Heterogeneous Ensemble Feature Selection and Multilevel Ensemble Approach to Machine Learning Phishing Attack Detection

[1]Gabriel O. Ogunleye, *[2] B.M. Olukoya, [2]A.T. Olusesi, [2]Patrick Olabisi,[2]Queen B. Sodipo and [3]Adekunle Osobukola
[1]Department of Computer Science, Federal University, Oye-Ekiti, Nigeria
[2]Department of Computer Engineering, Bells University of Technology, Ota, Nigeria
[3]Department of Electrical/Electronic & Telecommunication, Bells University of Technology, Ota, Nigeria
gabriel.ogunleye@fuoye.edu.ng|{bmolukoya|atolusesi|poolabisi|qbsodipo|aaosobukola}@bellsuniversity.edu.ng

Received: 08-SEP-2023; Reviewed: 17-OCT-2023; Accepted: 27-OCT-2023
http://doi.org/10.46792/fuoyejet.v8i4.1105

ORIGINAL RESEARCH

Abstract- Over the decade, technology has presented human facets with easiest means of accomplishing complex tasks seamlessly, especially in the area of communication. Malicious and vicious links are consciously doctored to resemble the original and sent through emails to millions of users at once at a lower price. Since the emergence of phishing and its cohorts, every solution and means to mitigate the attacks has proven unsuccessful due to the dynamic nature of the attacks. Meanwhile, machine learning (ML) is adopted as the right antidote to phishing detection, with its performance based on diverse steps, especially feature selection. Most studies in the problem domain concentrate more on model optimization than sourcing for a reliable feature selection system and fail to integrate a reliable feature selection along with the classification model. The systems are fed with low-quality data that hampers the performance of such models. The authors noticed the contribution of feature selection to the performance of machine learning models and developed a novel Heterogeneous Ensemble Feature Selection (HEFS) framework for multilevel ensemble machine learning-based phishing detection. In HEFS, three filter-based statistical techniques were exploited to produce a primary subset of phishing features, and the variable selected by each of the techniques was automatically aggregated to produce the baseline features. The selection of the techniques is to overcome each limitation since their ranking principles are different. The experiment revealed that the multilevel ensemble (stacked) on the baseline features outperformed others with an accuracy of 98.8%., including multilevel model on each filter-based method.

Keywords- Feature selection, Multilevel ensemble, Machine Learning, Phishing Attack, Uniform Resource Locator

1 INTRODUCTION

The rapid development of technologies has transfigured a lot of human conventional activities such as banking, booking, news, research, and commerce into cyberspace (Sahingoz *et al.,* 2019). This process made modern society solely dependent on the World Wide Web and the Internet, both serving as essential additives to human endeavours. This astronaut development gave way to hosting various beneficial web applications and also attracted unprecedented internet users globally to access the applications online (Abdul Samad *et al.,* 2023). The platform also creates an avenue for cyber-attacks to be perpetrated, creating severe security risks for both experienced and novice users due to the internet's open and unregulated architecture (Sahingoz *et al.,* 2019, Mao *et al.,* 2019).

The most notorious cyber-attack recognized among others challenging the network/cyberspace presently is phishing. Phishing is a form of cybercrime, cybercriminals perpetrated by presenting themselves as genuine individuals in an illegal attempt to steal sensitive information from potential users through spoofed websites and emails (Amusan et al., 2021).

*Corresponding Author

Section B- ELECTRICAL/COMPUTER ENGINEERING & COMPUTING SCIENCES
Can be cited as:
Ogunleye G. O., Olukoya B.M., Olusesi A.T., Olabisi P., Sodipo Q.B. and Osobukola A. (2023): Heterogeneous Ensemble Feature Selection and Multilevel Ensemble Approach to Machine Learning Phishing Attack Detection. *FUOYE Journal of Engineering and Technology* (FUOYEJET), 8(4), 438-447. http://doi.org/10.46792/fuoyejet.v8i4.1105

This attack applies the lure, hook, and catch methods to steal internet user's information and sensitive credentials via fake emails or URL links (Al-Sarem *et al.,* 2021). The user's awareness and skill cannot prevent them from falling into the phishing bait, due to the fact that attackers get prepared to conduct successful phishing attacks by studying the personality characteristics of the users (Azeez *et al.,* 2021, Sahingoz *et al.,* 2019).

The attacks have raised more alarm and over the years gained the attention of researchers (Lakshmi *et al.,* 2021). The attackers create phony, fraudulent websites that have a similar look to that of the original hosted sites to scam potential end-users (Niranjan *et al.,* 2020). The fact that both phony and legal sites have similar interfaces, but with different Uniform Resource Locators (URLs), one can easily differentiate between the legal and attacker's phony sites through a thorough study of the URLs (Le-Nguyen *et al.,* 2023). Most often people fall into this trap because they refuse to thoroughly check and compare the structure of the URL link received via email or other social media tools as demonstrated in Figure 1. The phishers mostly lure the victims to steal their vital personal information through spoofed and doctored URLs (Gangavarapu *et al.,* 2020, Niu *et al.,* 2018). Some of the reasons people consistently fall easily to attack include inadequate knowledge of URLs structure, the dynamic nature of URL webpages the Users can identify trusted webpages, Users' negligence in checking the URL while they only enter their information webpages and submitting, and the inability of the end-users to differentiate legitimate webpages from the phishing ones (Zhou *et al.,* 2023).

