

Received 1 November 2022, accepted 21 November 2022, date of publication 1 December 2022, date of current version 24 April 2023.

Digital Object Identifier 10.1109/ACCESS.2022.3225971

## RESEARCH ARTICLE

# Performance Evaluation of Phishing Classification Techniques on Various Data Sources and Schemes

RAHMAD ABDILLAH<sup>1</sup>, ZARINA SHUKUR<sup>1</sup>, MASNIZAH MOHD<sup>1</sup>,  
T. S. MOHD ZAMRI MURAH<sup>1</sup>, INSU OH<sup>2</sup>, AND KANGBIN YIM<sup>3</sup>

<sup>1</sup>Center for Cyber Security (CYBER), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi 43600, Malaysia

<sup>2</sup>Department of Information Security Engineering, Soonchunhyang University, Asan 31538, South Korea

<sup>3</sup>Department of Software Convergence Engineering, Soonchunhyang University, Asan 31538, South Korea

Corresponding author: Kangbin Yim (yim@sch.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) Grant by the Korean Government [Ministry of Science and ICT (MSIT)] under Grant NRF-2021R1A4A2001810, and in part by the Soonchunhyang University Research Fund.

**ABSTRACT** Phishing attacks have become a perilous threat in recent years, which has led to numerous studies to determine the classification technique that best detects these attacks. Several studies have made comparisons using only specific datasets and techniques without including the most crucial aspect, which is the performance evaluation of data changes. Hence, classification techniques cannot be generalized if they only use specific datasets and techniques. Therefore, this research determined the performance of classification techniques on changing data through a subset of schemes in a dataset. It was conducted using unbalanced and balanced phishing datasets, as well as subset schemes in ratios of 90:10, 80:20, 70:30, and 60:40. The thirteen most recent classification techniques used in preliminary phishing studies were compared and evaluated against ten performance measures. The results showed that the proposed schemes successfully uncover the maximum and minimum performance obtained by a classification technique. These comparisons can provide deeper insights into phishing classification techniques than related research.

**INDEX TERMS** Benchmark testing, classification algorithms, performance evaluation, phishing.

## I. INTRODUCTION

Phishing is a perilous threat to cybersecurity and according to The National Institute of Standards and Technology, it is attempts to get sensitive data, such as bank account numbers, or access to larger computerized systems by sending fraudulent requests through emails or websites. On average, the chances of being exposed to this attack in various sectors is 11% [1]. Phishing is also a socially engineered attack that tends to inflict physical or psychological harm on individuals and organizations [2]. The corporate sectors include technology, energy or utilities, retail, and financial services. These organizations are highly vulnerable to phishing. Therefore, cyber security-based measures are needed to prevent these attacks [3].

Several studies have been carried out on phishing prevention, one based on its identification and classification.

The associate editor coordinating the review of this manuscript and approving it for publication was Vlad Diaconita<sup>1</sup>.

Various techniques are used for the classification process, such as Random forest [4], [5], [6], [7], [8], [9], [10], support vector machine (SVM) [11], [12], [13], [14], Logistic regression [15], [16], [17], Multilayer perceptron (MLP) [18], C4.5 [19] and [20], and Naïve Bayes [21]. Each exhibits maximum performance according to the case it was applied. The results of the classification technique need not be generalized in all cases. Therefore, a comparative research must be carried out to resolve this gap.

However, only few studies have compared phishing classification techniques, such as [8], [18], [22], [23], and [24]. This comparative research is generally divided into four main parts, including phishing, the type of dataset, performance evaluation, and the techniques used. The data sources used by [8], [18], [22], [23], and [24] were obtained from a phishing website and URL, while [24] used raw emails sourced from Apache SpamAssassin and Nazario. The dominant performance evaluations are accuracy, precision, and F-measure. Random forest, SVM, and Naïve Bayes are the most widely

used techniques. This comparative research has a gap, which is how the existing techniques affect various public datasets, including the balanced and unbalanced ones.

Interestingly, this research is based on the performance evaluation of the classification technique when using a specific unbalanced dataset for certain phishing types. This is similar to the processes adopted by studies that did not compare these classification techniques. Vaitkevicius and Marcinkevicius [18] used two balanced and one unbalanced datasets. It was reported that they obtained better results than previous comparisons. Gana and Abdulhamid [23] only used unbalanced public datasets, and it was proven that the classification performance changes in accordance with its subset scheme.

This research is engineered by several studies that failed to prove how performance evaluation influences the techniques used to classify various subsets of dataset schemes. Some only described the limited impact of this performance on commonly used schemes, such as 90:10, 80:20, 70:30 and 60:40. Furthermore, performance evaluation and classification techniques are limited by the following measures, such as accuracy, F-Measure, Precision, True Positive Rate (TPR), Receiver Operating Characteristic (ROC), False Positive Rate (FPR), Precision-Recall Curve (PRC), Matthews Correlation Coefficient (MCC), Balanced Detection Rate (BDR), and Geometric Mean (G-Mean). It has been proven that each schema subset in both the balanced and unbalanced datasets affects the performance evaluation of the classification technique. This tends to significantly increase and decrease the performances of various subsets.

This research adopted three public datasets, namely, MDP-2018, UCI Phishing website, and Spambase. MDP 2018 is a balanced dataset, whereas the UCI Phishing website and Spambase datasets are unbalanced. The distribution of features in each dataset are as follows: MDP-2018, UCI Spambase, and Phishing website have 48, 58 and 31 features, respectively. In addition, thirteen of the most frequently used classification techniques, namely, Random forest, SVM, Logistic regression, MLP, C4.5, Bayesian Network, REP-Tree, Naïve Bayes, P.A.R.T, ABET (AdaBoost.M1 and Extra trees), ROFET (Rotation Forest and Extra trees), BET (Bagging and Extra-trees) and LBET (LogitBoost and Extra trees), were adopted. A subset scheme was established to ensure quality classification techniques were employed.

The subset scheme was derived from a proportion of each phishing and legitimate data class. This research utilized the 90:10, 80:20, 70:30, and 60:40 subset schemes, which were also applied to the legitimate and phishing data. For example, the UCI Phishing website dataset comprises 6157 phishing and 4898 legitimate websites, at a subset of 90:10 subset simply implies 90% of phishing and 10% of legitimate websites. The subset scheme was designed to match the actual conditions, and similar results were obtained from the experiment carried out, which was applied later. To ensure that the resulting classification model is excellent and reliable, a 10-fold cross-validation approach was adopted.

Relying only on accuracy as a performance evaluation measure is not advisable [18], [24]. This led to the use of ten performance evaluation measures, namely accuracy, F-measure, precision, TPR, ROC, FPR, PRC, BDR, MCC and G-Mean. Finally, a classification technique that excelled in all these tests was discovered.

This research focuses on a comprehensive performance evaluation of the technique used to classify various subset schemes and datasets. The following are the research contributions realized:

1. Performance evaluation using a subset scheme of 90:10, 80:20, 70:30, and 60:40 against several popular and recent classification techniques.
2. Identify the maximum and minimum performances of the 90:10, 80:20, 70:30, and 60:40 subset schemes.
3. Determine the decrease and increase in the performances of the subset scheme 90:10, 80:20, 70:30, and 60:40.
4. Identify the classification technique superior to all the subset schemes 90:10, 80:20, 70:30, and 60:40.

The remaining part of this research is organized as follows: Section II is a literature review on comparative phishing classification techniques. Section III describes the experimental methodology used. Conversely, the results were analysed in Section IV and conclusions were drawn in Section V.

## II. RELATED WORKS

Comparative research on phishing classification techniques is indispensable to determine the most appropriate procedure. Irrespective of the fact that several preliminary studies have been carried out, there are still gaps. One of such issues is the impact of balanced and unbalanced datasets and subset schemes on classification techniques. Therefore, there is a need to carry out comparative research that can resolve this gap. Generally, the most recent analysis comprises four main parts, namely, phishing, dataset type, performance evaluation, and the adopted techniques. This research creates opportunities for one to gain deeper insights into phishing detection.

The studies carried out by [18], [22], and [23] compared phishing websites' classification techniques as well as analysed its impact [24] on phishing emails and [8] on URLs. These were tested on an unbalanced dataset, however, only Vaitkevicius and Marcinkevicius [18] added a balanced dataset to their experiment. Karabatak and Mustafa [22] and Gana and Abdulhamid [23] used the Phishing website dataset from UCI machine learning, Vaitkevicius and Marcinkevicius [18] employed the UCI-2015, UCI-2016, and MDP-2018 datasets, Gangavarapu et al. [24] applied the SpamAssassin and Phishing Corpus dataset. While, Sahingoz et al. [8] used the Phishtank dataset and the Crawling URL results.

The numbers of classification techniques used in comparative research are stated as follows: 17 [23], 13 [22], eight [18], [24], and seven [8]. The comparison between Vaitkevicius and Marcinkevicius [18] shows that MLP, random forest, gradient tree boosting, and AdaBoost techniques were effective. Gana and Abdulhamid [23] also obtained

similar results that random forest was exceptional. On the contrary, Karabatak and Mustafa [22] stated that MLP, JRip, P.A.R.T., J48, random forest, and tree were ineffective. This was because the selected dataset features affected the performance of the classification technique. According to Karabatak and Mustafa [22], BayesNet, SGD, lazy.KStar, R.F.Classifier, LMT, and ID3 have the best performances. This contradicts the results obtained by Vaitkevicius and Marcinkevicius [18]. The naïve Bayes classification technique was ineffective in the experiments by Vaitkevicius and Marcinkevicius [18]. It was presumed that the difference in schemes can affect the performance of the classification technique. This simply indicates that sometimes, its performance is good, whereas, in some other circumstances, it tends to reduce [22]. It is also evidenced by the research carried out by Karabatak and Mustafa [22] that if random forest is not used on a dataset with feature selection, it turns out to be the most superior technique among all others used in the experiment. Special investigations are required to explore this gap further.

The use of features when evaluating classification performance on multiple datasets is also another point to consider. Gangavarapu et al. [24] used a feature extraction technique against a raw email. This generated 40 features without compromising the information contained in the raw email. The feature extraction technique was also employed by Sahingoz et al. [8] when constructing new datasets from URLs acquired from PhishTank. The website was crawled using a search engine with specified keywords. Sahingoz et al. [8] obtained a large number of features, relatively 102, besides some techniques were employed to obtain the optimal ones. References [18], [22], and [23] adopted an entirely different procedure than Sahingoz et al. [8] and Gangavarapu et al. [24]. They used the dataset as a medium to test their proposed method rather than the feature extraction techniques. Gana and Abdulhamid [23] and Karabatak and Mustafa [22] used a similar dataset with 31 features. However, Karabatak and Mustafa [22] evaluated feature reduction on classification performance. The dissimilarity discovered in the studies carried out by [8], [18], and [22] is the use of varying datasets and features. Vaitkevicius and Marcinkevicius [18] utilized the UCI 2015, UCI 2016 and MDP 2018 datasets with 30, nine, and 48 features, respectively. The varying dataset and features provide in-depth insights into the performance of the proposed classification techniques.

Comparative studies on classification techniques employed varying performance evaluation measures, such as, four [8], 10 [23], seven [24], and one [18], [22]. The more the performance measures, the more insights will be gained from these classification techniques. Gana and Abdulhamid [23] reported that random forest excels in all performance measures, namely, accuracy, precision, recall, F-measure, Area Under ROC Curve (AUC/AUROC), kappa statistics, root-mean-squared error, True Positive Rate (TPR), False Positive Rate (FPR), and root-relative-squared error. It is also effective in all performance evaluations (precision, sensitivity,

F-measure, and accuracy), especially when the natural language processing feature is used [8]. Random forest performance evaluation using accuracy, precision, recall, F1-measure, Matthews correlation coefficient (MCC), AUROC, and area under the precision-recall curve (AUPRC), was functional in the experiment [24]. The experiment conducted by Vaitkevicius and Marcinkevicius [18] and Karabatak and Mustafa [22] also exhibited good accuracy performance. However, not all classification techniques reported are effective in the diverse evaluation measures, especially the random forest. Gana and Abdulhamid [23], and Gangavarapu et al. [24], reported that it excels in all performance evaluations of the defined schemes.

