Journal of Informatics and Web Engineering

Vol. 3 No. 2 (June 2024) eISSN: 2821-370X

Ensemble Learning-Powered URL Phishing Detection: A Performance Driven Approach

Shougfta Mushtaq[1] *, Tabassum Javed[2] and Mazliham Mohd Su'ud[1]

[1] Faculty of Computing & Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Malaysia
[2] Faculty of Computing, Riphah International University, Islamabad Capital Territory, 44000, Pakistan
**corresponding author: (shouguftawajid@yahoo.com, ORCiD: 0009-0000-5794-8466)*

Abstract – With the rapid growth in the usage of the Internet, criminals have found new ways to engage in cyber-attacks. The most common and widespread attack is URL phishing. The proposed system focuses on improving phishing website detection using feature selection and ensemble learning. This model uses two datasets, DS-30 and DS-50, each with 30 and 50 features. Ensemble learning using a voting classifier was then applied to train the model, achieving more accuracy. The combination of HEFS with random forest distribution achieved 94.6% accuracy while minimizing the number of features used (20.8% of the base feature set). The classifier works in the proposed model, and the accuracy is 96% and 98% on the DS-30 and DS-50 datasets, respectively. The hybrid model uses a combination of different factors to distinguish phishing websites from legitimate websites.

Keywords—URL Phishing, Phishing Website Detection, Ensemble Learning, Feature Selection, AI-Powered Algorithm, Machine Learning Models

Received: 19 November 2023; Accepted: 11 March 2024; Published: 16 June 2024

I. INTRODUCTION

With the rapid development of technology, people need time to connect to the internet. Everyone in the world wants to be successful by any means necessary. People with a criminal mindset always try to keep others safe. It is easy for them to sit at their desk and breach security with cyber-attacks. Individuals, organizations, and governments attempt to gain authorized access to other computers or Internet connections. The purpose of a cyber-attack is to identify, change, delete or control confidential information. Organizations and governments are hiring computer experts to carry out cyber-attacks[1].

The most common and easiest way to commit a crime is to do social engineering while sitting at your job. Phishing attacks are very popular among criminals today because it is the best and easiest way to fool people. Criminals see this as the first step in another attack. Criminals manipulate people's desires, sorrows, needs, etc. Because it exploits them to express sympathy, it is a quick and easy way to obtain information through social engineering. It has better performance than other attacks. Among all other commercial attacks, phishing is the most common attack used by criminals to trick people and misuse their information. Many types of attacks fall under the umbrella of phishing attacks, including email, spoofed phone calls, ads, spam, and fake websites[2].

Criminal motives often lead to more than one type of phishing attack. URL phishing is the most common type because criminals just show the link and the probability is very high as people often unintentionally or accidentally click on

Journal of Informatics and Web Engineering
https://doi.org/10.33093/jiwe.2024.3.2.10

Published by MMU Press. URL: https://journals.mmupress.com/jiwe

the link. Sometimes they click quickly without thinking, doing anything else, or working on other tasks, which can lead to security breaches. Criminals expect people to make these kinds of mistakes. They understand human psychology. Phishing attacks are not new. A lot of research is being done on this topic, and companies, organizations, and governments are investing in improving the security of their data. Google has blacklisted phishing websites, but they still pose a big threat because many phishing websites are created by criminals every day. It may seem impossible to update your system every day to prevent phishing[3].

In response to a growing number of threats, there is a shift towards using artificial intelligence (AI) for online security. Intelligent tools and techniques have been developed to improve the detection and prevention of phishing attacks. However, while AI-powered URL phishing detection methods exhibit capability, they encounter boundaries and challenges that necessitate further research and development. The dynamic nature of cyber-criminal strategies and the constant creation of new phishing websites demand continuous innovation in AI-powered security measures to effectively counter these threats. Consequently, researchers are actively involved in evolving AI-driven solutions for improved online security despite the evolving cyber threats. The world's growing dependence on technology has unlocked the door to cyber-attacks, and criminals exploit this opportunity to breach security and steal valuable information. Phishing attacks, particularly through misleading web links, have become their preferred method. Many people click on these links without thinking, making them easy targets. Despite the efforts of companies and organizations to combat these attacks, cybercriminals continue to create new phishing websites. As technology advances, researchers are starting to turn to artificial intelligence to improve network security. But existing intelligence tools have limitations and need to be further developed to keep up with the changing methods of cybercriminals.

This article is divided into the following sections: Introduction (establishing the background, introducing the problem), Literature Review (providing an overview of existing studies) Problem Statements (introduction to the research competition), Artificial Intelligence Assisted URL Network Road Fishing (providing an overview of the proposed methods details), results (copy of details of findings and recommendations) and data (data used collectively).

II. LITERATURE REVIEW

In [4], the authors proposed a method that builds the CNN algorithm and called it CNN-Fusion to detect URL phishing. The method is designed to use different types of single-layer CNNs with different network parameters to eliminate the multilayer process. CNN-Fusion was tested using five publicly available databases containing 1.85 million phishing and legitimate URLs. Hostile attacks of artificial intelligence are also evaluated using this model. Experimental results show that the model requires five times less training time and uses less memory. The results show that the model achieved an average accuracy of over 99% across five datasets of AI-generated malicious attacks.

In [5], the authors aimed to find out the unique characteristics of phishing websites by analyzing them. These features include standards such as source code, security, page types, and URLs. This process includes feature extraction, training model learning, and performance evaluation. Support vector machines and random forest classifiers prove that random forests are more accurate than support transport vector machines. The proposed system supports machine learning to detect phishing websites, even new websites that have not yet been blacklisted. By analyzing URL features, the application achieved a classification accuracy of 96% using random forest classification. This is better than the 90% accuracy of the support vector machine. This study demonstrates the effectiveness of machine learning in phishing attacks and highlights the importance of URL metadata in classifying phishing websites.