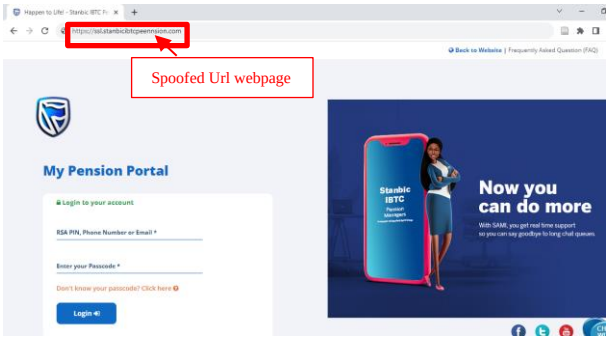

Fig. 1: Demonstration of a Spoofed Webpage

Lackadaisical attitudes allowed the number of phishing-related cases in 2019 and the assaults caused a financial loss amounting to over $21 billion(Le-Nguyen et al., 2023). However, in the first quarter of 2020, another overwhelming phishing assault that went beyond the previous years combined was reported. Zhou et al., (2023), reported that newly detected phishing websites rose by 30% between the third quarter and the second quarter of 2021. Phishing activities carry grave consequences on the network, and the more these activities, the more the network and users become insecure and prey to the attacks. The seamless access to smart devices has immensely contributed to the sporadic strive of phishing attacks which were envisaged to increase by 45% in 2025 (Zhou et al., 2023). Several studies monitored phishing trends and their strategies, and were able to give a valid account of the general behaviour of the existing phishing strategies. The technical approaches for conducting these attacks and various practical and combating solutions were also presented.

Among the various techniques employed to combat phishing assault by researchers include: Black/whitelist, Natural language processing, heuristic, webpage image processing, and artificial intelligence (machine learning) (Suryan *et al.,* 2020; Noureldien & Mohmoud, 2021; Mohamed *et al.,* 2022). Each of the given techniques has its pros and cons, but Artificial intelligence among others has had a great impact on almost every industry. Dada *et al.,* (2019) revealed that when new measure is developed to counterfeit phishing attacks, attackers also devise new powerful attacks to obfuscate the inventions. Researchers in the problem domain have applied machine/deep learning models to analyse and extract the patterns in data. Machine learning (ML) classifiers and deep learning (a branch of ML) are known as powerful tools for extracting new knowledge routinely from huge data.

The identified operational system of ML is based on several unique sub-steps: data collection and representation, feature selection, mapping (training), and making a good classification, but an essential sub-step is the process of selecting variable sets that will yield a model with good predictive performance on unseen data. The ML model's predictive performance depends on the quality of samples inputted to the classifier. A lot of research has been conducted to identify phishing URLs by using diverse feature sets collected from different datasets (Chen *et al.,* 2020; Khan *et al.,* 2020; Shin *et al.,* 2022).

The existing research hunt for model detection performance enhancement methods such as: crawling new indicators along with the existing variables in the dataset, categorizing the features/ranking the features based on their importance, and hybridising the classification classifiers in different ways to improve the predictive. These researchers pay inadequate attention to the ensemble feature selection and multilevel classification techniques, both could improve and expedite the performance of the existing anti-phishing systems if applied. By using ensemble feature selection and multilevel ensemble classifier technique, it is possible to develop a comprehensive and robust anti-phishing system that can detect and adapt to the dynamism of phishing attackers rather than a single approach by the security manager of the networks.

2 URL PHISHING OBFUSCATION TECHNIQUES

Another method usually used is the Uniform Resource Locator links. The links are sent to the target through email on the network. The URL is of two types, that is, http and https. The one with 's' is known to be a secure link to surfing because it has a security extension. Because of the security extension of HTTPs, most standard organizations adopt it over the HTTP, and the fact that it is secure, attackers find a variety of methods to masquerade the integrated security tools. Some of the methods applied by the attackers to obfuscate the HTTPS security apparatus include typo-squatting, injection of random special characters, a combination of words, and cybersquatting among others. The concept of the fundamental structure of URLs showing its elements is illustrated in Figure 2.

Fig. 2: Fundamental Structure of URLs

A URL usually starts with the name of the accessing protocol for a website. After the protocol, is the top-level domain name that presents the domains in the DNS root of the internet. The second level domain is responsible for frequent identification of the organization name and is mostly stationed immediately after the subdomain. These elements combine to produce the webpage domain name, but the page's path is represented by the black box address. The fact that the second level domain presents sorts of activity, the attackers usually find ways to doctor it. The attackers concentrate more on the inner address. They usually create an unlimited number of URL by increasing the path and file name of the second-level domain. What makes the uniqueness of URL is the integration of the Second-level domain and top-level domain. This makes it easy for cyber security companies to identify phony domains employed by the attackers through the name. This step could lead to the blocking of the IP address once the domain name is identified to be phishing to prohibit access to the web pages it contains.

http://doi.org/10.46792/fuoyejet.v8i4.1105 http://journal.engineering.fuoye.edu.ng/

3 Reviewed Related Works

As earlier mentioned, there is an array of research conducted on the topic of webpage phishing detection using different URL phishing datasets. Under this section, we present some of the reviewed related work based on machine learning. The study conducted by Khan *et al.,* (2020) tries to compare the intelligence of four different classical machine learning classifiers (SVM, NB, DT & ANN) and one ensemble classifier (RF) using multiple phishing datasets. The experiment was conducted using two different approaches: one was conducted using a 10-fold cross-validation, while the second approach was conducted on a Principal Component Analysis of reduced datasets. The result showed that both Random Forest and Artificial Neural Network classifiers surpassed the classical algorithms with 96.4% and 97.2% accuracy.

Chiew *et al.,* (2019) saw that much has not been done in the feature selection part as they recognized the importance of feature selection steps to machine learning performances. The authors developed a hybrid feature selection framework which was applied to the phishing variables, and the baseline features generated were inputted to the selected six (6) machine learning classifiers. The experiment shows Random Forest classifier outperformed other classifiers achieving a detection accuracy of 94.6%. Another related study is the work of Salihovic *et al.,* (2019) which was carried out to exclude human factors in security breaches with the invention of machine learning classifiers (i.e. RF, LR, SVM, k-NN, ANN & NB). These classifiers are exposed to spam and phishing datasets and their efficacy was observed based on the feature selection techniques applied. The outcome result revealed that each feature selection applied produced different outcomes, as RF with PCA + Ranker yielded an accuracy of 97.33% while RF with Correlation feature evaluation optimization + BestFirst achieved 94.24%. In another study that focused on feature selection Moedjahedy *et al.,* (2022), a combined correlation filter-based statistical technique was applied for the identification of phishing website variables for the machine learning models. Two datasets were used comprising 87 and 48 variables respectively. The best accuracy from the phishing detection models was achieved only on the 10 reduced variables from the two datasets recording 95.88%.

