International Journal of Scientific Research in Science, Engineering and Technology

Print ISSN - 2395-1990
Online ISSN : 2394-4099

OPEN ACCESS

Available Online at :www.ijsrset.com
doi : https://doi.org/10.32628/IJSRSET251211

An Ensemble Voting Classifier based on Machine Learning Models for Phishing Detection

Enas Mohammed Hussien Saeed

Department of Computer Science, College of Education, University of Mustansiriyah, Baghdad, Iraq
drenasmohammed@uomustansiriyah.edu.iq

ARTICLEINFO

Article History:

Accepted : 01 Jan 2025
Published: 03 Jan 2025

Publication Issue :
Volume 12, Issue 1
January-February-2025

Page Number :
15-27

ABSTRACT

The pervasive threat of phishing attacks has necessitated the development of more effective detection systems. This paper introduces a novel ensemble hard voting classifier that integrates the predictive capabilities of Logistic Regression, Gradient Boosting, and K-Nearest Neighbors for the identification of phishing websites with enhanced accuracy. Our methodology encompasses a comprehensive analysis starting with a rich dataset from Kaggle, consisting of over 11,000 websites, each described by 30 features. Through meticulous exploratory data analysis, we have discerned significant patterns and feature correlations, which informed the subsequent data preprocessing phase. We standardized feature scales using the StandardScaler and split the dataset into an 80-20 ratio for training and testing, ensuring both effective model learning and validation. The ensemble model capitalizes on the diversity of its constituent classifiers, outperforming individual models with an accuracy of 95.02%. Our approach demonstrates that an ensemble hard voting classifier not only improves the detection rate but also provides a balanced precision-recall performance, crucial for real-world applications.

Keywords : Phishing Detection, Ensemble Learning, Hard Voting Classifier, Machine Learning, Cybersecurity

I. INTRODUCTION

Phishing attacks (PA) are a persistent and sophisticated threat in the complicated and changing world of cyberse- curity. They erode the foundation of digital trust by exploiting human weaknesses and increasing in sophistication with each technical advancement. Social engineering has reached new heights in the digital era, with phishing methods developing in both subtlety and scope [1]. This has resulted in an unprece- dented surge in such instances, as indicated by the latest Anti- Phishing Working Group report, which shows a whopping 65% increase in phishing assaults in the last year alone [2]. This terrifying wave has left a trail of devastation, affecting millions worldwide and

resulting in significant financial and informational losses.

To put it into perspective, phishing email statistics show that approximately 1.20% of all e-mails sent are malicious, or to 3.4 billion phishing e-mails every day. Email phishing is thought to be the starting point for 90% of successful cyber-attacks. Phishing was the most common infection type among Asian organizations in 2021, accounting for 43% of all assaults on the region [2]. Moreover, phishing attacks' count has increased by 61% since the year 2021. With a ransomware or phishing attack occurring every 11 seconds, 2023 is expected to see theft of over 33 million records [2].

These figures highlight the great frequency of cyberattacks expected throughout 2023 and beyond [3]. It is imperative that we remain aware and knowledgeable as we negotiate such difficult challenges thereby strengthening our defenses against this always changing threat [4]. From heuristic as well as signature-based detection to blacklist-based detection, conventional protections against such attacks have long been like shields guarding our digital boundaries [5]. But when their static algorithms fail against the dynamic and polymorphic nature of contemporary phishing ploys, such conventional paradigms are progressively considered unsatisfactory. Their inherent rigidity lags behind the changing strategies of cyber adversaries, which frequently results in a significant amount of false positives compromising user trust and hence affecting user experience [6]. More importantly, such techniques exhibit a clear vulnerability in their inability to prevent and counteract zero-hour PAs, meticulously created to elude until they impact [7]. The growing complexity of threat actors requires a change in defense systems. The solution might be found in the field of ensemble techniques in machine learning (ML)—a domain where combined intelligence and adaptability reign supreme [8]. Ensemble techniques stand out by combining several learning algorithms, therefore leveraging their combined strengths to create a more powerful barrier against threats [9]. Among the ensemble approaches, most voting turns out as a sophisticated yet powerful approach. Under the consensus principle, which holds that the majority agreement among several classifiers determines the final conclusion, In addition to improving detection accuracy, less accurate predictions lower noise [10]. Our research article explores the creation regarding a unique majority voting-based ensemble framework designed especially to strengthen phishing protection. This new method not only combines several learning approaches but also improves their cooperation by means of a computed weighting mechanism [11]. Depending on their reliability and performance in practical environments, this mechanism is meant to assess and modify the impact of particular classifiers. By means of thorough investigation as well as performance benchmarking regarding every classifier in identifying phishing websites and emails, we build an adaptive system that is not static, yet develops in reaction to the changing phishing threat landscape. We provide in this work a thorough review of our approach and application. The data highlight a notable improvement in phishing detection rates, a drop in false positives, and a resistance against the most recent and advanced attacks. This methodological development marks a paradigm change in how phishing protection systems may adapt, learn, and preempt the several phishing expeditions that attack the digital sphere nowadays. It is not only incremental. Our contribution provides a peek into the future when cybersecurity systems are as dynamic and intelligent as the dangers they are meant to prevent, therefore laying the foundation for what might become a new benchmark in the proactive defense against phishing.