In [6], the authors proposed a new strategy to prevent phishing by using deep learning to improve online security. This method uses machine learning to predict whether a website is a phishing site by analyzing URLs (Universal Resource Locators) and URIs (Universal Resource Identifiers). The system uses Logistic Regression, principle analysis and prioritization techniques to create good models, and Logistic Regression emerges as the best choice. The system can detect phishing, a type of online fraud, with approximately 98% accuracy in detecting fake websites. As online threats become more prevalent, the techniques have proven effective in protecting online users from phishing attacks.

In [7], the authors proposed a meta-algorithm to improve the detection of phishing websites and proposed a different machine learning approach that combines three types of attacks. Random Forest, Decision Tree, and support vector machine. The main goal is to improve the accuracy of identifying phishing sites in web content. The proposed model was evaluated with individual models and the real quality was determined with augmented decision trees, random forest and support vector machine models. The results confirm the effectiveness of the combination and show a significant improvement in detection accuracy. Additionally, the study highlights the potential limitations of using URL signatures alone for phishing detection and underscores the need for more robust and comprehensive methods for effective detection and prevention of phishing attempts. The accuracy of the proposed algorithm is 98.52% higher than other algorithms.

The application process to distinguish the URL given in the text is phishing using Support Vector Machine (SVM) and natural or unnatural language[8]. Phishing processing (NLP) method. SVM is a supervised learning model for classification and regression. SVM works by finding a hyperplane that effectively separates the data into different groups. The application process evaluates URLs based on IP address, subdomain, URL length, web traffic, presence of the "https" symbol, etc. It focuses on classification according to various features such as. In addition to the NLP technique, SVM is also integrated to increase accuracy. The results show that this method can effectively detect phishing URLs, making it an important tool for Internet users to detect potentially malicious websites and reduce the risks associated with phishing attacks.

In [9], the authors proposed a system in which architecture focuses on phishing detection based on URL signature analysis. The process begins by collecting user input in the form of URLs and then processing the raw material using a variety of methods such as regular expressions, WHOIS and PageRank to extract important features. Feature selection based on the chi-square test determines the most important features for classification. In this study, two machine learning algorithms, Random Forest and Gradient Boosting Machine (GBM), are used to calculate the accuracy of URLs. Evaluate models using performance metrics such as precision, accuracy, precision, recall, and F-score. These results show that Gradient Boosting Machine (GBM) outperforms Random Forest in terms of accuracy, making it a good method for phishing detection where website legitimacy estimation is an important factor of change.

In [10], a feature selection process is proposed that combines business-related and data-related processes to achieve performance in terms of phishing website detection options. In this way, the filtering process is used to evaluate the features and the new CDF-g algorithm determines the cut-off level. The selected features form a baseline that, when combined with random forest splitting, reduces feature dimensionality while maintaining accuracy. The framework outperforms existing options and is adaptable to a wide range of data, making it useful for improving phishing detection and addressing emerging threats in the field. These results show the longevity of the best anti-phishing features. Additionally, the proposed system reduces the complexity of computing by adapting applications in real time, which is important in the context of big data and temperature complexity in cyber security. Overall, the findings support the performance of the HEFS framework in selecting details for phishing detection, reducing site specificity while maintaining high accuracy, making it an essential tool for updating the threat landscape in the cybersecurity field.

The effectiveness of AI learning techniques for phishing website detection was carefully evaluated and compared with four classification methods (benchmark i.e. Decision Tree, Naive Bays, Random Forest, Gradient Boosting, and Logistic Regression) [11]. Evaluation consists of training and testing using five feature subsets ranging from 5% to 100% of the features. The plan is designed to be able to distinguish legitimate websites from phishing websites and achieve a maximum accuracy of 95% while using 20% of the best features. Qualitative evaluation includes measurements such as accuracy, precision, recall, F1 score, and Kappa statistics. The proposed method outperforms other classifications with the highest Kappa number of 0.95, indicating a perfect match between the prediction and the real website. Moreover, the true recall curve and area under the curve (AUC) analysis further support the effectiveness of this method with an AUC value of 98.8%. Significance tests, significant difference plots, and comparisons with previous studies confirm the effectiveness of the proposed method in outperforming the existing process. This study recognizes the limitations of feature selection and dataset size and suggests future research on improved features and larger datasets. Together, machine learning techniques have proven effective in identifying phishing websites and improving online security and customer trust in e-commerce and marketing.

In [12], the authors proposed HELPHED, a multi-level phishing email detection system that creatively mixes text-based content producing similar features. The process includes email parsing, content-based feature extraction and pre-processing and text feature extraction using Word2Vec, feature selection, and integrated classification. The authors clearly show all the stages and solve the problems of managing various phishing attacks. Through rigorous analysis, HELPHED consistently applies the best ML/DL classifiers and learning methods, achieving high scores, precision, recall and F1 scores. Adherence to the evaluation process ensures transparency and reliability. This study highlights the importance of addressing compound properties using specific techniques and determining the effectiveness of HELPHED in real-world situations. Compared to existing studies, HELPHED's results appear encouraging and are considered a significant hurdle in detecting email phishing. Method 2 overall achieves test performance with a higher F1 score of 0.9942 and better accuracy. Rank, precision, recall, AUC and MMC.

