Journal of Telecommunication, Electronic and Computer Engineering
ISSN: 2180 – 1843 e-ISSN: 2289-8131 Vol. 16 No. 3 (2024) 9-17
jtec.utem.edu.my
DOI: https://doi.org/10.54554/jtec.2024.16.03.002

Impact Analysis of Filter and Wrapper-Based Feature Selection Techniques for Webpages Phishing Attacks Identification

Patrick Olabisi[1*], Gabriel Ogunleye[2], Bamidele Olukoya[3] and Adekunle Osobukola[1]

[1]*Department of Electrical/Electronic & Telecommunication Engineering, Bells University of Technology, Ota, Nigeria.*
[2]*Department of Computer Science, School of Sciences, Federal University, Oye-Ekiti, Nigeria.*
[3]*Department of Computing and Information Science, Bamidele Olomilua University of Education, Science and Technology, Ikere-Ekiti, Nigeria.*

Article Info

Article history:
Received Jun 21st, 2024
Revised Aug 11th, 2024
Accepted Sep 2nd, 2024
Published Sep 30th, 2024

Index Terms:
Phishing detection
Features selection
Ensemble classifiers
Phishing datasets

Abstract

Phishing, which involves fraudulently gaining access to sensitive assets of unsuspecting individuals through deceptive and malicious emails, is a major global threat to internet users. The proliferation of phishing sites and their operations is occurring at an alarming rate, raising significant concerns about how to forestall them. Numerous research efforts are underway to detect phishing attempts before they can compromise important information and cause damage. Compared to conventional methods, machine learning has proven highly effective at detecting phishing attacks by analyzing different features. This study analyzed the behaviors of seven classification data mining algorithms on optimal subset features selection using Wrapper (Boruta) and Filter-based (Mutual-Information). Real-life phishing webpage datasets were used for the analysis. Ensemble classifiers such as Voting, Gradient Boosting, and Random Forest were used in the experiments. Two experiments were conducted. In the first experiment, K-Nearest Neighbor (K-NN) yielded the highest accuracy among single classifiers, with a score of 94.1%, while Random Forest (RF) ensemble achieved 96.7%. In the second experiment, using another baseline feature set, RF performed excellently under the Boruta method with an accuracy of 97.25%, while K-NN retained the highest accuracy of 95.20% among single classifiers. This study provides empirical evidence that feature selection techniques have a great impact on the performance of ML models, for both single and ensemble classifiers, in the detection of phishing attacks.

*Corresponding Author: poolabisi@bellsuniversity.edu.ng

I. INTRODUCTION

Phishing has become one of the most prevalent threats in cyberspace and network services, gaining significant momentum over recent decades [1]. This fraudulent practice is typically carried out using electronic devices such as doctored web pages (hyperlinks), emails, texts and social networks (social engineering), with the aim of stealing sensitive personal information from both the novices and experienced users alike [2]. Given that the internet connects millions of computers globally, phishing threats can spread rapidly, causing severe damage to victims across various regions. The current global situation regarding phishing is alarming, with security threats increasing exponentially due to the growing number of internet-connected devices [3].

Technology advancements have allowed attackers to easily facilitate cybercrime, making networks more vulnerable to phishing attacks through a variety of methods. It was reported in [4], [5] that phishing incidents escalated significantly during the COVID-19 pandemic in 2020, as more people turned to the internet for transactions and communication [6]. It is confirmed that in March 2020 alone, over 60,000 phishing websites were discovered [4], a figure far exceeding previous records. Furthermore, the Anti-Phishing Working Group reported a total of 51,041 distinct phishing sites during this period, while an analysis by RSA revealed that phishing incidents in 2022 caused global corporate financial damages amounting to approximately $9 billion [5]. These devastating attacks have severely impacted economies worldwide, with millions of dollars lost every day.

The rising threat of phishing has prompted governments and other sectors to urgently seek solutions. In response, researchers, popular browsers, and email service providers have employed a range of countermeasures, including conventional and machine learning (ML) approaches [7]. While conventional approaches were initially effective, they have become inadequate due to their inability to detect evolving phishing techniques. Phishing websites frequently change their content, have short operational lifespans, and exploit these changes to evade traditional anti-phishing defenses [8]. Besides, conventional systems often require consistent technical input from users, such as regular updates and handling obfuscation techniques, which can be time-consuming. These limitations make traditional methods ineffective against the constantly evolving phishing threats.

In contrast, ML-based approaches have proven to be more effective and are widely regarded by researchers as a promising solution for phishing detection [9]. Although ML-techniques have significantly improved phishing detection compared to earlier methods, no single has been able to permanently identify every phishing site [10]. Issues such as misclassification, high computational costs, and overfitting remain challenges for ML models. ML techniques typically involve three crucial steps, namely data collection and representation, preprocessing (feature selection and dimensionality reduction), and model evaluation. In these steps, patterns are identified, and the relationship between the training and testing sets of the dataset is mapped [11]. Among these steps, feature selection is particularly crucial, as it has been highlighted in the literature as a pivotal aspect of ML process. The ML constructively detects both new and old phishing attacks through analysis of different inherent features.

Effective feature selection is essential for phishing detection because it helps identify and retain the most relevant features in a dataset, especially in cases where data is highly dimensional. High dimensional data, with numerous features, can negatively impact the performance of ML classifiers if not properly handled [12]. Therefore, features selection techniques are needed to remove redundant, correlated, and irrelevant features, ensuring that the selected subset contains only information correlated with the target class and uncorrelated with other features [13].

