Development of an Ensemble Classification Model Based on Hybrid Filter-Wrapper Feature Selection for Email Phishing Detection

R. B. Ibrahim, M. S. Argungu, I. M. Mungadi

Open Science Index, Computer and Systems Engineering Vol:17, No:9, 2023 publications.waset.org/10013250.pdf

Abstract**—**It is obvious in this present time, internet has become an indispensable part of human life since its inception. The Internet has provided diverse opportunities to make life so easy for human beings, through the adoption of various channels. Among these channels are email, internet banking, video conferencing, and the like. Email is one of the easiest means of communication hugely accepted among individuals and organizations globally. But over decades the security integrity of this platform has been challenged with malicious activities like Phishing. Email phishing is designed by phishers to fool the recipient into handing over sensitive personal information such as passwords, credit card numbers, account credentials, social security numbers, etc. This activity has caused a lot of financial damage to email users globally which has resulted in bankruptcy, sudden death of victims, and other health-related sicknesses. Although many methods have been proposed to detect email phishing, in this research, the results of multiple machine-learning methods for predicting email phishing have been compared with the use of filter-wrapper feature selection. It is worth noting that all three models performed substantially but one outperformed the other. The dataset used for these models is obtained from Kaggle online data repository, while three classifiers: decision tree, Naïve Bayes, and Logistic regression are ensemble (Bagging) respectively. Results from the study show that the Decision Tree (CART) bagging ensemble recorded the highest accuracy of 98.13% using PEF (Phishing Essential Features). This result further demonstrates the dependability of the proposed model.

Keywords**—**Ensemble, hybrid, filter-wrapper, phishing.

I. Introduction

EMAIL is a means of communication hugely accepted among individuals and organizations globally. Email is hugely accepted because of its unique attribute such as speed, effectiveness, inexpensive medium of sharing information over the network, and so on. The uniqueness that surrounded email has made it to rapid growth day by day in terms of the volumes of email messages sent and received over the internet [1]. The use of email continued to increase more than other interpersonal channels of communication. Reference [1] (2019) asserted that an average 269 billion of e-mails were sent per day during the first quarter of 2017. The adoption took another dimension during the COVID-19 lockdown era due to remote working scenarios for organizations, individuals, and government agencies. These huge messages sent via email platforms contained malicious Email also known as spam.

Spam can be defined as unsolicited messages usually sent to a large number of recipients. Spam can be in the form of advertisements, promotions, or similar explicit content which may contain malicious code embedded in them [8]. Despite the surplus of positive benefits that email offers, it is also answerable for security and privacy concerns. However, [9] pointed out some related catastrophes that surround email development such as spam email, email spoofing, email botnet, phishing email, and others.

Reference [2] identified that phishing email is recognized as a worrisome, dangerous, destructive, and deadly type of spam email. Phishing is further described as a malicious activity or means of social engineering spells often used to extract user data, including login credentials and credit card numbers. Spammers are populous, and the quantity of email phishing has become very high which exposes people to vulnerable to cybercrime practices. Phishing is purposely launched to harm users financially (cybercrime). Phishing is a worrisome attack that is mainly used or sent by the spammer to achieve two curious objectives: to make money from victims that respond to their emails, to glean sensitive information such as passwords, credit card numbers, bank account details, and so on and likewise to advertise malicious code [2]. Most of the cybercrime reported is usually carried out through email phishing (Yahoo). The dynamic nature of spam is the main reason why it is considered a difficult task. Spammers adopted obfuscation techniques to circumvent the spam filters. The methods employed by spammers to fool filters include; misplaced spaces, purposeful misspelling, embedded special characters, Unicode letters, transliteration, HTML redrawing, and so on. In addition, spammers used tokenization attacks to defeat the feature selection techniques by splitting and modifying the crucial message features [3]. Email phishing detection has received a lot of attention from researchers in the last decades.