Noureldien & Mohmoud (2021) performed feature importance on different phishing datasets using information gain. This is to identify a representative set of indicators. The authors used an intersection function on the top features to obtain 10 consistent baseline features. However, the authors did not establish the reason for picking only the top 10 information gain-ranked indicators. However, C5.0 decision tree classifier was evaluated on the feature sets and yielded 92.23%. The performance of the selected top 10 features is better than those features with lesser values. The work of Ojewumi *et al.,* (2022) proposed a means of detecting phishing attacks on web pages using machine learning tools. The study utilized a rule-based approach for detecting phishing with the use of three machine learning algorithms, namely: K-Nearest Neighbour, Support Vector Machine, and Random Forest. Out of the 1000 web pages, only 400 are legitimate, which are e-banking and financial web pages. After the implementation of the three algorithms employed, that is, SVM, KNN and RF, it was submitted that the Random Forest model delivered the best performance with 98.35%. Amusan et al., (2021) developed a Linkguard android anti-phishing system for quick response detection and prevention of known and unknown phishing attacks. The experiment was leveraged on 500 phishing and legitimate links collected from PhishTank and Alexa. The system evaluation was conducted and an accuracy of 96% was achieved by the proposed system. Despite that the system achieved an awesome accuracy, the dataset used is too small to generalize the system capacity.

A comparative analysis of ensemble and shallow classifiers was investigated by Igwilo & Odumuyiwa (2022). The study considered three shallow and five ensemble classifiers with four different imported features. The study reported that the stacking ensemble model outperformed other models on the duo dataset used with accuracy scores of 96% and 99.3% respectively. The quantity of the dataset is few and also feature selection phase was not performed which means that the models were inputted with irrelevant and redundant features.

4 Proposed Feature Selection

The illustration of the automated Heterogeneous Ensemble Feature Selection (HEFS) Method proposed for the selection of optimal URL phishing features used three filter-based statistical techniques, that is, Gain Ratio (GR), Chi-Square (CHI-2), and Pearson Correlation Coefficient as presented in Figure 3. The benefits and limitations of each of these techniques are considered to achieve robust results.

Let Z denote the original URL phishing dataset: $Q = \{q_1, q_2, \dots q_n\}$, $J = \{r_1, r_2, \dots r_n\}$ be the class target and filter measures are denoted as $FP_k, FP_t, \& FP_j$ (GR PCC, & CHI-2) respectively. The FP_k feature measure is applied homogeneously to data Z, and the features are ranked based on their importance. FP_k generates these values using information theories as given in (1), filter measure values $\{\emptyset_1 k, \emptyset_2 k, \dots, \emptyset_j k,\}$ are generated with respect to FP_k. Secondly, FP_t (PCC) filter measure is also applied to the same dataset to measure the dataset variables using its method. The FP_t applied Eqn.2 to measure the relationship between the independent variables and the target class.

$$FP_k = \frac{IG(T, f_k)}{SplitInfo_k(T)} \quad (1)$$

where, $SplitInfo_{fk}(T)$ represent 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 $SplitInfo_{fk}(T)$.

$$FP_t\ (k,t) = \frac{cov(q,r)}{\sigma q \sigma r} \quad (2)$$

http://doi.org/10.46792/fuoyejet.v8i4.1105 http://journal.engineering.fuoye.edu.ng/

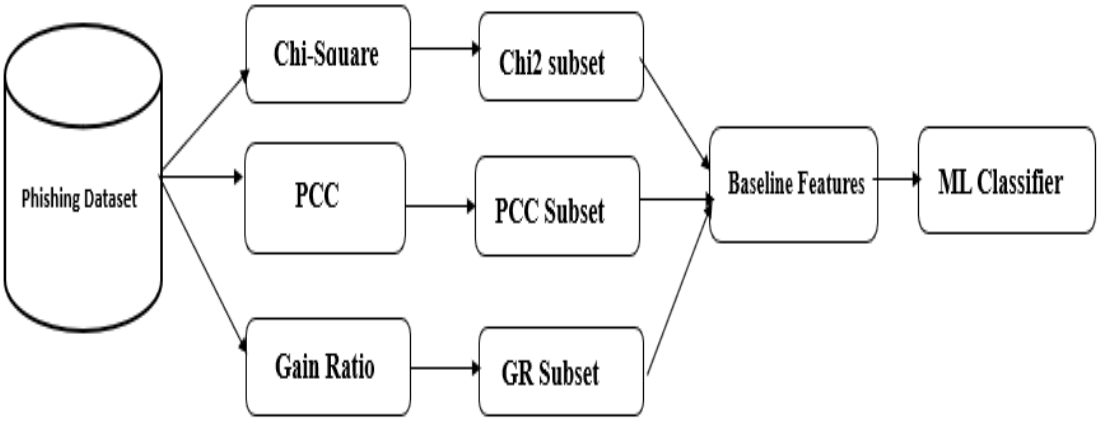

Fig. 3: The Proposed HEFS Framework.

where, q denotes the phishing independent variables, t is the phishing target class, $cov(k,t)$ present the covariance of q and r, $\sigma q, \sigma r$ is the standard deviation of q and r. Any two independent variables that correlate are dropped, after dropping the correlated features a set of ranked filter measure values are generated: $\{\varphi_1 k, \varphi_2 k, \ldots, \varphi_j k,\}$. The third filter predictor FP_j (CHI-2), is also a filter-based statistical technique that measures the divergence from the distribution. FP_j is applied on the Z and J and a set of $\{\alpha_1 k, \alpha k, \ldots, \alpha_j k,\}$ values are generated using Eqn. 3. The three filter measures generate lists of feature ranking $\{\emptyset_n k, \varphi_n k, \alpha_j k,\}$ using FP_k, FP_t, FP_j respectively from the original dataset: $Z = \{q_1, q_2, \ldots q_n\}$, $J = \{r_1, r_2, \ldots r_3\}$. An automatic threshold is used to select the important features from $\{\emptyset_1 k, \emptyset_2 k, \ldots, \emptyset_j k,\}$, $\{\varphi_1 k, \varphi_2 k, \ldots, \varphi_j k,\}$, and $\{\alpha_1 k, \alpha k, \ldots, \alpha_j k,\}$.

