.72	0.83	0.72

Naïve Bayes	0.90	0.98	0.94	0.97
Support Vector Machine	0.96	0.98	0.97	0.98
Voting Classifier	0.91	0.98	0.98	0.97

6. Discussion of Results.

This section discusses the experimental findings of the creation of a phishing detection system utilizing an ensemble machine learning method comprised of Logistics Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and |Nave Bayes (NB). The accuracy of the approaches was determined to be 96% for the LR model, 98% for the SVM, 72% for the RF, 97% for the NB, and the voting classifier is 97% accurate. Support vector model shows superior phishing detection accuracy among the models employed for testing and training, followed by Random Forest and the proposed ensemble voting classifier. The distribution of Phishing Websites is greatest in the Count plot. Support Vector Machine, Naive Bayes, and Logistic Regression are good fit models, but Random Forest has low accuracy, precision, and recall when compared to other classifiers. The proposed voting classifier combined the strength of the four base classifiers and shielded the weakness of the weakest of them all.

6.1 System Use case

The link to check is entered into the system using the web interface built to interact with our developed model. The link is then uploaded into the web application where it is sent to the model. The machine learning model is a pretrained model which classifies whether the link is safe to use or not. Figure 9 shows the phishing URL detector system.

If the link is safe to use as in Figure 10, the interface indicates the link is safe to use with a “Website is safe to use” with a green button to proceed to the resource.

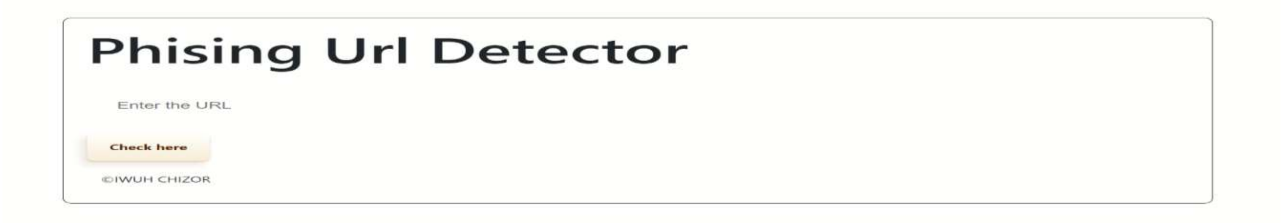
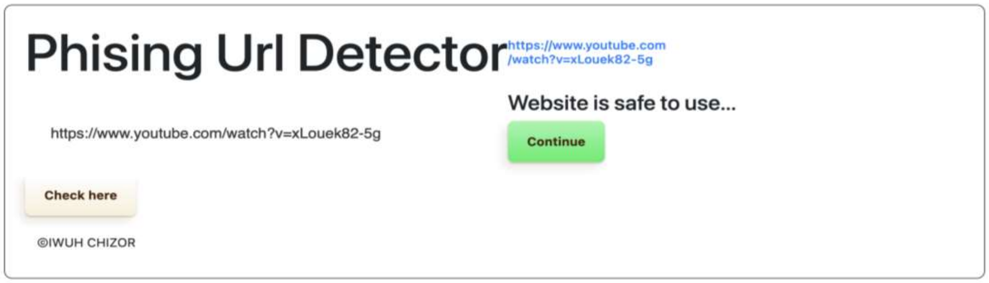


Figure 9: Phishing URL detector system



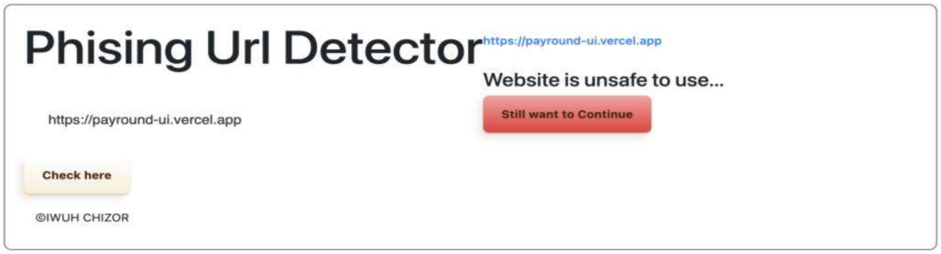
https://www.youtube.com/watch?v=xLouek82-5g

Website is safe to use...

https://www.youtube.com/watch?v=xLouek82-5g

Figure 10: Good link input with result

When the link is not safe to use as shown in Figure 11, it indicates that with a "Website is unsafe to use" and a red button to proceed to the resource if the user still wishes to go ahead.