Three different categories of approaches have been identified [4] in filtering phishing emails. These approaches include listed-based i.e. black and white lists approach, heuristic rule-based approach, and machine learning classification approach.

R. B. Ibrahim is with the Nigeria Customs Service, No. 2 Lake Taal Street, Maitama, Abuja (phone: +234 0806 458 5724; e-mail: bellorashida92@gmail.com).

M. S. Argungu is with the Department of Computer Science, Kebbi State University of Science and Technology Aliero, Kebbi State, Nigeria (phone: +2348100543186; e-mail: sm279arg@gmail.com)

I. M. Mungadi is with the Federal University Birnin Kebbi, and Argungu Emirate Schools, Argungu Kebbi State, Nigeria (phone: +234 8100887932; e-mail: ibrahimmusamgd@gmail.com).

In the black-and-white list approach, spammers' email accounts are blacklisted, and future email messages from the black-listed accounts are filtered as spam. All legitimate accounts are kept on the white list. The rule-based heuristic approach series of rules or patterns are stored in the Email system to track every message's probability of being spam. Every message that passed certain thresholds is considered spam and filtered out. The truth is that the traditional, conventional, and listed-based methods do not have the strength and power to detect current email phishing except the machine learning approach.

The machine learning classification approach is the state-of-the-art approach used for phishing email detection. The approach uses a sort of artificial intelligence to learn about previous messages categorized as phishing. The machine learning classification approach has witnessed tremendous application in email phishing filtering. Machine learning approaches are more efficient, a set of training data is used, and these samples are the set of emails that are pre-classified. Machine learning approaches have a lot of algorithms that can be used for email filtering. These algorithms include “Naïve Bayes, support vector machines, Neural Networks, K-nearest neighbour, Random Forests, etc.” State-of-the-art approaches in email phishing detection have focused on tackling the high dimensionality problem inherently present in the real word email phishing dataset and class imbalance problem with the majority employing ensemble classification models. Towards improving the performance accuracy of the email spam detection model, this study proposes a hybrid feature selection-based multi-level ensemble classification model for detecting email phishing messages.

Most data for Phishing are structured data for supervised learning approaches. However, these data are characterized by both relevant and redundant features. The direct use of these data can affect classification performance as well as introduce both time and space complexities. Hence, there is a need for feature selection or reduction. Email phishing detections are required to be scalable, lightweight, and effective since they are meant to function as a real-time system. Therefore, feature selection is needed to reduce the dimensionality of the phishing attack dataset to improve performance and speed.

Feature selection is a technique for selecting a subset of relevant features while the redundant ones are rejected. Reference [5] classified feature selection methods into three categories; filter, wrapper, and embedded. Each of the feature selection techniques has its merits and demerits while their performance relies on the type of dataset they are fed with. Filter models select features based on certain ranking criteria without considering the learning algorithm. The filter feature selection approach is seen to be efficient and fast, however, it may miss some relevant features that are relevant for the learning classifier to be used [6]. Wrapper feature selection models specifically use learning algorithms to evaluate and select features. While embedded feature selection combines the advantages and qualities of both filter and wrapper feature selection approaches. Embedded models are computationally inexpensive since they do not require running on a learning algorithm before they select their features. Nevertheless, feature selection approaches are not limited to the aforementioned methods as more dynamic selection approaches have been proposed. A typical example is the hybridization of the feature selection approach which at the time of this research has not been employed on email phishing datasets as it is in this research. Hence, this research will be applying a combination of filter and wrapper feature selection techniques with a machine-learning model for effective and efficient email phishing detection.