Several recent studies are similar to the research carried out by Priya et al. [25], Indrasiri et al. [26], Ozcan et al. [27], Bu and Kim [28], Zeng et al. [29], and Aassal et al. [17], which evaluated the performance of classification techniques and their impact on various datasets. However, these were limited to phishing websites, in contrast to this research, which involved email and website phishing. Various datasets were evaluated to ensure that the proposed technique or method is known for its performance. Diverse studies employed different performance evaluations, such as Priya et al. [25] used TPR, MCC, Recall, Precision, f-measurements, Indrasiri et al. [26] adopted Precision, Accuracy, F1-Score, Recall, Ozcan et al. [27], utilized TPR, FPR, Precision, Accuracy, F1-Score, Bu and Kim [28], only used Accuracy and Recall. El Aassal et al. [17], and Zeng et al. [29], utilized performance accuracy, precision, recall, F1-Score, Geometric Mean, Balanced Detection Rate, Area Under Curve and Matthew's Correlation Coefficient. Meanwhile, this research used Accuracy, F-Measure, Precision, TPR, ROC, FPR, PRC, MCC to obtain a detailed performance evaluation of the classification technique.

However, certain studies have successfully described the proposed technique's performance, while others, such as Indrasiri et al. [26], and Bu and Kim [28] evaluated the impact of feature selection on various datasets. It reduces the dimensions of the dataset because there is a process of selecting the relevant key features in each category [30]. Indrasiri et al. [26] and Bu and Kim [28] used cross-validation to ascertain the authenticity of the model generated from the dataset that had undergone the feature selection process. Bu and Kim [28] also used the same method as the one obtained from feature extraction. Ozcan et al. [27], used cross-validation directly on the model formulated from the proposed technique and the dataset.

Ozcan et al. [27] evaluated the parameters to obtain maximum performance against the proposed technique. The studies carried out by Priya et al. [25], Indrasiri et al. [26], and Bu and Kim [28] were centred on improving its performance. Ozcan et al. [27] failed to state the performance of the proposed technique before and after the parameters were evaluated. Therefore, when parameter evaluation is employed, the process of increasing or reducing its performance is unknown. More detailed information is needed, such

as the performance before and after the proposed method was applied.

Generally, the experiments carried out in this research are similar to those of Indrasiri et al. [26] namely, comparing the performance before and after using various parameters as well as analyzing the proposed technique. Indrasiri et al. [26], evaluated the performance of accuracy and ROC\_AUC on models with various cross-validation values, such as 10, 20, 30, 40 and 50. Hyper-parameter tuning and feature selection were carried out to boost the performance of the proposed technique.

However, the difference between this research and that conducted by Indrasiri et al. [26], lies in using datasets, proposed techniques, performance evaluation, data retrieval, and subset schemas. This research adopted three datasets, namely MDP-2018, UCI Phishing Website, and UCI Spambase. Indrasiri et al. [26], performed feature selection and hyper-parameter tuning to obtain the maximum performance, while this research used a subset of schemes such as 90% Phishing:10% Legitimate, 80% Phishing:20% Legitimate, 70% Phishing:30% Legitimate, 60% Phishing:40% Legitimate, and 50% Phishing:50% Legitimate (balance), 90% Legitimate:10% Phishing, 80% Legitimate:20% Phishing, 70% Legitimate:30% Phishing, 60% Legitimate:40% Phishing, and 50% Legitimate:50% Phishing (balance) to find maximum performance. It simply shows that if phishing data is distributed more than the legitimate ones or vice versa, then there is a need to pay attention to the impact of the resulting performance. This research contributed to the performance evaluation by altering the data distribution, thereby significantly affecting its increase or decrease. It aids future studies to better understand data distribution, enabling them to perform hyper-parameter tuning to get maximum performance in detecting phishing attacks [26].

Several studies, such as El Aassal et al. [17] and Zeng et al. [29], adopted a similar concept. They used PhisBench, whereas this research used Weka to test the classification techniques against the proposed one, which involves reducing the datasets sourced from both websites and emails to 75%, 50% and 25% of their original sizes. Meanwhile, this research is based on the comparison between (subset scheme) 90:10, 80:20, 70:30 and 60:40. For example, the comparison at 90:10 simply implies 90% and 10% of the data are from phishing and legitimate, respectively, data from the MDP-2018 dataset, UCI Phishing website and UCI Spambase. The order of these datasets, such as 90% Legitimate and 10% Phishing, were also compared to ascertain the Engineering performance under various conditions.

El Aassal et al. [17], and Zeng et al. [29], reduce data from the dataset regardless of the performance quality of the discarded ones. On the contrary, this research evaluated the performance of unused data. It is believed that any of them can potentially affect the detection of a phishing attack. El Aassal et al. [17], and Zeng et al. [29], adopted various legitimate and phishing data sources, such as Enron, Wikileaks, Nazario, Bluefin, SpamAssassin, PhishTank,

OpenPhish, Alexa, DMOZ and Yahoo Directory. They also generated several new datasets due to the complexity of their sources. The use of the development made by Zeng et al. [29], as a data source does not allow comparisons to be made with related studies. Similarly, El Aassal et al. [17], also encountered certain problems building models from a combination of their datasets and was only able to report the comparison made with the results of related studies.

The apparent difference between the research carried out by El Aassal et al. [17], and Zeng et al. [29], and the present one is the adoption of standardized public datasets and commonly used performances. This present research employed datasets from the UCI Phishing website and Spambase, including MDP-2018. Several studies widely used these to test the performance of the proposed technique. To compare standardized datasets to performance. The techniques proposed by Alsariera et al. [31], namely ABET (AdaBoost.M1 and Extra trees), ROFET (Rotation Forest and Extra trees), BET (Bagging and Extra-trees) and LBET (LogitBoost and Extra trees) were tested for the subset schema of the dataset. Therefore, the present research explained the performance when the proposed subset scheme is used.

Some studies reported that any unbalanced dataset needs to be balanced because it is bound to affect performance [32]. Therefore, this analysis proves how unbalanced data is converted into balance and vice versa. The research aims to determine the extent the performance of phishing attack detection techniques increases or decreases.

Therefore, it is crucial to uncover gaps that have not been resolved by previous studies [8], [18], [22], [23], [24]. Vaitkevicius and Marcinkevicius [16] and Karabatak and Mustafa [20] used accuracy to evaluate the performance of classification techniques, which only verifies its ability to classify the acquired data. There is a need for more performance evaluation measures to get better insights. This includes the use of four [8], 10 [23], and seven measures [24]. Although, these are limited to the use of unbalanced datasets, thereby causing the classification techniques' performance on the balanced datasets to remain unknown. Coincidentally, how do the various subset schemes employed by Gana and Abdulhamid [23] affect phishing classification techniques?

### III. METHODOLOGY

This section describes the experimental research methodology, selection of datasets, subset schemes, classification techniques, and performance evaluation.

Several studies used public datasets as the benchmarks for the proposed technique. A variety of metrics were used to measure its performance. However, the evaluation performance is limited to the use of the entire dataset. It was further stated that the proposed technique results are better than those dependent on the dataset.

Some studies also employed additional techniques such as feature selection to improve the performance of the proposed one. This only focuses on improving technical performance, while the features in the dataset provide a solid



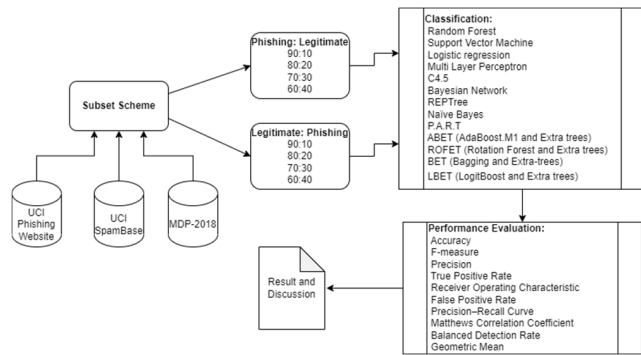


FIGURE 1. Experimental design.

relationship. Its function significantly affects performance, especially in terms of detecting phishing attacks. These features are adjustable, especially the ones generated from the extraction technique, and their importance tends to differ from the various studies. Therefore, the role of public and standardized datasets serves as a bridge to measure the performance of the proposed technique. The use of both standard and public datasets makes it easier for one to compare the proposed technique.

Therefore, this research evaluates the dataset's quality, openness, difference, and evaluation matrix. Its quality is evaluated by dividing each of the acquired data into a subset scheme, namely 90:10, 80:20, 70:30 and 60:40. This also includes the conversion of the unbalanced dataset into a balanced one, and its performance is generated. The openness of the dataset makes it easier to obtain maximum results. Assuming the dataset used is private, obstacles are bound to be encountered compared to the proposed technique. The UCI Spambase and Phishing Website, including the MDP-2018 datasets, were selected because several studies tend to use them to test the performance of the proposed technique. Innumerable matrices such as accuracy, TPR, precision, F-measure, FPR, PRC, ROC, BDR, MCC and G-Mean, were also used for performance evaluation.

### A. SELECTION OF DATASET

Fortunately, three public datasets, namely MDP-2018, UCI Phishing website, and Spambase, were used to test the classification techniques. The UCI Phishing website and Spambase datasets have an imbalanced class distribution, whereas that of the MDP-2018 is balanced. It [33] comprises 5000 phishing and legitimate websites, respectively. The MDP-2018, has 48 features, while the UCI Spambase comprises 58 features with distributed records, namely, 2,788 legitimate and 1,813 phishing emails. The UCI Phishing website comprises 31 features with distributed records of 6,157 phishing and 4,898 legitimate websites.

### B. PROPOSED SUBSET SCHEMES

The proposed subset schemes were established by dividing the acquired datasets by the available ones, thereby obtaining

the following sizes 90:10, 80:20, 70:30, and 60:40. Furthermore, the under-sampling technique was used to generate the schema subsets, including balancing the unbalanced datasets such as UCI Spambase and Phishing websites. This procedure reduces the sample to a specific size [34], a subset scheme. For example, in a balanced dataset (MDP-2018) with a total of 5000 phishing and legitimate records, the 90:10 subset scheme that was established comprises 90% of phishing and 10% of legitimate records. The under-sampling technique is used because it is free from overfitting problems as experienced by oversampling because oversampling duplicates data in minority data classes [35]. The present research tested how the subset scheme was constituted of 90% and 10% legitimate and phishing records, respectively. This is also applicable to balanced and unbalanced datasets. A cross-validation approach was employed to ensure that the resulting model is of high quality and to avoid overfitting the subset schemes [13]. It is also used to ensure that the performance of the classification technique is reliable [36]. And, The performance evaluation uses the Iteration value against the cross-validation technique, which is 100. This follows the recommendations from [17] and [19] to obtain maximum performance results accurately. The experimental setting of the subset scheme is shown in Table 1.

The experiments conducted by Gana and Abdulhamid [23] showed a change in the performance of the classification technique when the 70% data-taking test scheme was utilized. This led to the proposition of different subset schemes and datasets used to prove that schema changes affect the classification techniques' performance.

### C. SELECTION OF PHISHING CLASSIFICATION TECHNIQUES

Meanwhile, 9 of the most recent classification techniques employed by preliminary studies on phishing detection include Random forest, SVM, Logistic regression, MLP, C4.5, Bayesian Network, REPTree, Naive Bayes, and P.A.R.T. Random forests performed exceptionally, as reported on phishing websites [4], [5], [6], [7], URLs [8], and tweets [9]. It is also effective on balanced (websites) and unbalanced datasets (websites, URLs, and tweets). SVM is exceptional in terms of detecting phishing webpages [11] and URLs [12], [13], [14]. It is also used to detect phishing with balanced (and unbalanced (URLs) datasets. Logistic regression exhibits the best performance on both balanced (webpages) and unbalanced (websites and webpages) datasets [15], [16], [17]. MLP is only exceptional on an unbalanced dataset of phishing websites [18]. C4.5 performs best when a phishing webpage balanced and URL unbalanced datasets are used [19], [20]. Similarly, Naive Bayes exhibits its best performance on a phishing SMS unbalanced dataset [21].

This research further employed some classification techniques that are rarely used. These include the Bayesian Network [37], decision tree, and P.A.R.T. The essence is to prove that these less frequently used techniques are effective.