II. LITERATURE REVIEW

In the critical domain of cybersecurity, the research presents a whole picture of the changing threat

landscape and the continuous initiatives to create robust defense systems. This section analyzes a series of studies investigating several approaches to combat phishing, a main type of cyberattack taking advantage of internet user weaknesses. With an emphasis on phishing as a means of allowing cybercriminals to access sensitive data and systems, the study [12] explores the growing cybersecurity risks resulting from growing digital transformation. It underlines the need of in-depth protection measures and admits the shortcomings of present anti-phishing techniques. The main focus of the work is the creation of ML model to identify PAs by use of Decision Tree (DT) and Random Forest (RF) algorithms. Tested on a Kaggle dataset utilizing Principal Component Analysis (PCA) for feature selection, the model showed a notable accuracy of 97% in spotting phishing attempts with the use of RF technique. The development and application of a model for PAs detection using supervised ML methods is presented in paper [13]. It describes following standard ML cycles doing a literature analysis to identify characteristics of phishing-infected emails and the construction of a model integrating Naive Bayes (NB) and DT algorithms. With the model tested in a controlled setting with the use of PhishTank, tools employed include the Jupyter framework and Python. Based on the literature review, the research contains a validation regarding the accuracy of the model against recognized techniques as RF, Fictitious Classifier, and Logistic Regression (LR). An important security step for safe internet browsing, the research [14] looks at many ML approaches to find phishing websites. It seeks the best way to identify common cyberattacks, therefore enabling faster iden-tification and blacklisting of risky websites and improving general web security. The paper presents thorough explanations of the examined approaches together with evaluation approaches to visually show their effectiveness. The most successful method for phishing website deteciton, it is concluded, the RF Classifier.

The paper [15] tackles the prevalent issue of PAs on the internet, where attackers deceive users into divulging sensi- tive information through fraudulent means like emails and deceptive webpages. The study explores the use of three ML algorithms focused on URL-based features to detect phishing websites and prevent Zero-Day attacks. The proposed model, which operates solely on URL analysis without requiring additional resources, demonstrated high effectiveness in differ- entiating between legitimate and phishing sites. The Random Forest classifier, in particular, showed high precision (97%), recall (99%), and an F1 Score of 97%, indicating a fast and efficient system for phishing detection compared to previous studies.

Especially relevant in the context of increasing remote working throughout COVID-19 pandemic, the research [16] offers a new ensemble model for PAs. It describes how k-nearest neighbors (KNN), artificial neural network (ANN), and decision tree (C4.5)—integrated with a Random Forest Classifier (RFC) in an ensemble method to increase the accuracy of phishing detection on websites. With the KNN and RFC ensemble scoring an accuracy of 97.33%, the model shows better performance than current techniques. One of the main ideas of the suggested model is the combination of the classifiers with RFC as the basis using a voting technique.

The paper [17] addresses the increased cybersecurity risks associated with the shift to remote work during the COVID- 19 pandemic, with a specific focus on phishing—a prevalent cybercrime that involves deceiving individuals into giving up their credentials via fake webpages. Despite existing defenses like blacklists, whitelists, and antivirus software, attackers continue to find new ways to breach security. The study introduces a data-driven deep learning framework for the detection of phishing web-pages,

utilizing a multilayer perceptron (feed- forward neural network) for prediction. The model has been trained and tested on a dataset from Kaggle, featuring ten thousand webpages with ten attributes, and achieved a high level of accuracy: 95% in training and 93% in testing scenarios.