Feature selection techniques play a critical role in enhancing model performance, reducing computational costs, minimizing variance, provided that informative and relevant features are selected [14]. The effectiveness of a machine learning-based phishing identification system is largely determined by the quality of features selected during the selection process. There are three main categories of feature selection techniques: filter, wrapper, and embedded methods. However, none of these can be universally regarded as the best. The filter method is known for its speed, low cost, and simplicity, using statistical measures for information theory to evaluate features. However, it is not as powerful as the wrapper method, which uses machine learning algorithms to identify the optimal features. Although the wrapper method is more accurate, it is also more time-consuming and resource-intensive compared to the filter method [15]. Many previous studies in this domain have focused on fine-tuning, enhancing, and modifying the applied machine learning algorithms, often neglecting the impact of feature selection. As a result, these studies sometimes expose the enhanced or modified models to low-quality training sets, reducing the model’s effectiveness.

It has been demonstrated that feature selection is the most important step in improving the detection accuracy of anti-phishing systems. To highlight the importance of feature selection in classical machine learning algorithms, this study compares two feature selection techniques: one from the filter category (Mutual Information) and one from the wrapper category (Boruta). This study examines the behavior of seven classification algorithms on phishing webpage datasets, utilizing the optimal feature subsets identify by these two methods. The experimental findings indicate that applying feature selection methods significantly improve the accuracy of traditional machine learning-based anti-phishing systems. Additionally, the number of the selected optimal features is comparatively lower, which results in reduced computational costs and processing time for pattern recognition tasks.

The remainder of this paper is structured as follows: Section II presents the comparative studies, Section III outlines the methodologies adopted, and Section IV presents an analysis of the experimental results. Finally, Section V concludes the paper and discusses potential directions for future research.

II. RELATED WORKS

Several studies have been published in this problem domain with some applying feature selection techniques, while others did not. The studies that omitted feature selection often viewed it as a waste of time, given that the number of features in earlier datasets was relatively small compared to current datasets that contain a much larger volume of features.

In this study, several phishing detection publications that employed recent ML techniques alongside feature selection were reviewed. For example, the work of [16] presents a flexible and responsive phishing detection system that employs case-based reasoning (CBR) to accurately detect phishing attempts. By integrating a hybrid methodology that combines Information Gain and Genetic Algorithms, their system, CBR-PDS, improves detection accuracy while reducing processing time. This represents a notable advancement in combating phishing, as it emphasizes improved result reliability through meticulous feature selection. Similarly, the research conducted by [17] addresses the growing challenges of cybersecurity in today's digital landscape. The researchers introduce a novel approach for identifying malicious webpages, highlighting the critical role of dataset features in improving detection efficacy. Their findings make an important contribution to continuous efforts to safeguard users from online threats.

Likewise, [7] proposes a novel feature selection approach for phishing detection, utilizing filter techniques to generate initial feature subsets, which are then refined through a data perturbation ensemble. The study aims to enhance the efficacy of machine learning techniques in detecting phishing threats, thereby contributing to a safer digital environment. In contrast, [4] performed performance analysis of various classical and ensemble ML algorithms for phishing detection, without utilizing feature selection. The focus was on observing the behavior of these classifiers in the context of phishing attack classification, based on a dataset containing 11,054 records and 31 features. The study used logistic regression, voting, decision, and random forest algorithms, recoding the performances of 0.86, 0.91, 0.87, and 0.91, respectively. However, the performance of the classical classifiers was undermined due to the noise present in the dataset. The low detection rates of these classical algorithms were attributed to the noise in the training sets.

The study conducted by a team of researchers in [15], investigated the effectiveness of machine learning models designed for identifying phishing URLs, a crucial component of cybersecurity. The findings are anticipated to deepen our understanding of how optimizing these models can significantly improve their ability to detect malicious links, ultimately fostering a safer online environment. The research utilized the selectKbest filtering method, which employs various scoring techniques, including correlations and mutual information gain, to select a targeted subset of features from the overall dataset. The study demonstrated that optimizing machine learning models through careful feature selection and tuning techniques markedly increases their ability to detect phishing URLs, thereby bolstering cybersecurity initiatives. However, [13] explored the performance and feature selection in machine learning models by applying filter-based methods to three different datasets, with the goal of identifying optimal features. The researchers used intersection functions across the dataset and identified nine key features. The C5.0 decision tree classifier was then evaluated on these selected features, with results showing that the model performed better using these nine features compared to other variables with lower values. Nonetheless, the study has limitations, as it only considered and generalized results based on a single filter-based feature selection method.

In [15] three feature selection methods were applied, namely Information Gain, Chi-Square and CFS to select important features for classification data mining algorithms. They identified a drastic reduction in feature importance between the 20th and the 21st features when using Information Gain and Chi-Square techniques. Despite the reduced feature set, the results showed that the detection accuracy of the classifiers remained stable. In [18], a novel approach to spam email filtering was investigated to highlight the prominent role of feature selection in improving the classification accuracy of ML algorithms. The study explored the effectiveness of different feature selection measures, yielding remarkable outcomes with both RF and SVM classifiers. A total of twelve essential feature selection methods were utilized to extract significant attributes from different email categories, thereby enhancing spam detection capabilities. The performance of RF and SVM classifiers was assessed on the optimal features obtained from each technique. Notably, the study recorded an overall F1- score of 0.978 using RF classifier with a streamlined feature set, showcasing its effectiveness in spam detection. This success was attributed to the classifier's ability to identify and utilize informative features.

The review work in [19] focuses on the importance of feature selection methods in medical training sets, particularly due to the challenges associated with obtaining optimal subsets of relevant features while minimizing complexity. This article highlights the needs for increased awareness of effective feature selection techniques and also identifies the limitations of current methods, thereby paving the way for future research in this vital area. Existing feature selection methods often face a trade-off between classification accuracy and the number of features used. The paper points out that many existing methods rely on univariate ranking, which fails to account for interactions between variables. This can lead to selecting too many features, which introduces noise and decreases classification accuracy, or selecting too few features, which may result in the loss of important information. Moreover, while some methods achieve good accuracy, they often use a larger number of features, which complicates the modeling task and reduces interpretability of the model. The challenge remains to develop a universal feature selection method that can provide optimal classification accuracy with fewer features. This area remains an open field for research.