$$FP_j^{\,2} = \sum_{i=1}^{m}\sum_{j=1}^{k}\left(\frac{A_{i,j} - \left(\frac{R_i * c_j}{N}\right)^2}{\frac{R_i * c_j}{N}}\right) \qquad (3)$$

where, m is the attributes 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 the size of patterns in the i^{th} attribute, C_j is the size of patterns in the J^{th} class, and A_{ij} is the size of patterns in the i^{th} internal and the J^{th} class. The features that obtained higher Chi-square scores were fetched.

Feature values below the threshold are marked as not important while those within and above the threshold are selected. The obtained features at this level are known to be the primary informative feature subsets. To obtain the baseline features, a novel Borda count algorithm was applied. The Borda count algorithm combined the three primary informative features and selected the secondary informative feature subsets known as Baseline features. Borda count aggregator algorithm is a consensus-based voting system given by:

$$b_i = \sum_{v|} N_f - P_v \qquad (4)$$

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

5 Method and Materials

The proposed anti-phishing detection system is based on an ensemble feature selection and multilevel ensemble classification model. The aim is to efficiently select optimal (important) phishing features and also improve the classical ML algorithms (i.e. SVM, NB, & LR) that are usually adapted in the existing studies. Phishing detection comprises three stages: the first stage is the collection of data and pre-processing, followed by the Heterogeneous Ensemble Feature Selection (HEFS) phase, and the last is the evaluation phase. The flow of the proposed system is shown in Figure 4 respectively.

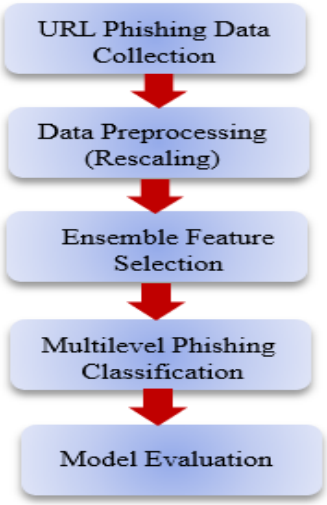

Fig. 4: Proposed Workflow

5.1 Data Collection / Pre- Processing

The URL phishing dataset used for experimentation in this study was extracted and prepared in a comma-separated format (CSV). This dataset was downloaded from www.kaggle.com/datasets/shashwatwork/phishing-dataset-for-machine-learning and was also used in Chiew *et al.*, (2019). It is a balanced dataset with 5,000 phishing and legitimate records and 48 URL attributes with the characteristics given in Table 1.

Table 1. Data Observation/Characteristics

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

http://doi.org/10.46792/fuoyejet.v8i4.1105 http://journal.engineering.fuoye.edu.ng/

The dataset was discovered to have value variation during Exploratory Data Analysis (EDA). Value variation is a crucial problem in a machine-learning environment, as it tends to affect the model's performance. The value variation was corrected through the rescaling technique given in Eqn. 5, to keep the values between 0 and 1.

$$V' = \frac{V - V_{min}}{V - V_{max}} \quad (5)$$

where, V is the new value to be converted, $V_{min}$is the minimum value, and, $V_{max}$is the maximum value in the dataset.

5.2 URL Phishing Detection Classifiers

The selected algorithms for the task fall into two groups: ensemble and shallow. Three shallow classifiers (SVM, NB & LR) were selected, and two ensemble classifiers Bagging (level-0) and Stacking (level 1) were also selected.

6 Results and Discussion

After performing all the necessary pre-processing steps, the next step is to conduct the phishing model implementation and evaluation. Several experiments were conducted to establish the potency and efficacy of the developed feature selection framework and the multilevel ensemble classification model. The developed phishing models were evaluated on four standard metrics and confusion matrix parameters. The experiment was conducted in a Python 3.7 environment. The developed phishing models were evaluated on four standard metrics as given in Eqn.6-9 and the confusion table presented in Table 2.

Table 2. Confusion Matrix

		Predicted	
		Phishing	Legitimate
Actual	Phishing	TP	TN
	Legitimate	FP	TN

True Positive (TP), True Negative (TN), False Positive (FP) & False Negative (FN).

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

$$Prec = \frac{TP}{TP + KP} \quad (7)$$

$$Recall = \frac{TP}{TP + FP} \quad (8)$$

$$FS = \frac{2 \times Prec + Recall}{Prec + Recall} \quad (9)$$

6.1 Model Evaluation Using Features Filter-Based Technique (Chi-square, Pearson & Gain Ratio)

For the authors to prove and establish the potency of the proposed anti-phishing for this study, the shallow, bagging, and multilevel ensemble classifiers were evaluated on the features obtained by the individual statistical techniques and the proposed ensemble feature selection methods. The models are evaluated on the features obtained and the results obtained are presented in tables and figures.

A. The Result for Chi-Square-Based Features

The Chi-square was applied to the dataset and 14 features were selected based on their importance out of the 48 attributes available in the dataset. The features selected by the Chi-square are presented in Table 3, the classification algorithms are evaluated on the variables, and their results are presented in Table 4 and Figure 5.

B. The Result for Gain Ratio-Based Features

The Gain Ratio (GR) was also applied to the dataset as a standalone statistical technique, the process ranked the features, and the best 25 features selected are presented in Table 5. The phishing classification was conducted on the Gain Ratio features, and their results are presented in Table 6 and Figure 6.