In [13], the authors highlight the important limitations of data in monitoring studies for phishing investigation, often considering the nature of the attack. Phishing websites. Given the significant difficulties in obtaining a comprehensive list of such sites and the lack of independent data without the work of third parties, the authors collected tagged data from many platforms containing more than 50,000 phishing and legitimate website URLs. This work uses a 21-element feature extraction process that considers parameters such as length, unique characters, and domain names, and employs a variety of learning models including Random Forest, Gaussian Naive Bayes, Decision trees, AdaBoost, KNN, XGBoost, and others. logistic regression. The results show the performance of XGBoost with the highest accuracy, precision, recall, F1 score, and AUC values of the benchmarks. This study demonstrates the effectiveness of semantic and overhead-based features of machine learning algorithms in real phishing detection. Additionally, he supports further research on parent-based features to improve performance and praises future research that may include deep learning to track the relationship between features.

In [14], the authors reported a malicious URL detection (CTI-MURLD) model based on network threats. Seven stages; data collection, feature preprocessing, feature extraction, feature representation, feature selection and joint learning-based prediction and decision making. It uses three classes of learning-based predictions, each using a different forest algorithm that trains on different features: URL-based, Google-based CTI, and WHOIS-based features. The model improves URL classification by combining different variables and using learning methods. In the evaluation, it was seen that the CTI-MURLD model outperformed traditional URL and WHOIS-based features, especially in terms of true, false positive, false negative price, amount of precision, return rate and F1 score. The accuracy of the random forest (RF) algorithm is as high as 96.8%, which outperforms deep learning models. Further advances have been found through feature selection, custom RF-based feature selection, and hyper parameter tuning via grid search. This work demonstrates the effectiveness of collaborative network threat detection capabilities for malicious URL detection and the advantages of collaborative learning in processing high-quality data, loud noise. The results contribute to the existing literature on threat detection techniques by highlighting the effectiveness of the model and its ability to precisely detect the phishing URL.

In [15], the authors proposed a phishing detection based on neural networks (CNN) to create a Broad base that can classify URLs as legitimate or phishing. . The program includes pre-processing, including URL concatenation, tokenization, padding, and sorting, followed by the embedding input layer. The test combines globally recognized data on 73,575 URLs to demonstrate the effectiveness of the system by balancing phishing and legitimate samples. CNN-based deep neural networks combined with dense neural networks achieve high precision, recall, accuracy, and F1 scores for both legitimate and phishing URLs. Performance metrics include 98% accuracy, recall, and F1 score, highlighting the robustness of the proposed system. The confusion matrix can provide additional information regarding the distribution of model values. Overall, this study demonstrates the success of the CNN-based phishing system with outstanding accuracy and instant detection performance.

In [16], authors present a method to use to interpret machine learning models (especially with native model translation) Phishing Detection System (LIME) and Translation Boosting Machine (EBM) This work contains legitimate and phishing URLs Ebbu2017 dataset and follows the same pattern as previous studies. The author introduced the LIME method, which uses random forest and SVM algorithms to achieve an accuracy of over 97.9%. The CPA model was chosen for its transparency and competitive reporting. The most important factor was the presence of top-level domains (TLDs) in Alexa's top 1 million websites. This article concludes that these translation models not only classify URLs but also provide insights into the decision-making process. This will make it easier to spot and understand phishing.

[17] write on the development of effective machine learning models to identify malicious URLs (mostly phishing websites). The database contains 3,000 URLs collected by Phish Tank, divided into malicious and legitimate groups. Perform multiple data preprocessing steps, including handling null values and merging data. Custom extraction involves extracting 15 features divided into URL column-based and column-based features to provide insight into URL patterns and features. Machine learning algorithms such as Random Forest Classifier, Decision Tree, Light GBM, Logistic Regression and SVM were used for custom classification. The lightweight GBM algorithm has better performance, with 89.5% training accuracy and 86.0% testing accuracy. Factor analysis shows the impact of certain

characteristics on improving accuracy. The study concluded that machine learning models, specifically the GBM model, can effectively detect phishing URLs.

In [18], the authors talk about the threat of phishing threats in various industries on the web, the need to measure security stability. This study uses data containing phishing and legitimate URLs and employs a variety of machine learning algorithms, including decision trees, naive Bayes, linear regression, K-neighborhood classifier, support vector machines, random forests, gradient boosting, and Hybrid model. Self-testing performs differently, with decision trees, unknown Bayes, and support vector machines performing well in many respects. Basically, Random Forest is the most efficient and outperforms other methods in terms of accuracy, precision, recall, and F1 score. In addition, the proposed LSD model combines hybrid models as well as linear, support vector machine and decision trees. Comparative analysis shows the best clothing and tree models based on methods and results, provides better insight into the data of phishing detection systems and encourages sharing of the combination of these models to increase efficiency and accuracy. This work discusses how future phishing detection systems should combine instruction-based and machine learning-based methods to prevent and detect phishing URLs.

In [19], the important role of URL phishing detection vision, awareness of the nature of cyber threats and advanced technologies. It is necessary to investigate. The article introduces the PILU-90K dataset, highlights its importance and novel content, and discusses the shortcomings of relying on regular data to show patterns. This study collects a literature review, evaluates existing methods, highlights the limitations of traditional classification methods, and encourages the adoption of machine learning (ML) and deep learning (DL) techniques. Various methods such as handcrafted, TF-IDF and data analysis using symbolic N-grams and CNN models are reviewed to provide a comprehensive understanding of their advantages and limitations. This study contributes to the literature by highlighting the importance of training models that use existing data and makes an interesting contribution to improving phishing detection.