**TABLE 1. Subset schemes configuration used in this experiment.**

Dataset	Number of records	Feature	Distribution records in the subset schemes (Phishing:Legitimate)				Distribution records in the subset schemes (Legitimate:Phishing)			
			90:10	80:20	70:30	60:40	90:10	80:20	70:30	60:40
MDP-2018	10000	48	4500:500	4000:1000	3500:1500	3000:2000	4500:500	4000:1000	3500:1500	3000:2000
UCI Spambase	4601	58	1632:279	1450:558	1269:837	1088:1115	2510:181	2230:363	1952:544	1673:725
UCI Phishing Website	11055	31	5541:490	4925:980	4310:1469	3694:1959	4408:616	3918:1231	3429:1847	2939:2463

Furthermore, these were implemented using the Waikato Environment for Knowledge Analysis (Weka) version 3.8.4 with default parameters [38].

#### D. EVALUATION METHOD

The five most frequently employed performance evaluation procedures, such as accuracy, F-measure, precision, TPR, and ROC, were used to carry out the experiments. FPR and PRC, were included to obtain more information regarding the classification performance. Moreover, PRC displays precision and recall information in different probabilities. It is better than ROC and tends to measure the performance of classification techniques in a dataset with an imbalanced class distribution [39]. Based on these evaluations, the best and worst techniques were selected during each experiment. Then, the ones that performed best in all evaluations were ranked. The seven most widely used measures in the phishing classification technique's performance evaluation includes accuracy, TPR, precision, F-measure, FPR, and PRC and ROC, was adopted. PRC is the precision value for the corresponding sensitivity (recall) [39], while ROC involves the plotting of TPR against FPR using various threshold settings [40].

Furthermore, several evaluation performances such as Geometric Mean (G-Mean), Balanced Detection Rate (BDR), and Matthew's Correlation Coefficient (MCC) were used to compare recent research. These tend to add insight to the conducted experiments. G-Mean is the geometric mean of True Negative Rate (TNR) and Recall [17]. BDR Performance is a Metric used to measure the number of correctly classified minority class instances and to appropriately penalize them [41]. The MCC considers both positive and negative or false values, generally regarded as an unbalanced procedure that can be used even if the classes have diverse measures [25].

The following is the performance formula for BDR, MCC and G-Mean:

$$BDR = \frac{TP}{1 + FP} \quad (1)$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TN + FN)(P)(N)}} \quad (2)$$

$$G - Mean = \sqrt{\frac{TP}{TP + FN} * \frac{TN}{TN + FP}} \quad (3)$$

#### IV. EXPERIMENTAL RESULT AND DISCUSSION

The results of the experiment carried out based on the methodology are presented in this section.

First, the datasets were selected and classified as described in the previous section. Then, the 90:10, 80:20, 70:30, and 60:40 schemes were generated. Furthermore, the previously balanced dataset was made unbalanced and vice versa. This was realized by adjusting the number of records for the smallest class. The essence is to show how this dataset schema affects the performance of the classification technique. Afterward, it was tested on the scheme and dataset using Weka. A 10-fold cross-validation procedure was adopted to ensure that the model generated by this classification technique remains profitable.

Next, a training and testing session was performed on all these datasets, which had been schematically assigned. It was executed using seven classification techniques, namely, accuracy, F-measure, precision, TPR, ROC, FPR, and PRC. Table 2 shows that random forest and P.A.R.T. are the most favoured approaches in the MDP-2018 dataset, UCI Phishing website, and Spambase. Random forest performance is better in the MDP-2018 dataset than in the UCI Phishing website and Spambase. It has an accuracy, F-measure, precision, TPR, and ROC of 98.37%, 0.984, 0.984, 0.984, and 0.999, respectively in the MDP-2018 dataset. Meanwhile, Naïve Bayes did not get the best performance in every evaluation, except for ROC. The lowest ROC lies in the SVM, 0.939, 0.936 and 0.891 for the MDP-2018, UCI Phishing website, and Spambase datasets, respectively.

Afterwards, experiments were conducted on the UCI website Phishing and Spambase datasets. As is well known, these two datasets have an imbalanced number of data classes. Therefore, the data on the most prominent class was adjusted to suit the smallest one. The data generated are on the UCI Phishing website to 4898 and 1812 for each class of legitimate and phishing websites as well as legitimate and phishing emails on the UCI Spambase.

However, when Tables 3 and 2 were compared, it was discovered that the random forest's performance had changed significantly. On the imbalanced UCI Phishing website, its performance had an initial accuracy, F-measure, precision, TPR and ROC of 97.259%, 0.973, 0.973, 0.973, and 0.996, which were altered to 97.396%, 0.974, 0.974, 0.974, and 0.996, respectively. It is interesting that only the ROC

**TABLE 2. The results of performance evaluation with model validation using ten cross-validations on an imbalanced and balanced dataset.**

Classifier	MDP-2018-Balanced Dataset										UCI Phishing Website-Imbalanced Dataset					UCI SpamBase – Imbalanced Dataset									
	Accuracy	F-Measure	Precision	TP R	ROC	FP R	PR C	MCC	Accuracy	F-Measure	Precision	TP R	ROC	FP R	PR C	MCC	Accuracy	F-Measure	Precision	TP R	ROC	FP R	PR C	MCC	
RF	98.37%	0.984	0.984	0.9	0.9	0.0	0.9	0.9	97.2%	0.974	0.974	0.9	0.9	0.0	0.9	0.9	95.5%	0.955	0.955	0.9	0.9	0.0	0.9	0.9	0.9
SVM	93.87%	0.939	0.939	0.9	0.9	0.0	0.9	0.8	93.8%	0.939	0.939	0.9	0.9	0.0	0.9	0.8	90.415%	0.903	0.903	0.9	0.8	0.1	0.8	0.7	
LR	94.49%	0.945	0.945	0.9	0.9	0.0	0.9	0.8	93.9%	0.940	0.940	0.9	0.9	0.0	0.9	0.8	92.41%	0.924	0.924	0.9	0.9	0.1	0.8	0.8	
MLP	96.71%	0.967	0.967	0.9	0.9	0.0	0.9	0.9	96.9%	0.969	0.969	0.9	0.9	0.0	0.9	0.9	91.219%	0.912	0.912	0.9	0.9	0.1	0.9	0.8	
C4.5	97.31%	0.973	0.973	0.9	0.9	0.0	0.9	0.9	95.8%	0.958	0.958	0.9	0.9	0.0	0.9	0.9	92.979%	0.930	0.930	0.9	0.9	0.0	0.9	0.8	
BN	95.79%	0.958	0.958	0.9	0.9	0.0	0.9	0.9	92.9%	0.929	0.929	0.9	0.8	0.9	0.8	89.8%	0.899	0.899	0.8	0.9	0.1	0.9	0.7		
RT	96.67%	0.967	0.967	0.9	0.9	0.0	0.9	0.9	95.3%	0.953	0.953	0.9	0.9	0.0	0.9	0.9	92.89%	0.929	0.929	0.9	0.9	0.0	0.9	0.8	
NB	85.15%	0.854	0.860	0.8	0.8	0.1	0.9	0.3	92.9%	0.929	0.930	0.9	0.9	0.0	0.9	0.8	79.287%	0.793	0.840	0.7	0.9	0.1	0.9	0.6	
P	97.66%	0.976	0.976	0.9	0.9	0.0	0.9	0.9	96.7%	0.967	0.966	0.9	0.9	0.0	0.9	0.9	93.58%	0.935	0.933	0.9	0.9	0.0	0.9	0.8	

Note: RF = Random forest, LR = Logistic regression, BN = Bayesian network, RT = REPTree, NB = Naïve bayes, P = P.A.R.T.

**TABLE 3. The results of performance evaluation with model validation using ten cross-validations on a balanced dataset.**

Classifier	MDP-2018 -Balanced Dataset										UCI Phishing Website-Balanced Dataset					UCI SpamBase – Balanced Dataset								
	Accuracy	F-Measure	Precision	TP R	ROC	FP R	PR C	MCC	Accuracy	F-Measure	Precision	TP R	ROC	FP R	PR C	MCC	Accuracy	F-Measure	Precision	TP R	ROC	FP R	PR C	MCC
RF	98.37%	0.984	0.984	0.9	0.9	0.0	0.9	0.9	97.40%	0.974	0.974	0.9	0.9	0.0	0.9	0.9	96.03%	0.960	0.960	0.9	0.9	0.0	0.9	0.9
SVM	93.87%	0.939	0.939	0.9	0.9	0.0	0.9	0.8	93.88%	0.938	0.938	0.9	0.9	0.0	0.9	0.8	91.34%	0.913	0.913	0.9	0.9	0.0	0.8	0.8
LR	94.49%	0.945	0.945	0.9	0.9	0.0	0.9	0.8	93.83%	0.938	0.938	0.9	0.9	0.0	0.9	0.8	93.96%	0.940	0.940	0.9	0.9	0.0	0.9	0.8
MLP	96.59%	0.966	0.966	0.9	0.9	0.0	0.9	0.9	96.94%	0.969	0.969	0.9	0.9	0.0	0.9	0.9	92.74%	0.927	0.927	0.9	0.9	0.0	0.9	0.8
C4.5	97.31%	0.973	0.973	0.9	0.9	0.0	0.9	0.9	95.97%	0.959	0.959	0.9	0.9	0.0	0.9	0.9	93.82%	0.938	0.938	0.9	0.9	0.0	0.9	0.8
BN	95.79%	0.958	0.958	0.9	0.9	0.0	0.9	0.9	92.62%	0.926	0.926	0.9	0.9	0.0	0.9	0.8	91.04%	0.910	0.910	0.9	0.9	0.0	0.9	0.8
RT	96.67%	0.967	0.967	0.9	0.9	0.0	0.9	0.9	94.89%	0.949	0.949	0.9	0.9	0.0	0.9	0.8	93.10%	0.931	0.931	0.9	0.9	0.0	0.9	0.8
NB	85.15%	0.854	0.860	0.8	0.8	0.1	0.9	0.3	92.62%	0.926	0.926	0.9	0.9	0.0	0.9	0.8	86.10%	0.861	0.870	0.8	0.9	0.1	0.9	0.7
P	97.66%	0.976	0.976	0.9	0.9	0.0	0.9	0.9	96.32%	0.963	0.963	0.9	0.9	0.0	0.9	0.9	93.52%	0.935	0.933	0.9	0.9	0.0	0.9	0.8

Note: RF = Random forest, LR = Logistic regression, BN=Bayesian network, RT = REPTree, NB = Naïve bayes, P = P.A.R.T.

remained the same when balancing data. In contrast to the balanced UCI Spambase, the Random forest’s performance of each measure was increased. Its classification technique’s performance is presumed to handle both balanced and imbalanced data classes in the dataset.

Naïve Bayes remains the classification technique with the lowest performance in both balanced and imbalanced datasets. It has reduced accuracy, F-measure, precision, and TPR in the two balanced datasets. On the contrary, ROC Naïve Bayes has increased performance only on the UCI Spambase balanced dataset, whereas the Phishing website dataset does not progress without experiencing certain changes. ROC Naïve Bayes on the UCI Spambase dataset is not balanced and balanced at 0.937 and 0.951. Both balanced and imbalanced datasets influence the performance of the classification technique.

A performance evaluation was carried out on a subset scheme of 90% phishing:10% legitimate (90:10), 80% phishing:20% legitimate (80:20), 70% phishing:30% legitimate

(70:30), and 60% phishing:0% legitimate (60:40) in the MDP-2018 dataset. Table 4 shows the maximum performance of random forest at the 90:10 subset schemes is 98.84%, whereas the lowest accuracy at 60:40 is 98.68%. Tables 3 and 4 show that Naïve Bayes have the lowest performance, although this tends to increase based on the subset scheme. Naïve Bayes produced an accuracy of 93.2% with the 90:10 subset scheme, which was higher than the accuracy of 85.15% obtained when the MDP-2018 balanced dataset was used.