The paper [18] outlines the persistent challenge of PAs, particularly those targeting email systems, and the inadequacy of existing anti-phishing methods. It presents a ML-based tech- nique for detecting PAs, developed after analyzing more than 4,000 phishing e-mails that were aimed at the University of North Dakota's email service. The authors created a dataset with 10 significant features for training and testing the ML models. The performance of these models was evaluated using 4 metrics, which are: probability of detection, probability of miss-detection, probability of false alarm, and accuracy. The results indicate that artificial neural networks offer improved detection capabilities for PAs.

The paper [19] introduces a new method for detecting PAs by analyzing hyperlinks in the HTML source code of websites. It identifies unique hyperlink-specific features, categorized into 12 groups, to train ML algorithms. The performance of this detection approach was tested on datasets of phishing and legitimate websites using various classification algorithms. The solution operates entirely on the client side, without reliance on third-party services, and is language independent, capable of detecting phishing on websites in any language. The method notably achieved over 98.4% accuracy using the LR classifier, indicating a higher detection rate compared to other existing approaches.

The paper [20] addresses the shift from traditional crimes like bank or shop robbery to cybercrimes, with a particular focus on phishing, where attackers use fake websites to steal sensitive user information like account IDs and passwords. The challenge lies in distinguishing legitimate web pages from phishing ones, given the sophisticated nature of these semantic-based attacks that exploit user vulnerabilities. De- spite new anti-phishing software utilizing blacklists, heuristics, and ML, none have been fully successful in preventing PAs. The paper proposes a novel real-time anti-phishing system using seven different classification algorithms and natural language processing (NLP) features. The system boasts lan- guage independence, the ability to work with large datasets, real-time functionality, detection of new sites without third- party services, and feature-rich classifiers. A new dataset was constructed for testing, and the Random Forest algorithm, using NLP features, achieved the highest accuracy at 97.98% for detecting phishing URLs.

The paper [21] discusses the dual nature of the internet as a vital resource and a platform for anonymous malicious activ- ities, focusing on phishing—where attackers deceive victims to steal sensitive information. It acknowledges the evolution of phishing techniques to evade detection and posits ML as a successful method for identifying common phishing characteristics. The paper presents a comparative analysis of various ML methods for the prediction and detection of phishing websites.

III. PROPOSED APPROACH

Our methodical approach to creating a phishing website de- tector involves a comprehensive process from data acquisition to evaluation as shown in Figure 1. We commence with a detailed Data Description of over 11,000 websites, each with

30 features, forming the basis for our binary classification model. Through Exploratory Data Analysis (EDA), we analyze patterns and feature correlations, utilizing visual tools like Class Distribution and Correlation Heatmaps. In the Data Preprocessing step, we employ StandardScaler to normalize features and split the dataset into an 80-20 training-testing ratio to ensure model effectiveness and avoid overfitting.

The Modeling phase involves a hard voting ensemble classifier that synergizes the strengths of KNN, Gradient Boosting (GB), and LR to enhance prediction accuracy. The Evaluation of the model's performance is thorough, encom- passing accuracy, precision, recall, f1-score, and analysis via the confusion matrix. This structured approach is crafted to yield a sophisticated detection tool to combat phishing threats.

A. Data Descripton

Retrieved from Kaggle [22], the dataset used for the Phishing Website Detector project is a complete collection of data points related to over 11,000 sites. Every website in the dataset is labeled as either a phishing site (1) or not a phishing site (-1) together with thirty distinct parameters. Both text as well as CSV file formats of such dataset enable simplicity of usage in model building procedures. The dataset comes with a code template that loads the data alongside thorough definitions of input and output variables and helps import required modules. For project scoping, in which the functional and non-functional prerequisites of the system to be developed are laid down, this is especially helpful. Moreover, the dataset is set up to enable the development of a Python Scikit-Learn binary classification model. The model seeks to precisely ascertain whether a website is phishing site. Every attribute in the dataset—from the use of IP addresses to the presence of SSL certificates—is classified using categorical signed numerical values, therefore offering a rich and complex input for strong model training.

B. Exploratory Data Analysis (EDA)

This section explores the Exploratory Data Analysis (EDA) for our phishing detection research, a foundational stage that lets us dissect and understand the intrinsic patterns as well as associations in the dataset. Two classes' frequency distribution within a dataset is shown in the bar chart Class Distribution in Phishing Data in Figure 2: "Legitimate" and "Phishing". While the "Phishing" class, shown by the green bar, relates to the number of phishing website instances in the dataset, the "Legitimate" class—shown by the blue bar—indicates the number of valid website instances in the dataset. The chart clearly shows that both classes have a significant frequency, implying a balanced dataset that would be perfect for training ML models since it offers enough instances of both types of websites.

