Heterogeneous Borda Ensemble Feature Selection: An Enhancement to Machine Learning Approach for Uniform Resource Locator Phishing

P. O. Olabisi[1*], B. M. Olukoya[2], G. O. Ogunleye[3], O. J. Adetunji[4]

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

ABSTRACT: Phishing attacks, which involve deceptive methods to harvest sensitive user information, pose significant threats to online transactions. Despite ongoing efforts to combat these cyberattacks, traditional solutions have struggled to address the increasing sophistication of phishing schemes. Machine learning (ML) techniques as emerging tools for detecting these threats are particularly more effective when combined with robust feature selection methods. The authors therefore introduced a novel approach called Heterogeneous Borda Ensemble Feature Selection (HBEFS), wherein three filter-based statistical techniques generate primary subsets of phishing indicators. The variables selected by each technique are aggregated to form the final baseline features. The proposed method efficiently identifies essential phishing features. The use of a random forest (RF) model with HBEFS indicators achieved an impressive performance of 97.4%. Also, classical models exhibit significant improvement when evaluated on the baseline features. Comparing the performance of models using HBEFS variables, individual statistical techniques, and other studies, the HBEFS consistently outperforms them. These research findings necessitate the need for innovative feature selection strategies to enhance potency of ML classification for phishing detection.

KEYWORDS: Borda count, Feature selection, Machine learning, Cyberattack, Phishing detection

[Received Nov. 24, 2021; Revised May 15, 2022; Accepted May 26, 2022] Print ISSN: 0189-9546 | Online ISSN: 2437-2110

I. INTRODUCTION

The efficiency and capacity of communication technologies to transmit information at the speed of light and at a very minimal cost made email a major globally accepted channel for transferring voluminous digital content (Alshingiti *et al.*, 2023; Gualberto *et al.*, 2020; Osho *et al.*, 2019). However, it is disheartening to discover that despite the stack of benefits these technologies offer, their security and privacy are seriously challenged. The security challenge of internet-based platforms started with spam, which later metamorphosed in*to dreadful attacks* known as phishing. Phishing is the act of using social engineering devices to deceive people to reveal their personal and corporate secret information (Khan *et al.*, 2020; Amusan *et al.*, 2021). Over the years, phishing attacks, which were first perpetrated through the phone in what is regarded as vishing, have been used to conduct heinous crimes seamlessly, and this has made it a prime target in cybercrimes. This evil act has an array of effects which include enormous economic losses, massive operational disruptions, data breaches, reputation damages, and untimely deaths (some victims commit suicide) (Ali *et al*., 2022). Trending phishing activities presently come in the form of uniform resource locator (URL) links disseminated across millions of targets through an email that ushers the victims to spoofed sites (Basit *et al.*, 2020).

Phishing attacks and their resilient effects have continued to increase sporadically due to seamless access to smart kits to conduct them. These attacks are envisaged to increase by 45% in 2025 if it is not decisively addressed (Alshammari *et al*., 2023). However, diverse efforts made over the years to curb these real-world impacts of phishing attacks are frustrated as phishers source more sophisticated means to masquerade the anti-phishing solutions provided by well-known browsers, email service providers, and researchers (Chen *et al*., 2020; Baig et al., 2021). Likewise, it is revealed through the review that no single approach or strategy has accurately and seamlessly defended the targets against phishing attacks. The efforts at tackling phishing attacks have been a continuous cyberwar because website content is subject to change, attackers are becoming more insidious and the website's operating lifespan is also limited (Mohammad, 2020; Abutair *et al*., 2019).

Anti-phishing solutions, either origin-based or content-based filtering (Larasati *et al*., 2019), need to develop into real-time, more reliable, and effective anti-phishing agents capable of optimally detecting phishing attacks to safeguard users from being caught unaware (Osanaiye *et al*., 2016). With various

*Corresponding author: poolabisi@bellsuniversity.edu.ng
doi: http://dx.doi.org/10.63746/njtd.v22i3.

efforts at identifying and preventing phishing attacks, Safi and Singh (2023) in their study, extensively reviewed and discussed five different approaches on which anti-phishing methods are based, which are: the heuristic techniques, visual similarity-based technique, list-based technique, machine learning techniques and deep learning technique.

Beyond conventional phishing solutions, which have become ineffective due to the continuously evolving phishing activities, a machine learning (ML) approach is being deployed for overwhelming performance (Zamir *et al*., 2020; Abdul Samad *et al*., 2023). Machine learning techniques have the potency to explore the morphology and semantic contents of the email or URL structure of a webpage to optimally differentiate between phishing and legitimate. Though ML techniques have shown striking efficiency in detection challenges, some obstacles are still sabotaging the performance compared to their subset (Deep learning) (Al-Shalabi & Jazyah, 2024).

Of the steps involved in ML: data collection and representation, feature selection, mapping (training), and making a good classification, feature selection is recognized to be an important phase, where redundant, irrelevant, and superfluous features are expunged (Bellapu *et al*., 2021). Therefore, the benefits of improved feature selection include increased model accuracy and training time, complexity cost, interpretability, space, and decreased over-fitting.

Most anti-phishing researchers have focused more on optimising classification algorithms (Chiew *et al*., 2019), which are later exposed to poor-quality data that hampers the detection performance of such models. Some other adopted approaches to improve ML algorithm performance include: hybridization, ensemble, modification, and so on, which are inefficient when the feature selection is not given adequate attention in the whole process (Kumar *et al*., 2020).

In order to find a collection of efficient features, a reliable feature selection framework is required with three kinds of such feature selection techniques generally revealed in literature, namely: wrapper, filter, and embedded techniques (Bellapu *et al*., 2021). The time complexity of the wrapper method is always high due to the involvement of an algorithm in the attribute selection process. This limits its usage, as it is computationally tensed for real-time applications. Likewise, the embedded techniques require special hyperparameter tuning for performance improvement, which is also computationally higher than filter methods.

In the course of this work, the authors discovered that studies in this field generally applied one feature selection method or compared the efficacy of two feature selection methods independently in order to select features for their models. Thess were recognised to have been carried out in the problem domain and not significant for feature generalization. This study focused on expediting the classification prowess of the classical algorithms to compete with deep learning models for phishing detection. To address this gap, the authors devised a technique to heterogeneously combine the selection power of three individual filter-based techniques based on their application in prior studies for feature generalization for the classification model. The developed technique known as Heterogeneous Borda Ensemble Feature Selection (HBEFS), serves for the optimal generation of an indicator cut-off selection to secure the network against URL phishing attacks.

Relevant literature reviewed include the work of Siva *et al*.. (2024). The work contributes to phishing detection by proposing PhishPatro, which is an integration of classical machine learning algorithms (SVM and Decision tree) with deep learning models (RNN and GNN). The approach leveraged lexical and host-based URL features to classify them as either legitimate or malicious. Despite that, the study reported that the proposed system can enhance detection accuracy and maintain low False Positive value, but the system computation is high. Chiew *et al*. (2019) recognized the importance of feature selection steps to ML performance and developed a hybrid feature selection framework, which was applied to the phishing variables. They evaluated the baseline features generated using six ML classifiers, which showed the Random Forest (RF) classifier outperformed other classifiers by achieving a detection accuracy of 94.6%. In the study conducted by Fagbola *et al*., (2012), an optimization approach was applied to the development of a phishing detection system to improve on the model's weakness. A Support Vector Machine (SVM) was optimised with a Genetic Algorithm (GA) to overcome inaccurate classification due to the data dimensionality in terms of variables, and the model achieved accuracy of 93.5%.