In [20], the authors provide a description of the method for identifying faulty URLs using machine learning. This study used historical data published by detailed authors labeling URLs classified as "non-malicious" and "malicious". Data preprocessing involved removing duplicate data and checking for missing values; As a result, a data set containing 1,057,151 URLs was obtained. Feature extraction is done by lexical analysis, which produces 22 features such as URL length, hostname length, content, and various token-related features. To solve the class bias problem, random under sampling technique was used and features were further reduced based on variance and multivariate analysis. The final feature selection includes features such as URL length, title length, content, number of tokens, and content availability. Testing includes classification using Gradient Boosting Classifier (GBC), Support Vector Machine (SVM), Random Forest and Neural Network (ADAS). Statistical tests, including exact, regression, and F test, are used to evaluate model performance. The results show that the accuracy of each classification is very high, with the accuracy of SVM being as high as 99.896%. The paper provides a better understanding of malicious URL detection, demonstrating the effectiveness of the proposed method in handling inconsistent data and preventing overloading.

In [21], the authors proposed the FMPED and FMMPED algorithms and proposed a six-step process method to improve phishing e-mail on random files -mail distribution. These algorithms are based on the HELPHED framework and include segmentation, content removal, text processing, feature selection, background knowledge, and segmentation learning. Using a combination of decision trees and support vector machines in the soft learning process, the algorithm resolves inconsistencies in the data by applying advanced techniques such as removing overlapping patterns, poor quality, etc. Extensive experiments were conducted on the HELPHED dataset, and FMPED and FMMPED consistently outperformed existing methods across a variety of metrics, demonstrating their effectiveness in improving phishing email detection. These hybrid learning methods perform well, demonstrating the potential for practical use in improving email security.

In [22], the authors demonstrate the advantages of machine learning by addressing the long-standing problem of detecting malicious websites. The superiority of based methods over other dominance methods. They often highlight the effectiveness of support vector machine (SVM) and K-Nearest Neighbor (KNN) techniques. Additionally, the study investigates the impact of COVID-19 on smart people using fake information, fake URLs, and phishing attacks. The authors divide existing solutions into content-driven, URL-based, and machine learning. The proposed method integrates machine learning including AdaBoost, CatBoost, and Gradient Boosting Classifier to identify phishing websites based on various URL features. The dataset, attained from GitHub, undergoes examining analysis, preprocessing, and Stratified K-Fold cross-validation. AdaBoost arises as the most accurate model, achieving a ROC AUC Score of 99%. The study recommends practical applications in browser extensions for recognizing suspicious websites and proposes further exploration of URL and webpage content analysis.

[23] is the literature survey exposes assorted methodologies for phishing URL identification, with approaches extending from decision trees and Bayesian classifiers to ensemble learning. The suggested methodology introduces a hybrid ensemble model, combining classifiers like random forest, decision tree, MLP, and SVM, achieving an impressive 85.37% accuracy on a dataset of 20,000 URLs. Comparative analysis demonstrates its dominance over previous studies, outstanding accuracies of 82.6% and 90%. The model's prospective application as a browser extension for real-time phishing detection is highlighted, and future work could involve organizing the model and discovering deep learning techniques for further enhancement.

In [24], experimental teaching methods to fine-tune machine learning model performance , uses eight classification algorithms and evaluates three key areas of optimization: dataset balancing, hyperparameter optimization, and feature selection. This study uses data from the UCI and Mendeley repositories to standardize, balance and classify them for training and testing purposes. Reducing Kbest filtering options and performance metrics such as accuracy, precision, recall, and F1 score. Cross-validation confirmed the stability of the algorithm. This experiment determined the effect of data measurement, hyperparameter transformation, and specific choice of classification algorithm, with a slight improvement in data measurement varying the degree of improvement obtained from hyperparameter tuning and the number of features on the performance of the algorithm. The optimization plan is based on the existing system and derives a competitive percentage from different data and algorithms. Future directions include exploring technologies such as software-defined networking and blockchain for cybersecurity and addressing cloud security issues to detect threats and ensure risk mitigation for risk owners.

In [25], the authors demonstrated the use of machine learning (ML) and deep learning (DL) techniques for phishing detection, popular UCI-2015, Mendeley-2018 and Mendeley-2020 data analysis. Three deep learning models – Convolutional Neural Network (FCNN), Short Term Network (LSTM), and Convolutional Neural Network (CNN) – are compared with Random Forest (RF) integration into ML. Experiments conclude that RF outperforms DL models in terms of testing accuracy and learning time, demonstrating the performance of common ML, especially when dealing with data-critical texts. This study presents a unique selection process that provides a significant reduction of 87.6% with low accuracy. The discussion focused on the advantages of combining machine learning techniques with deep learning in phishing detection to solve problems such as changes in features and activity patterns. Future work includes different data authentication and real-world deployments against effective phishing attacks.

The EnLeM model was developed in the general context of machine learning (ML) and deep learning [26]. Phishing approved websites. The UCI Phishing Dataset serves as a benchmark for evaluating various models, including ML-based models (Decision Trees, Random Forests, AdaBoost, SVM, and k-NN) and DL-based models (1D-CNN and LSTM). The data contains 30 features extracted from legitimate and phishing pages and subjected to complex data preprocessing, including normalization, denoising, and extraction. Feature selection is then used using a random method (specifically Select-k-Best with shared data) to identify 20 key factors important for classification. This article introduces the new EnLeM method, a complex voting based on learning models, describes its architecture and demonstrates its advantages in achieving high accuracy (97.51%) and measurement performance. The comparative analysis demonstrates the performance of EnLeM over traditional ML and DL models and highlights its future importance in phishing web detection. The literature review thus forms the basis of the proposed EnLeM model and demonstrates the state's contribution of this model to the existing knowledge in the field.