Such balanced datasets are crucial for developing robust models that can accurately classify new, unseen websites as either phishing or legitimate. The actual numbers are not visible, but the relatively equal height of the bars suggests that the dataset likely contains a near-equal number of samples from each class, thus mitigating the risk of a classification bias towards the more frequent class.

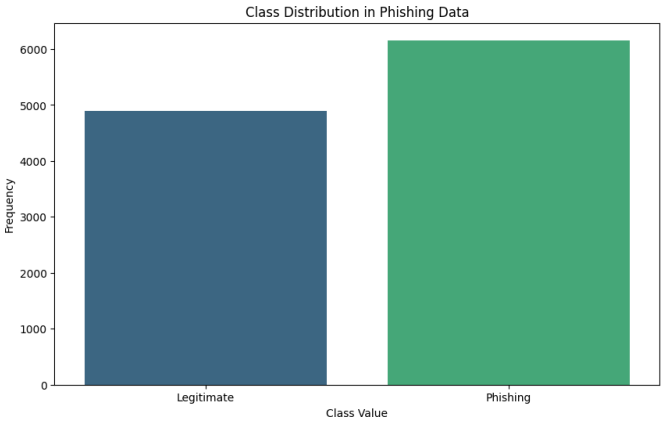

Fig. 2. Class Distribution in Phishing Data

The "Correlation Heatmap of Phishing Data Features" pro- vides a visual representation of the relationships between different features within a phishing dataset (Figure 3). This heatmap uses color intensities to indicate the degree of corre- lation between pairs of features; darker red signifies a stronger positive correlation, darker blue indicates a stronger negative correlation, and lighter colors denote weaker relationships. The diagonal, a line of perfect positive correlation, represents the relationship of each feature with itself. Notably, some features exhibit a marked

correlation with the 'class' variable, suggesting a stronger predictive power for those features in determining whether a website is phishing or legitimate. The heatmap serves as an analytical tool to identify which features might be redundant (highly correlated with each other) or most informative for building a classification model. Understanding these correlations is critical for feature selection and engineer- ing, which in turn, can considerably impact performance of ML models tasked with detecting phishing activities.

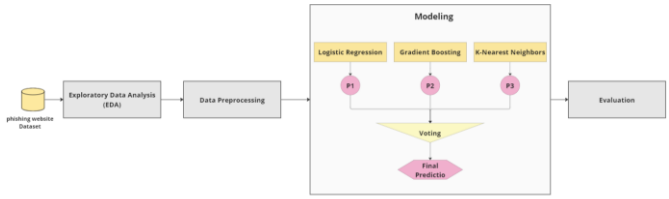

Fig. 1. Proposed Approach

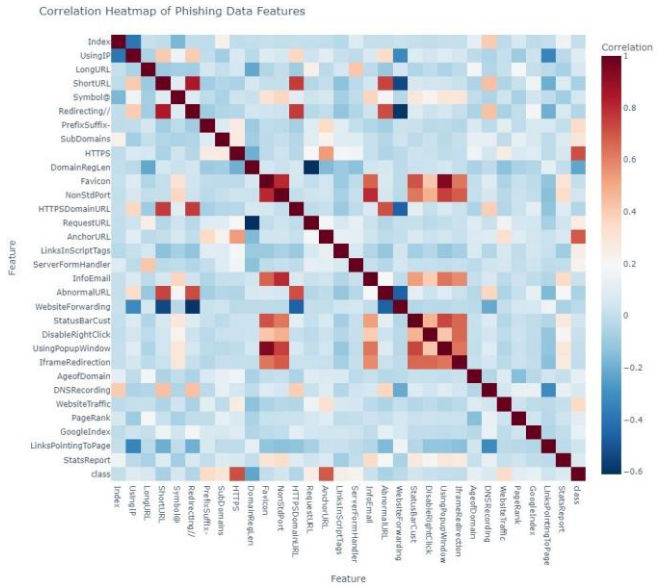

Fig. 3. Correlation Heatmap of Phishing Data Features

C. Data Preprocessing

In the data preprocessing phase of our ML pipeline, a critical step undertaken was the normalization of feature scales using the StandardScaler. This step is pivotal as it ensures that each feature contributes equally to the distance calculations in our model, preventing any single feature with a larger scale from dominating the learning process. StandardScaler standardizes the features through the removal of mean and scaling to unit variance, a process that is particularly beneficial for algo- rithms that are sensitive to feature scaling. The equation for the StandardScaler, which standardizes features through removing the mean and scaling to unit variance, is given by:

$$z = \frac{(x - \mu)}{\sigma} \quad (1)$$

In this equation σ is the standard deviation of the data, x is the original data, and μ is the mean of the data. The standardized score that results, z, is Following normalisation, we split our dataset in two separate sets: 20% for testing and 80% for training. This split is a well-known method that guarantees a distinct dataset to impartially assess the model's performance and lets a significant volume of data be utilized in training the model, therefore assuring it learns successfully. Since the model is tested on unknown data and offers a consistent estimate of its generalization ability, this method aids in reducing overfitting. Developing a strong model that can effectively anticipate phishing attempts when used in real-world situations depends much on the careful balance of training and test data.

D. Modeling

In the realm of ensemble learning, the voting model ar- chitecture stands out for its robustness and simplicity [23]. It functions on the principle of democracy, where each individual classifier in the ensemble casts a 'vote' for a particular class label for a given input *x*. These classifiers, which are an as- semblage of diverse algorithms with distinct decision-making strategies, ensure a well-rounded approach to classification tasks. In our model, we have *C*, the ensemble of classifiers, comprising c_1, c_2, and c_3, where c_1 represents K-Nearest Neighbors (KNN) [24], c_2 Gradient Boosting (GB) [25], and c_3 Logistic Regression (LR) [26].

The process is as follows: for a given input x, each classifier ci within C generates a prediction pi(x). These predictions can be viewed as votes. The ensemble hard voting classifier *V* then synthesizes these votes to deliver a consensus prediction *P* (*x*). The function *P* (*x*) is defined as the mode of the

predictions from the individual classifiers, mathematically expressed as:

$$P(x) = \text{mode}\{p_1(x), p_2(x), p_3(x)\}$$

In this expression, the mode is the statistical measure that identifies the most frequently occurring prediction made by the classifiers for the input x.

This 'hard' voting mechanism eschews the probabilistic nuances of 'soft' voting, where predictions are based on the probability estimates of class membership. Instead, it relies solely on the class labels predicted by each classifier, thus the term 'hard'. The aggregation of predictions in hard voting is intended to amplify the collective accuracy of the ensemble, under the premise that while individual classifiers might err, the ensemble as a whole is more likely to converge on the correct classification through majority rule.

The strength of this ensemble method is that it capitalizes on the diversity of the classifiers involved. Each classifier in the ensemble might be making its predictions based on different heuristics or patterns it has learned from the data. By combining them, we are essentially aiming to smooth out their individual biases and variances, ideally leading to a more accurate and stable prediction. This is particularly effective in complex domains where different models capture different aspects of the data, and their combined votes can lead to a more reliable decision than any single model could achieve on its own.

E. Evaluation

The performance of classification models is commonly evaluated using a suite of metrics that capture various aspects of predictive accuracy and error. These metrics are derived from the confusion matrix, a tabular representation of Actual vs Predicted classifications [27].

Accuracy: This measures the overall correctness of the model and is calculated as the ratio of correctly predicted observations to the total observations.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: Also known as the positive predictive value, this metric assesses the proportion of positive identifications that were actually correct.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall: This measure, also known as the true positive rate, calculates the proportion of actual positives that were identified correctly.

F1-Score: This is the harmonic mean of precision and recall, providing a balance between the two when a model may favor one over the other.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Confusion Matrix: A matrix that visualizes the performance of an algorithm. Every matrix row represents instances in an actual class while every column represents instances in a predicted class, or vice versa. The matrix is composed of:

- True Positives (TP): Correctly predicted positive obser- vations.
- True Negatives (TN): Correctly predicted negative obser- vations.
- False Positives (FP): Incorrectly predicted positive obser- vations (Type I error).
- False Negatives (FN): Incorrectly predicted negative ob- servations (Type II error).

Each of these metrics encapsulates different dimensions of the model's predictions. Accuracy is a useful overall measure, but it can be misleading in imbalanced datasets. Precision and recall are particularly informative in such scenarios, where the costs of false positives and false negatives may vary significantly. The F1-score is often more beneficial than accuracy, especially in the case of an uneven class distribution.