A portion of the data, 90% legitimate:10% phishing (90:10), 80% legitimate:20% phishing (80:20), 70% legitimate:30% phishing (70:30), and 60% legitimate:40% phishing (60:40), on the MDP-2018 dataset was selected. The essence is to show how the performance was affected when legitimate dominant data were used rather than that phishing. Table 4 shows that random forest had the best accuracy of 98.84% at the 60:40 subset. The resulting value is similar to the subset scheme of 90% phishing and 10% legitimate in the MDP-2018 dataset. Naïve Bayes produced an accuracy

**TABLE 4. MDP-2018 using data selection (phishing: legitimate / legitimate: phishing).**

Classifier	Accuracy		F-Measure		Precision		TPR		ROC		FPR		PRC		MCC	
	90:10	60:40	90:10	60:40	90:10	60:40	90:10	60:40	90:10	60:40	90:10	60:40	90:10	60:40	90:10	60:40
RF	<b>98.84%</b>	<b>98.84%</b>	<b>0.988</b>	<b>0.988</b>	<b>0.988</b>	<b>0.988</b>	<b>0.988</b>	<b>0.988</b>	<b>0.997</b>	<b>0.999</b>	0.083	<b>0.014</b>	<b>0.998</b>	<b>0.999</b>	<b>0.934</b>	<b>0.976</b>
SVM	96.88%	95.12%	0.968	0.951	0.968	0.951	0.969	0.951	0.883	0.95	0.203	0.052	0.949	0.929	0.818	0.898
LR	97.1%	95.76%	0.97	0.958	0.97	0.958	0.971	0.958	0.98	0.989	0.169	0.044	0.987	0.987	0.834	0.912
MLP	97.68%	97.46%	0.976	0.975	0.976	0.975	0.977	0.975	0.984	0.995	0.141	0.027	0.992	0.995	0.864	0.941
C4.5	97.44%	97.52%	0.974	0.975	0.974	0.975	0.974	0.975	0.928	0.975	0.156	0.025	0.96	0.962	0.853	0.948
BN	97%	96.64%	0.971	0.966	0.973	0.966	0.97	0.966	0.993	0.994	<b>0.064</b>	0.039	0.995	0.995	0.848	0.93
RT	97.56%	96.52%	0.975	0.965	0.975	0.965	0.976	0.965	0.965	0.982	0.143	0.039	0.983	0.976	0.861	0.927
NB	93.2%	90.44%	0.937	0.904	0.932	0.904	0.932	0.904	0.948	0.962	0.12	0.112	0.956	0.957	0.697	0.8
P	98%	97.48%	0.98	0.975	0.98	0.975	0.98	0.975	0.98	0.979	0.1	0.027	0.966	0.971	0.888	0.948

Note: RF = Random forest, LR = Logistic regression, BN=Bayesian network, RT=REPTree, NB=Naïve bayes, P=P.A.R.T.

**TABLE 5. UCI Phishing website dataset using data selection (Phishing:Legitimate/legitimate-Phishing).**

Classifier	Accuracy		F-Measure		Precision		TPR		ROC		FPR		PRC		MCC	
	70:30	90:10	70:30	90:10	70:30	90:10	70:30	90:10	70:30	90:10	70:30	90:10	70:30	90:10	70:30	90:10
RF	<b>98.31%</b>	98.39%	<b>0.983</b>	0.983	<b>0.983</b>	0.984	<b>0.983</b>	0.984	0.983	0.983	<b>0.022</b>	0.985	<b>0.997</b>	<b>0.982</b>	<b>0.956</b>	0.888
SVM	94.74%	98.07%	0.947	0.98	0.947	0.98	0.947	0.981	0.94	0.98	0.067	0.969	0.923	0.956	0.886	0.865
LR	94.54%	98.02%	0.945	0.98	0.945	0.98	0.945	0.98	0.99	0.98	0.067	0.966	0.99	0.955	0.883	0.863
MLP	97.68%	<b>98.59%</b>	0.977	<b>0.986</b>	0.977	<b>0.986</b>	0.977	<b>0.986</b>	<b>0.996</b>	0.986	0.029	0.983	0.996	0.981	0.95	<b>0.902</b>
C4.5	96.30%	97.77%	0.963	0.977	0.966	0.977	0.966	0.978	0.99	0.977	0.049	0.974	0.988	0.969	0.918	0.842
BN	92.47%	97.29%	0.925	0.973	0.927	0.973	0.925	0.973	0.982	0.973	0.077	<b>0.958</b>	0.983	0.946	0.856	0.816
RT	95.43%	98.39%	0.954	0.973	0.954	0.973	0.954	0.974	0.985	0.983	0.059	0.985	0.982	0.982	0.894	0.816
NB	92.45%	98.07%	0.925	0.973	0.925	0.973	0.925	0.973	0.982	0.98	0.077	0.969	0.983	0.956	0.858	0.816
P	96.60%	98.02%	0.966	0.982	0.966	0.982	0.966	0.983	0.99	0.98	0.042	0.966	0.987	0.955	0.927	0.878

Note: RF = Random forest, LR = Logistic regression, BN=Bayesian network, RT=REPTree, NB=Naïve bayes, P=P.A.R.T.

of 93.12% at the 90:10 subset scheme. This is greater than the value obtained when the MDP-2018 balanced dataset (85.15%) was used. However, Naïve Bayes experienced a decrease in the 90:10 subset scheme compared to the values shown in Table 4 (90.44%).

The random forest has the highest subset scheme’s accuracy on the UCI Phishing website, as shown in Table 5. This outperforms the results generated from both balanced and imbalanced datasets. The random forest has the highest accuracy value of 98.31% on the subset scheme of 70% phishing:30% legitimate. This is compared to 97.396% and 97.259% UCI Phishing website balanced and imbalanced dataset. Simultaneously, the Bayesian network has the lowest accuracy of 92.15% (60:40) in all subset schemes. Compared to the initial accuracy, this measure produces better accuracy on imbalanced (92.989%) and balanced (92.62%) datasets from the UCI Phishing website. The Bayesian network and Naïve Bayes are less effective when used on the UCI Phishing website dataset subset scheme.

Meanwhile, when the data portions, namely, legitimate and phishing on the UCI Phishing website, were changed, the random forest was unable to outperform 90% legitimate and 10% phishing data selection techniques, as shown in Table 5. MLP, with 98.59% accuracy, was able to outperform random forest (98.39%). This has the highest accuracy on the UCI Phishing website dataset with the legitimate:phishing scheme. Compared to the balanced and imbalanced datasets, the MLP accuracy of the legitimate:phishing scheme is much better. It is presumed to have increased accuracy, starting from imbalanced (96.9%), and balanced datasets (96.927%), including phishing:legitimate scheme (97.72%) UCI Phishing websites.

Table 6 shows that the Random forest has the highest accuracy of 96.96% for the 90% phishing scheme and 10% legitimate on the UCI Spambase dataset. This is better than the UCI Spambase balanced and imbalanced datasets. Random forest’s accuracy in the UCI Spambase balanced and imbalanced datasets are 96.0287% and 95.5%, respectively. This proves there is an increase in accuracy when the 90:10 scheme is carried out on the UCI Spambase dataset. Naïve Bayes had a similar experience, it encountered a significant increase of 86.97 % in the accuracy of the UCI Spambase, especially the 60:40 scheme. The accuracy of the UCI Spambase imbalanced and balanced datasets is 79.2871% and 86.1%, respectively.

Table 8 shows that Random forest had the best performance when the UCI Spambase dataset with 80% legitimate and 20% phishing schemes was used. Its accuracy is 97.14%, which is higher than the balanced (96.03%), imbalanced UCI Spambase datasets (95.50%) and the 90:10 UCI Spambase scheme (96.96%). However, the maximum and minimum accuracies of Naïve Bayes are 83.23% (60:40) and 76.58% (90:10), respectively. This implies that the best performance was only detected in UCI Spambase with the phishing:legitimate scheme. It had a maximum and minimum accuracies of 94.19% (90:10) and 86.97% (60:40), respectively. Table 7 shows that a total of 8 classifications improved their precision performances in all the subset schemes for the legitimate:phishing class sequence. However, only Naïve Bayes have a partially increased precision performance in the subset scheme. All classification techniques experienced partial performance improvements in ROC and FPR. Decision tree and P.A.R.T experienced partial performance improvements in virtually all the subset schemes. They also



**TABLE 6. UCI Spambase using data selection (Phishing:Legitimate/Legitimate:Phishing).**

Classifier	Accuracy		F-Measure		Precision		TPR		ROC		FPR		PRC		MCC	
	90:10	80:20	90:10	80:20	90:10	80:20	90:10	80:20	90:10	80:20	90:10	80:20	90:10	80:20	90:10	80:20
RF	<b>96.96%</b>	<b>97.1%</b>	<b>0.969</b>	<b>0.971</b>	<b>0.969</b>	<b>0.971</b>	<b>0.97</b>	<b>0.971</b>	<b>0.969</b>	<b>0.969</b>	0.964	0.971	<b>0.962</b>	<b>0.961</b>	<b>0.875</b>	<b>0.879</b>
SVM	92.88%	94.42%	0.921	0.924	0.93	0.925	0.929	0.929	0.921	0.932	0.928	0.924	0.922	0.925	0.686	0.677
LR	95.55%	95.72%	0.955	0.946	0.955	0.946	0.956	0.947	0.955	0.956	0.948	0.946	0.938	0.939	0.818	0.776
MLP	94.92%	96.358%	0.948	0.941	0.948	0.941	0.949	0.943	0.948	0.963	0.927	0.941	0.936	0.93	0.785	0.732
C4.5	95.86%	96.58%	0.958	0.962	0.958	0.961	0.959	0.962	0.958	0.964	0.937	0.962	0.93	0.94	0.832	0.839
BN	91.784%	95.8%	0.924	0.951	0.941	0.951	0.918	0.95	0.924	0.96	0.924	0.951	0.918	0.936	0.743	0.796
RT	94.61%	96.24%	0.945	0.947	0.945	0.946	0.946	0.947	0.945	0.961	0.927	0.947	0.923	0.933	0.727	0.786
NB	94.19%	76.58%	0.941	0.819	0.941	0.9	0.942	0.789	0.941	0.823	<b>0.908</b>	<b>0.819</b>	0.892	0.826	0.762	0.517
P	95.7%	96.99%	0.957	0.962	0.956	0.962	0.957	0.962	0.957	<b>0.969</b>	0.949	0.962	0.937	0.945	0.825	0.843

Note: RF = Random forest, LR = Logistic regression, BN=Bayesian network, RT=REPTree, NB=Naïve bayes, P=P.A.R.T.

**TABLE 7. Improved performance against the subset schemes in the MDP-2018 dataset.**

Classifier	Accuracy		F-Measure		Precision		TPR		ROC		FPR		PRC		MCC	
	P:L	L:P	P:L	L:P	P:L	L:P	P:L	L:P	P:L	L:P	P:L	L:P	P:L	L:P	P:L	L:P
RF	FS	FS	FS	FS	FS	FS	FS	FS	X	X	X	X	X	X	X	X
SVM	FS	FS	FS	FS	FS	FS	FS	FS	X	X	X	X	FS	FS	X	X
LR	FS	FS	FS	FS	FS	FS	FS	FS	X	X	X	X	FS	FS	X	X
MLP	FS	FS	FS	FS	X	FS	FS	FS	X	X	X	X	X	X	X	X
C4.5	X	FS	X	FS	FS	FS	X	FS	X	X	X	X	X	X	X	X
BN	FS	FS	FS	FS	FS	FS	FS	FS	X	X	X	X	FS	FS	X	X
RT	X	X	X	X	FS	FS	X	X	X	X	X	X	X	X	X	X
NB	FS	FS	FS	FS	X	X	FS	FS	X	X	FS	X	FS	FS	X	X
P.A.R.T	X	X	X	X	FS	FS	X	X	X	X	X	X	X	X	X	X

Note: RF = Random forest, LR = Logistic regression, BN=Bayesian network, RT=REPTree, NB=Naïve bayes, P = Phishing, L = Legitimate, FS= Full Scheme, x=Partial Scheme.

**TABLE 8. Improved performance against the subset schemes in the UCI Phishing website.**

Classifier	Accuracy		F-Measure		Precision		TPR		ROC		FPR		PRC		MCC	
	P:L	L:P	P:L	L:P	P:L	L:P	P:L	L:P	P:L	L:P	P:L	L:P	P:L	L:P	P:L	L:P
RF	FS	FS	FS	FS	FS	FS	FS	FS	X	X	X	X	X	X	X	X
SVM	FS	FS	FS	FS	FS	FS	FS	FS	X	X	X	X	FS	FS	X	X
LR	X	FS	X	FS	X	FS	X	FS	FS	FS	X	X	FS	FS	X	X
MLP	FS	FS	FS	FS	FS	FS	FS	FS	X	X	X	X	X	X	X	X
C4.5	FS	FS	FS	FS	FS	FS	FS	FS	X	X	X	X	X	FS	X	X
BN	X	FS	X	FS	X	FS	X	FS	X	X	X	X	X	FS	X	X
RT	X	FS	X	FS	X	FS	X	FS	X	X	X	X	X	X	X	X
NB	X	FS	X	FS	X	FS	X	FS	X	FS	X	X	X	FS	X	X
P.A.R.T	X	FS	X	FS	X	FS	X	FS	X	X	X	X	X	FS	X	X

Note: RF = Random forest, LR = Logistic regression, BN=Bayesian network, RT=REPTree, NB=Naïve bayes, P = Phishing, L = Legitimate, FS= Full scheme, x=Partial scheme.

experienced an overall improvement only in terms of the performance evaluation precision. This differs from the SVM, logistic regression, and Bayesian network, which can only increase virtually all performances except precision.