II. Methodology

A. Proposed System Architecture

The architecture for the proposed ensemble-based phishing detection model is presented in Fig. 1. The proposed architecture comprised five stages namely: Data collection/acquisition, Data pre-processing, Feature selection, Model Training, and Evaluation. Due to the nature of the dataset obtained, it is passed to the pre-processing stage where data normalization using the min-max method is applied to scale data values into 0 to 1 scale. Thereafter, the normalized data are fed into the feature selection stage. Under the feature selection stage, two feature selection algorithms were applied to select the optimal subset of features; i.e. first of all, the whole dataset was sent to filter (information gain), while its output is fed as input to the wrapper to select the final features. The SMOTE (Synthetic Minority Oversampling Technique) is applied to overcome the problem of class imbalance in the dataset. The selected features are passed to the model training stage where the ensemble learning method is employed to learn the patterns of the features and make a prediction when fed with a validation or test set. The evaluation stage shows the results of the tested model.

B. Data Collection/Description

The dataset used for this research was collected from the Kaggle repository database [7]. The mail text format has been prepared, converted, and readily available for classification purposes. The dataset contains records of 525,577 email phishing. This is the email phishing dataset with the highest records with 21 attributes and class status: Phishing (1) and Not Phishing (0). The dataset is in the comma-separated value (CVS) format, which makes it suitable for the experiment.

There are 525754 records in the dataset with all numeric data values. A sample of the phishing dataset which includes 15 instances, five column-wise attributes, and their corresponding phishing class is shown in Table I, and the attribute description is unveiled in Table I.

C. Dataset Pre-processing

Data pre-processing is a crucial task in the machine-learning environment, it ensures the maximum availability of the dataset for the experiment. During this step redundant, irrelevant, and variant features are treated to suit machine learning processes. Several pre-processing steps can be applied to machine learning tasks, depending on the available dataset. The characteristics of the dataset for this particular research required normalization, feature selection due to the value variation, and, over-sampling due to the class imbalance.

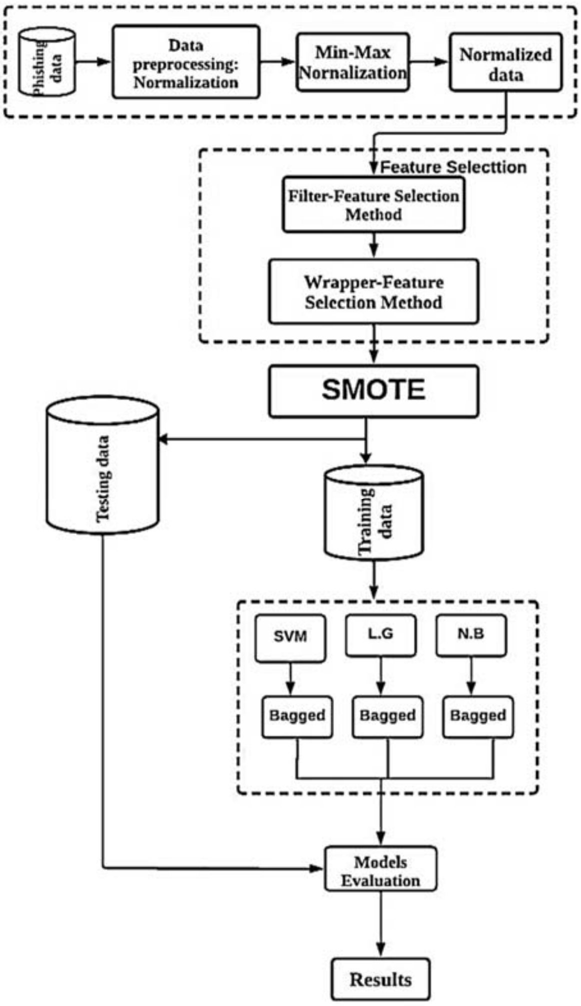

Fig. 1 Framework for the proposed system architecture

D.Data Normalization

The dataset collected required only three pre-processing stages as mentioned above. The first step is normalization, the goal is to normalize the skewed values present in data to fall between 0 and 1. Several methods are available to carry out the task, but for this research work, Min-Max normalization is used to transform the values between 0 and 1.