As stated in [8], it is neither advisable nor practical to rely solely on filter-based statistical techniques for optimal features. This is because statistical techniques alone do not have the capacity to critically analyze the information of websites features.

III. MATERIALS AND METHODS

The primary material required for this study is the phishing dataset and the necessary system hardware devices. The methodology is discussed in the following sub-sections:

A. Dataset and Process of Collection

The phishing website dataset used in this study was crawled by Chiew et al. [5] and made available for researchers. It can be downloaded from the Kaggle ML community under the title " Phishing Dataset for Machine Learning | Kaggle." The dataset consists of 48 features extracted from 5,000 phishing web pages and 5,000 legitimate web pages, which were downloaded over two years. The phishing web pages were collected from Phish-Tank and Open-Phish, while legitimate web pages were collected from Alexa and Common Crawl.

The dataset's features were classified into three categories: Address Bar-based, Abnormal-based, and HTML/JavaScript-based feature. The address bar in web browsers is a versatile tool that offers several features beyond merely entering web addresses. It also makes the address bar a powerful and multifunctional tool for enhancing the browsing experience. These features include search engine interpretation. It helps to search for results of every query. In addition, Abnormal-based features play a crucial role in various fields by providing an additional layer of security and monitoring, helping to detect and respond to potential issues before they become critical problems. For instance, User Behavior Analytics (UBA) deals with analyzing user behavior to detect anomalies that could indicate malicious activity, such as insider threats. Similarly, HTML and JavaScript are fundamental technologies for creating web applications. They offer various features that enhance the functionality, interactivity, and user experience of web pages.

The dataset consists of 10,000 records, with all features represented as numeric values. Both the training and test datasets are provided in comma-separated value (CSV) format, where each row represents a phishing or legitimate instance, and the columns represent the features describing those instances. The detailed characteristics of the dataset are presented in Table 1.

Table 1
Data Observation/Characteristics

Observation	Value
Missing values	No
Input features	Numeric
Target Class	Binary
No of records	10,000
No of attributes	48

B. Exploratory Data Analysis (EDA)

This scientific method is commonly applied to investigate datasets to harvest the key characteristics of such dataset. EDA was conducted on the phishing dataset, and it was discovered that many of the feature values are skewed. As shown in Figure 1a, features such as NumDots, UrlLength exhibit skewed distributions, with values ranging from 1 to 20 and 12 to 253, respectively. This skewness poses a challenge for the ML algorithms, as significant variation in the training set can lead to problems in model performance. Specifically, skewness refers to the asymmetry in the distribution of data values, which can be either positive (right-skewed) or negative (left-skewed), as shown in Figure 1b. This could lead to issues such as overfitting and negatively impact the model's prediction accuracy. To mitigate the effects of skewness, data scientists often apply normalization techniques to the training set to enhance the robustness and predictive capabilities of machine learning models. The min-max rescaling technique was applied in this study. This technique was chosen because it ensures that all features are treated equally in terms of scale, which can enhance the performance of various machine learning algorithms. The formula for min-max rescaling is shown in equation (1).

$$R' = \frac{Z - V_{min}}{V_{max} - V_{min}} \qquad (1)$$

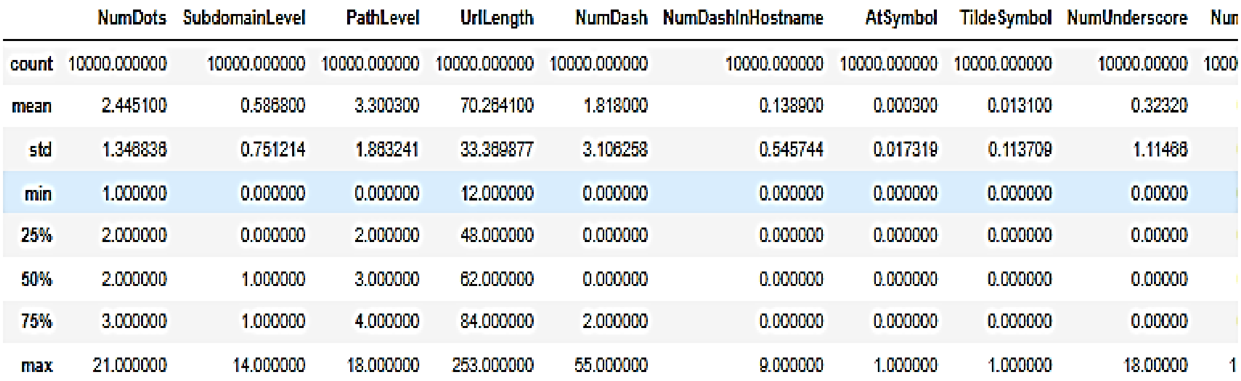
where: R'= the new normalized value of each feature data in the form of 0 and 1
Z = the old value of each attribute data
V_{min}= the minimum absolute value of Z
V_{max}= the maximum absolute value of Z

After applying the rescaling method, the phishing website data was normalized between 0 and 1. It is essential for the researchers to detect and remove features that are highly correlated with one another. This was accomplished using a Pearson Correlation matrix, which is a widely used technique in data mining for removing redundant features. The formular for the Pearson correlation is shown in equation (2). The 2-D heatmap feature correlation is presented in Figure 2.

$$\emptyset(x, y) = \frac{cov(x, y)}{\sigma x \sigma y} \qquad (2)$$

where: x = the phishing independent variables,
y = the phishing target class,
cov(x, y) = the covariance of x and y,
σx and σy = the standard deviation of x and y.