C. The Result of Pearson Correlation-Based Features

The Pearson Correlation Feature Selection (PCFS) is the third filter-based technique applied to the phishing dataset and the best 16 features selected are presented in Table 7. The phishing classification was conducted on the outcome of the Pearson Correlation predictions and the results are presented in Table 8 and Figure 7.

Table 3. Features Selected Through Chi-Square

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

Table 4. Results of the Classical ML and the Bagged Ensemble Models on Chi-Square Features

Parameters	**Single Classifier**			**Ensembled**		
	SVM	LR	NB	**Bagged_SVM**	**Bagged_LR**	**Bagged_NB**
TP	825	933	935	901	747	828
FN	160	55	53	87	241	160
FP	137	318	377	130	81	137
TN	875	694	635	882	931	875
ACCURACY	0.852	0.814	0.785	0.892	0.839	0.852
PRECISION	0.858	0.746	0.713	0.874	0.902	0.858
RECALL	0.838	0.944	0.946	0.912	0.756	0.838
FI-SCORE	0.848	0.833	0.813	0.893	0.823	0.848

http://doi.org/10.46792/fuoyejet.v8i4.1105 http://journal.engineering.fuoye.edu.ng/

Table 5. Feature Selected Through Gain Ratio

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

Table 6. Results of the Classical ML and the Bagged Ensemble Models on GR Features

Parameters	**Single Classifier**			**Ensembled**		
	SVM	**LR**	**NB**	**Bagged_SVM**	**Bagged_LR**	**Bagged_NB**
TP	849	964	913	892	849	897
FN	139	24	75	96	139	91
FP	262	359	309	208	262	362
TN	750	653	703	804	750	650
ACCURACY	0.800	0.809	0.808	0.848	0.799	0.774
PRECISION	0.764	0.729	0.747	0.811	0.764	0.712
RECALL	0.859	0.976	0.924	0.903	0.860	0.908
FI-SCORE	0.809	0.834	0.826	0.854	0.809	0.798

Table 7. Feature Selected Through PCFS

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

Table 8. Results of the Classical ML and the Bagged Ensemble Models on PC Features

Parameters	**Single Classifier**			**Ensembled**		
	SVM	LR	NB	**Bagged_SVM**	**Bagged_LR**	**Bagged_NB**
TP	892	747	849	853	873	828
FN	96	241	139	135	115	160
FP	208	81	262	155	112	137
TN	804	931	750	857	900	875
ACCURACY	0.848	0.839	0.800	0.855	0.887	0.852
PRECISION	0.811	0.902	0.764	0.846	0.886	0.858
RECALL	0.903	0.756	0.859	0.863	0.884	0.838
FI-SCORE	0.854	0.823	0.809	0.855	0.885	0.848

Table 9. The Baseline Features

S/N	**Features**
1	FrequentDomainNameMismatch
2	SubmitInfoToEmail
3	PctExtNullSelfRedirectHyperlinksRT
4	ExtMetaScriptLinkRT
5	PctExtNullSelfRedirectHyperlinks
6	PctExtHyperlinks
7	NumDashInHostname

http://doi.org/10.46792/fuoyejet.v8i4.1105 http://journal.engineering.fuoye.edu.ng/

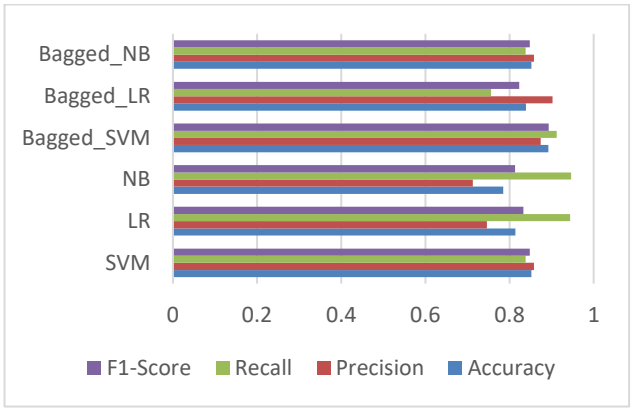

Fig. 5: Result of Classical ML and the Bagged Ensemble Models on Chi-Square Features

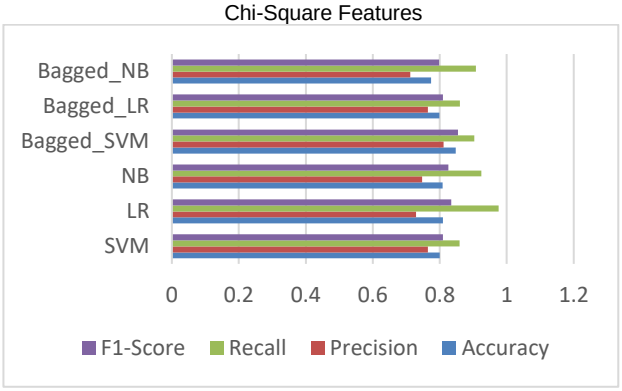

Fig. 6: Result of Classical ML and the Bagged Ensemble Models on Gain Ratio Features

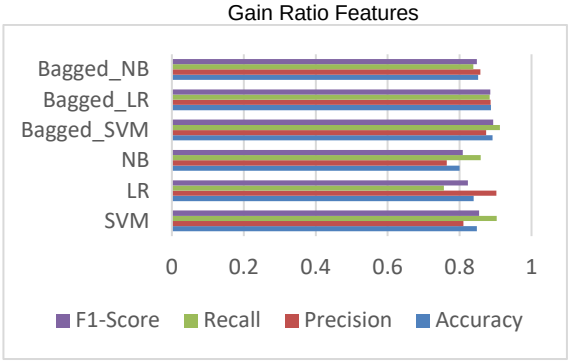

Fig. 7: Result of Classical ML and the Bagged Ensemble Models on PCFS Features

6.2 URL Phishing Classification Model Evaluation Based on HEFS Features

The classical classification and the multilevel ensemble models were implemented on the filter-based techniques and were repeatedly implemented on the features obtained by the proposed ensemble feature selection method. The proposed ensemble feature framework was applied to the URL phishing dataset and the process produced 7 features that represent the benchmark features for this study. The selected features of the HEFS method are presented in Table 9. The results generated are presented in Tables 10 & 11 and Figures 8 & 9.