The experimental study of Rashid *et al*. (2024) was carried out as part of efforts at addressing significant online threats posed by phishing. Their work explored the potential of machine learning algorithm, namely: Multilayer Perceptron Neural Network, to identify malicious websites. The model was exposed to two different datasets for thorough evaluation. The developed model was compared against two classical learners: the decision tree and Naïve Bayes. The results revealed that the developed model was significantly effective for the detection of phishing attacks.

The study by Moedjahedy *et al*. (2022) focused on feature selection, where a combined correlation filter-based statistical technique was considered for the identification of phishing website variables for the ML models. Two datasets were used, comprising 87 and 48 variables, respectively. With 10 reduced variables from the two datasets a best accuracy of 95.88% was achieved. The selection of features from 40 URL indicators available in the crawled dataset was investigated (Igwilo and Odumuyiwa, 2022). The indicator analysis was carried out with ReliefF algorithm, and the features ranked based on their importance. The selected URL features were tested and evaluated against different ML classifiers. The results obtained show that RF performed better on the variables selected.

The study conducted by Zhou *et al*. (2020) proposed a feature selection method termed Weight Composition of Feature Relevancy (WCFR). The algorithm proposed used standard deviation to allocate weights to relationships existing in the features and sub-features. Likewise, in (Larasati *et al*., 2019; Anandita *et al*., 2021), a feature selection strategy majorly used in their study for high-dimensional records, termed fast correlation-based filter, was proposed. This feature selection algorithm is capable of removing irrelevant and redundant data from the dataset used. They evaluated the features obtained from their proposed feature selection

framework on two ML models, that is, C4.5 and Naïve Bayes. In (Nam *et al*., 2023), the authors proposed ML-based models for predicting phishing attacks, with five ML classifiers employed for building the models. The researchers used a phishing dataset sourced from the UCI ML repository and three major standard evaluation metrics, namely: accuracy, sensitivity, and specificity.

The overview of this study's contributions is stated as follows:

1. The novel Heterogeneous Borda Ensemble Feature Selection technique can extract frequent and correlated phishing features, feature-to-feature, and highly correlated phishing features with the target class.
2. The selection of features using a filter-based statistical analysis technique is achieved.
3. Development of a novel Heterogeneous Borda Ensemble Feature Selection to detect malicious webpage features is achieved.
4. Significant reduction was achieved in the number of selected features by individual techniques.
5. Elimination of unnecessary and redundant phishing features through the novel HBEFS method significantly improved the classification performances of the traditional machine learning algorithms.

The remaining contents of this paper are sectioned as follows: discussion of the methodologies and model architecture, experimental setup, results and discussion, and finally, the conclusion and recommendation of future works.

II. METHODOLOGY

This research is aimed at strengthening existing cybersecurity traditional models and providing more reliable protection against increasingly sophisticated phishing attacks through the developed novel feature selection technique. Efforts were deployed at identifying an efficient set of features in the phishing dataset in order to attain a commendable accuracy and reduce detection computation. The proposed phishing detection architecture shown in Figure 1 comprises three modules: the first module deals with collection and pre-processing of data, the second module is the HBEFS phase, while the last is the evaluation phase. These modules are discussed in subsections that follow.

A. Data Collection/Pre-processing

The URL phishing dataset used for conducting experiments in this paper is accessible at www.kaggle.com/datasets/ shashwatwork/phishing-dataset-for-machine-learning. It was extracted and prepared in comma-separated version (CSV) format by Chiew et al (2019). It has 48 indicators, with 5000 phishing and legitimate records, respectively, and is classified into three groups, namely: Address bar-based, Abnormal-based, and HTML/JavaScript-based features. The characteristics of the dataset are presented in Table 1. Conducting a detailed Exploratory Data Analysis (EDA) on the dataset, the phishing values were found to be skewed.

Value variation is a crucial problem in ML environments. It distorts the model's performance. It was handled with a rescale as in (1) (Ali and Malebary, 2020), to keep the varied value range between 0 and 1.

$$V' = \frac{V - V_{min}}{V_{max} - V_{min}} \tag{1}$$

where, V is the rescaled value, V_{min} is the minimum and V_{max} is the maximum value present in the dataset.

Table 1 Data observation/characteristics

Observation	Value
Missing values	No
Input features	Numeric
Target Class	Binary
Records	10,000
Attributes	48

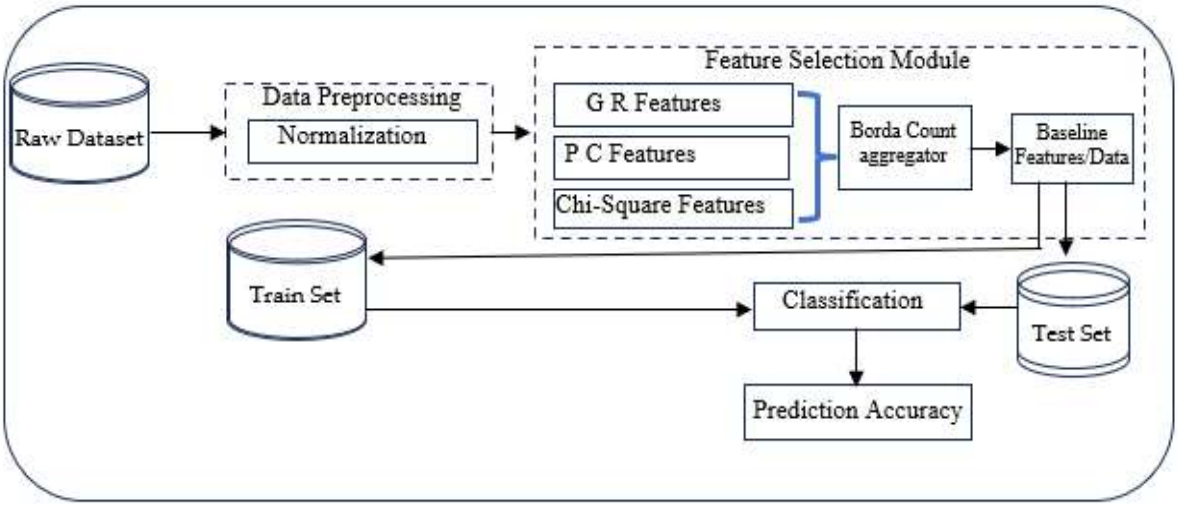

Figure 1 Overall Proposed Phishing Mechanism

B. Proposed Ensemble Feature Selection

ML tasks require an appropriate selection of a feature subset that reduces the feature space and ensures detection accuracy is high. To achieve this, an optimal cut-off rank, which determines the threshold rank in an organised order of filter-measured values was identified. Most of the existing studies (Chiew *et al*., 2019; Bellapu *et al*., 2021; Basnet *et al*., 2012) had challenges in determining the cut-off, with their cut-off ranks found to be unstable and not robust. This study devises a new strategy that is more flexible in selecting an optimal feature from different filter measures. The description of the Heterogeneous Borda Ensemble Feature Selection (HBEFS) method proposed for the selection of optimal phishing features relies on three statistical techniques: Gain Ratio (GR), Chi-Square ($\chi 2$) and Pearson Correlation Coefficient (PCC), as presented in Figure 2, with their benefits and limitations also considered.