The Pearson correlation matrix generated five highly correlated features: HostnameLength, NumAmpersand, NumNumericChars, PathLength, and QueryLength. These features were removed from the original dataset to avoid redundancy. Two experiments were then conducted using different baseline feature selections from two distinct variable selection techniques, allowing for a comparative analysis of their performance.



	NumDots	SubdomainLevel	PathLevel	UrlLength	NumDash	NumDashInHostname	AtSymbol	TildeSymbol	NumUnderscore	Nun
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	1000
mean	2.445100	0.586800	3.300300	70.264100	1.818000	0.138900	0.000300	0.013100	0.32320	
std	1.346836	0.751214	1.863241	33.369877	3.106258	0.545744	0.017319	0.113709	1.11468	
min	1.000000	0.000000	0.000000	12.000000	0.000000	0.000000	0.000000	0.000000	0.00000	
25%	2.000000	0.000000	2.000000	48.000000	0.000000	0.000000	0.000000	0.000000	0.00000	
50%	2.000000	1.000000	3.000000	62.000000	0.000000	0.000000	0.000000	0.000000	0.00000	
75%	3.000000	1.000000	4.000000	84.000000	2.000000	0.000000	0.000000	0.000000	0.00000	
max	21.000000	14.000000	18.000000	253.000000	55.000000	9.000000	1.000000	1.000000	18.00000	1

Figure 1(a). Summary of EDA Phishing webpage Dataset

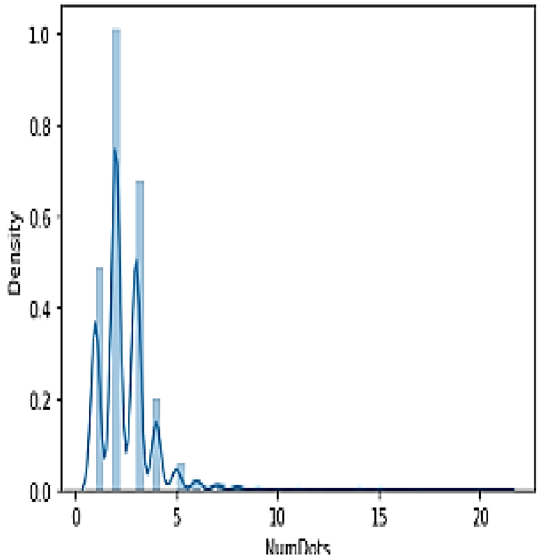


Figure 1(b). Plot of EDA Phishing Webpage Dataset

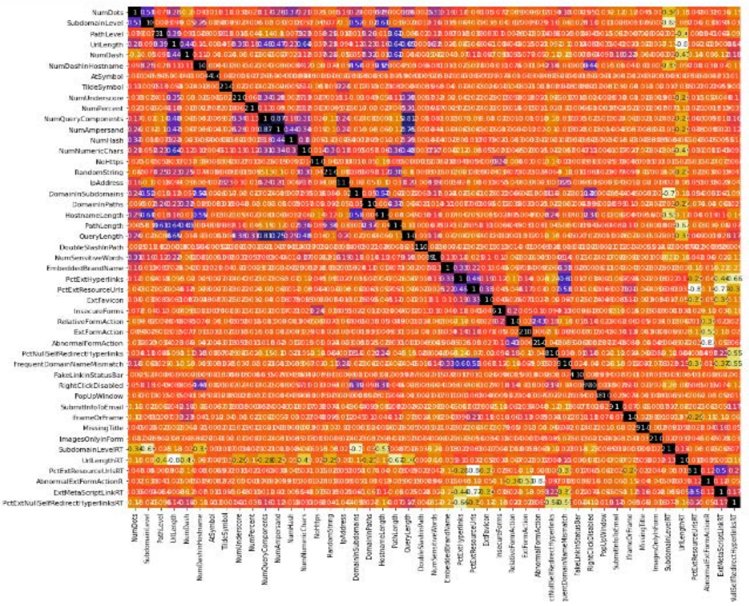


Figure 2. A 2-D Phishing correlation Matrix.

C. Boruta Feature Selection Algorithm

Boruta is a well-known wrapper feature selection method that utilizes the RF classifier as its foundation. It leverages RF as the base learner due to its relatively fast training process and the fact that it does not require parameter tuning. The relevance of features is determined by measuring the decrease in classification accuracy that results from the randomly shuffling feature values among instances. This assessment is conducted for each tree in the forest that utilizes specific features for classification.

Additionally, the Z-score is employed to evaluate the significance of variance as shown in equation (3).

$$Z = \frac{\alpha - \emptyset}{\theta} \qquad (3)$$

where α is the phishing training instances,
$\emptyset$ represent the average, and
θ stand for standard deviation of the accuracy loss.

However, relying solely on the Z-score is inadequate for identifying meaningful correlations between features and the target variables. Additional references are required to distinguish between truly important features and irrelevant ones. To address this, Boruta enhances the information system by introducing features that are randomly generated. For each original feature, Boruta creates a corresponding 'shadow' attribute, which is generated by randomly shuffling the values of the original feature across instances. Classification is then performed using all attributes from this augmented system, including the original and shadow features. The importance of each attribute is calculated, and the shadow attributes' importance values, which should only reflect random variations, serve as baseline. Only those original features with importance scores exceeding those of shadow attributes are deemed important. This process is repeated for a set number of iterations or until all features are either classified as important or rejected , depending on which occurs first. In situations where neither outcome is achieved, some attributes remain unclassified and are referred to as undetermined. Table 2 displays the results of the Boruta algorithm applied to the phishing training datasets.