Table 8 shows that all classification techniques experienced significant performance improvements in the subset scheme using accuracy, F-measure, precision, and TPR for the legitimate:phishing data class sequence. SVM and C.45 are capable of superior performance for all schemes, except ROC and FPR. Meanwhile, the decision tree is a classification

technique that got the most performance improvements in some subset schemes. Overall, all the others experienced a decline in FPR performance for all subset schemes.

Similar to those in Table 7, all classification techniques in Table 9 were partially increased when the FPR performance evaluation was used. However, the decision tree and C4.5 experienced the most partial increases in the subset schemes' performance. The Bayesian network and MLP experienced an increase in the overall performances of all subset schemes, except for the FPR, which received

**TABLE 9. Improved performance against the subset schemes in the UCI Spambase.**

Classifier	Accuracy		F-Measure		Precision		TPR		ROC		FPR		PRC		MCC	
	P:L	L:P	P:L	L:P	P:L	L:P	P:L	L:P	P:L	L:P	P:L	L:P	P:L	L:P	P:L	L:P
RF	FS	X	FS	X	FS	X	FS	X	FS	X	X	X	X	X	X	X
SVM	FS	FS	FS	FS	FS	FS	FS	FS	X	X	X	X	FS	FS	X	X
LR	FS	FS	FS	FS	FS	FS	FS	FS	X	X	X	X	FS	X	X	X
MLP	FS	FS	FS	FS	FS	FS	FS	FS	FS	FS	X	X	FS	FS	X	X
C4.5	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
BN	FS	FS	FS	FS	FS	FS	FS	FS	FS	FS	FS	X	FS	FS	X	X
RT	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
NB	FS	X	FS	FS	FS	FS	FS	X	FS	X	X	FS	FS	X	FS	X
P.A.R.T	X	FS	X	FS	X	FS	X	FS	X	X	X	X	X	FS	X	X

Note: RF = Random forest, LR = Logistic regression, BN=Bayesian network, RT=REPTree, NB=Naïve bayes, P = Phishing, L = Legitimate, FS= Full scheme, x=Partial scheme.

a partial increase. Generally, all classification techniques tend to increase overall performances using accuracy, F-measure, precision, and TPR.

Table 10 shows that the evaluation that does not experience a decrease in performance is accuracy, F-measure, TPR, and precision. This involved the use of the UCI Phishing website dataset with the class order legitimate:phishing. The PRC's minimum performance experienced a decline, as much as 0.1% using MLP with the 90:10 subset scheme (legitimate:phishing) on the UCI Phishing website dataset. Meanwhile, the highest decrease of 56.4% was experienced in the FPR using SVM with the 90:10 subset scheme (legitimate:phishing) in the UCI Spambase dataset. The subset scheme produces performance improvements, especially accuracy. The minimum accuracy performance was 0.02% using C4.5, whereas the maximum was 14.9% using Naïve Bayes. The UCI Spambase dataset with phishing:legitimate class order produced better performance when using the 90:10 subset than the 60:40 subset.

The classification techniques were tested on different datasets and schemes. This research is aimed at determining whether their performances increased or decreased. The random forest classification technique is superior to balanced and imbalanced datasets. Naïve Bayes experienced poor performance, similar to Sahingoz et al.'s findings.

The UCI Phishing website and Spambase are imbalanced datasets. Therefore, the classification technique's performance on the conversion of an initially imbalanced dataset into a balanced one was tested. Adjustments were made to the two UCI datasets by modifying the most negligible class to balance each data. The schema model shows that random forest tends to improve classification performance.

Irrespective of the fact that the previous dataset was balanced, a new test scheme was created by generating an imbalanced one. Meanwhile, the following balanced datasets, 90:10, 80:20, 70:30, and 60:40, were used to generate the MDP-2018 under actual conditions similar to the experiments conducted. Acquiring the actual phishing data is

difficult because collaborating with the phishing victims is of paramount importance. There is a high probability that data imbalance is bound to occur, for example, 90% legitimate and 10% phishing data or 30% phishing and 70% legitimate data.

Random forest produces the best performance on the schemes that were used. The lowest accuracy is 90.78% on the UCI Spambase dataset, with a 60% legitimate and 40% phishing scheme. Conversely, the highest accuracy, which is 98.84%, was realized on the MDP-2018 dataset with the 60:40 (legitimate:phishing) and 90:10 schemes (phishing:legitimate).

Accuracy is not the final measure of classification techniques' performance because it only accumulates identifiable amounts [24]. This led to several performance measures such as TPR, FPR, precision, F-measure, ROC, and PRC. According to Saito and Rehmsmeier [39], these have their respective advantages and disadvantages in balanced and imbalanced datasets. The distribution of data classes needs to be analysed, intending to cover each other's deficiencies. Therefore, it was ensured that the best classification technique was superior to all performance measures. Based on the experiments carried out, random forest has the best performance on any existing measure.

However, random forest underperformed only on the UCI Phishing website dataset with the 90% legitimate and 10% phishing scheme, as shown in Table 7. It only excelled at PRC 0.995 when compared with MLP PRC 0.994. The accuracy obtained by random forest is 98.39% when compared to that of MLP, which is 98.59%. This is because it has not been able to fully identify 89 of the 490 legitimate data, whereas MLP identified 71 of them. Random forest was able to identify 5533 phishing when compared with MLP, which was only able to identify 5527. It is presumed that its performance is still the best in terms of identifying phishing under imbalanced dataset conditions, such as 90% legitimate and 10% phishing.

On the basis of the experiments that were carried out, it was concluded that differences in the subset schemes

**TABLE 10. Classification performance improvement and reduction using subset schemes.**

Performance evaluation	Dataset	Range of performance improvements	Performance reduction range
Accuracy	MDP-2018-Phishing:Legitimate	0.04% - 8.05%	0.04% - 0.51%
	MDP-2018-Legitimate:Phishing	0.07% - 7.97%	0.12% - 0.4%
	UCI-Phishing Website- Phishing:Legitimate	0.1% - 3.03%	0.05% - 0.84%
	UCI-Phishing Website- Legitimate:Phishing	0.57% - 4.33%	Null
	UCI-Spambase- Phishing: Legitimate	0.02% - 14.9%	0.01% - 0.59%
	UCI-Spambase- Legitimate:Phishing	0.16% - 6.97%	0.35% - 4.72%
F measure	MDP-2018-Phishing:Legitimate	0.1% - 8.7%	0.2% - 0.5%
	MDP-2018-Legitimate:Phishing	0.1% - 8.1%	0.1% - 0.4%
	UCI-Phishing Website- Phishing:Legitimate	0.1% - 3%	0.1% - 0.8%
	UCI-Phishing Website- Legitimate:Phishing	0.5% - 4.3%	Null
	UCI-Spambase- Phishing:Legitimate	0.1% - 14.7%	0.2% - 0.6%
	UCI-Spambase- Legitimate:Phishing	0.2% - 6.9%	0.6% - 5.5%
Precision	MDP-2018-Phishing:Legitimate	0.2% - 4.3%	0.2% - 3%
	MDP-2018-Legitimate:Phishing	0.1% - 4.2%	2.6% - 4.1%
	UCI-Phishing Website- Phishing:Legitimate	0.1% - 3%	0.1% - 0.8%
	UCI-Phishing Website- Legitimate:Phishing	0.5% - 4.3%	Null
	UCI-Spambase- Phishing:Legitimate	0.1% - 9.9%	0.2% - 0.6%
	UCI-Spambase- Legitimate:Phishing	0.2% - 10.1%	0.6% - 5.5%
TPR	MDP-2018-Phishing:Legitimate	0.1% - 8%	0.2% - 0.5%
	MDP-2018-Legitimate:Phishing	0.2% - 7.9%	0.1% - 0.4%
	UCI-Phishing Website- Phishing:Legitimate	0.1% - 3%	0.1% - 0.8%
	UCI-Phishing Website- Legitimate:Phishing	0.5% - 4.3%	Null
	UCI-Spambase- Phishing:Legitimate	0.1% - 14.9%	0.2% - 0.6%
	UCI-Spambase- Legitimate:Phishing	0.1% - 7%	0.4% - 5.5%
ROC	MDP-2018-Phishing:Legitimate	0.1% - 0.9%	0.1% - 5.6%
	MDP-2018-Legitimate:Phishing	0.2% - 1.3%	0.1% - 12.7%
	UCI-Phishing Website- Phishing:Legitimate	0.1% - 0.6%	0.2% - 3.6%
	UCI-Phishing Website- Legitimate:Phishing	0.1% - 0.6%	0.1% - 3.5%
	UCI-Spambase- Phishing:Legitimate	0.1% - 2.9%	0.4% - 12.1%
	UCI-Spambase- Legitimate:Phishing	0.1% - 1.4%	0.2% - 26.2%
FPR	MDP-2018-Phishing:Legitimate	0.1% - 5.6%	0.1% - 14.2%
	MDP-2018-Legitimate:Phishing	0.2% - 3.7%	0.3% - 27.1%
	UCI-Phishing Website- Phishing:Legitimate	0.1% - 1.1%	0.1% - 12.2%
	UCI-Phishing Website- Legitimate:Phishing	0.1% - 0.7%	0.1% - 18.9%
	UCI-Spambase- Phishing:Legitimate	0.1% - 6.9%	0.2% - 26.8%
	UCI-Spambase- Legitimate:Phishing	0.1% - 8.9%	0.6% - 56.4%
PRC	MDP-2018-Phishing:Legitimate	0.1% - 3.7%	0.1% - 1.4%
	MDP-2018-Legitimate:Phishing	0.1% - 2.7%	0.1% - 1%
	UCI-Phishing Website- Phishing:Legitimate	0.1% - 4%	0.1% - 2.1%
	UCI-Phishing Website- Legitimate:Phishing	0.1% - 5.6%	0.1%
	UCI-Spambase- Phishing:Legitimate	0.2% - 3.8%	0.2% - 2.6%
	UCI-Spambase- Legitimate:Phishing	0.1% - 4%	0.1% - 1.2%
MCC	MDP-2018-Phishing:Legitimate	0.2%-0.88%	0.3-9.3%
	MDP-2018-Legitimate:Phishing	0.2%-0.85%	0.4-14.4%
	UCI-Phishing Website- Phishing:Legitimate	0.1%-0.16%	0.1-8.9%
	UCI-Phishing Website- Legitimate:Phishing	0.1%-0.19%	0.1-7.5%
	UCI-Spambase- Phishing:Legitimate	0.3%-14.6%	0.4-7.4%
	UCI-Spambase- Legitimate:Phishing	0.6%-0.48%	0.4-37.6%

tend to affect the classification technique's performance. In this subset scheme, the UCI Phishing website dataset contributed to the increase and decrease in performances. However, the UCI Spambase dataset with legitimate data class sequence:phishing significantly increased and decreased performance.

Some performance evaluations were either increased or decreased when this subset scheme was used. The accuracy, TPR, F-measure, and precision performances were significantly improved within the range of 0.01% to 14.9%. Meanwhile, FPR and ROC experienced a decrease in their performances ranging from 0.1% to 56.4%. In FPR, all

classification techniques experienced a significant increase in several subset schemes except the Bayesian network. The majority occurred because of the legitimate class order:phishing with the 90:10 and 80:20 subset schemes. This is because the subset scheme was used to generate innumerable new data that differ regarding phishing and legitimacy, especially in 90:10 and 80:20. The classification performance decreases because the FPR value is high.

#### A. COMPARISON OF THE RESULTS WITH OTHER WORKS

This research compared the results of preliminary research on the basis of datasets, schemes, and classification techniques.