EXPERIMENTAL RESULTS

A. Results of logistic regression

The application of LR to our classification task yielded promising results, as evidenced by the obtained accuracy score of approximately 93.31%. This metric signifies that the LR model correctly predicted whether a website was phishing or legitimate in over 93% of the cases in the testing set.

The confusion matrix for the LR classifier provides a detailed breakdown of the model's predictions for classify- ing websites as legitimate (class 0) or phishing (class 1) (Figure 4). From the matrix, we observe that the model correctly identified 886 legitimate websites (True Negatives) and 1177 phishing websites (True Positives). However, there were instances where the model made errors: 90 legitimate websites were misclassified as phishing (False Positives), and 58 phishing websites were incorrectly labeled as legitimate (False Negatives).

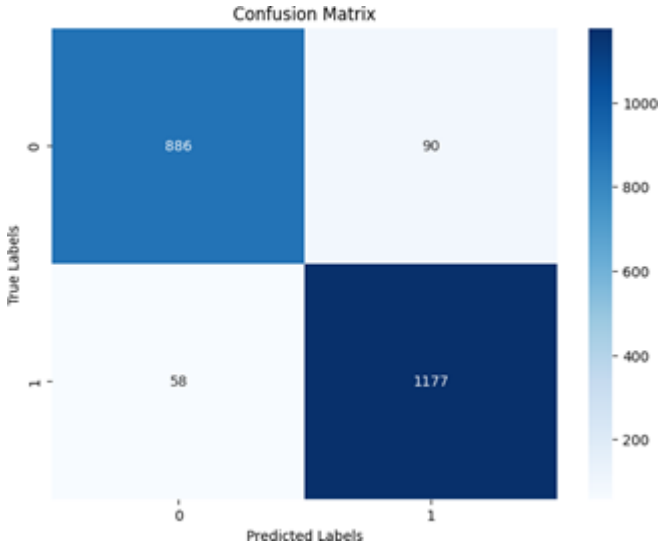

Fig. 4. Confusion matrix of logistic regression

As shown in Figure 5, the classification report for the LR model provides a comprehensive view of its performance in distinguishing between legitimate (class 0) and phishing (class 1) websites. The model's precision for the lowest level (legitimate) was 0.94, meaning that 94% among the websites its predicted to be valid actually were. With a recall of 0.91 for this class, 91% of real, legitimate websites were successfully recognized by the model. Class 0's F1-score of 0.92 indicates that recall and precision have a balanced relationship in this class.

For class 1 (phishing), the precision is slightly lower at 0.93, denoting that 93% of websites classified as phishing were correct, while the recall is higher at 0.95, signifying that the model successfully identified 95% of all phishing websites. The F1-score for phishing websites is 0.94, which is slightly higher than that for legitimate sites, indicating a slightly better precision-recall balance for phishing site predictions.

The model exhibits a commendable accuracy of 0.93 across the total of 2211 instances it was tested on. The macro average for precision, recall, and F1-score is 0.93, which demonstrates that the model is equally adept at identifying both classes. Additionally, 0.93 is the W.A, which takes into consideration the support (the quantity of real cases for each label), suggesting consistent performance across both classes, weighted according to their prevalence in the dataset. These metrics collectively demonstrate that the LR model has a high and balanced classification ability for both legitimate and phishing websites.

	precision	recall	f1-score	support
0	0.94	0.91	0.92	976
1	0.93	0.95	0.94	1235
accuracy			0.93	2211
macro avg	0.93	0.93	0.93	2211
weighted avg	0.93	0.93	0.93	2211

Fig. 5. Classification report of logistic regression

B. Results of Gradient Boosting

The GB classifier's performance in our analysis yielded an accuracy of approximately 94.93%. This suggests that the model can accurately distinguish among legitimate and phishing websites. An accuracy score close to 95% suggests that the GB model has learned the distinctions between the two classes effectively, utilizing the boosting technique to sequentially improve upon the classification by correcting previous errors. The confusion matrix in Figure 6 for the GB classifier in the context of classifying websites as legitimate or phishing presents a detailed performance break- down. The matrix shows that the model correctly classified 911 legitimate websites (True Negatives) and 1188 phishing websites (True Positives), demonstrating a strong ability to identify both classes accurately. However, the model was not infallible; it incorrectly

classified 65 legitimate websites as phishing (False Positives) and 47 phishing websites as legitimate (False Negatives).