Table 2
Boruta Algorithm Generated Features

Feature	Status	Features	Status
NumDots	1	AbnormalExtFormAct-ionR	6
IframeOrFrame	1	RelativeFormAction	8
SubmitInfoToEmail	1	IpAddress	9
FrequentDomainNameMismatch	1	EmbeddedBrandName	10
PctNullSelfRedirectHyperlinks	1	RandomString	11
InsecureForms	1	NumPercent	12
ExtFavicon	1	MissingTitle	13
PctExtResourceUrls	1	DomainInPaths	14
ExtMetaScriptLinkRT	1	DomainInSubdomains	14
NumSensitiveWords	1	AbnormalFormAction	16
PctExtHyperlinks	1	ExtFormAction	17
PctExtNullSelfRedirectHyperlinksRT	1	NoHttps	18
NumDash	1	SubdomainLevelRT	19
NumQueryComponents	1	RightClickDisabled	20
UrlLength	1	ImagesOnlyInForm	21
PathLevel	1	TildeSymbol	22
NumDashInHostname	2	NumHash	23
PctExtResourceUrlsRT	3	PopUpWindow	24
NumUnderscore	4	FakeLinkInStatusBar	25
SubdomainLevel	5	DoubleSlashInPath	25
UrlLengthRT	6	AtSymbol	27

The steps of the Boruta process are outlined below:

Input: Phishing dataset
Output:

1. Phishing feature with higher Z-score.
2. Extend the duplicate (shadow attributes) of all phishing features
3. Shuffle and permutate the original attribute to remove feature correlations.
4. Apply a random forest classifier on the extended dataset and permutated phishing information to calculate the Z-scores for phishing attributes.
5. Identify the maximum Z-score among the shadow attributes, and designate a hit to each phishing attributes that performs better than the maximum Z-score among shadow attributes (MZSA).
6. Perform a two-sided quality test with the MZSA for each attribute that is undermined.

7. Phishing attributes with lower importance than MZSA are identified and removed from the information system.
8. Phishing attributes with significantly higher importance than MZSA are selected and labeled as ‘important’.
9. Remove all shadow phishing attributes from the information system.
10. Repeat the process iteratively until all phishing attributes are classified as important or rejected.

End

D. *Mutual Information Feature Selection*

Mutual information is a filter-based statistical tool applied to select an optimal subset of features for this study. Mutual information statistical tool measures the relationship between random variables that are shuffled simultaneously, providing insight into the information shared between attributes.

According to information theory, two variables are considered statistically independent when their mutual information is 0. The mutual information of two random variables Q and J, with a joint distribution defined by P(Q, J) is given by equation (4):

$$I(Q;J) = \sum_{x\in X}\sum_{y\in y} P(q,j) log \frac{P(q,j)}{P(q)P(y)} \quad (4)$$

where: $P(Q)$ and $P(\mathrm{Y})$ = the marginal distributions of Q and J, respectively, obtained through marginalization process.

The baseline features that scored above the threshold are selected and presented, as shown in Figure 3.

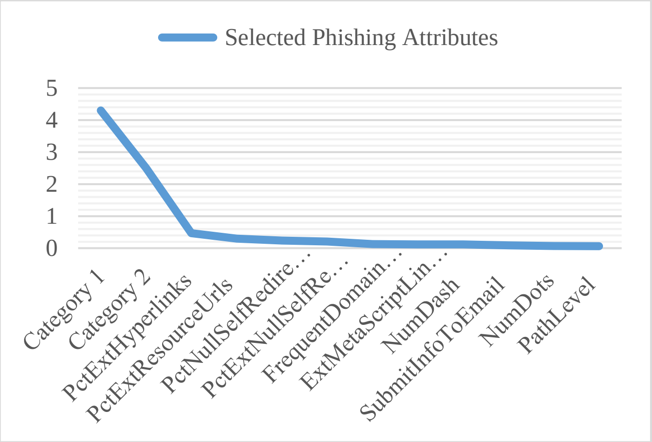


Figure 3: Phishing features selected through mutual information tool

E. *Phishing Classification based on Selected Algorithm*

The ML-based classifiers used in this study are categorized into two groups: shallow (single) and ensemble classifiers. These classifiers are described in this sub-section based on their behaviors.

1) Shallow (single) ML Classification Algorithms

The empirical analysis in this work utilizes several shallow classifiers, including Logistic Regression (LR), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), and K-Nearest Neighbor. These classifiers are known as supervised ML learners and are commonly used for classification and regression tasks [20]. These four classifiers have demonstrated consistent performance in the phishing detection domain.

2) Ensemble (combined) ML Classifiers

Ensemble learning classifiers, also known as combined classifiers, work by combining multiple base (single model) estimators into a unified model that generally improves the performance [21]. This approach forms the basis of ensemble classifiers, often known as “strong classifiers,” while the individual base models are known as “weak learners.” Ensemble approach aims to develop models with better predictive performance [22]. Examples of ensemble classifiers include Gradient Boosting (GB), Voting, Random Forest, Bagging, AdaBoost and Stacking. Gradient Boosting classifier is a boosting-based classifier, while Random Forest and Voting are based on bagging techniques. The Gradient Boosting algorithm combines multiple weak learners to create strong learners, thereby enhancing the model’s overall performance. The new models are trained to minimize the loss functions such as mean squared error or cross-entropy, using gradient descent to optimize performance [3].

At the phishing classification stage, the pre-processed dataset (after applying min-max scaling and feature selection techniques) is used as input to the classifiers, one at a time, and their behavioral performance is recorded. Tables 3 and 4 shows the classification results obtained from the two feature selection methods.