In [27], a study on phishing activity was proposed using eight machine learning-based algorithms to classify web pages as legitimate or malicious. A special selection method improves the performance of the classification model by focusing on the classification and use of the most important features. The file contains 30 features extracted from legitimate and phishing pages on public websites. Experimental results show that Random Forest (RF) always provides the highest accuracy, reaching 96.52% before feature selection and 96.3% after feature selection. Information gain (IG) is used as a feature selection method, showing the 15 most important features for classification. This study also calculates the physical model showing that the time decreases after the feature is selected. Future work is to investigate the overall framework, examine the effectiveness of deep learning models, and extend the study to a larger phishing dataset containing IoT-based phishing attack data.

In [28], the phishing web detection method adopts custom selection and machine learning using DS-30 and DS-50 data. These techniques include data collection, preprocessing, wrapper removal, embedding, correlation coefficient, data boosting, and chi-square methods, as well as machine learning algorithms using random forest, optical gradient boosting, and class boosting. The final model Random Forest with coefficient-based feature selection achieved an accuracy of 97.47%. The implementation extends to the UI using React Native and Flask servers and provides a real-life implementation for URL lookup. Comparative analysis shows that the system has better performance than traditional methods with 97.47% accuracy. The API allows users to enter a URL and receive a notification indicating whether the URL is potentially malicious or insecure.

In [29], the authors conduct an evaluation of various machine learning techniques for phishing URL detection. They review existing studies and propose different approaches such as hybrid models, inference, and integration. The database used in the research consists of 48 features of legitimate and phishing websites, including lexical features, host features and related features. Phishing URL detection architecture strategy includes data collection, transformation, feature selection, visualization, and design. Use a variety of machine learning algorithms, including logistic regression, KNN, support vector machines, pruning trees, random forests, and ensemble methods. The results show that the LGBM classifier has an accuracy rate of 98.4%, which is better than other algorithms and proves its effectiveness in identifying phishing URLs. This study also includes hyperparameter changes to improve the LGBM model. Overall, this study provides insight into effective strategies for detecting phishing URLs using machine learning.

In [30], the paper proposes a comprehensive approach to URL classification, combining various models and datasets to achieve a high accuracy of 95.3%. Data collection involves standard datasets and web scraping, with derived datasets covering lexical analysis, HTML tags, domain information, web page text, DOM tree, BERT, Alexa, and ensemble outputs. Models include LSTM for text and DOM tree, BERT for URL classification, and decision trees for special symbols, HTML tags, and domain features. An ensemble model integrates these for a collective prediction displayed through a Flask-based frontend. Observations highlight LSTM's high accuracy (99.98%), and decision trees' significant contributions. The paper acknowledges the change in negative URLs, advocates for continued research, and emphasizes that strong integration, change, and error impact other models such as SVM and KMeans provide a less efficient basis for URL classification.

III. AI-DRIVEN URL PHISHING DETECTION

A. Enhancing Prediction Accuracy Through Feature Selection and Ensemble Learning

The solution improves accuracy by choosing a strategy that filters various features of the original data based on their relevance. This process selects key points for determining the prediction as shown in two datasets: DS-30 with 30 features and DS-50 with 50 features. By focusing on the best features, random features have less impact on the model's prediction and accuracy.

We also use the above information to model the learning process. Predictions from multiple models were combined equally across each model, eliminating bias for each model. Therefore, the system shows the majority of votes obtained by adding up the results of each model. For example, the ensemble's final prediction confirms a website as phishing if the majority of models predict it as such.

B. Phishing Website Datasets

The proposed system obtains DS-1 and DS-2 tagged phishing website data from UCI machine learning repository and Kaggle. DS-1 has 30 features and DS-2 has 50 features and consequences. To optimize the dataset, feature selection techniques are used to improve detection accuracy and performance by removing irrelevant features. This step ensures that only important features help determine whether a website is genuine or phishing.

C. Cross Validation of The Models

To assess their mean accuracy, an initial selection of 10 widely recognized classifiers undergoes evaluation through a stratified k-fold cross-validation procedure. The chosen classifiers include Gradient Boosting, Random Forest, Support Vector Classifier (SVC), Decision Tree, Ada Boost, Extra Trees, Multiple Layer Perceptron (Neural

Network), K-Nearest Neighbors (KNN), Logistic Regression and Linear Discriminant Analysis. The outcomes of this validation process are presented in the following Tables 1-2.

Table 1. Cross Validation Results DS-30

Algorithm	Cross Val Means	Cross Val errors
Random Forest	0.971644	0.006727
Extra Trees	0.970563	0.007782
Multiple Layer Perceptron	0.966514	0.006524
Decision Tree	0.956385	0.004848
Ada Boost	0.955844	0.006981
Gradient Boosting	0.946803	0.008312
SVC	0.946803	0.005837
K Nearest Neighbors	0.938158	0.007199
Logistic Regression	0.92925	0.007638
Linear Discriminant Analysis	0.921552	0.006087

In evaluating the performance of different machine learning algorithms on the DS-30 dataset, Table 1 presents mean accuracy scores under "Cross Val Means," and the standard errors for each algorithm are denoted by "Cross Val Errors." Notably, Random Forest and Extra Trees emerged as the top-performing algorithms, demonstrating the highest accuracy results.

Table 2. Cross Validation Results DS-50

Algorithm	Cross Val Means	Cross Val errors
Random Forest	0.982687	0.004438
Extra Trees	0.981642	0.004862
Gradient Boosting	0.977015	0. 004823
Decision Tree	0.965821	0.006206
Ada Boost	0.964925	0.010238
Multiple Layer Perceptron	0.963284	0.007847
Linear Discriminant	0.938806	0.007669
Logistic Regression	0.934776	0.009488
K Nearest Neighbors	0.844627	0.012042
SVC	0.844627	0.012042

Table 2 illustrates the performance of different machine learning algorithms on the DS-50 dataset, with "Cross Val Means" indicating mean accuracy scores. Notably, Random Forest and Extra Trees exhibited the highest accuracy, as depicted in the "Cross Val Errors," which represent the standard errors for each algorithm.