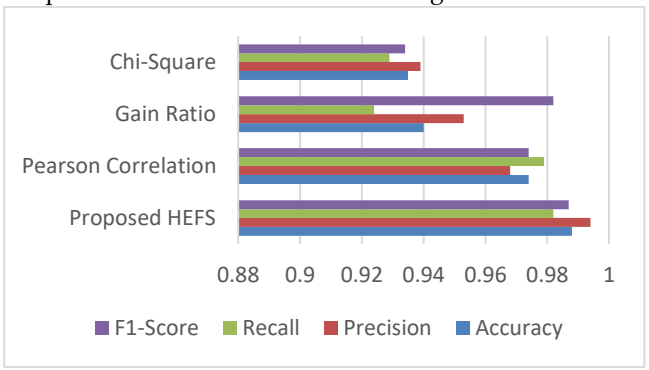

Fig. 8: Multilevel Ensemble Model Performances on the Individual Techniques and HEFS

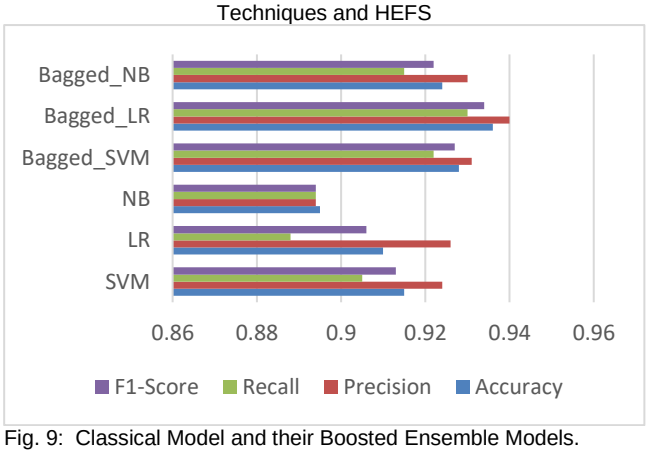

Fig. 9: Classical Model and their Boosted Ensemble Models.

Table 10. Results of the Classical ML and the Bagging Ensemble Models on HEFS Features

Parameters	Single Classifier			Ensembled		
	SVM	LR	NB	Bagged_SVM	Bagged_LR	Bagged_NB
TP	894	877	803	911	918	907
FN	94	111	105	77	70	84
FP	77	70	105	67	59	68
TN	935	942	907	945	953	944
ACCURACY	0.915	0.910	0.895	0.928	0.936	0.924
PRECISION	0.924	0.926	0.894	0.931	0.940	0.930
RECALL	0.905	0.888	0.894	0.922	0.930	0.915
FI-SCORE	0.913	0.906	0.894	0.927	0.934	0.922

Table 11. Multilevel Ensemble Model on Chi-Square, PC, GR & HEFS Features

Parameters	Proposed Multilevel Ensemble			
	Chi-Square	Pearson Correlation	Gain Ratio	HEFS
TP	918	968	913	970
FN	70	20	75	18
FP	59	32	45	6
TN	953	980	967	1006
ACCURACY	0.935	0.974	0.94	**0.988**
PRECISION	0.939	0.968	0.953	0.994
RECALL	0.929	0.979	0.924	0.982
FI-SCORE	0.934	0.974	0.938	0.987

http://doi.org/10.46792/fuoyejet.v8i4.1105 http://journal.engineering.fuoye.edu.ng/

6.3 Discussion

The results of the phishing classification models obtained from the feature selection techniques are presented in both tables and figures. The results show that bagged_SVM and shallow SVM had the highest accuracy of 89.2% and 85.2% using the chi-square features. Under the features obtained by the GR technique, bagged_SVM, and Logistic regression models outperformed others with 84.8% and 80.9%. The model evaluated on the Pearson correlation features revealed that bagged_LR and SVM single classifiers recorded the highest accuracy of 88.7% and 84.8% as shown in Table 8.

However, the results of the models on the benchmark features (baseline) unveiled that bagged_LR and SVM outperformed the rest of the models having accuracy scores of 93.6% and 91.5% respectively. In addition to these results of single and bagged ensembled models, the proposed multilevel ensemble model was implemented and evaluated on the features obtained by the base feature selection methods and the HEFS methods. The results show that the proposed multilevel ensemble (stacked) model achieved 93.5% on chi-square, 97.4% on Pearson correlation, 94.8% on Gain Ratio, and 98.8% on HEFS features. The results show that the proposed model outperformed the single and their bagged ensembled models by achieving an accuracy of 98.8%. The proposed multilevel ensemble (stacked) model improved the detection performance of the single and bagged ensembled models by 28%. The proposed model had 970 true positives, 1006 true negatives, 0.994 precision, 0.982 recall, and 0.987 F1 score.

6.4 Comparison of the Proposed Framework

The performance of the classifier under the proposed framework is compared to the individual statistical techniques and that of HEFS as shown in Figure 10-Fig.13. The comparison is based on the four major metrics, i.e., accuracy, precision, recall, and F1-Score. However, the performance of the proposed multilevel ensemble (stacked) and HEFS features is compared with studies like Chiew et al., (2019), Abdul Samad *et al* (2023), and Amusan *et al.,* (2021) as shown in Table 12 and Figure 14. The findings of this work established that the performance of the proposed multilevel ensembled under HEFS outperformed the existing studies. This revealed that the existing feature selection and classification systems are not effective like the HEFS. Furthermore, the mind-blowing performance obtained in this work implies that the proposed HEFS is reliable and hence adaptable to both webpage and email semantic datasets.