IV. RESULTS AND DISCUSSION

The experiments were conducted in a Python environment, which was chosen due to its extensive libraries that support machine learning. The aim was to evaluate the impact of feature selection on shallow and ensemble algorithms without tuning their parameters. The classification evaluation metrics used include True Positive, True Negative, False Positive, False Negative, accuracy, precision, recall, and F1-Score. These metrics were computed using the following equations:

$$\textbf{Accuracy} = \frac{MP\ + CN}{MP + CN + KP + XP} \quad (5)$$

$$\textbf{Precision} = \frac{MP}{MP + KP} \quad (6)$$

$$\textbf{Recall} = \frac{MP}{MP + XP} \quad (7)$$

$$\textbf{F1 - Score (FS)}: = \frac{2\ \times Prec. + Recall}{Prec. + Recall} \quad (8)$$

where: MP = the true positive value
CN = the true negative value
KP = the false positive value
XP = the false negative value

The experiments were conducted on a laptop with the following specifications: Intel(R) Core (TM) i5-7200U CPU @ 2.50GH 2.70 GHz Laptop, CPU, 16.0 GB RAM, and Windows 10 Pro (64-bit operating system).

A. *Discussion of Results*

The data exploration revealed that the input features were numerical, while the target class was categorical. Min-max scaling was used to normalize the skewed values, and the two feature selection techniques (Mutual Information and Boruta Algorithm) were applied to the dataset. The experimental analysis involved the use of seven ML algorithms, programmed in a Python 3.7 environment. The dataset was divided into training and test sets in an 80:20 ratio, with a larger portion allocated to training to enable the models to learn from the data effectively. The results of the phishing classification experiments, based on the evaluation metrics, are presented in Tables 3 and 4. The performance of each model was compared using the baseline features. Table 3 summarizes the experimental results based on the baseline features shown in Figure 3. The analysis revealed that KNN yielded the highest accuracy of 94.1% among the single classifiers, while the RF ensemble classifier obtained 96.7%. The results for all the ML algorithms, including both shallow and ensemble classifiers' performance on Mutual Information Features, are shown in Figure. 4.

Table 3
Performances of the Selected ML Algorithms from Experiment 1: Shallow and Ensemble classifier's behaviours on Mutual Information Features

Parameters	Single Classifier				Ensemble Classifier		
	LR	SVM	LDA	KNN	GB	Simple Voting	RF
TP	814	855	883	915	900	932	957
TN	915	930	876	967	969	917	977
FP	89	74	128	37	35	87	27
FN	122	141	113	81	96	64	39
ACCURACY	89.5	89.25	87.95	94.1	93.45	92.45	96.7
PRECISION	0.906	0.920	0.873	0.961	0.963	0.915	0.973
RECALL	0.876	0.858	0.887	0.919	0.904	0.936	0.961
FI-SCORE	0.892	0.888	0.880	0.939	0.932	0.925	0.967
PRECISION	0.937	0.934	0.929	0.954	0.958	0.939	0.970
RECALL	0.913	0.912	0.899	0.949	0.913	0.959	0.975
FI-SCORE	0.925	0.923	0.914	0.952	0.935	0.949	0.973

Following the result from experiment 2, which focused on the baseline features presented in Table 2, the performance results shown in Table 4 demonstrate that the RF classifier performed exceptionally well under the Boruta Algorithm, achieving an accuracy of 97.25%. This performance surpassed that of the Mutual Information method, using the same training and testing distributions. Additionally, it was observed that the performances of the single classifiers improved significantly when using the Boruta-selected features. Although KNN still retained the highest accuracy among the single classifiers, with a recorded accuracy of 95.20%.

Table 4
Performances of the Selected ML Algorithms from Experiment 1: Shallow and Ensemble classifier's behaviors on Boruta Algorithm

Parameters	Single Classifier				Ensemble Classifier		
	LR	SVM	LDA	KNN	GB	Simple Voting	RF
TP	909	908	895	945	909	955	971
TN	943	940	986	959	964	942	974
FP	61	64	68	45	40	62	30
FN	87	88	101	51	87	41	25
ACCURACY	92.60	92.4	91.55	95.20	93.65	94.85	97.25

The results for all the ML algorithms for both shallow and ensemble classifiers using Boruta Algorithm-selected features, are shown in Figure 5. Overall, this study provides empirical evidence that feature selection techniques have great impact on the performance of ML models. This holds true for both on single and ensemble classifiers in the context of phishing detection.