III. Class Balancing

Class balancing is the process of adjusting the skew class imbalance in the dataset. Class imbalance is a critical problem in machine learning, this is a situation where there is no equality in the data classes. A dataset with skewed class distribution is called imbalanced, in such a dataset there will be majority and minority classes. The classes with a larger proportion of the dataset are called the majority while the lesser ones are called the minority. The email phishing dataset provided for the experiment in this research is skewed and there is a need to balance the data before applying learning algorithms to them. Out of the 525754 total records in the dataset, 8,353 records are under "Class 1" while 517,401 are under "Class 0". This distribution shows that this particular dataset is having class imbalance, which is not good enough for classification tasks.

To overcome the problem posed by the class imbalance, SMOTE over-sampling technique is applied to modify the skew data class occurrences. SMOTE approach synthetically creates examples rather than over-sampling with duplication. SMOTE over-sampling approach is detailed in the following algorithm.

A. Algorithm SMOTE(T, N, K)

```
Input: Integer of splinter group class samples T; Amount of SMOTE N%; Integer of nearest neighbors k.
Output: (N/100) * T synthetic splinter group class samples
  (* If N is less than 100%, randomize the splinter group class samples as only a random percent of them will be SMOTEd. *)
  if N < 100
      then Randomize the T splinter group class samples
          T = (N/100)*
          N = 100
  endif
  N = (int) (N/100) (* The amount of SMOTE is assumed to be in integral multiples of 100. *)
  K = Number of nearest neighbors
  Numattrs = Integer of attributes
  Sample [ ] [ ]: array for original splinter group class samples
  new index: keeps a count of the number of synthetic samples generated, initiated to 0
  Synthetic [ ] [ ]: array for synthetic samples
(* Compute k nearest neighbors for each splinter group class sample only. *)
  for I←1 to T
          Compute k nearest neighbors for I, and save the indices in the nnarray
          Populate (N,i,nnarray)
  Endfor
  Populate (N,i,nnarray) (*Function to generate the synthetic samples.*)
  While N≠0
          Choose a random number between 1 and k, call it nn. This step chooses one of the k nearest neighbors of i.
          For attr←1 to numattrs
              Compute: dif=Sample[nnarray[nn]][attr]-Sample[i][attr]
              Compute: gap = random number between 0 and 1
              Synthetic[newindex][attr]=Sample[i][attr]+gap*dif
        endfor
        newindex++
      N = N – 1
      endwhile
     return (* End of Populate, *)
   End of Pseudo-Code.."
```

IV. Result and Discussion

A. Experimental Setup

For the ease of machine learning, and programming, and the fact that it is much easier to work with numbers, the varied values in the dataset were normalized to fall between 0 and 1 respectively. After both the training and testing sets were formatted into acceptable format, classification experiments were then carried out. To achieve the stated objectives of the work, the experimental setup was broken into various steps and the results obtained are illustrated in the forms of tables and charts.

The steps are predefined as follows:

Step1. Each of the two feature selection algorithms (Information Gain and Recursive Feature Selection) was used independently, and the outcome of information gain is passed for the wrapper to select the final features for the experiment. To identify relevant features among the initially identified variables in the data set collected from the Kaggle machine learning repository regarding Phishing, the two-feature selection algorithm was individually tested on the testing data set. The features selected by the Recursive feature selection algorithm were used to develop the predictive models for Phishing detection.

Step2. The bagging of each three base learners was used to generate the predictive models using a 10-fold cross-validation technique. The three base-level algorithms are Naïve Bayes, Decision tree (CART), and Logistic Regression.

Step3. Results from Step 1 to Step 2 are compared and the best ensemble learning classification model that gives the highest performance accuracy was selected.

B. Results of Feature Selection Methods

Following the process of identification, collection, and description of the dataset explaining the detection of phishing, the next important step was the selection of the most optimal features among the identified factors that will improve the prediction accuracy of phishing detection better. As earlier stated, two filter-based feature section methods were used in this work to identify the most relevant features in the dataset.