The phishing dataset, $Q = \{q_1, q_2, \ldots q_n\}$. $R = \{r_{1,} r_{2,} \ldots r_{n,}\}$ represent the class target, and the three statistical techniques are denoted as FP_k, FP_t & FP_j (GR, PCC, & CHI2). FP_k, is applied to the data to generate value for each of the features available in the set, and applies information theories as in (2–5) (Chiew et al, 2019) to generate/rank the attributes. The values $\{\emptyset_{1,}k, \emptyset_{2,}k, \ldots., \emptyset_{j,}k,\}$ are generated according to their importance with respect to FP_k as an improvement over the information gain method. Secondly, the Correlation feature selection method is one of the filter-based techniques that could be used to minimize the search space dimensionality and to select the predictive power of features in a classification dataset.

Also, FP_t (PCC) filter measure is applied to the same dataset to measure the variables using its method. The FP_t (6) is applied to measure the relationship between the independent variables and the target class (Aljofey *et al*., 2021). Any two independent variables that correlate are dropped, and a set of ranked filter measure values $\{\varphi_{1,}k, \varphi_{2,}k, \ldots., \varphi_{j,}k,\}$ is generated. The third filter predictor FP_j (CHI2), is also applied. It measures the divergence from the distribution. FP_j is applied on the Q and R with a set of $\{\alpha_{1,}k, \alpha k, \ldots., \alpha_{j,}k,\}$ values generated using (7).

$$E(T_v) = -\sum_{i=1}^{m} P_{Tv} \log_2 P_{Tv} \tag{2}$$

$$IG(f_k) = Ent(T) - info_{fk}(T) \tag{3}$$

$$Splitinfo_{fk}(T) = -\sum_{v=1}^{n} \frac{|T_v|}{T} \log_2 \left(\frac{|T_v|}{|T|}\right) \tag{4}$$

$$FP_k = \frac{IG(T, f_k)}{SplitInfo_k(T)} \tag{5}$$

where, $SplitInfo_{fk}(T)$ represents split information value generated by splitting the sample set T into p partitions corresponding to p distinct subsets on the feature f_k, and $G.R$ represents the Gain ratio, which is the fraction of $IG(T, f_k)$ and $SplitInfo_{fk}(T)$.

$$FP_t\,(k,t) = \frac{cov(q,r)}{\sigma q \sigma r} \tag{6}$$

where, q is the phishing independent variable, t is the phishing target class, $cov(q,r)$ is for the covariance of q and r, σq, σr are the standard deviations of q and r, respectively.

$$FP_j{}^2 = \sum_{i=1}^{m}\sum_{j=1}^{k} \left(\frac{A_{i,j} - \left(\frac{R_i * c_j}{N}\right)^2}{\frac{R_i * c_j}{N}} \right) \tag{7}$$

where, m is the attribute magnitude in the phishing dataset, k is the size of classes in the dataset, N is the total size of samples in the dataset, R_i is the size of the patterns in the i^{th} attribute, C_j is the size of the patterns in the J^{th} class, and A_{ij} is the size of the patterns in the i^{th} attribute and the J^{th} class. The features that obtained higher Chi-square scores were fetched.

The three filter measures generate lists of feature rankings $\{\emptyset_{n,}k, \varphi_{n,}k, \alpha_{j,}k,\}$ using FP_k, FP_t, FP_j respectively from the original dataset, $Q = \{q_1, q_2, \ldots q_n\}$, $R = \{r_{1,} r_{2,} \ldots r_{3,}\}$. A threshold is used to differentiate high from low-impact features on the phishing dataset.

This study used ranker's feature selection methods, and it is paramount to set a cutoff point to find relevant features from the individual rankers. Most of the existing studies applied a percentage as the threshold to retain features. The approach adopted in this study obtains the individual complexity measures for features of the phishing dataset, and the final subset of features is established according to Eqn. (8):

$$e = \alpha \times CM + (1\text{-}\propto) \times \rho \tag{8}$$

where α with value interval [0,1] regulates the error and the features, CM is the complexity measure, and ρ is the attributes retained with value in the interval [0,1].

The feature percentages are calculated for batches of $\log_2(n)$ features and e is calculated for the batch of $2 \times \log_2(n)$ to select the best feature for both. Features that measure values below the threshold are marked as not important, whereas those within and above the threshold are selected. The features obtained at this level are known to be the primary informative feature subsets. To obtain the baseline features, a Borda count algorithm was applied. It combines the three primary informative features and selects the secondary informative feature subsets known as Baseline features. The Borda count aggregator algorithm is computed using (9):

$$b_i = \sum\nolimits_{v]} N_f - P_v \tag{9}$$

where, b_i denotes Borda count, N_f represent the entire quantity of features, P_v is the position of the i^{th} attribute in an ordered list produced by the v^{th} ranker and $i = 1, \ldots. N_f$.

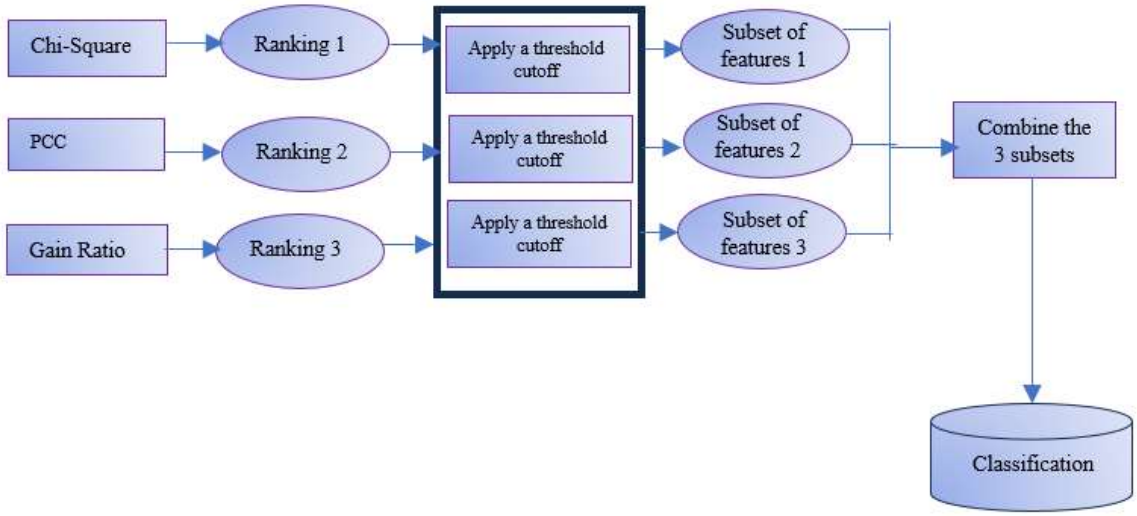

Figure 2 Filter-Based Ensemble Feature Selection Framework

Algorithm 1 Heterogeneous Borda Ensemble Feature Selection (HBEFS) Algorithm

Algorithm: Ensemble feature selection

Input:

FSM = $\{FSM_i | i = 1,2, \ldots . . q\}$; // where q = {Gain Ratio, Chi-Square, PCC}

Train D = $\{x_i y_i | i = 1,2, \ldots . . t\}$; // t denotes the number of phishing samples of train data

$F = \{f_1, f_2, \ldots f_{k=48}\}$; //k denotes the number of phishing features

Output: SubFeatures (SubFs)

Begin

for t = 1, 2, …, q do

Step 1: RankFs$_i$ = Feature selector (TrainD, FSM_i, k)

Step 2: Compute borda point = $b_i = \sum_{v|} N_f - P_v$

end for

Step 3: RankFs$_i$ based on Borda count in decreasing order

Step 4: Select final features based on a threshold value (SubFs)

End

C. Phishing Attacks Classification Algorithms

Phishing detection as a classification problem, necessitates the use of four supervised classification algorithms, namely: Naïve Bayes, Support Vector Machine, Logistic Regression, and Random Forest. These classification models were selected based on their popularity for classification tasks and their prior performance as observed in literature (Wibowo *et al*., 2024; Berliana and Riasetiawan, 2023; Putri *et al*., 2024; Munmun *et al*., 2025). The authors intended to use the feature selection phase to improve their classification performances to compete with the deep learning models. The classification algorithms create and approximate a map function (f) from phishing training data points (P) to discrete target labels (y).

D. Experimentation and Performance Evaluation

For the model evaluation, two experiments were conducted to validate the performance of the proposed feature selection framework. The dataset was divided into 80:20 ratio of train and test data subsets. This data subset ratio was chosen contrary to other ratios like the 70:30 or 75:25 (Masih and Kushwaha, 2023) for the models to be familiar with the data for better understanding and performance. The first experiment was conducted with the phishing variables selected by individual filter-based statistical techniques, that is, Chi-square, Pearson correlation coefficient, and Gain ratio, while the second experiment was conducted on the baseline features selected by the HBEFS method. The experiments were conducted using Python 3.7, which contains robust machine-learning algorithm libraries. The performance of the phishing detection technique being proposed is evaluated with the use of the K by K dimensional confusion matrix parameters, namely: True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). These confusion matrix parameters are used for classification tasks based on standard metrics, namely: recall, accuracy, precision, and F1-score.

III. RESULTS AND DISCUSSION

Results of the model evaluation conducted with the two experiments are hereby presented.

A. Experiment 1: Model Evaluation Using Variables from Each Filter-Based Technique.

The results of using the three identifiers to select variables, and the selected traditional classification algorithms evaluated on the variables are presented as follows. The Chi-square was applied individually to the dataset, and features were selected based on their ranking importance. Out of the 48 attributes available in the dataset, the best 14 selected features are presented in Table 2. The phishing models were applied to the features, and their prediction efficacies are presented in Table 3.

To determine the significance of the proposed feature selection technique for this work, the Gain Ratio was applied to the dataset as a standalone statistical technique, based on the ranking, The best 25 features out of the 48 features recognised based on their ranking are presented in Tables 4, while the prediction outcomes are presented in Table 5.

The third feature selection method, Pearson Correlation Feature Selection (PCFS) technique, was also applied directly to the raw phishing dataset, and 16 features recognised to be the best to represent the entire dataset based on their ranking order are as presented in Table 6. The classification capacity of the phishing models conducted on the best-selected features are presented in Table 7.

Table 2 Best Features Selected through Chi-Square

S/N	Features
1	PctExtHyperlinks
2	PctExtResourceUrls
3	FrequentDomainNameMismatch
4	PctExtNullSelfRedirectHyperlinks
5	NumDashInHostname
6	PctNullSelfRedirectHyperlinks
7	FrequentDomainNameMismatch
8	EmbeddedBrandName
9	HostnameLength
10	UrlLengthRT
11	PathLevel
12	InsecureForms
13	UrlLength
14	SubmitInfoToEmail

Table 3 Performance of the models using Chi-square variables

PARAMETERS	SVM	LR	NB	RF
TP	901	747	828	892
FN	87	241	160	96
FP	130	81	137	74
TN	882	931	875	938
ACCURACY	0.892	0.839	0.852	0.918
PRECISION	0.874	0.902	0.858	0.923
RECALL	0.912	0.756	0.838	0.903
FI-SCORE	0.893	0.823	0.848	0.913

Table 4 Features Selected by Gain Ratio Feature Se1ection Technique

S/N	Features
1	SubdomainLevel
2	NumDashInHostname
3	TildeSymbol
4	IpAddress
5	PctExtNullSelfRedirectHyperlinksRT
6	ExtMetaScriptLinkRT
7	PctExtResourceUrlsRT
8	ExtFormAction
9	SubdomainLevelRT
10	PopUpWindow
11	FrequentDomainNameMismatch
12	PctExtHyperlinks
13	FrequentDomainNameMismatch
14	SubmitInfoToEmail
15	NumSensitiveWords
16	DoubleSlashInPath
17	RandomString
18	DomainInPaths
19	IframeOrFrame
20	InsecureForms
21	AbnormalFormAction
22	TildeSymbol
23	NumHash
24	DomainInSubdomains
25	EmbeddedBrandName

Table 5 Performance of the Phishing Models using Gain Ratio Features

PARAMETERS	SVM	LR	NB	RF
TP	849	964	897	898
FN	139	24	91	90
FP	262	359	362	81
TN	750	653	650	931
ACCURACY	0.800	0.809	0.774	0.915
PRECISION	0.764	0.729	0.712	0.917
RECALL	0.859	0.976	0.908	0.969
FI-SCORE	0.809	0.834	0.798	0.913

Table 6 Features Selected through PCFS

S/N	Features
1	NumDots
2	NumDash
3	InsecureForms
4	PctNullSelfRedirectHyperlinks
5	FrequentDomainNameMismatch
6	SubmitInfoToEmail
7	PctExtNullSelfRedirectHyperlinksRT
8	EmbeddedBrandName
9	NumUnderScore
10	NumDashInHostname
11	ExrFormAction
12	ExtMetaScriptLinkRT
13	SubmitInfoToEmail
14	FrequentDomainNameMismatch
15	NumDashInHostname
16	PathLevel

Table 7 Performance of the models using PCFS variables

PARAMETERS	SVM	LR	NB	RF
TP	892	873	849	971
FN	96	115	139	17
FP	208	112	262	123
TN	804	900	750	889
ACCURACY	0.848	0.887	0.800	0.93
PRECISION	0.811	0.886	0.764	0.886
RECALL	0.903	0.884	0.859	0.983
FI-SCORE	0.854	0.885	0.809	0.933

B. Experiment 2: HBEFS Performance Evaluation Based on Baseline Features

The proposed HBEFS method was applied to the dataset to pick the best informative features, noting that the method is a combination of the three filter-based methods applied individually. The Borda Count aggregator was later applied to the predictions of the individual predictors to select the final baseline features, and it eliminates the problem of cut-off values encountered in previous studies. Out of the 48 features in the dataset, the proposed HBEFS technique selected 8 as the baseline features to represent the entire dataset, and are presented in Table 8 and shown in Figure 3. The performance of the phishing classification models validated on the baseline features is presented in Table 9 and shown in Figure 4.