**TABLE 11.** Performance evaluation accuracy of ABET, ROFET, BET and LBET using the subset scheme on the MDP-2018 and UCI SpamBase datasets in Phishing:Legitimate.

Classifier	MDP-2018-Phishing:Legitimate Performance evaluation (ACC)					UCI SpamBase -Phishing:Legitimate Performance evaluation (ACC)				
	Full dataset	90:10:00	80:20:00	70:30:00	60:40:00	Full dataset	90:10:00	80:20:00	70:30:00	60:40:00
ABET	95.850%	96.760%	95.860%	95.780%	95.780%	94.175%	95.186%	94.273%	95.542%	94.235%
ROFET	98.400%	<b>98.660%</b>	98.440%	98.560%	98.560%	<b>95.718%</b>	<b>96.703%</b>	<b>95.916%</b>	<b>96.106%</b>	<b>96.096%</b>
BET	<b>98.420%</b>	<b>98.660%</b>	98.360%	<b>98.620%</b>	<b>98.620%</b>	95.414%	96.023%	95.518%	<b>96.106%</b>	95.915%
LBET	96.870%	97.240%	<b>96.880%</b>	96.680%	96.680%	<b>95.718%</b>	95.971%	95.518%	94.913%	94.508%

Note: ABET= AdaBoost.M1 and Extra trees, ROFET= Rotation Forest and Extra trees, BET= Bagging and Extra-trees, LBET= LogitBoost and Extra trees

Although, this was limited because the proposed schemes are the dataset's newest mechanisms. Not all studies have in common the attributes that were intended to be compared. Meanwhile, only [22], [18], and [23] have slight similarities, that is, the use of the UCI phishing website dataset and MDP-2018. However, [23] is almost similar to the present research regarding testing techniques or data acquisition.

The following are the results of comparisons with [22], [18], and [23]:

- Karabatak and Mustafa [22] succeeded in presenting a performance evaluation of the website phishing classification technique with an unbalanced dataset. A fivefold cross-validation was used to ensure that the model built by the classification technique is better. However, the measurement performance relies only on accuracy, and the classification technique is not necessarily optimal when using different performance evaluations.
- Gana and Abdulhamid [23] proposed a view different from that of Karabatak and Mustafa [22], involving 10 performance evaluation measures on an unbalanced phishing website dataset. The results obtained by Gana and Abdulhamid [23] are better than those acquired by Karabatak and Mustafa [22]. One of the factors of the model performance generated by the classification technique is the use of a 10-fold cross-validation procedure. Coincidentally, Gana and Abdulhamid [23] and Karabatak and Mustafa [22] adopted fold cross-validation. Although the number of folds aids in producing better performances, Gana and Abdulhamid [23] only used it on an unbalanced dataset, making it difficult to prove whether or not they used a balanced dataset.
- Vaitkevicius and Marcinkevicius [18] identified the best classification technique for the MDP-2018 balanced dataset using 30-fold cross-validation. However, despite being similar to the test results of Karabatak and Mustafa [22], including Gana and Abdulhamid [23], the performance of the classification techniques is difficult to prove when different datasets are used.
- The present research aims to determine the performance of classification techniques on balanced and unbalanced

datasets across multiple subset schemes. The present research shows its impact on the MDP-2018 balanced dataset, as reported by Vaitkevicius and Marcinkevicius [18], and on the UCI Phishing website unbalanced dataset, as stated by Karabatak and Mustafa [22], including Gana and Abdulhamid [23]. Meanwhile, better insights were extracted from Gana and Abdulhamid [23] in the 90:10, 80:20, 70:30, and 60:40 subset schemes, thereby enabling the results obtained from this research to be used in resolving crucial gaps as well as providing directives for the development of studies on phishing, especially classification techniques.

Based on the comparative results, the classification techniques proved better than those applied in related research. Some studies analysed the impact of these approaches on unbalanced and balanced datasets concerning certain phishing types. The present research provides more in-depth insights into the impact of classification techniques on balanced and unbalanced datasets using various subset schemes. Finally, [22], [18], and [23] stated the performance of classification techniques in some instances and schemes. At the same time, the present research tends to resolve certain limitations, such as classification techniques' performance in various datasets and subset schemes.

The concept of the proposed subset schema was tested based on a recent research by Alsariera et al. [31]. Incidentally, Alsariera et al. [31] combined meta and base-learners to obtain maximum performance. These two techniques are based on their weaknesses and strengths [42] to achieve maximum performance. The research by Alsariera et al. [31] was selected because it employed Weka techniques. These include ABET (AdaBoost.M1 and Extra trees), ROFET (Rotation Forest and Extra trees), BET (Bagging and Extra-trees) and LBET (LogitBoost and Extra trees). In Tables 11 and 12, it is evident that the ROFET performance has a maximum accuracy of 98.660% for MDP-2018 in phishing:legitimate and 98.9% in that of legitimate: phishing. At the same time, Alsariera et al. [31] obtained maximum accuracy on LBET, ROFET and ABET techniques of 97.5758%, 97.4491%, and 97.4853%, respectively. The maximum LBET performance was 97.8%, higher than the 97.5758% realized by Alsariera et al. [31].



**TABLE 12. Performance evaluation accuracy of ABET, ROFET, BET and LBET using the subset scheme on the MDP-2018 and UCI SpamBase datasets in Legitimate:Phishing.**

Classifier	MDP-2018-Legitimate:Phishing Performance evaluation (ACC)					UCI SpamBase - Legitimate:Phishing Performance evaluation (ACC)				
	Full dataset	90:10	80:20	70:30	60:40	Full dataset	90:10	80:20	70:30	60:40
	ABET	95.850%	97.020%	95.880%	95.620%	95.900%	94.175%	97.324%	95.989%	95.192%
ROFET	98.400%	<b>98.900%</b>	<b>98.560%</b>	<b>98.720%</b>	<b>98.700%</b>	<b>95.718%</b>	97.176%	<b>97.146%</b>	<b>96.514%</b>	95.955%
BET	<b>98.420%</b>	98.860%	98.540%	98.620%	98.620%	95.414%	97.547%	97.031%	<b>96.514%</b>	<b>96.165%</b>
LBET	96.870%	97.800%	97.160%	96.680%	96.720%	<b>95.718%</b>	<b>97.659%</b>	96.876%	96.194%	95.371%

Note: ABET= AdaBoost.M1 and Extra trees, ROFET= Rotation Forest and Extra trees, BET= Bagging and Extra-trees, LBET= LogitBoost and Extra trees

**TABLE 13. Performance evaluation MCC of ABET, ROFET, BET and LBET using the subset scheme on the MDP-2018 and UCI SpamBase datasets in Phishing:Legitimate.**

Classifier	MDP-2018-Phishing:Legitimate Performance evaluation (MCC)					UCI SpamBase -Phishing:Legitimate Performance evaluation (MCC)				
	Full dataset	90:10:00	80:20:00	70:30:00	60:40:00	Full dataset	90:10:00	80:20:00	70:30:00	60:40:00
	ABET	0.917	0.821	0.869	0.899	0.899	0.878	0.802	0.856	0.865
ROFET	<b>0.968</b>	<b>0.924</b>	<b>0.951</b>	0.966	0.966	0.910	<b>0.864</b>	0.897	<b>0.919</b>	<b>0.922</b>
BET	<b>0.968</b>	<b>0.924</b>	0.948	<b>0.967</b>	<b>0.967</b>	0.904	<b>0.864</b>	<b>0.898</b>	<b>0.919</b>	0.918
LBET	0.939	0.838	0.901	0.921	0.921	<b>0.910</b>	0.833	0.887	0.894	0.891

Note: ABET= AdaBoost.M1 and Extra trees, ROFET= Rotation Forest and Extra trees, BET= Bagging and Extra-trees, LBET= LogitBoost and Extra trees

In terms of accuracy performance (table 12), UCI Spam-Base ABET, ROFET, BET, and LBET obtained maximum performance in the 90:10 subset scheme with legitimate:phishing order. However, in contrast to the MDP-2018 with the legitimate:phishing sequence, it was increased to the 90:10 and 60:40 subsets of the scheme. Based on Table 12, the ROFET technique has the best performance in each subset 90:10, 80:20, 70:30 and 60:40 on the MDP-2018 dataset in the order legitimate:phishing. In Table 11, ROFET excels in all subset schemes on UCI Spambase in the order phishing:legitimate.

In contrast to the performance of MCC, several subset schemes experienced significant changes. The UCI Spambase, ROFET and BET techniques have improved performance in the 70:30 and 60:40 subset schemes in the order phishing:legitimate (Table 13). In the MDP-2018 dataset only ROFET had increased performance in the 70:30 and 60:40 for the legitimate:Phishing sequence (table 14). Therefore, 94% of the subset schemes used succeeded in reducing the performance of the ROFET, BET, LBET, and ABET techniques on the UCI Spambase dataset in the legitimate:phishing order. In the MDP-2018, 100% of the subset schemes reduced the performance of the ROFET, BET, LBET, and ABET techniques, in the order of phishing:legitimate.

Table 14, shows that ROFET excelled in all subset schemes in MDP-2018 in the order of legitimate:phishing. It has the highest performance of 0.973, in the 60:40 subset scheme for the MCC. However, ROFET has reduced performance from

0.910 to 0.753 for MCC on UCI SpamBase in the order of legitimate:phishing. This also occurred in the 90:10 subset scheme on UCI Spambase and MDP-2018 in the order legitimate:phishing. In general, ABET, ROFET, BET and LBET in the subset scheme of MDP-2018 and UCI Spambase with legitimate:phishing order experienced a significant increase.

Several other performance measures such as BDR, G-Mean and MCC were included in Alsariera et al’s research, to get more insight. MDP 2018 and UCI Spambase were selected because of their significant performances on the adopted technique and that proposed by Alsariera et al. [31]. Based on Table 15, Random forest produces maximum performance in BDR, G-Mean and MCC on MDP 2018 in the order phishing:legitimate. ROFET only excelled at MDP 2018 in the order of legitimate:phishing for G-mean and MCC performance. LBET obtained the highest BDR value at MDP 2018 in the order of legitimate:phishing.

Random forest also produces the best G-Mean, MCC and BDR performances on UCI Spambase in the order phishing:legitimate (table 16). Meanwhile, LBET generates the best G-Mean, MCC and BDR performance on UCI Spambase in the order legitimate:phishing. In Table 15, it generated the highest BDR, while in XVI, it produced maximum performance for G-Mean, MCC and BDR.

Based on the various schemes carried out, whether it is the comparison with the latest research (Alsariera et al. [31]), the inclusion of the most recent measures (G-Mean, MCC and BDR), or the significant performance of each technique,

**TABLE 14. Performance evaluation MCC of ABET, ROFET, BET and LBET using the subset scheme on the MDP-2018 and UCI SpamBase datasets in Legitimate: Phishing.**

Classifier	MDP-2018-Legitimate: Phishing Performance evaluation (MCC)					UCI SpamBase - Legitimate: Phishing Performance evaluation (MCC)				
	Full dataset	90:10:00	80:20:00	70:30:00	60:40:00	Full dataset	90:10:00	80:20:00	70:30:00	60:40:00
ABET	0.917	0.837	0.872	0.897	0.915	0.878	0.770	0.831	0.857	0.876
ROFET	<b>0.968</b>	<b>0.938</b>	<b>0.955</b>	<b>0.969</b>	<b>0.973</b>	<b>0.910</b>	0.753	<b>0.879</b>	<b>0.897</b>	0.904
BET	<b>0.968</b>	0.935	0.954	0.967	0.971	0.904	0.789	0.874	<b>0.897</b>	<b>0.909</b>
LBET	0.939	0.889	0.917	0.925	0.934	<b>0.910</b>	<b>0.805</b>	0.868	0.887	0.890

Note: ABET= AdaBoost.M1 and Extra trees, ROFET= Rotation Forest and Extra trees, BET= Bagging and Extra-trees, LBET= LogitBoost and Extra trees