Based on these results, Random Forest, Decision Tree, Extra Trees, Ada Boost, and Neural Network were selected as classifiers for the proposed approach. These classifiers underwent network search optimization to enhance their performance. The study focused on assessing overfitting tendencies using learning curves during the calculation process. Particularly, the Neural Network demonstrated minimal vulnerability to overfitting when applied to the DS-50 dataset, showcasing its robustness. Similarly, for the DS-30 dataset, Random Forest displayed the lowest overfitting tendency, making it a suitable choice. In contrast, Ada Boost exhibited a higher tendency for overfitting in the DS-50 dataset, while Decision Tree showed similar behavior in the DS-30 dataset. These findings offer valuable insights into the behavior and performance of classifiers in diverse dataset scenarios, aiding in their selection and application. Visual representation of learning curves further enhances comprehension of these characteristics.

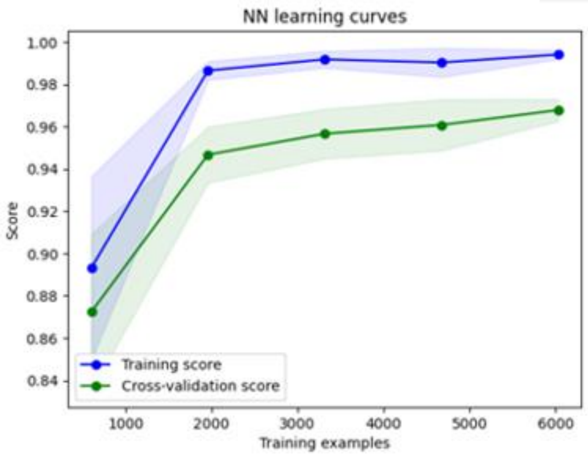

Figure 1. NN Learning Curve (DS-50)

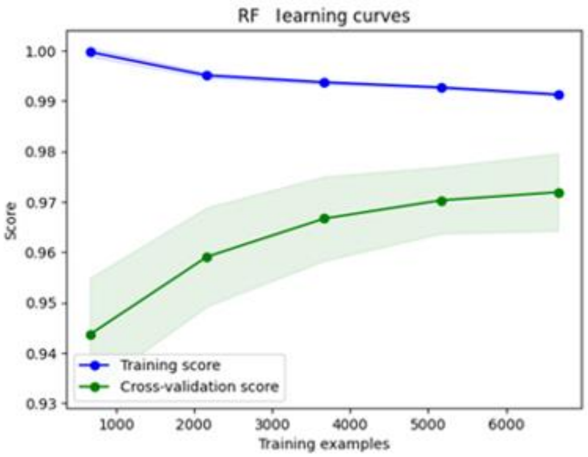

Figure 2. RF Learning Curve (DS-30)

The results in Figure 1 indicate that the Neural Network exhibited the least overfitting in the context of the DS-50 dataset, while in Figure 2 Random Forest (RF) displayed the lowest overfitting tendency when considering DS-30.

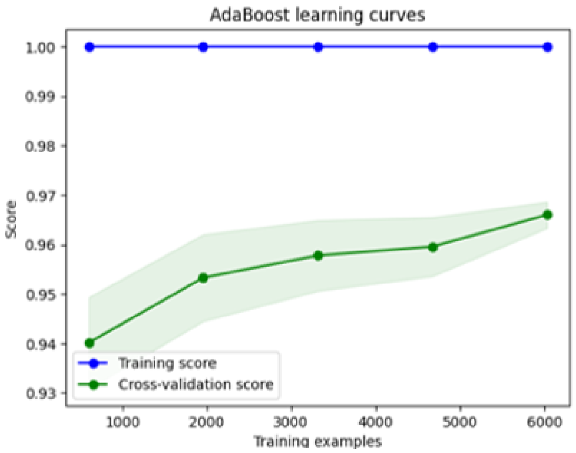

Figure 3. Ada Boost Learning Curve (DS-50)

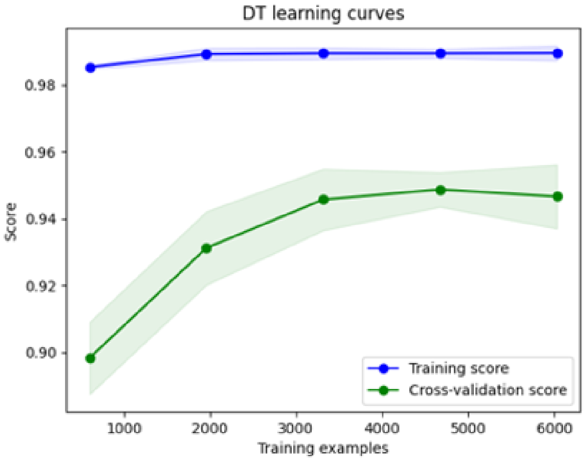

Figure 4. DT Learning Curve (DS-30)

In Figure 3 and 4 Ada Boost exhibited the highest overfitting tendency when evaluating the DS-50 dataset, while Decision Tree showed the highest overfitting rate for the DS-30 dataset. The presented figures depict the learning curves for both the best- and worst-case scenarios, providing insights into the overfitting tendencies of the classifiers.

D. Ensemble Learning

The combined results of the selected classifiers were used to train the model with a voting classifier. Five classifiers were passed as estimators, and a 'soft' voting strategy was applied to consider the probability of each vote. The datasets were split into 0.3 and 0.7 ratios for testing and training, respectively.