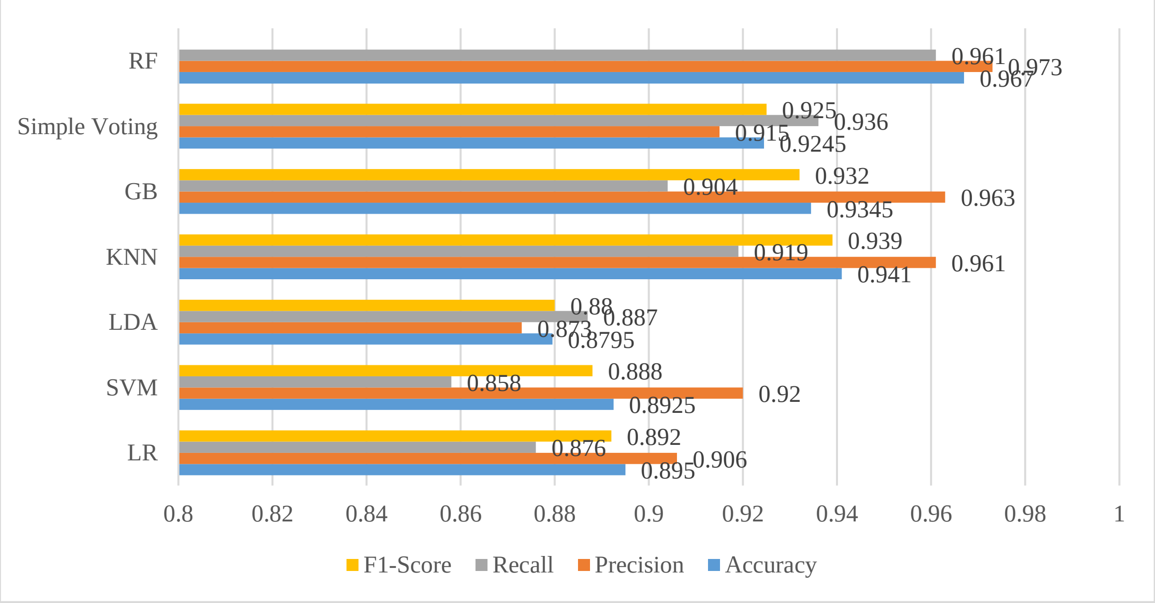


Figure 4. Shallow and Ensemble classifier's behaviours on Mutual Information Features.

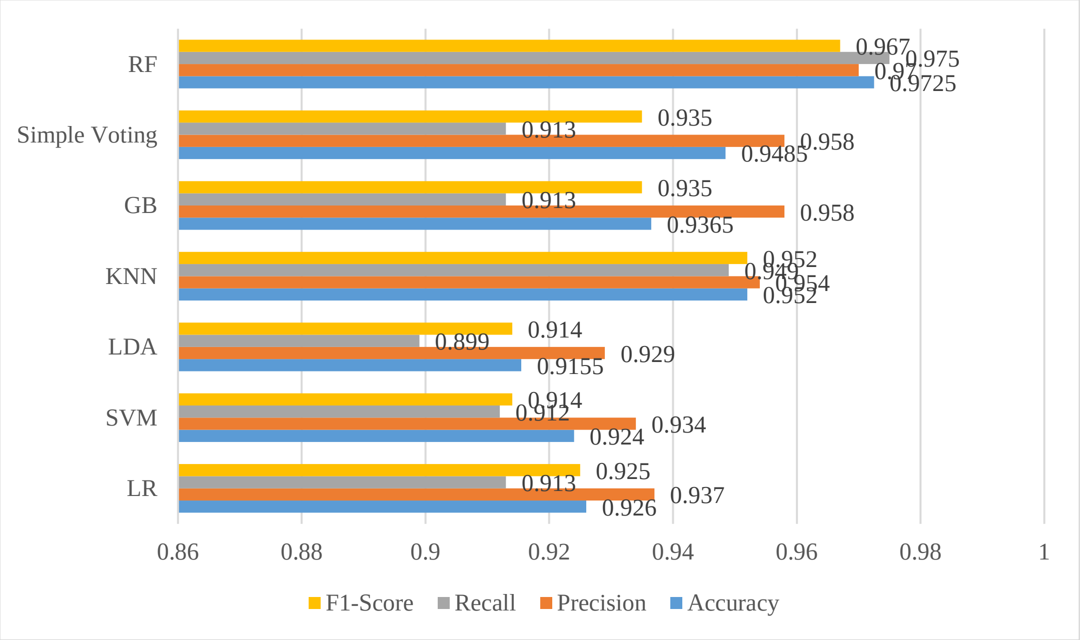


Figure 5: Shallow and Ensemble classifier's behaviors on Boruta Algorithm Features.

V. CONCLUSION

The research aimed to explore how the feature selection phase can assist in addressing cybersecurity challenges, particularly in phishing detection, through a machine learning approach. Identifying informative features that signify phishing is crucial for developing effective ML-based anti-phishing solutions. This is especially important given the ever-evolving nature of websites, which frequently introduce new features to enhance their resilience. The study generated phishing-related cyber-attack features using two distinct feature selection methods: the filter (Mutual Information) and the wrapper (Boruta Algorithm). Both methods influenced the quality of data input into the phishing detection models. The selected ML algorithms were trained and evaluated using the chosen phishing dataset, and the results revealed that the feature selection methods yielded different sets of features. This suggests that relying on a single feature selection technique is insufficient for identifying the most important webpage features for phishing detection.

Additionally, the findings showed that the algorithms performed better with features selected by the Boruta Algorithm compared to those identified by the Mutual Information method. The Boruta method identified features that not only improved detection accuracy but also reduce computational costs and processing time in detecting phishing attacks.

It can be concluded that feature selection is a valuable and effective step for enhancing the performance of ML-based phishing detection systems. Integrating the Boruta Algorithm with classifiers, particularly Random Forest, could serve as an effective framework for developing or improving anti-phishing tools for major browsers, email service providers, and governmental and corporate organizations. This would enable quicker and more efficient responses to phishing threats.

Future research should focus on utilizing more real-world data to validate the findings of this study. Emphasis should be placed on employing cross-validation techniques rather than the hold-out method for training algorithms. Additionally, exploring other feature selection methods and feature extraction techniques that can consolidate features into a new, reduced set will be beneficial. Finally, further efforts should be made to optimize the algorithm, with the goal of realizing real-time phishing detection performance.

CONFLICT OF INTEREST

Authors declare that there is no conflict of interests regarding the publication of the paper.

AUTHOR CONTRIBUTION

The authors confirm contribution to the paper as follows: study conception and design: Bamidele Olukoya, Patrick Olabisi; data collection: Gabriel Ogunleye; analysis and interpretation of findings: Bamidele Olukoya, Adekunle Osobukola; draft manuscript preparation: Patrick Olabisi, Adekunle Osobukola. All authors had reviewed the findings and approved the final manuscript.