**TABLE 15. Performance evaluation MCC, BDR and G-Mean using the subset scheme 90:10 on the MDP-2018.**

Classifier	MDP 2018 –Phishing:Legitimate (90:10)					MDP 2018 –Legitimate:Phishing (90:10)				
	F-measure	Accuracy	G-Mean	BDR	MCC	F-measure	Accuracy	G-Mean	BDR	MCC
ABET	0.968	96.760%	0.968	51.434	0.821	0.97	97.020%	0.97	59.85	0.837
ROFET	0.986	98.660%	0.987	73.7	0.924	<b>0.989</b>	<b>98.900%</b>	<b>0.989</b>	89.258	<b>0.938</b>
BET	0.986	98.660%	0.987	78.96	0.924	0.988	98.860%	<b>0.989</b>	92.997	0.935
LBET	0.97	97.240%	0.972	41.682	0.838	0.979	97.800%	0.978	<b>233.768</b>	0.889
RF	<b>0.988</b>	<b>98.840%</b>	<b>0.988</b>	<b>89.433</b>	<b>0.934</b>	0.987	98.740%	0.987	83.755	0.929

Note: ABET= AdaBoost.M1 and Extra trees, ROFET= Rotation Forest and Extra trees, BET= Bagging and Extra-trees, LBET= LogitBoost and Extra trees, RF= Random forest

**TABLE 16. Performance evaluation using different subset scheme on the UCI SpamBase.**

Classifier	UCI Spambase –Phishing:Legitimate (90:10)					UCI Spambase –Legitimate:Phishing (90:10)				
	F-measure	Accuracy	G-Mean	BDR	MCC	F-measure	Accuracy	G-Mean	BDR	MCC
ABET	0.951	95.186%	0.952	24.792	0.802	0.973	97.324%	0.973	41.994	0.77
ROFET	0.966	96.703%	0.967	31.235	0.864	0.97	97.176%	0.972	37.488	0.753
BET	0.959	96.023%	0.96	28.131	0.836	0.974	97.547%	0.975	43.745	0.789
LBET	0.959	95.971%	0.96	26.352	0.833	<b>0.976</b>	<b>97.659%</b>	<b>0.977</b>	<b>55.663</b>	<b>0.805</b>
RF	<b>0.969</b>	<b>96.964%</b>	<b>0.97</b>	<b>32.993</b>	<b>0.875</b>	0.969	97.101%	0.971	37.954	0.748

Note: ABET= AdaBoost.M1 and Extra trees, ROFET= Rotation Forest and Extra trees, BET= Bagging and Extra-trees, LBET= LogitBoost and Extra trees, RF= Random forest

further investigation was conducted to ascertain how this phenomenon tend to either decrease or increase. Samples were collected based on the highest performance, namely the UCI Spambase dataset with a 90:10 subset scheme in the order legitimate:phishing and the MDP-2018 dataset with a 90:10 subset scheme in the order phishing:legitimate order.

The unused dataset, which includes 90% of the phishing data on UCI Spambase and 90% of the legitimate on MDP-2018, were proven. The data on the UCI Spambase in the order legitimate:phishing was labelled as follows, 90:10a, 90:10b and 90:10c is 10% of the first, second and third data records, while the next 10% data record was not used. Based on Table 17, the 90:10b subset scheme has maximum performance on F-Measure, Accuracy, BDR, G-Mean, and MCC. This was also detected in Random forest and BET techniques. However, BET was able to excel at F-Measure, Accuracy, G-Mean and MCC for the 90:10b subset scheme, while Random forest exceeded at F-Measure, Accuracy, G-Mean, BDR and MCC for the 90:10c subset scheme.

In MDP-2018 with phishing:legitimate, the same process carried out on the UCI Spambase, was also performed. Furthermore, 10% of each valid data record was labelled like 90:10a, 90:10b and 90:10c. The subsequent 10% of records was not used as was the case with the previous experiments. Based on Table 18, the ROFET experienced maximum performance with a subset of the 90:10c scheme, namely the F-Measure, Accuracy, G-Mean and MCC. ROFET accuracy reached 99.08% and F-Measure 0.991 in the 90:10c subset scheme, while Random forest was able to excel in the 90:10b and 90:10c subset schemes in the performance of F-Measure, Accuracy, G-Mean, BDR and MCC.

Based on the investigation of data that were not used during the performance evaluation, it was concluded that they significantly influenced the performance of the classification technique. Therefore, each subset scheme has a superior technique for detecting phishing attacks. The proposed Technical Development needs to be able to adapt to changes in

**TABLE 17. Performance evaluation using different subset scheme on the UCI SpamBase.**

Classifier	F-measure			Accuracy			G-Mean			BDR			MCC		
	90:10 a	90:10 b	90:10 c	90:10a	90:10b	90:10c	90:10 a	90:10 b	90:10 c	90:10 a	90:10 b	90:10 c	90:10 a	90:10 b	90:10 c
ABET	0.973	0.957	0.957	97.324 %	95.65 %	95.88 %	0.973	0.957	0.959	41.99 4	33.99 6	0.77	0.669	0.652	
ROFET	0.97	0.977	0.968	97.176 %	97.88 %	97.10 %	0.972	0.979	0.971	37.48 8	46.16 4	0.753	0.819	0.743	
BET	0.974	<b>0.981</b>	0.969	97.547 %	<b>98.14</b> %	97.18 %	0.975	<b>0.981</b>	0.972	43.74 5	56.48 4	0.789	<b>0.843</b>	0.751	
LBET	<b>0.976</b>	0.971	0.968	<b>97.659</b> %	96.88 %	96.84 %	<b>0.977</b>	0.969	0.968	<b>55.66</b> <b>3</b>	<b>142.4</b> <b>6</b>	<b>48.26</b> <b>7</b>	<b>0.805</b>	0.789	0.743
RF	0.969	0.979	<b>0.972</b>	97.101 %	98.03 %	<b>97.36</b> %	0.971	0.98	<b>0.974</b>	37.95 4	54.98 6	0.748	0.834	<b>0.77</b>	

Note: ABET= AdaBoost.M1 and Extra trees, ROFET= Rotation Forest and Extra trees, BET= Bagging and Extra-trees, LBET= LogitBoost and Extra trees, RF= Random forest

**TABLE 18. Performance evaluation MCC, BDR and G-Mean using the subset scheme on the MDP-2018.**

Classifier	F-measure			Accuracy			G-Mean			BDR			MCC		
	90:10 a	90:10 b	90:10 c	90:10a	90:10b	90:10c	90:10 a	90:10 b	90:10 c	90:10 a	90:10 b	90:10c	90:10 a	90:10 b	90:10 c
ABET	0.968	0.967	0.969	96.760 %	96.74 %	96.96 %	0.968	0.967	0.97	51.43 4	42.55 8	45.012	0.821 %	0.815	0.827
ROFET	0.986	0.987	<b>0.991</b>	98.660 %	98.72 %	<b>99.08</b> %	0.987	0.987	<b>0.991</b>	73.7	77.52 7	110.99 6	0.924 %	0.927	<b>0.948</b>
BET	0.986	0.987	0.99	98.660 %	98.76 %	99.02 %	0.987	0.988	0.99	78.96	86.92 5	<b>115.25</b>	0.924 %	0.93	0.945
LBET	0.97	0.972	0.972	97.240 %	97.38 %	97.38 %	0.972	0.974	0.974	41.68 2	43.50 7	49.409	0.838 %	0.847	0.847
RF	<b>0.988</b>	<b>0.989</b>	0.989	<b>98.840</b> %	<b>98.94</b> %	98.92 %	<b>0.988</b>	<b>0.989</b>	0.989	<b>89.43</b> <b>3</b>	<b>97.77</b> <b>2</b>	91.061	<b>0.934</b>	<b>0.94</b>	0.939

Note: ABET= AdaBoost.M1 and Extra trees, ROFET= Rotation Forest and Extra trees, BET= Bagging and Extra-trees, LBET= LogitBoost and Extra trees, RF= Random forest

existing data, thereby ensuring that the phishing attack detection technique becomes optimal.

**B. INSIGHTS AND FUTURE RESEARCH DIRECTIONS**

Based on the experiments performed, the following are some contributions of the classification techniques for phishing attacks:

- There is no accuracy-capable classification technique that is tamper-resistant against publicly available datasets.

The results of the evaluation process proves that there is no superior classification technique for performance testing on various datasets. Therefore, its use has to be adjusted to the immediate conditions or data held. According to Japkowicz and Shah [43], an experiment that involves the use of a specific dataset need not generalize the results of different data. This is in line with an experiment carried out by Rao et al. [4], that the performance obtained from public datasets is not the same as that of private (owned) ones.

- Disclosure of detailed information on parameter settings in classification techniques

This research was unable to find detailed information on the parameters used by several others, therefore, it was difficult to

make performance comparisons. Weka, including its default parameters, was used to test classification techniques on the three datasets, namely, MDP-2018, UCI Phishing website, and Spambase. This research aims to prove that the use of default parameters can be used to realize better performance measures.

- There is no standard value or cutoff range for performance evaluation.

The classification technique performance assessment was not found in any category. Generally, preliminary studies used a value of approximately 1, indicating the best performance [44]. Several studies used alternative measurements besides accuracy, for instance, detecting a higher TPR or lower FPR value.

Each classification technique performed effectively in some of the tests. Based on the SLR applied, not all classification techniques excelled in all the performance tests. Therefore, its measurement is carried out through accuracy, although there is need for more insight into the classification technique’s performance on the experimental model formulated.

- The selection of a subset scheme tends to affect the classification performance.

Various classification techniques produce different performances in the subset scheme. The 90:10, 80:20, 70:30, and 60:40 subset schemes were selected in the order of legitimate/phishing. The balanced dataset was also changed to imbalanced using the schema.

The formulated scheme reflects the original conditions. Besides, when observed, it is unlikely that one is bound to get legitimate and phishing data in a balanced state. Therefore, there is a need for a classification technique that can deal with these data.

The implemented scheme has been proven to affect the performance of classification techniques. For example, although Naïve Bayes was ranked the lowest, it tends to increase the performance values.

## V. CONCLUSION

This research explores diverse classification techniques to explain the maximum performance using a subset scheme. The objectives realized are based on the fact that the use of a subset scheme can affect the performance of classification techniques on various datasets. Therefore, the challenge addressed was the ability of the classification technique to perform when using a subset scheme on balanced and imbalanced datasets.

The classification technique was tested for performance against the subset schemes of 90:10, 80:20, 70:30, and 60:40. In addition, ten performance measures, namely Accuracy, F-Measure, Precision, TPR, ROC, FPR, PRC, MCC, BDR, and G-Mean, were utilized. The scheme is applied to the data in the following sequence phishing:legitimate and legitimate:phishing. Its users tend to produce a significant increase and decrease with respect to performance. Moreover, each classification technique excels at specific performance measures. Not many of them excel on all performance measures.

The performance of unused data was investigated during testing, such as 90% of the data was legitimate for the phishing:legitimate sequence with the distribution of data of 90% phishing and 10% legitimate etc. The findings of this research prove that unused data significantly affects performance during the classification process. Therefore, further investigation of such data is required.

In addition to the under-sampling technique, there is also an over-sampling technique that researchers often use. Therefore, we are interested in trying the over-sampling approach as a technique used to form new datasets that are sourced from balanced and imbalanced datasets. And we will evaluate the performance of the classification technique on the newly formed dataset.

Recent research that adopted a mix of meta and base-learners to detect phishing attacks was also compared. This is intended to prove that the recent detection techniques also encountered certain problems associated with performance when faced with the proposed scheme. It tends to make the recent detection techniques experience a significant performance change based on the results obtained.

There was a significant performance increase and decrease in the subset scheme. This includes a decreased and increases from 0.01% to 56% and 0.04% to 14.9%, respectively. Random forest classification technique excelled in some of the proposed schemes. Meanwhile, the highest performance in the subset scheme is the ROFET technique with an accuracy of 99.08%, while the lowest was detected in the SVM with an FPR of 0.686.

The selection of the dataset also has a significant impact on classification performance. When the subset scheme was applied to the UCI Phishing website dataset, it contributed the least. However, the UCI Spambase dataset with a legitimate data class sequence:phishing significantly increases or decreases performance.

Many researchers use the hyper-parameter technique to find the best parameters for the best performance. Therefore, we are interested in implementing the hyper-parameter method to determine the performance of the classification technique on the subset scheme that we propose in future studies.

The proposed schemes are better than those of related research. The present research revealed the weaknesses of classification techniques by using datasets and subset schemes. This also includes being able to discern which techniques are capable of being superior to the established schemes. In future, this subset scheme needs to be applied to confirmed cases to compare its performances. Moreover, certain recommendations were proposed for developing research on improving phishing classification techniques. For example, evaluating phishing or legitimate data labelled as unused or bad, specifying a standard performance value for a technique to detect phishing attacks, such as accuracy greater than 90% is acceptable, and creating a phishing attack detection concept that is adaptive to the data provided.