IV. RESULTS AND DISCUSSION

The experiment revealed that Random Forest Classifier, Extra Trees, Ada Boost, Multilayer Neural Network, and Decision Tree achieved accuracies of 96% and 98% with the ensemble method using both datasets, DS-30 and DS-50, respectively. Consequently, as the number of features increased, ensemble learning demonstrated the best accuracy. The performance measures for both datasets are provided in the following Table 3-4.

Table 3. Summary of tests (DS-30)

DS-30 (dataset with 30 features)				
Class	**Precision**	**Recall**	**F1-Score**	**Support**
Phishing	0.98	0.95	0.96	1565
Legitimate	0.96	0.98	0.97	2084
Accuracy	0.97	None	None	3649
Macro Avg	0.97	0.97	0.97	3649
Weighted Avg	0.97	0.97	0.97	3649

Table 4. Summary of tests (DS-50)

DS-50 (dataset with 50 features)				
Class	**Precision**	**Recall**	**F1-Score**	**Support**
Phishing	0.98	0.99	0.98	1622
Legitimate	0.99	0.98	0.98	1678
Accuracy	0.98	None	None	3300
Macro Avg	0.98	0.98	0.98	3300
Weighted Avg	0.98	0.98	0.98	3300

The ensemble method, employing an elective classifier, attains an accuracy of 96% for DS-30 and 98% for DS-50 as represented in Table 3 and Table 4, underscoring the effectiveness of ensemble learning in enhancing accuracy. The results suggest that as the number of features increases, ensemble learning becomes progressively advantageous for improving accuracy.

V. CONCLUSION

In conclusion, the proposed model, combining feature selection and ensemble learning, has shown promising results in detecting phishing websites. By selecting the most significant features from datasets DS-30 and DS-50 and employing ensemble learning with cross-validation, the model achieved accuracy levels of 96% and 98%, respectively. While demonstrating reasonable precision, recall, and F1-scores, the model did exhibit a notable number of false negatives in the DS-30 dataset, indicating potential for refinement. Future enhancements could involve merging related datasets and utilizing machine learning to extract crucial features. Overall, this model provides a robust foundation for advancing cybersecurity efforts with high accuracy and offers potential for further research and development in the field.

ACKNOWLEDGEMENT

The authors received no funding from any party for the research and publication of this article.

AUTHOR CONTRIBUTIONS

Shougfta Mushtaq : Established the research concept, designed the methodology, literature reviewed and edited the manuscript.
Tabassum Javed: Conducted the research experiments, compiled the data, and implemented the formal analysis.
Mazliham Mohd Su'ud: Supervised the research project, guide in all aspects of paper

CONFLICT OF INTERESTS

No conflict of interests were disclosed.

REFERENCES

[1] K. M. Pratt, “What is a Cyber Attack? Definition, Examples and Prevention TechTarget,” *TechTarget*. 2022. [Online]. Available: https://www.techtarget.com/searchsecurity/definition/cyber-attack

[2] A. A. Alsufyani and S. M. Alzahrani, “Social engineering attack detection using machine learning: Text phishing attack,” *Indian J. Comput. Sci. Eng.*, vol. 12, no. 3, pp. 743–751, 2021, doi: 10.21817/indjcse/2021/v12i3/211203298.

[3] D. He, X. Lv, S. Zhu, S. Chan, and K.-K. R. Choo, “A Method for Detecting Phishing Websites Based on Tiny-Bert Stacking,” *IEEE Internet Things J.*, vol. PP, p. 1, 2023, doi: 10.1109/JIOT.2023.3292171.

[4] M. A. Musarat Hussain, Chi Cheng, Rui Xu, “CNN-Fusion: An effective and lightweight phishing detection method based on multi-variant ConvNet,” *Inf. Sci. (Ny).*, vol. 631, no. July 2022, pp. 328–345, 2023, doi: 10.1016/j.ins.2023.02.039.

[5] K. S. N. Sushma, M. Jayalakshmi, and T. Guha, “Deep Learning for Phishing Website Detection,” *MysuruCon 2022 - 2022 IEEE 2nd Mysore Sub Sect. Int. Conf.*, pp. 1–6, 2022, doi: 10.1109/MysuruCon55714.2022.9972621.

[6] Pavansai and G. G. sai Ziaul Haque Choudhury, “Classification of Phishing Website Using Hybrid Machine Learning Techniques,” vol. 8, no. 7, pp. 1385–1390, 2023.

[7] H. Abusaimeh and Y. Alshareef, “Detecting the Phishing Website with the Highest Accuracy,” *TEM J.*, vol. 10, no. 2, pp. 947–953, 2021, doi: 10.18421/TEM102-58.

[8] G.Ravi Kumar, Dr.S.Gunasekaran and Nivetha.R, “Url Phishing Data Analysis and Detecting Phishing Attacks Using Machine Learning in Nlp,” *Int. J. Eng. Appl. Sci. Technol.*, vol. 3, no. 10, pp. 26–31, 2019, doi: 10.33564/ijeast.2019.v03i10.007.

[9] S. Dangwa and A.-N. M. School, “Feature Selection for Machine Learning-based Phishing Websites Detection,” *2021 Int. Conf. Cyber Situational Awareness, Data Anal. Assessment, CyberSA 2021*, pp. 1–6, 2021, doi: 10.1109/CyberSA52016.2021.9478242.

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

[11] A. Taha, “Intelligent ensemble learning approach for phishing website detection based on weighted soft voting,” *Mathematics*, vol. 9, no. 21, 2021, doi: 10.3390/math9212799.