Phising Url Detector

https://payround-ui.vercel.app

Website is unsafe to use...

https://payround-ui.vercel.app

©IWUH CHIZOR

Figure 11: Bad link input with result

7. CONCLUSION

This work employed machine learning to differentiate between dangerous and lawful web pages. A phishing detection system based on an ensemble of four different machine learning algorithms: Logistics Regression, Support Vector Machine, Random, Forest, and Naive Bayes. A dataset consisting of 549347 URLs from the Kaggle repository was used to evaluate the performance of the suggested URL detector. Lexical and domain-based features were extracted from the data. The proposed ensemble classifier combines the individual outcome of the four algorithms and uses a soft voting technique to make the final prediction based on the mean probability of the four base learners.

The ensemble technique using soft voting has proven to be dependable and effective in detecting phishing links since it combined the strengths of the models used to develop it while balancing out their weaknesses. The evaluation of the methodologies revealed that the LR model has an accuracy of 96%, SVM has an accuracy of 98%, RF has an accuracy of 72%, NB has an accuracy of 97%, and the voting classifier has an accuracy of 97%. Support vector models shows superiority in terms of phishing detection accuracy among the models employed for testing and training. The system was tested and built with only HTTP links and hence might not be well equipped to handle other URL links. Time constraint was another limitation as there was not enough to properly train model to better detect complex URLs. Phishing URLs might pass with URL shortening services because of the age of the domains

REFERENCES

[1] Abdelhamid, N., Thabtah, F., & Abdel-Jaber, H. (2017). Phishing detection: A recent intelligent machine learning comparison based on models content and features. *2017 IEEE International Conference on Intelligence and Security Informatics: Security andBig Data, ISI 2017*, 72–77. https://doi.org/10.1109/ISI.2017.8004877

[2] Adepetun, A. (2019). Microsoft Report: Phishing Attacks Increase by 250%. Cybersecurity Journal, 15(3), 45-62.

[3] Alazab, M., & Fellow, S. (2020). *Malicious URL Detection using Deep Learning*. 1–9.Alkhalil, Z., Hewage, C., Nawaf, L., & Khan, I. (2021). Phishing Attacks: A Recent Comprehensive Study and a New Anatomy. *Frontiers in Computer Science*, *0*, 6.https://doi.org/10.3389/FCOMP.2021.563060

[4] Anti-Phishing Working Group. (2020). Phishing Activity Trends Report 3rd Quarter 2020. *Apwg*, *November*, 1–12. https://docs.apwg.org/reports/apwg_trends_report_q3_2020.pdf

[5] APWG. (2020). APWG Trends Report Q2 2020. *Phishing Activities Trends Report*, *Q2 2020*(August), 1–13. www.apwg.org

[6] Baykara, M., & Gürel, Z. Z. (2018). Detection of phishing attacks. *6th International Symposium on Digital Forensic and Secughigrity, ISDFS 2018 - Proceeding*. https://doi.org/10.1109/ISDFS.2018.8355389

[7] Bahaghighat M., Ghasemi M., & Ozen F. (2023). A high-accuracy phishing website detection method based on machine learning. Journal of Information Security and Application (77). https://doi.org/10.1016/j.jisa.2023.103553 (https://www.sciencedirect.com/science/article/pii/S2214212623001370)

[8] Bahnsen, A. C., Bohorquez, E. C., Villegas, S., Vargas, J., & Gonzalez, F. A. (2017). Classifying phishing URLs using recurrent neural networks. *ECrime Researchers Summit, ECrime*, 1–8. https://doi.org/10.1109/ECRIME.2017.7945048