Table 8 Baseline Features Selected by the Proposed HBEFS Technique

S/N	Features
1	FrequentDomainNameMismatch
2	SubmitInfoToEmail
3	PctExtNullSelfRedirectHyperlinksRT
4	ExtMetaScriptLinkRT
5	PctExtNullSelfRedirectHyperlinks
6	PctExtHyperlinks
7	Insecure form
8	PctExtResourceUrlsRT

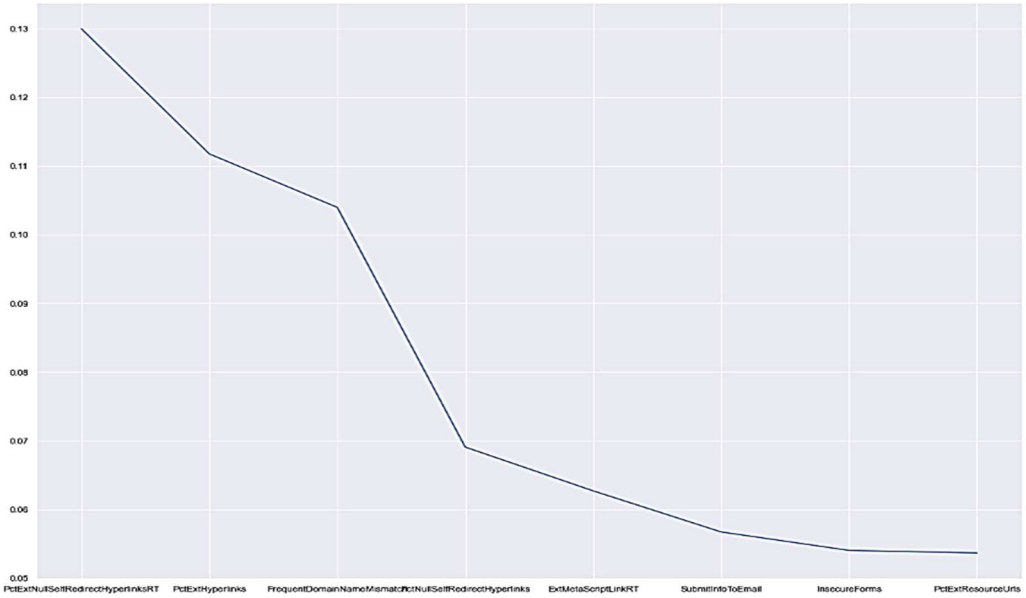

Figure 3 Final Baseline Phishing Features Selected by the HBEFS

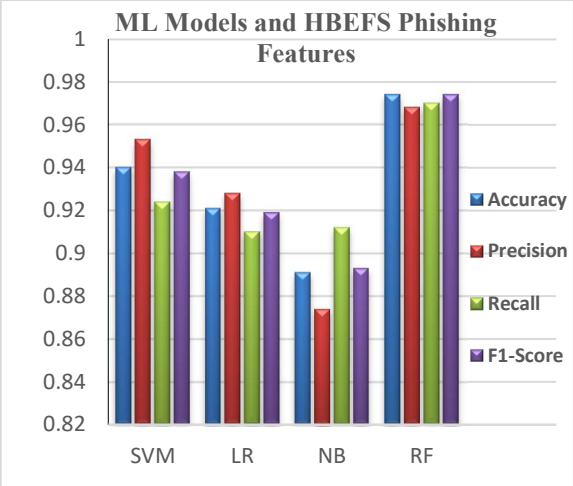

Figure 4 ML Classifiers using HBEFS

Table 9 Performance of the models using HBEFS features

PARAMETERS	SVM	LR	NB	RF
TP	913	899	901	968
FN	75	89	87	20
FP	45	70	130	32
TN	967	942	882	980
ACCURACY	0.94	0.921	0.891	0.974
PRECISION	0.953	0.928	0.874	0.968
RECALL	0.924	0.910	0.912	0.970
FI-SCORE	0.938	0.919	0.893	0.974

C. Discussion

This study aimed at enhancing the detection performance of the selected traditional algorithms through the proposed HBEFS technique in order to assist in the selection of optimal phishing features. The proposed technique encased three filter-based statistical techniques, and its performance was

established by applying the individual methods to the phishing dataset to select the best features based on their ranking. The model evaluation results presented in Table 3 revealed that the Random Forest ensemble classifier under the Chi-Square variables outperformed the single models with 91.8% accuracy, followed by the Support Vector Machine classifier recording 89.2%. Looking at other metrics apart from accuracies, SVM enjoyed the best True Positive of 901 and the highest Recall of 91.2% followed by RF with 90.3%, whereas RF has the highest Precision of 92.3% and F1-Score of 91.3%. RF could be seen as having outperformed the other classifiers.

Based on the Gain Ratio theory, the model evaluation performance on the features in Table 4 shows that among the four phishing classifiers, RF still recorded the highest accuracy of 91.5% but dropped compared to the previous performance enjoyed under chi-square features. There is a slight difference between the accuracy of SVM and LR with NB scoring the least value. LR has the highest True Positive rate of 964. Likewise, for the evaluation analysis conducted using the PCFS features, the results as shown in Table 7 reveals that the performance of RF moved further up to 93.0%. This is followed by Logistic regression and Support Vector Machine with 88.7% and 84.8% accuracy, respectively.

However, Table 9 presents the performance outcome of the phishing learning schemes validated on the baseline phishing features selected by the proposed HBEFS technique. The results revealed that all the classifier performances increased drastically. However, RF yielded outstanding performance compared to the other classifiers, recording a detection accuracy of 97.4%. The baseline features have a great impact on the shallow classifiers' performance with SVM and LR recording 94.0% and 92.1% accuracy respectively. These results established the fact and provided empirical evidence that the HBEFS technique has high potency to identify optimal phishing features on the network. Random Forest performed significantly under both the features obtained from individual statistical techniques and the proposed feature selection (HBEFS) technique. This is because Random Forest is an ensemble of decision trees, so it can capture complex, non-linear relationships in the data. Also, the model is enclosed with feature importance scores, which help in the interpretation and identification of features that can influence predictions. The proposed feature selection mechanism neutralises the computational challenge that the model could have encountered if not applied.