[12] P. Bountakas and C. Xenakis, “HELPHED: Hybrid Ensemble Learning PHishing Email Detection,” *J. Netw. Comput. Appl.*, vol. 210, Jan. 2023, doi: 10.1016/j.jnca.2022.103545.

[13] B. Maini, A., Kakwani, N., B, R., M K, S., & R, "Improving the Performance of Semantic-Based Phishing Detection System Through Ensemble Learning Method," *2021 IEEE Mysore Sub Sect. Int. Conf. MysuruCon 2021*, pp. 463–469, 2021, doi: 10.1109/MysuruCon52639.2021.9641614.

[14] M. Alsaedi, M., Ghaleb, F. A., Saeed, F., Ahmad, J., & Alasli, "Model Using Ensemble Learning," *Sensors*, pp. 1–20, 2022.

[15] R. Singh, S., Singh, M. P., & Pandey, "2nd Phishing Detection from URLs Using Deep Learning."

[16] K. A. Galego Hernandes Jr., P. R., Floret, C. P., Cardozo de Almeida, K. F., Camargo da Silva, V., Papa, J. P., & Pontara da Costa, "Phishing Detection Using URL-based XAI Techniques," in *2021 IEEE Symposium Series on Computational Intelligence, SSCI 2021 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/SSCI50451.2021.9659981.

[17] M. D. K. Hasane Ahammad, S. K., Kale, S. D., Upadhye, G. D., Pande, S. D., Venkatesh Babu, E., Dhumane, A. V., & Jang Bahadur, "Phishing URL detection using machine learning methods," *Adv. Eng. Softw.*, vol. 173, no. January, p. 103288, 2022, doi: 10.1016/j.advengsoft.2022.103288.

[18] S. R. K. Karim, A., Shahroz, M., Mustofa, K., Belhaouari, S. B., & Joga, "2nd Phishing Detection System Through Hybrid." IEEE, pp. 36805–36822, 2023. doi: 10.1109/ACCESS.2023.3252366.

[19] V. Sanchez-Paniagua, M., Fidalgo Fernandez, E., Alegre, E., Al-Nabki, W., & Gonzalez-Castro, "Phishing URL Detection: A Real-Case Scenario Through Login URLs," *IEEE Access*, vol. 10, pp. 42949–42960, 2022, doi: 10.1109/ACCESS.2022.3168681.

[20] K. M. and S. A.-H. B. M. Abutaha, M. Ababneh, "2nd URL Phishing Detection using Machine Learning," in *URL Phishing Detection using Machine Learning Techniques based on URLs Lexical Analysis*, Spain, 2021, pp. 147–152. doi: doi: 10.1109/ICICS52457.2021.9464539.

[21] C. Qi, Q., Wang, Z., Xu, Y., Fang, Y., & Wang, "Enhancing Phishing Email Detection through Ensemble Learning and Undersampling," *Appl. Sci.*, vol. 13, no. 15, 2023, doi: 10.3390/app13158756.

[22] M. Kaibassova, D., Saginov, A., Nurtay, M., Tau, A., & Kissina, "Solving the Problem of Detecting Phishing Websites Using Ensemble Learning Models," *Sci. J. Astana IT Univ.*, pp. 55–64, 2022, doi: 10.37943/12oyrs4391.

[23] A. Pandey and J. Chadawar, "Phishing URL Detection using Hybrid Ensemble Model," *Artic. Int. J. Eng. Tech. Res.*, vol. 11, no. 04, pp. 479–482, 2022, [Online]. Available: https://www.researchgate.net/publication/360412387

[24] A. Raja, S. A., Balasubaramanian, S., Al-Kaabi, A. S., Sharma, B., Chowdhury, S., Mehbodniya, A., Webber, J. L., & Bostani, "Analysis of the Performance Impact of Fine-Tuned Machine Learning Model for Phishing URL Detection," *Electron.*, vol. 12, no. 7, 2023, doi: 10.3390/electronics12071642.

[25] Y. Wei and Y. Sekiya, "Sufficiency of Ensemble Machine Learning Methods for Phishing Websites Detection," *IEEE Access*, vol. 10, no. November, pp. 124103–124113, 2022, doi: 10.1109/ACCESS.2022.3224781.

[26] M. A. Yeasmin, M. N., Refat, M. A. R., Singh, B. C., Alom, Z., Aung, Z., & Azim, "EnLeM : An Ensemble Learning-based Model for Detecting Phishing Websites," 2023.

[27] E. Gandotra and D. Gupta, "An Efficient Approach for Phishing Detection using Machine Learning," pp. 239–253, 2021, doi: 10.1007/978-981-15-8711-5_12.

[28] M. S. Khatun, M., Mozumder, M. A. I., Polash, M. N. H., Hasan, M. R., Ahammad, K., & Shaiham, "An Approach to Detect Phishing Websites with Features Selection Method and Ensemble Learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 8, pp. 768–775, 2022, doi: 10.14569/IJACSA.2022.0130888.

[29] T. Agarwal, G., Goel, C., Jindal, K., & Subbulakshmi, "Visualisation and Classification of Phishing URL using Ensemble Learning Algorithms and Hyper-Parameter Tuning," *ICSCCC 2023 - 3rd Int. Conf. Secur. Cyber Comput. Commun.*, pp. 13–18, 2023, doi: 10.1109/ICSCCC58608.2023.10176642.

[30] U. Venugopal, S., Panale, S. Y., Agarwal, M., Kashyap, R., & Ananthanagu, "Detection of Malicious URLs through an Ensemble of Machine Learning Techniques," *2021 IEEE Asia-Pacific Conf. Comput. Sci. Data Eng. CSDE 2021*, pp. 1–6, 2021, doi: 10.1109/CSDE53843.2021.9718370.