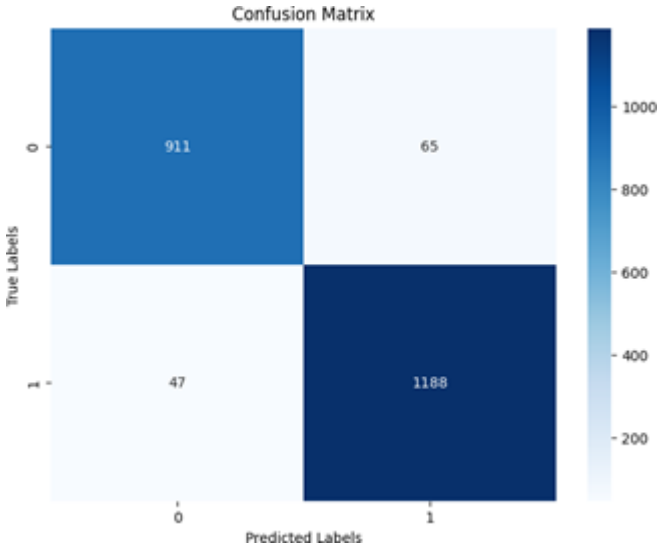

Fig. 6. Confusion matrix of Gradient Boosting

As shown in Figure 7, the classification report for the GB model in the task of classifying websites as legitimate or phishing reveals strong performance metrics across the board. For legitimate websites (class 0), the model achieved a precision of 0.95, which means that 95% of the websites it predicted as legitimate were indeed legitimate. The recall for legitimate websites is 0.93, indicating that the model correctly identified 93% of all actual legitimate websites. The F1-score, which balances precision and recall, is 0.94 for this class, indicating a high degree of accuracy and a balanced performance between precision and recall.

Similarly, for phishing websites (class 1), the model also achieved a precision of 0.95, suggesting that when it predicts a website to be phishing, it is correct 95% of the time. The recall for phishing websites is slightly higher at 0.96, showing that the model is able to identify 96% of all phishing websites correctly. The F1-score for this class is 0.95, indicating a very strong performance and a slight edge over the legitimate class in terms of recall.

The model exhibits an excellent accuracy of 0.95 for the total of 2211 instances that it was tested on, which suggests that the GB model is exceptionally well-tuned for this par- ticular classification task. Regarding recall, F1-score, and precision the weighted average and the macro average are both 0.95, underscoring the model's consistent and balanced classification ability for both legitimate and phishing websites.

	precision	recall	f1-score	support
0	0.95	0.93	0.94	976
1	0.95	0.96	0.95	1235
accuracy			0.95	2211
macro avg	0.95	0.95	0.95	2211
weighted avg	0.95	0.95	0.95	2211

Fig. 7. Classification report of Gradient Boosting

C. Results of K-Nearest Neighbors (KNN)

The KNN algorithm, known for its simplicity and ef- fectiveness, has demonstrated commendable results in our classification task, achieving an accuracy of approximately 94.35%. This reflects the algorithm's capability to accurately discern between legitimate and phishing websites with a high degree of reliability. An accuracy level above 94% indicates that the model is well-calibrated and that the feature space is effectively capturing the relevant information needed for classification.

The confusion matrix in Figure 8 for the KNN classifier in the task of discerning legitimate websites from phishing ones reveals a robust predictive performance. The matrix shows that the model accurately identified 909 legitimate websites as legitimate (True Negatives) and 1177 phishing websites as phishing (True Positives), which underscores its effectiveness in correctly classifying both types of websites. However, the model misclassified 67 legitimate websites as phishing (False Positives) and 58 phishing websites as legitimate (False Negatives).

	precision	recall	f1-score	support
0	0.94	0.93	0.94	976
1	0.95	0.95	0.95	1235
accuracy			0.94	2211
macro avg	0.94	0.94	0.94	2211
weighted avg	0.94	0.94	0.94	2211

Fig. 9. Classification report of K-Nearest Neighbors

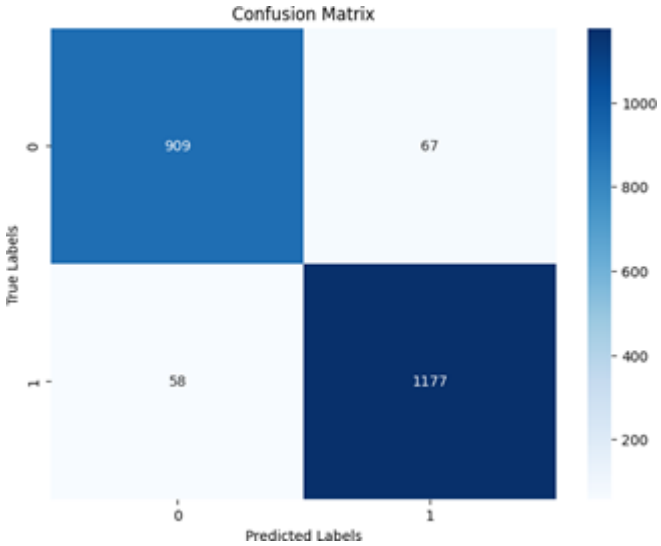

Fig. 8. Confusion matrix of K-Nearest Neighbors (KNN)