Table 12. Proposed Technique vs. Other Studies

Method	Accuracy	Precision	Recall	F1-Score
Chiew et al., 2019	93.55	0.939	0.929	0.934
Abdul Samad *et al* (2023)	89.5	0.894	0.894	0.894
Amusan et al., (2021)	96%	0.97	0.966	0.957
HEFS	**98.8%**	**0.994**	**0.982**	**0.987**

7 Conclusion

Phishing is a deceptive mechanism adopted to deceive internet users into revealing their sensitive information. Most phishing agents are majorly distributed through email or other media which is presented to the users like email coming from a genuine source. The existence and detection of phishing attack is challenging, the machine-learning approach is seen as the right antidote as it offers automatic detection. The existing studies switched attention towards the enhancement of ML model accuracy using series methods such as optimization, hybridization, ensemble and feature ranking, and so on. Feature selection sub-steps of ML and multilevel ensemble techniques are given less consideration.

The development of an ensemble feature selection and multilevel ensemble classification methods has a significant improvement in the underlined phishing detection. The results revealed that the variables selected by HEFS perform excellently when combined with the multilevel ensemble model. This study significantly established that when appropriate features are inputted into the rightly combined classical models, such a phishing detection model will yield better performance. The HEFS helped to pick efficient features (baseline) inputted to the model to actively detect phishing attacks with an accuracy of 98.8%. The synergy of multilevel phishing model and HEFS features outperformed single, bagging (level-0) and multilevel models on the individual filter-based features. The rule of the proposed ensemble feature selection and that of the phishing detection models could be integrated into browsers or deployed to email servers to actively detect phishing attacks.

Future researchers can still look for optimal means; this particular system could be enhanced to achieve 100% accuracy. Likewise, hyper-parameters of the algorithms can be fine-tuned to improve the detection rate. However, it is worth extending the application of the system to a large dataset to investigate the performance and its computation time. Generally, the study provided empirical evidence that the Heterogenous Ensemble Feature Selection (HEFS) framework has a positive impact on classification performances.

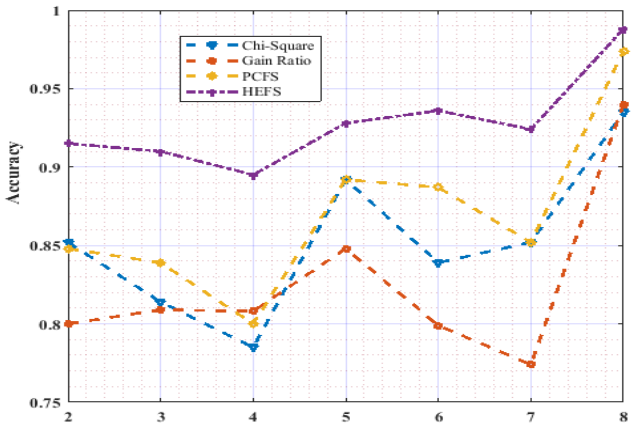

Fig. 10: Accuracy of the Individual Statistical Techniques and HEFS

http://doi.org/10.46792/fuoyejet.v8i4.1105 http://journal.engineering.fuoye.edu.ng/

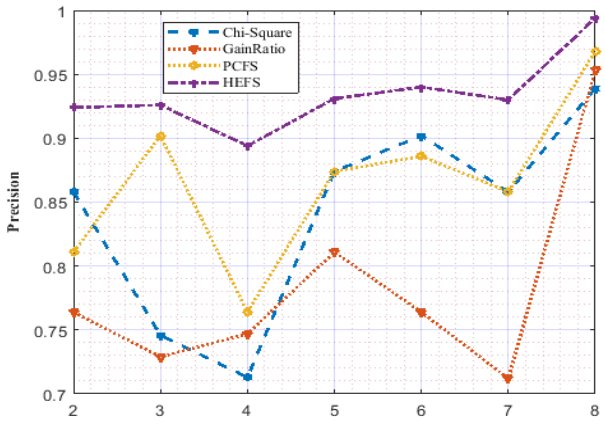

Fig. 11: Precision of the Individual Statistical Techniques and HEFS

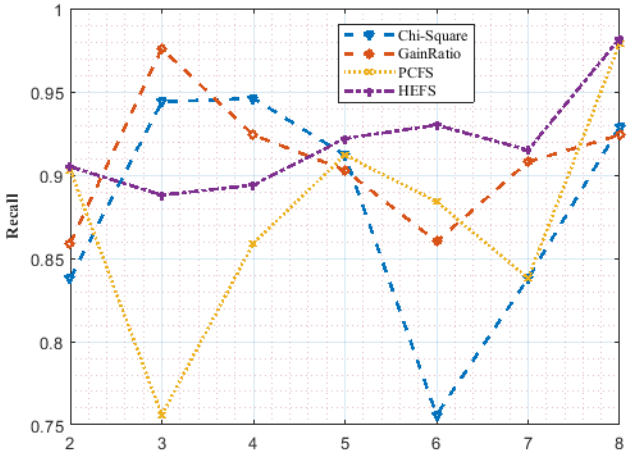

Fig. 12: Recall of the Individual Statistical Techniques and HEFS

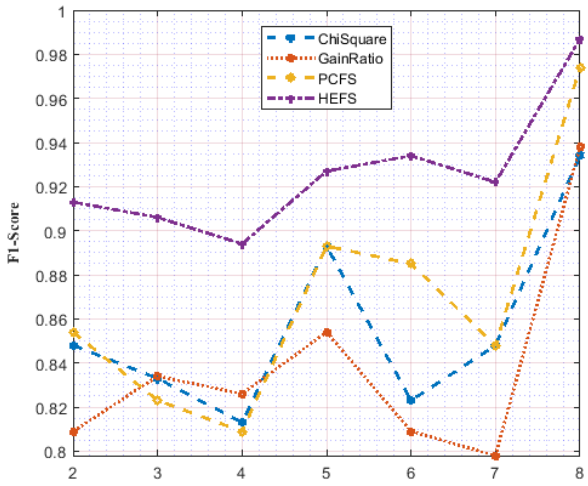

Fig. 13: F1-Score of the Individual Statistical Techniques and HEFS

REFERENCES

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

Al-Sarem, M., Saeed, F., Al-Mekhlafi, Z. G., Mohammed, B. A., Al-Hadhrami, T., Alshammari, M. T., Alreshidi, A., & Alshammari, T. S. (2021). An optimized stacking ensemble model for phishing websites detection. *Electronics (Switzerland)*, *10*(11). https://doi.org/10.3390/electronics10111285