REFERENCES

[1] O. Ogunleye, B. M. Olukoya, A. T. Olusesi, P. Olabisi, Q. B. Sodipo, and A. Osobukola, “Heterogeneous ensemble feature selection and multilevel ensemble approach to machine learning phishing attack detection,” FUOYE Journal of Engineering and Technology, vol. 8, no. 4, pp. 438–447, 2023.

[2] G. Alshammari, M. Alshammari, T. S. Almurayziq, A. Alshammari, and M. Alsaffar, “Hybrid phishing detection based on automated feature selection using the chaotic dragonfly algorithm,” Electronics, vol. 12, no. 13, 2023, doi: 10.3390/electronics12132823.

[3] R. Vinayakumar, K. P. Soman, Prabaharan Poornachandran, S. Akarsh, and M. Elhoseny, “Deep learning framework for cyber threat situational awareness based on email and URL data analysis,” in Advanced Sciences and Technologies for Security Applications, Springer, 2019, pp. 87–124, doi: 10.1007/978-3-030-16837-7_6.

[4] A.M. Oyelakin, O. M. Alimi, I. O. Mustapha, and I. K. Ajiboye, “Analysis of single and ensemble machine learning classifiers for phishing attacks detection,” Int. J. Softw. Eng. Comput. Syst., vol. 7, no. 2, pp. 44–49, 2021, doi: 10.15282/ijsecs.7.2.2021.5.0088.

[5] G. O Ogunleye, B. A. Ayogu, O. Tolulope, and P. O. Ogunrekun, “Comparative analysis of some machine learning techniques for uniform resource locator-based phishing detection,” Journal of Engineering, Technology and Innovation, vol. 3, no. 2, pp. 1–12, 2024.

[6] H. M. Farghaly, A. A. Ali, and T. A. El-Hafeez, “Building an effective and accurate associative classifier based on support vector machine,” Sylwan, vol. 164, no. 3, 2020.

[7] K. L. Chiew, C. L. Tan, K. S. Wong, K. S. C. Yong, and W. K. Tiong, "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system," Inf. Sci. (Ny)., vol. 484, pp. 153–166, 2019, doi: 10.1016/j.ins.2019.01.064.

[8] J. Zhou, H. Cui, X. Li, W. Yang, and X. Wu, "A novel phishing website detection model based on lightgbm and domain name features," Symmetry (Basel)., vol. 15, no. 1, 2023, doi: 10.3390/sym15010180.

[9] J. Tanimu, S. Shiaeles, and M. Adda, "A comparative analysis of feature eliminator methods to improve machine learning phishing detection," J. Data Sci. Intell. Syst., vol. 00, no. 00, pp. 1–13, 2023, doi: 10.47852/bonviewjdsis32021736.

[10] N. A. Azeez, S. Misra, I. A. Margaret, L. Fernandez-Sanz, and S. M. Abdulhamid, "Adopting automated whitelist approach for detecting phishing attacks," Comput. Secur., vol. 108, 2021, doi: 10.1016/j.cose.2021.102328.

[11] A. Zamir et al., "Phishing web site detection using diverse machine learning algorithms," Electron. Libr., vol. 38, no. 1, pp. 65–80, 2020, doi: 10.1108/EL-05-2019-0118.

[12] G. H. Lokesh and G. BoreGowda, "Phishing website detection based on effective machine learning approach," J. Cyber Secur. Technol., vol. 5, no. 1, pp. 1–14, 2021, doi: 10.1080/23742917.2020.1813396.

[13] E. Hokijuliandy, H. Napitupulu, and Firdaniza, "Application of svm and chi-square feature selection for sentiment analysis of Indonesia's national health insurance mobile application," Mathematics, vol. 11, no. 17, 2023, doi: 10.3390/math11173765.

[14] H. F. Atlam and O. Oluwatimilehin, "Business email compromise phishing detection based on machine learning: a systematic literature review," Electronics, vol. 12, no. 1, 2023, doi: 10.3390/electronics12010042.

[15] S. R. A. Samad et al., "Analysis of the performance impact of fine-tuned machine learning model for phishing url detection," Electronics, vol. 12, no. 7, 2023, doi: 10.3390/electronics12071642.

[16] H. Abutair, A. Belghith, and S. AlAhmadi, "CBR-PDS: a case-based reasoning phishing detection system," J. Ambient Intell. Humaniz. Comput., vol. 10, no. 7, pp. 2593–2606, 2019, doi: 10.1007/s12652-018-0736-0.

[17] S. S. Shin, S. G. Ji, and S. S. Hong, "A heterogeneous machine learning ensemble framework for malicious webpage detection," Appl. Sci., vol. 12, no. 23, 2022, doi: 10.3390/app122312070.

[18] J. Thomas, P. Vinod, and N. S. Raj, "Towards spam mail detection using robust feature evaluated with feature selection techniques," Int. J. Eng. Technol., vol. 6, no. 5, pp. 2144–2158, 2014.

[19] M. W. Mwadulo, "A review on feature selection methods for classification tasks," International Journal of Computer Applications Technology and Research, vol. 5, pp. 395-402, 2016.

[20] V. Mhaske-Dhamdhere and S. Vanjale, "A novel approach for phishing emails real time classification using k-means algorithm," International Journal of Engineering & Technology, vol. 7, no. 12, pp. 96-100, 2018.

[21] T. O. Omotehinwa and D. O. Oyewola, "Hyperparameter optimization of ensemble models for spam email detection," Appl. Sci., vol. 13, no. 3, 2023, doi: 10.3390/app13031971.

[22] N. Noureldien and S. Mohmoud, "The efficiency of aggregation methods in ensemble filter feature selection models," Trans. Mach. Learn. Artif. Intell., vol. 9, no. 4, pp. 39–51, 2021, doi: 10.14738/tmlai.94.10101.