[9] Basit, A., Zafar, M., Javed, A. R., & Jalil, Z. (2020). A Novel Ensemble Machine Learning Method to Detect Phishing Attack. *Proceedings - 2020 23rd IEEE International Multi-Topic Conference, INMIC 2020, November*. https://doi.org/10.1109/INMIC50486.2020.9318210

[10] Blum, A., Wardman, B., Solorio, T., & Warner, G. (2010). Lexical feature based phishing URL detection using online learning. *Proceedings of the ACM Conference on Computer and Communications Security*, 54–60. https://doi.org/10.1145/1866423.1866434

[11] Choudhary T., Mhaphankar S., Bhddha R., Kharuk A., & Patil R. (2023). A Machine Learning Approach for Phishing Detection. Journal of Artificial Intelligence and Technology (3), pp 108 -113. https://doi.org/10.37965/jait.2023.0197

[12] Christou, O., Pitropakis, N., Papadopoulos, P., McKeown, S*., & Buchanan, W*. J. (2020). Phishing URL detection through top-level domain analysis: A descriptive approach. *ICISSP 2020 - Proceedings of the 6th International Conference on Information Systems Security and Privacy*, *March*, 289–298. https://doi.org/10.5220/0008902202890298

[13] Dinesh, P. M., Mukesh, M, Navaneethan, B, Sabeenian, R. S., Paramasivam, M. E & Manjunathan, A (2023). Proceedings of E3S Web of Conferences 399, 04010, International Conference on Newer Engineering Concepts and Technology (ICONNECT-2023). https://doi.org/10.1051/e3sconf/202339904010

[14] FireEye. (2020). *Cyber Trendscape 2020*. https://www.fireeye.com/offers/rpt-cyber-trendscape.html

[15] Han, W., Cao, Y., Bertino, E., & Yong, J. (2012). Using automated individual white-list toprotect web digital identities. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2012.02.020

[16] Hota, H. S., Shrivas, A. K., & Hota, R. (2018). An Ensemble Model for Detecting PhishingAttack with Proposed Remove-Replace Feature Selection Technique. *Procedia Computer Science*, *132*, 900–907. https://doi.org/10.1016/j.procs.2018.05.103

[17] Jain, A. K., & Gupta, B. B. (2018). Towards detection of phishing websites on client-side using machine learning based approach. *Telecommunication Systems*, *68*(4), 687–700.https://doi.org/10.1007/S11235-017-0414-0

[18] Jang-Jaccard, J., & Nepal, S. (2014). A survey of emerging threats in cybersecurity. *Journalof Computer and System Sciences*. https://doi.org/10.1016/j.jcss.2014.02.005

[19] Jayaraji R., Pushpalatha A., Sangeetha K., Kamaleshwar T., Udhaya Shree S. (2024). Intrusion Detection Based On Phishing Detection With Machine Learning. Measurement: sensors (31), (2024) 101003. Elsevier. www.science direct.com/journal/measurement-sensor. http://doi.org/101016/j.measen.2023.101003

[20] Joshi, A., & Pattanshetti, P. T. R. (2019). Phishing Attack Detection using Feature SelectionTechniques. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3418542

[21] KUALA LUMPUR. (2020). *91% of all cyber attacks begin with a phishing emailto an unexpected victim | Deloitte Malaysia | Risk Advisory | Press releases*. https://www2.deloitte.com/my/en/pages/risk/articles/91-percent-of-all-cyber-attacks- begin-with-a-phishing-email-to-an-unexpected-victim.html

[22] Madhuri, M., Yeseswini, K., & Sagar, U. V. (2013). Intelligent Phishing Website Detection and Prevention System by Using Link Guard Algorithm. *International Journal of Communication Networks and Security*. https://doi.org/10.47893/ijcns.2013.1083

[23] Mao, J., Bian, J., Tian, W., Zhu, S., Wei, T., Li, A., & Liang, Z. (2019). Phishing page detection via learning classifiers from page layout feature. *EURASIP Journal on Wireless Communications and Networking*, *2019*, 43. https://doi.org/10.1186/s13638-019-1361-0