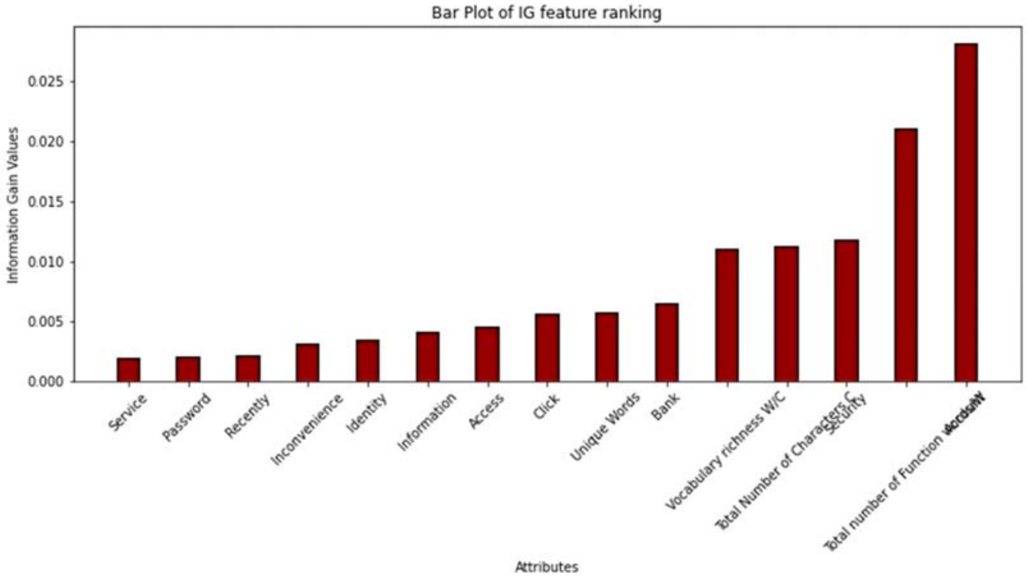

Fig. 2 The Selected 15 Ranked Attributes by the Information Gain

TABLE I
RANKED FEATURES BASED ON THE INFORMATION GAIN VALUES

Symbol	Features	Information Gain Values
1	Service	0.0018936042378105977
2	Password	0.0019977250649640954
3	Recently	0.002170042699291219
4	Inconvenience	0.0031195020605985224
5	Identity	0.003448731840018704
6	Information	0.0040978139962317695
7	Access	0.0045887054297472085
8	Click	0.005629425905341212
9	Unique words	0.005752465490833636
10	Bank	0.006450748475699375
11	Vocabulary richness W/C	0.011105733636406767
12	Total Number of Characters C	0.011323997391195495
13	Security	0.01186044761824645
14	Total number of Function words/W	0.02110117550031676
15	Account	0.028171424578432913

C. Result of Recursive Feature Selection Method (Wrapper)

The RFS algorithm is a wrapper technique that dependently selects email phishing attributes based using an algorithm, it was employed to select the optimal features from the 15 ranked features selected by the prior technique. Table I presents the final attributes selected for the experiment by the RFS.

After the selection of the PEF, the features are confronted with class imbalances. However, the class distribution was mentioned earlier, where 8,353 records are under "Class 1" while 517,401 are under "Class 0", as shown in the class distribution in Fig. 1 The occurrence of class imbalance in the data will prone the system to overfit, to relieve the machine learning algorithm from this challenge, a SMOTE technique was employed to augment for the lesser class.

D. Discussion of Results

The interpretation of mail contents is estimator (machine learning) dependent. Thus, the effort was to design a detective model that is capable of detecting email phishing to abate the

Open Science Index, Computer and Systems Engineering Vol:17, No:9, 2023 publications.waset.org/10013250.pdf

menace surrounding the usage of email services. Three experiments were carried and classification performance has been compared to determine optimal algorithms for detecting email phishing. The experiments were designed for two basic purposes; to investigate how feature selection affects the detection accuracy, and to compare ensemble bagging of three base classifiers; CART Decision tree, Logistic regression, and Naïve Bayes.