D. Comparison of the Proposed Framework

The performance of the classifier under the proposed framework is compared to the individual statistical techniques (Chi-square, Pearson & Gain ratio) as shown in Figures 5 – 9. The comparison is based on the four major metrics: accuracy, precision, recall, and F1-score. The performance of the Random Forest classifier under the pr Managing Editor oposed framework is compared with studies like (Masih and Kushwaha, 2023; Kumar *et al*., 2020; Osho *et al*., 2019; Azeez *et al*., 2021) as shown in Table 10 and Figure 8. The findings of this work established that the performance of Random Forest under BHEFS outperformed the existing studies. This revealed that the existing feature selection and classification systems are not as effective as the HBEFS. Furthermore, the significant performance obtained in this work implies that the proposed HBEFS is reliable and hence adaptable to both webpage and email semantic datasets.

The performance of the HBEFS could be seen from the evaluation metrics considered presented in Figures 5-8. The evaluation metrics for the HBEFS outperformed others by seeing the curves under the accuracy, precision, and F1-score, while the recall revealed that the curve for HBEFS dropped.

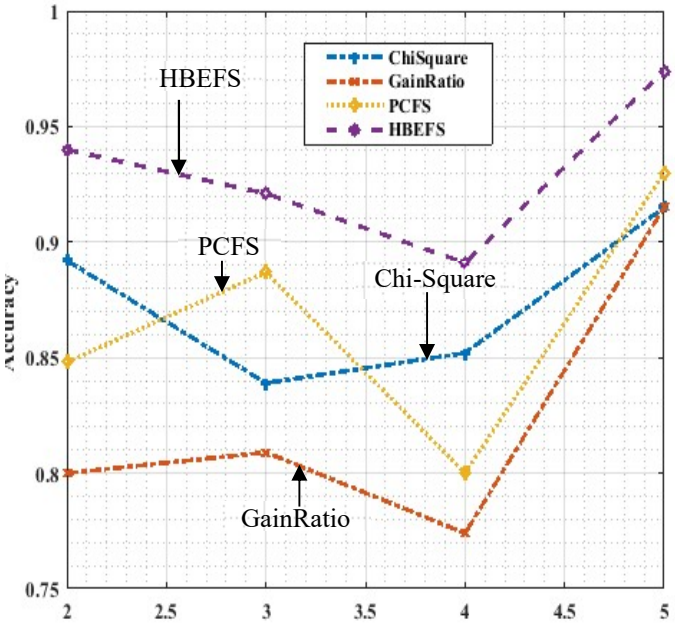

Figure 5 Plot of Accuracy parameter values

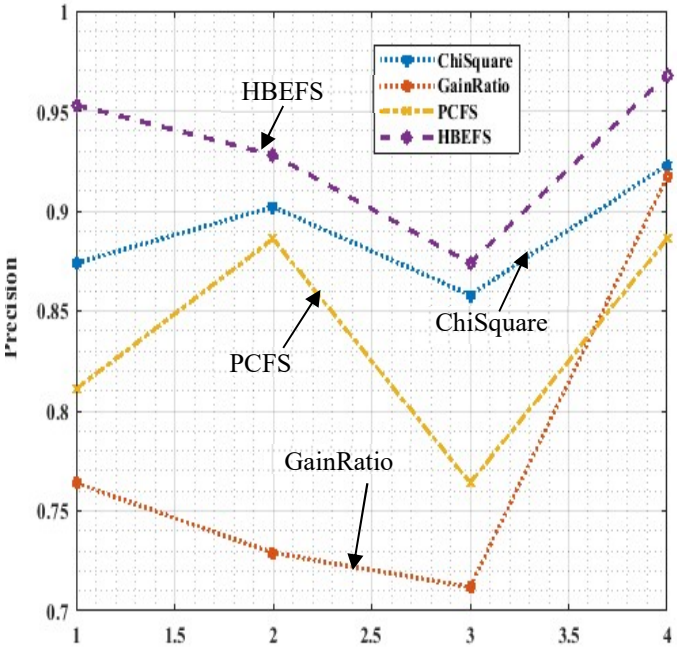

Figure 6 Plot of Precision parameter values

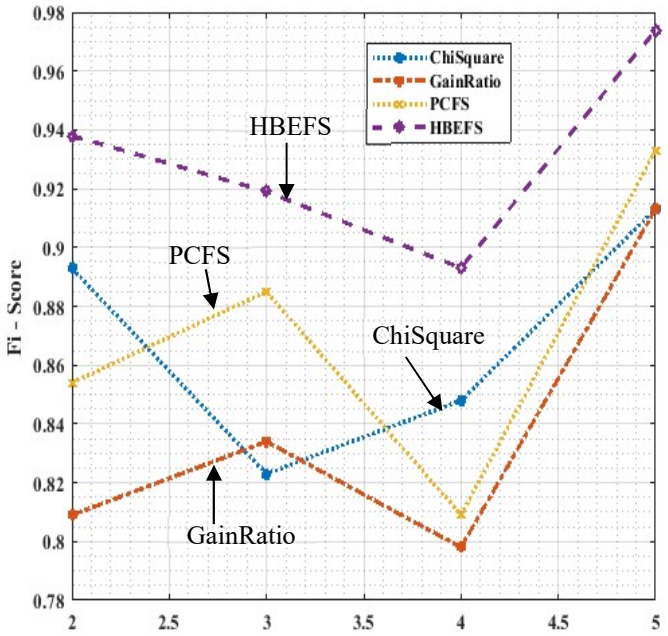

Figure 7 Plot of F1-Score parameter values

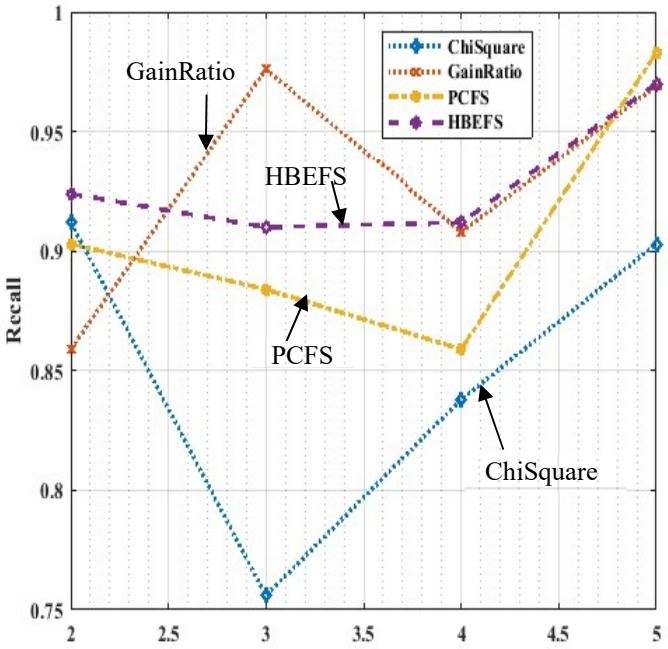

Figure 8 Plot of Recall parameter values

Table 10 Performance of Random Forest of HBEFS and other studies

	Accuracy	Precision	Recall	F1-Score
[19]	93.55	0.939	0.929	0.934
[14]	89.5	0.894	0.894	0.894
[4]	92.4	0.927	0.919	0.923
[24]	96.55	0.959	0.968	0.964
HBEFS	**97.4**	**0.968**	**0.970**	**0.974**