## REFERENCES

- [1] Proofpoint. (Jun. 2022). *2021 State of the Phish*. [Online]. Available: <https://www.proofpoint.com/sites/default/files/gtd-pfpt-us-tr-state-of-the-phish-2020.pdf>
- [2] C. Naksawat, S. Akkason, and C. K. Loi, "Persuasion strategies: Use of negative forces in scam E-mails," *GEMA Online J. Lang. Stud.*, vol. 16, no. 1, pp. 1–17, 2016.
- [3] M. A. Pitchan, S. Z. Omar, and A. H. A. Ghazali, "Amalan keselamatan siber pengguna internet terhadap buli siber, pornografi, e-mel phishing dan pembelian dalam talian (cyber security practice among internet users towards cyberbullying, pornography, phishing email and online shopping)," *Jurnal Komunikasi, Malaysian J. Commun.*, vol. 35, no. 3, pp. 212–227, Sep. 2019.
- [4] R. S. Rao, T. Vaishnavi, and A. R. Pais, "CatchPhish: Detection of phishing websites by inspecting URLs," *J. Ambient Intell. Hum. Comput.*, vol. 11, no. 2, pp. 813–825, Feb. 2020.
- [5] W. Ali and S. Malebary, "Particle swarm optimization-based feature weighting for improving intelligent phishing website detection," *IEEE Access*, vol. 8, pp. 116766–116780, 2020.
- [6] R. S. Rao and A. R. Pais, "Detection of phishing websites using an efficient feature-based machine learning framework," *Neural Comput. Appl.*, vol. 31, no. 8, pp. 3851–3873, Aug. 2019.
- [7] K. L. Chiew, C. L. Tan, K. 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.*, vol. 484, pp. 153–166, May 2019.



- [8] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," *Expert Syst. Appl.*, vol. 117, pp. 345–357, Mar. 2019.
- [9] S. W. Liew, N. F. M. Sani, M. T. Abdullah, R. Yaakob, and M. Y. Sharum, "An effective security alert mechanism for real-time phishing tweet detection on Twitter," *Comput. Secur.*, vol. 83, pp. 201–207, Jun. 2019.
- [10] V. Muppavarapu, A. Rajendran, and S. K. Vasudevan, "Phishing detection using RDF and random forests," *Int. Arab J. Inf. Technol.*, vol. 15, no. 5, pp. 817–824, 2018.
- [11] A. S. Bozkir and M. Aydos, "LogoSENSE: A companion HOG based logo detection scheme for phishing web page and E-mail brand recognition," *Comput. Secur.*, vol. 95, Aug. 2020, Art. no. 101855.
- [12] S. E. Raja and R. Ravi, "A performance analysis of software defined network based prevention on phishing attack in cyberspace using a deep machine learning with CANTINA approach (DMLCA)," *Comput. Commun.*, vol. 153, pp. 375–381, Mar. 2020.
- [13] M. Sameen, K. Han, and S. O. Hwang, "PhishHaven—An efficient real-time AI phishing URLs detection system," *IEEE Access*, vol. 8, pp. 83425–83443, 2020.
- [14] R. S. Rao, T. Vaishnavi, and A. R. Pais, "PhishDump: A multi-model ensemble based technique for the detection of phishing sites in mobile devices," *Pervasive Mobile Comput.*, vol. 60, Nov. 2019, Art. no. 101084.
- [15] Y. Ding, N. Luktarhan, K. Li, and W. Slamun, "A keyword-based combination approach for detecting phishing webpages," *Comput. Secur.*, vol. 84, pp. 256–275, Jul. 2019.
- [16] A. K. Jain and B. B. Gupta, "A machine learning based approach for phishing detection using hyperlinks information," *J. Ambient Intell. Hum. Comput.*, vol. 10, no. 5, pp. 2015–2028, May 2019.
- [17] A. E. Aassal, S. Baki, A. Das, and R. M. Verma, "An in-depth benchmarking and evaluation of phishing detection research for security needs," *IEEE Access*, vol. 8, pp. 22170–22192, 2020.
- [18] P. Vaitkevicius and V. Marcinkevicius, "Comparison of classification algorithms for detection of phishing websites," *Informatica*, vol. 31, no. 1, pp. 143–160, Mar. 2020.
- [19] Y.-H. Chen and J.-L. Chen, "AI@ntiPhish—Machine learning mechanisms for cyber-phishing attack," *IEICE Trans. Inf. Syst.*, vol. E102.D, no. 5, pp. 878–887, May 2019.
- [20] C. L. Tan, K. L. Chiew, K. S. C. Yong, S. N. Sze, J. Abdullah, and Y. Sebastian, "A graph-theoretic approach for the detection of phishing webpages," *Comput. Secur.*, vol. 95, Aug. 2020, Art. no. 101793.
- [21] S. Mishra and D. Soni, "Smishing detector: A security model to detect smishing through SMS content analysis and URL behavior analysis," *Future Gener. Comput. Syst.*, vol. 108, pp. 803–815, Jul. 2020.
- [22] M. Karabatak and T. Mustafa, "Performance comparison of classifiers on reduced phishing website dataset," in *Proc. 6th Int. Symp. Digit. Forensic Secur. (ISDFS)*, Mar. 2018, pp. 1–5.
- [23] N. N. Gana and S. M. Abdulhamid, "Machine learning classification algorithms for phishing detection: A comparative appraisal and analysis," in *Proc. 2nd Int. Conf. IEEE Nigeria Comput. Chapter (NigeriaComput-Conf)*, Oct. 2019, pp. 1–8.
- [24] T. Gangavarapu, C. D. Jaidhar, and B. Chanduka, "Applicability of machine learning in spam and phishing email filtering: Review and approaches," *Artif. Intell. Rev.*, vol. 53, no. 7, pp. 5019–5081, Oct. 2020.
- [25] S. Priya, S. Selvakumar, and R. L. Velusamy, "Evidential theoretic deep radial and probabilistic neural ensemble approach for detecting phishing attacks," *J. Ambient Intell. Hum. Comput.*, vol. 14, no. 3, pp. 1951–1975, Jul. 2021.
- [26] P. L. Indrasiri, M. N. Halgamuge, and A. Mohammad, "Robust ensemble machine learning model for filtering phishing URLs: Expandable random gradient stacked voting classifier (ERG-SVC)," *IEEE Access*, vol. 9, pp. 150142–150161, 2021.
- [27] A. Ozcan, C. Catal, E. Donmez, and B. Senturk, "A hybrid DNN-LSTM model for detecting phishing URLs," *Neural Comput. Appl.*, vol. 35, no. 7, pp. 4957–4973, Aug. 2021.
- [28] S.-J. Bu and H.-J. Kim, "Optimized URL feature selection based on genetic-algorithm-embedded deep learning for phishing website detection," *Electronics*, vol. 11, no. 7, p. 1090, Mar. 2022.
- [29] V. Zeng, S. Baki, A. E. Aassal, R. Verma, L. F. T. De Moraes, and A. Das, "Diverse datasets and a customizable benchmarking framework for phishing," in *Proc. 6th Int. Workshop Secur. Privacy Anal.*, Mar. 2020, pp. 35–41.
- [30] A. Ihsan and E. Rainarli, "Optimization of  $k$ -nearest neighbour to categorize Indonesian's news articles," *Asia-Pacific J. Inf. Technol. Multimedia*, vol. 10, no. 1, pp. 43–51, Jun. 2021.
- [31] Y. A. Alsariera, V. E. Adeyemo, A. O. Balogun, and A. K. Alazzawi, "AI meta-learners and extra-trees algorithm for the detection of phishing websites," *IEEE Access*, vol. 8, pp. 142532–142542, 2020.
- [32] E. Sukawai and N. Omar, "Corpus development for Malay sentiment analysis using semi supervised approach," *Asia-Pacific J. Inf. Technol. Multimedia*, vol. 9, no. 1, pp. 94–109, Jun. 2020.
- [33] C. L. Tan, "Phishing dataset for machine learning: Feature evaluation," *Mendeley Data*, V1, 2018, doi: 10.17632/h3cgnj8hft.1.
- [34] X.-Y. Lu, M.-S. Chen, J.-L. Wu, P.-C. Chang, and M.-H. Chen, "A novel ensemble decision tree based on under-sampling and clonal selection for web spam detection," *Pattern Anal. Appl.*, vol. 21, no. 3, pp. 741–754, Aug. 2018.
- [35] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Trans. Syst., Man C, Appl. Rev.*, vol. 42, no. 4, pp. 463–484, Jul. 2012.
- [36] E. S. Gualberto, R. T. De Sousa, T. P. D. B. Vieira, J. P. C. L. Da Costa, and C. G. Duque, "From feature engineering and topics models to enhanced prediction rates in phishing detection," *IEEE Access*, vol. 8, pp. 76368–76385, 2020.
- [37] H. Zhang, G. Liu, T. W. S. Chow, and W. Liu, "Textual and visual content-based anti-phishing: A Bayesian approach," *IEEE Trans. Neural Netw.*, vol. 22, no. 10, pp. 1532–1546, Oct. 2011.
- [38] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Amsterdam, The Netherlands: Elsevier, 2017.
- [39] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, pp. 1–21, Mar. 2015.
- [40] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Dec. 2006.
- [41] A. E. Aassal, L. Moraes, S. Baki, A. Das, and R. Verma, "Anti-phishing pilot at ACM IWSPA 2018: Evaluating performance with new metrics for unbalanced datasets," in *Proc. Anti-Phishing Shared Task Pilot 4th ACM IWSPA*, 2018, pp. 2–10.
- [42] H. A. Alshalabi, S. Tiun, and N. Omar, "A comparative study of the ensemble and base classifiers performance in Malay text categorization," *Asia-Pacific J. Inf. Technol. Multimedia*, vol. 6, no. 2, pp. 53–64, Dec. 2017.
- [43] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2011.
- [44] R. Gowtham and I. Krishnamurthi, "PhishTackle—A web services architecture for anti-phishing," *Cluster Comput.*, vol. 17, no. 3, pp. 1051–1068, Sep. 2014.



**RAHMAD ABDILLAH** received the S.T. degree from the Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia, in 2010, and the M.T. degree in informatics from the Institut Teknologi Bandung (ITB), Indonesia, in 2013. He is currently pursuing the Ph.D. degree with the Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM). He has been working as a Lecturer with the Department of Informatics, Universitas Islam Negeri Sultan Syarif Kasim Riau, since 2010. His research interests include cybersecurity, artificial intelligence, and social engineering.



**ZARINA SHUKUR** received the Ph.D. degree from the University of Nottingham, in 1999. She is currently a Professor at the Center for Cyber Security Studies, Universiti Kebangsaan Malaysia. Her research interests include formal methods and cybersecurity.



**MASNIZAH MOHD** received the B.I.T. and M.I.T. degrees in information science from Universiti Kebangsaan Malaysia (UKM), in 1999 and 2002, respectively, and the Ph.D. degree in computer and information sciences from the University of Strathclyde, in 2010. She is currently an Associate Professor at UKM. Her current research interests include information retrieval, natural language processing, and cyber intelligence.



**T. S. MOHD ZAMRI MURAH** received the B.Sc. and M.Sc. degrees in statistics from the University of Iowa, IA, USA, in 1987 and 1989, respectively. He is currently a Senior Lecturer at the Center for Cyber-Security, Universiti Kebangsaan Malaysia. His current research interests include the development of deep learning models for cybersecurity, automated penetration testing, and cyber range.



**INSU OH** received the B.S. and M.S. degrees from the Department of Information Security Engineering, Soochunhyang University, Asan, South Korea, in 2018 and 2020, respectively, where he is currently pursuing the Ph.D. degree with the Department of Information Security Engineering. His research interests include vulnerability analysis, mobile baseband security, automotive security, and V2X security.



**KANGBIN YIM** received the B.S., M.S., and Ph.D. degrees from the Department of Electronics Engineering, Ajou University, Suwon, South Korea, in 1992, 1994, and 2001, respectively. He is currently a Professor with the Department of Information Security Engineering, Soochunhyang University. His research interests include malware analysis, vulnerability identification, code obfuscation, secure architecture, and mobile baseband and automotive security.

...