Comparing the results obtained with previous studies that use single feature selection or without feature selection [1] revealed that reducing the number of attributes has improved the classification accuracy. Feature engineering has played a major role in the models by increasing classification accuracy and decreasing model complexity by expunging irrelevant and redundant attributes from the dataset. And also, having a reduced population of attributes has the additional benefit of fast execution time.

E. Model Comparison

After performing the experiments, the next step was comparing the models and selecting the best model. The experiments were conducted on three setups: Bagging Decision tree, Bagging Naïve Bayes, and Bagging Logistic Regression. The models were compared using different performance measures like accuracy, Sensitivity (TP Rate), Specificity (TN Rate), Precision, F-Measure (F-1 score), and False Alarm Rate (FAR).

V. Conclusion

Email platforms throughout the world today are confronted with the problem of phishing. Email is one of the services supported by the Internet for sending bulky digital distribution. Interesting knowledge can be extracted from the email contents to enhance the security of emails. In this study, email phishing detection models based on machine learning are proposed with optimal features called PEF. PEF are related to email. The performance of email phishing detection models is evaluated by a set of classifiers, namely; Naïve Bayes, CART, and Logistic Regression. Consequently, an ensemble method is applied to improve the performance of the single classifiers. Bagging among the array of most frequently used ensemble methods as reported in different studies is employed. The accuracy of the email phishing detective model using PEF in the case of CART as a bagging ensemble achieved 98.13% detection accuracy.

References

[1] Abdulrahaman, M. D., Alhanssan, J. K., Adebayo, Oyeniyi, J. A., and Olalere, M. (2019). Phishing Attack Detection Based on Random Forest with Wrapper Feature Selection Method. International Journal of Information Processing and Communication (IJIPC) Vol. 7 No. 2, Pp. 209-224

[2] Aggarwal, S., Kumar, V., & Sudarsan, S. D. (2015). Identification and Detection of Phishing Emails using Natural Language Processing Techniques. In Proceedings of the 7th International Conference on Security of Information and Networks (Pp. 217-222).

[3] Ahmed, D. S., Hussein, K. Q and Allah, H. A (2022). Phishing Websites Detection Model based on Decision Tree Algorithm and Best Feature Selection Method. Turkish Journal of Computer and Mathematics Education. Vol.13 No. 01(2022), 100-107

[4] Akarsh, T. and Elhoseny, P. E (2019). Phishing Email Detection Based on Structural Properties. In NYS Cyber Security Conference, Pp. 1-7).

[5] Akinyelu, A. A, and Adewumi, A. O., (2014). Classification of Phishing Email Using Random Forest Machine Learning Technique. Journal of Applied Mathematics. Volume 2014, Article ID 425731, 6 pages http://dx.doi.org/10.1155/2014/425731

[6] Alauthman, j. K. (2020). A Framework for Big Data Analysis in Smart Cities. In: International Conference on Advanced Machine Learning Technologies and Applications. Springer, Cham, Pp. 405–414

[7] Al-Saaidah, K. J. (2017). Leveraging Machine Learning Techniques for Windows Ransomware Network Traffic Detection. In Cyber Threat Intelligence (pp. 93-106). Springer, Cham.

[8] Mohammad, S. M. (2020). Sentiment "Analysis of Mail and Books". Technical report, National Research Council Canada.

[9] Fariska, H. F. (2019). "Phishing Attacks: Information Flow and Chokepoints," in *Phishing and Countermeasures: Understanding the Increasing Problem of Electronic Identity Theft*, M. Jakobsson and S. Myers, Eds., pp. 31–64, John Wiley & Sons, New York, NY, USA.