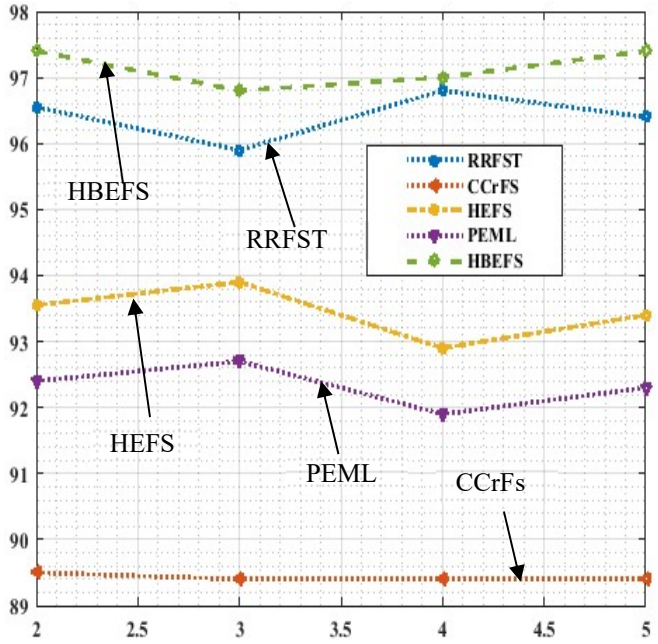

Figure 9 Comparison of the RF Classifiers

IV. CONCLUSION

The authors proposed a new feature selection framework, termed HBEFS, as a result of the resilience of phishing attacks. The existing filter indicator techniques are saddled with identifying effective subsets of attributes for utilization in ML-based phishing detection. As part of the feature selection framework proposed, a novel algorithm called Borda count was integrated to automate the aggregation process of the individual selected features to determine the optimal (baseline) features. Four phishing classification models were evaluated on each of the indicators identified through the individual statistical techniques, namely: chi-square, gain ratio, Pearson Correlation Coefficient, and baseline features. The proposed HBEFS approach helps to curtail the stress in identifying the cut-off limit usually faced in filter-based methods, and also to select optimal phishing features that can represent the whole features without reducing the quality of the dataset. The result of the experiments suggests that the optimal features produced by the HBEFS perform excellently when evaluated with RF and SVM classifiers, with an accuracy of 97.4% and 94.0%, respectively, using only 14.9% of the whole features present in the initial data, which is known as baseline features.

The authors compared the performance of the Random Forest phishing detection model of this study with the other four studies. It is shown that the result Random Forest obtained under this study using the baseline features outperformed the results obtained under those studies. Likewise, the traditional models' detection performance was also improved when compared to their performances in the previous studies. This shows that when models especially the traditional algorithms were exposed to redundant and uninformative features it hampered the efficacy of such models.

The rule of RF classifier and the HBEFS can be integrated into browsers or deployed to email servers to detect phishing attacks. The following are some of the importance of this

research: (i) this experiment has benchmarked a set of baseline phishing webpage features for machine learning phishing detection; (ii) enhanced the detection efficiency of machine learning phishing models; (iii) provided a flexible and automated feature selection system that can extract relevant features when operated on different large phishing datasets.

In the future, it may be beneficial to study the impact of feature selection using different techniques for classification, for example, associative classification. Also, feature transformation may be investigated as a feasible feature reduction strategy. Furthermore, anti-phishing research efforts in the future could disregard the use of outmoded features. There are several other ways this particular system can be enhanced, and the researchers are looking forward to applying the HBEFS to solve multi-variable problems in ensemble classification.

The results demonstrate promising potential for using HBEFS in phishing detection, though further refinement may be needed to address emerging attack patterns. The hyperparameters of the SVM algorithm and logistic regression could also be tuned to cope with outliers and scalability in high dimensions that can be easily handled by Random Forest. More phishing-related datasets could also be tested on the model for generalisability into various concerns which include dataset biases, multiple datasets, and non-URL phishing. Lastly, the models could also be validated through the use of a cross-validation approach.

AUTHORS CONTRIBUTIONS

P. O. Olabisi: conceptualisation, writing, review and editing of draft, preparation of manscript; **B. M. Olukoya:** literature review, methodology design, software, writing of draft; **G. O. Ogunleye:** conceptulisation and supervision; **O. J. Adetunji:** analysis, curation of dataset.

ACKNOWLEDGEMENT

This research was not supported or sponsored by any individual, institution or research granting body, other than the self-efforts of the authors.

REFERENCES

Abdul Samad, S. R.; S. Balasubaramanian; A. S. Al-Kaabi; B. Sharma; S. Chowdhury; A. Mehbodniya; J. L. Webber and A. Bostani. (2023). Analysis of the Performance Impact of Fine-Tuned Machine Learning Model for Phishing URL Detection. Electronics, 12(7): 1642, doi: 10.3390/electronics12071642.

Abutair, H.; A. Belghith and S. AlAhmadi. (2019). CBR-PDS: a case-based reasoning phishing detection system, J. Ambient Intell. Humaniz. Comput., 10(7): 2593–2606.

Al-Shalabi, L. and Jazyah, Y. H. (2024). Phishing detection using clustering and machine learning. IAES International Journal of Artificial Intelligence, 13(4): 4526–4536.

Ali, W. and Malebary, S. (2020). Particle Swarm Optimization-Based Feature Weighting for Improving Intelligent Phishing Website Detection. IEEE Access, 8: 116766–116780.

Ali, S.; Rehman, S. U.; Imran, A.; Adeem, G.; Iqbal, Z. and Kim, K. Il. (2022). Comparative Evaluation of AI-Based Techniques for Zero-Day Attacks Detection. Electronics (Switzerland), 11(23).

Aljofey, A.; Q. Jiang; Q. Qu; M. Huang and J. P. Niyigena. (2020). An effective phishing detection model based on character level convolutional neural network from URL, Electronics, 9(9): 1–24.

Alshammari, G.; M. Alshammari; T. S. Almurayziq; A. Alshammari and M. Alsaffar. (2023). Hybrid Phishing Detection Based on Automated Feature Selection Using the Chaotic Dragonfly Algorithm. Electronics, 12(13): 2823.

Alshingiti, Z.; R. Alaqel; J. Al-Muhtadi; Q. E. U. Haq; K. Saleem and M. H. Faheem. (2023). A Deep Learning-Based Phishing Detection System Using CNN, LSTM, and LSTM-CNN. Electronics. 12(1): 232.

Amusan, E. A.; O. T. Adedeji; O. Alade; F. A. Ajala and K. O. Ibidapo. (2021). A Mobile Anti-Phishing System Using Linkguard Algorithm. FUOYE Journal of Engineering Technology, 6(3): 10–14, doi: 10.46792/fuoyejet.v6i3.666.

Anandita, Yadav, D. P.; Paliwal, P.; Kumar, D. and Tripathi, R. (2021). A Novel Ensemble Based Identification of Phishing E-mails. ACM International Conference Proceeding Series.