[24] Naresh, U. (2013). Intelligent Phishing Website Detection and Prevention System by Using Link Guard Algorithm. *IOSR Journal of Computer Engineering*, *14*(3), 28–36. https://doi.org/10.9790/0661-1432836

[25] Ogbonnaya, m. (2020). *Cybercrime in Nigeria demands public- private action - ISS Africa*. https://issafrica.org/iss-today/cybercrime-in-nigeria- demands-public-private-action

[26] Sahingoz, O., Buber, E., Demir, Ö., & Diri, B. (2019). Machine learning based phishingdetection from URLs. *Expert Systems with Applications*, *117*, 345–357. https://doi.org/10.1016/J.ESWA.2018.09.029

[27] Serianu. (2017). *Nigeria Cyber Security Report 2017 Demystifying Africa`s CyberSecurity Poverty Line*. 1–80. https://www.serianu.com/downloads/NigeriaCyberSecurityReport2017.pdf

[28] Srinivasan, S., Vinayakumar, R., Arunachalam, A., Alazab, M., & Soman, K. (2021). DURLD: Malicious URL Detection Using Deep Learning-Based Character Level Representations. *Malware Analysis Using Artificial Intelligence and Deep Learning*,535–554. https://doi.org/10.1007/978-3-030-62582-5_21

[29] Ubing, A. A., Kamilia, S., Jasmi, B., Abdullah, A., Jhanjhi, N. Z., & Supramaniam, M. (2019). Phishing Website Detection: An Improved Accuracy through Feature Selection and Ensemble Learning. *IJACSA) International Journal of Advanced Computer Scienceand Applications*, *10*(1). www.ijacsa.thesai.org

[30] Vinayakumar, R., Soman, K. P., Prabaharan Poornachandran, Akarsh, S., & Elhoseny, M. (2019). Deep learning framework for cyber threat situational awareness based on emailand URL data analysis. *Advanced Sciences and Technologies for Security Applications*,87–124. https://doi.org/10.1007/978-3-030-16837-7_6

[31] Wenyin, L., Fang, N., Quan, X., Qiu, B., & Liu, G. (2010). Discovering phishing target basedon semantic link network. *Future Generation Computer Systems*. https://doi.org/10.1016/j.future.2009.07.012

Author's Brief Profile

Oguntunde, Bosede Oyenike holds a B.Tech (Hons) degree in Computer Engineering from LAUTECH, Ogbomosho, Nigeria. She obtained her M.Sc and PhD degree in Computer Science from the University of Ibadan, Nigeria. She is a Senior Lecturer in Department of Computer Science at the Redeemer's University, Ede, Nigeria. Her research interests are in Data Communication, Networking, Data mining and Machine Learning. She has published articles in learned journals both at local and international levels and presented papers at conferences. She can be reached on phone by +2347058938585 and E-mail oguntunden@run.edu.ng

Iwuh Chizor Samson holds BSc in Computer Science from the Redeemer's University, Ede. He is currently undergoing the mandatory National Youth Service Corps, his interests are in the area of cyber security and machine learning. He can be reached by phone on +2348142242545 and through E-mail iwuh55548905gb@run.edu.ng

Ojewumi Theresa Omolayo holds a B. Tech and PhD degree from LAUTECH, Ogbomosho, Oyo State, Nigeria. She got her MSc in Computer Science from University of Ibadan, Ibadan, Oyo State, Nigeria. Her research areas include Machine Learning, Data Science and Computational Biology. She can be reached by phone on +2348032751268 and through E-mail ojewumit@run.edu.ng.

Abolarinwa Michael Oluwagbenga has his BSc in Computer Science from University of Ilorin, Ilorin, Kwara State, his Msc in Computer Science from University of Ibadan, Ibadan, Oyo State, and PhD from Kwara State University, Malete, Kwara State all in Nigeria. His research interests include Mobile Agent Technology, Cyber Security, and Machine Learning. He can be reached by phone on +2348036749999 and through E-mail gbenga1abolarinwa@gmail.com.