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

Azeez, N. A., Misra, S., Margaret, I. A., Fernandez-Sanz, L., & Abdulhamid, S. M. (2021). Adopting automated whitelist approach for detecting phishing attacks. *Computers and Security*, *108*. https://doi.org/10.1016/j.cose.2021.102328

Chen, J. L., Ma, Y. W., & Huang, K. L. (2020). Intelligent visual similarity-based phishing websites detection. *Symmetry*, *12*(10), 1–16. https://doi.org/10.3390/sym12101681

Chiew, K. L., Tan, C. L., Wong, K. S., Yong, K. S. C., & Tiong, W. K. (2019). A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. *Information Sciences*, *484*, 153–166. https://doi.org/10.1016/j.ins.2019.01.064

Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon*, *5*(6). https://doi.org/10.1016/j.heliyon.2019.e01802

Gangavarapu, T., Jaidhar, C. D., & Chanduka, B. (2020). Applicability of machine learning in spam and phishing email filtering: review and approaches. *Artificial Intelligence Review*, *53*(7), 5019–5081. https://doi.org/10.1007/s10462-020-09814-9

Igwilo, C. M., & Odumuyiwa, V. T. (2022). Comparative Analysis of Ensemble Learning and Non-Ensemble Machine Learning Algorithms for Phishing URL Detection. *FUOYE Journal of Engineering and Technology*, *7*(3), 305–312. https://doi.org/10.46792/fuoyejet.v7i3.807

Khan, S. A., Khan, W., & Hussain, A. (2020). Phishing Attacks and Websites Classification Using Machine Learning and Multiple Datasets (A Comparative Analysis). *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12465 LNAI*, 301–313. https://doi.org/10.1007/978-3-030-60796-8_26

Lakshmi, L., Reddy, M. P., Santhaiah, C., & Reddy, U. J. (2021). Smart Phishing Detection in Web Pages using Supervised Deep Learning Classification and Optimization Technique ADAM. *Wireless Personal Communications*, *118*(4), 3549–3564. https://doi.org/10.1007/s11277-021-08196-7

Le-Nguyen, M. K., Nguyen, T. C. H., Le, D. T., Nguyen, V. H., Tôn, L. P., & Nguyen-An, K. (2023). Phishing Website Detection as a Website Comparing Problem. *SN Computer Science*, *4*(2). https://doi.org/10.1007/s42979-022-01544-9

Mao, J., Bian, J., Tian, W., Zhu, S., Wei, T., Li, A., & Liang, Z. (2019). Phishing page detection via learning classifiers from page layout feature. *Eurasip Journal on Wireless Communications and Networking*, *2019*(1). https://doi.org/10.1186/s13638-019-1361-0

Moedjahedy, J., Setyanto, A., Alarfaj, F. K., & Alreshoodi, M. 022). CCrFS: Combine Correlation Features Selection for Detecting Phishing Websites Using Machine Learning. *Future Internet*, *14*(8). https://doi.org/10.3390/fi14080229

Mohamed, G., Visumathi, J., Mahdal, M., Anand, J., & Elangovan, M. (2022). An Effective and Secure Mechanism for Phishing Attacks Using a Machine Learning Approach. *Processes*, *10*(7). https://doi.org/10.3390/pr10071356

Niranjan, A., Sakhamuri, V. K., Deepa Shenoy, P., & Venugopal, K. R. (2020). ERCRFS: Ensemble of random committee and random forest using stackingc for phishing classification. *International Journal of Emerging Trends in Engineering Research*, *8*(1), 79–86. https://doi.org/10.30534/ijeter/2020/13812020

Niu, W., Zhang, X., Yang, G., Ma, Z., & Zhuo, Z. (2018). Phishing emails detection using CS-SVM. *Proceedings - 15th IEEE International Symposium on Parallel and Distributed Processing with Applications and 16th IEEE International Conference on Ubiquitous Computing and Communications, ISPA/IUCC 2017*, 1054–1059. https://doi.org/10.1109/ISPA/IUCC.2017.00160

Noureldien, N., & Mohmoud, S. (2021). The Efficiency of Aggregation Methods in Ensemble Filter Feature Selection Models. *Transactions on Machine Learning and Artificial Intelligence*, *9*(4), 39–51. https://doi.org/10.14738/tmlai.94.10101

Ojewumi, T. O., Ogunleye, G. O., Oguntunde, B. O & Folorunsho, O. (2022). *Performance Evaluation of Machine Learning Tools for Detection of Phishing Attacks on the Webpage*. African Institute of Mathematical Sciences. doi.org/10.1016/j.sciaf.2022.e01165

Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications*, *117*, 345–357. https://doi.org/10.1016/j.eswa.2018.09.029

Shin, S. S., Ji, S. G., & Hong, S. S. (2022). A Heterogeneous Machine Learning Ensemble Framework for Malicious Webpage Detection. *Applied Sciences (Switzerland)*, *12*(23). https://doi.org/10.3390/app122312070

Suryan, A., Kumar, C., Mehta, M., & A.Sinha, R. J. (2020). Learning Model For Phishing Website Detection. *EAI Endorsed Transactions on Scalable Information Systems*, *7*(27), 1–9. https://doi.org/10.4108/eai.13-7-2018.163804

Zhou, J., Cui, H., Li, X., Yang, W., & Wu, X. (2023). A Novel Phishing Website Detection Model Based on LightGBM and Domain Name Features. *Symmetry*, *15*(1). https://doi.org/10.3390/sym15010180