Azeez, N. A.; S. Misra; I. A. Margaret; I. Fernandez-Sanz and S. M. Abdulhamid. (2021). Adopting automated whitelist approach for detecting phishing attacks. Computer Security 108 Sep. 2021.

Baig, M. S.; F. Ahmed and A. M. Memon. (2021). Spear-Phishing campaigns: Link Vulnerability leads to phishing attacks, Spear-Phishing electronic/UAV communication-scam targeted. In Proceedings of IEEE 4th International Conference on Computing and Information Sciences, (ICCIS 2021). In book: Advances in Computing and Data Sciences.

Basit, A.; M. Zafar; A. R. Javed and Z. Jalil. (2020). A Novel Ensemble Machine Learning Method to Detect Phishing Attack. Proc. - 2020 23rd IEEE Int. Multi-Topic Conf. (INMIC), 05-07 November, 2020, Bahawalpur, Pakistan.

Basnet, R. B.; A. H. Sung and Q. Liu. (2012). Feature selection for improved phishing detection. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 7345: 252–261.

Bellapu, R. P.; R. Tirumala and R. N. Kurukundu. (2021). Evaluation of homogeneous and heterogeneous distributed ensemble feature selection approaches for classification of rice plant diseases. In Proc. of 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India. 06-08 May 2021: 1086–1094.

Berliana, E. V. and Riasetiawan, M. (2023). Comparative Analysis of Naïve Bayes Classifier, Support Vector Machine and Decision Tree in Rainfall Classification Using Confusion Matrix. International Journal of Advanced Computer Science and Applications (IJACSA), 15(7): 560-567.

Chen, J. L.; Y. W. Ma and K. L. Huang. (2020). Intelligent visual similarity-based phishing websites detection. Symmetry (Basel). 12(10): 1–16.

Chiew, K. L.; C. L. Tan; K. S. Wong; K. S. C. Yong

and W. K. Tiong. (2019). A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. Information Science. 484: 153–166,

Fagbola, T.; S. O. Olabiyisi; A. Abimbola Baale; O. Stephen and A. Abimbola. (2012). Hybrid GA-SVM for Efficient Feature Selection in E-mail Classification Machine Learning View project Development of Neuro-Fuzzy Model for Evaluation E-Learning Activities View project Hybrid GA-SVM for Efficient Feature Selection in E-mail Classification. Online, 2012. [Online]. Available: www.iiste.org

Gualberto, E. S.; R. T. De-Sousa; T. P. B. De-Vieira; J. P. C. L. Da-Costa and C. G. Duque. (2020). From Feature Engineering and Topics Models to Enhanced Prediction Rates in Phishing Detection. IEEE Access, 8: 76368–76385.

Igwilo, C. M. and Odumuyiwa, V. T. (2022). Comparative Analysis of Ensemble Learning and Non-Ensemble Machine Learning Algorithms for Phishing URL Detection. FUOYE J. Eng. Technol., 7(3): 305–312,

Khan, S. A.; W. Khan and A. Hussain. (2020). Phishing Attacks and Websites Classification Using Machine Learning and Multiple Datasets (A Comparative Analysis), Lect. Notes Comput. Sci. 12465: 301–313.

Kumar, N.; S. Sonowal and Nishant (2020). Email Spam Detection Using Machine Learning Algorithms. Proc. 2nd 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 15-17 July 2020. Coimbatore, India. 108–113.

Larasati, U. I.; M. A. Muslim; R. Arifudin and A. Alamsyah. (2019). Improve the Accuracy of Support Vector Machine Using Chi Square Statistic and Term Frequency Inverse Document Frequency on Movie Review Sentiment Analysis. Scientific Journal Informatics, 6(1): 138–149,

Masih, P. and Kushwaha, S. (2023). Comparison of Logistic Regression, Naive Bayes and Random Forest Classifier Methods for Drug Review. Educational Administration: Theory and Practice 2023, 29(3): 381-388 ISSN: 2148-2403 https://kuey.net/

Moedjahedy, J.; A. Setyanto; F. K. Alarfaj and M. Alreshoodi. (2022). CCrFS: Combine Correlation Features Selection for Detecting Phishing Websites Using Machine Learning, Futur. Internet, 14(8), Aug. 2022,

Mohammad, H. T. (2020) Intelligent Rule-based Phishing Websites Classification. IET Inf. Secur., 8: 153–160

Munmun, Z.S., Akter, S. and Parvez, C.R. (2025) Machine Learning-Based Classification of Coronary Heart Disease: A Comparative Analysis of Logistic Regression, Random Forest, and Support Vector Machine Models. Scientific Research Publishing Open Access Library Journal, 12: e13054.

Nam, S. G.; Y. Jang; D. G. Lee; and Y. S. Seo. (2022). Hybrid Features by Combining Visual and Text Information to Improve Spam Filtering Performance. Electronics, 11(13): 2053,

Osanaiye, O., H. Cai; K. K. R. Choo; A. Dehghantanha; Z. Xu and M. Dlodlo. (2016). Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing, Eurasip J. Wirel. Commun. Netw., 1.

Osho, O.; A. Oluyomi; S. Misra; R. Ahuja; R. Damasevicius and R. Maskeliunas. (2019). Comparative evaluation of techniques for detection of phishing URLs, CCIS. Springer, Cham. 1051: 385-394.

Putri, A. I.; Nur A.; Husna, N. M.; Cia, M. A.; Arba, N. R.; Aisyi, C. H.; Pramesthi, A. S. and Irdayusman (2024). Implementation of K-Nearest Neighbors, Naïve Bayes Classifier, Support Vector Machine and Decision Tree Algorithms for Obesity Risk Prediction. Public Research Journal of Engineering, Data Technology and Computer Science. 2(1) July 2024: 26-33. Institute of Research and Publication Indonesia (IRPI) Public Research

Rashid, M. K.; Mousa, A. A. and Hassan, S. A. H. (2024). 2024 - Virtual Conference in Person Edition. 996–998.

Safi A. and Singh S. (2023). A systematic literature review on phishing website detection techniques. Journal of King Saud University – Computer and Information Sciences, 35 (2023): 590–611.

Siva, N.; Sivaiah, B. V.; Reddy, S. S. K.; Irfan, S.; Kumar, U. P. and Nayagara, S. N. (2024). Phishing Detection System through Hybrid Machine Learning Based on URL. 2024 5th International Conference for Emerging Technology, INCET 2024, 13(4).

Wibowo, J. S.; Wahyudi, E. N. and Listiyono, H. (2024). Performance Comparison of SVM, Naive Bayes, and Random Forest Models in Fake News Classification. Engineering and Technology Journal e-ISSN: 2456-3358 August-2024. 9(8): 4799-4804.

Zamir A.; Khan, H. U.; Iqbal, T.; Yousaf, N.; Aslam, F.; Anjum, A. and Hamdani M. (2020). Phishing web site detection using diverse machine learning algorithms, Electron. Libr., 38(1): 65–80.

Zhou, H.; X. Wang and Y. Zhang. (2020). Feature selection based on weighted conditional mutual information. Applied Computer and Informatics. 20(1/2): 55-68.