As shown in Figure 9, the classification report for the KNN classifier reveals a high degree of accuracy in distinguishing between legitimate (class 0) and phishing (class 1) websites. With an accuracy of 0.94 for class 0, 94% among the websites that were predicted to be legitimate actually were. Impressively, the recall rate for class 0 is 0.93, meaning that 93% of all real, legitimate websites were properly identified by the algorithm. This yields an F1-score of 0.94, which is a weighted average of precision and recall for class 0, signifying a well- balanced classification performance for legitimate websites.

For class 1, the model performs equally well with a pre- cision of 0.95, meaning 95% of the websites classified as phishing were correctly identified. The recall matches the precision at 0.95, reflecting the model's effectiveness in de- tecting phishing websites. When recognizing phishing websites, the classifier appears to maintain a solid balance among precision and recall, as indicated by the matching F1-score of 0.95.

Overall, the model's accuracy is at 0.94 across the 2211 instances evaluated, demonstrating the classifier's consistent performance. The macro average and weighted average for precision, recall, and F1-score are all 0.94, further indicating that the KNN classifier has a uniform classification strength for both classes.

D. Results of pour proposed model

The ensemble hard voting classifier, which combines the insights of multiple ML models, has achieved an impressive accuracy of approximately 95.02%. This indicates that the ensemble model has effectively harnessed the strengths of its constituent classifiers to deliver a combined predictive power that surpasses that of the individual models. By integrating the decision-making capabilities of diverse algorithms, the ensemble method has capitalized on their collective intelli- gence, leading to a more accurate and robust classification of websites as either legitimate or phishing. An accuracy score just above 95% is a strong testament to the efficacy of the ensemble approach, particularly in tasks where the cost of misclassification can be high. The ensemble hard voting system's high accuracy level suggests that it has been able to compensate for individual classifiers' weaknesses, reducing variance and bias, and improving overall performance.

The confusion matrix for the ensemble hard voting classifier in the task of classifying websites as legitimate or phishing reveals a strong predictive performance (Figure 10. In the matrix, we observe that the model has correctly classified 906 legitimate websites (True Negatives) and 1195 phishing websites (True Positives). These numbers are indicative of the model's high capability to correctly identify the nature of the websites. Conversely, there were 70 instances where legitimate websites were misclassified as phishing (False Positives) and 40 instances where phishing sites were mistaken for legitimate sites (False Negatives).

The low number of False Positives and False Negatives suggests that the ensemble model has achieved a commendable balance between precision and recall.

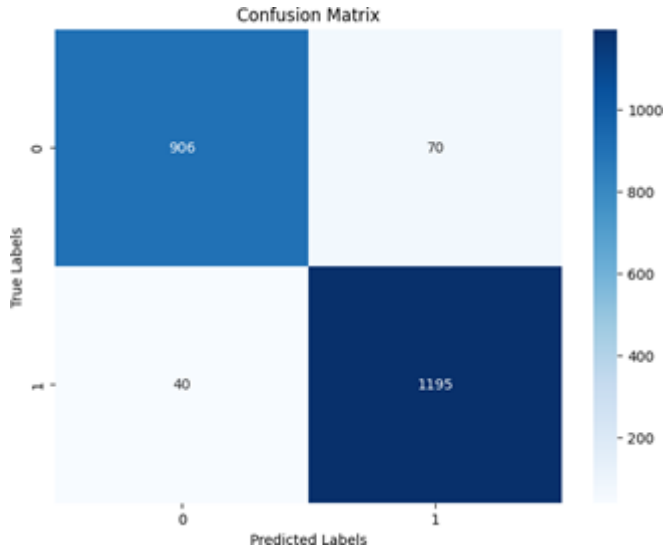

Fig. 10. Confusion matrix of ensemble hard voting

The classification report for the ensemble hard voting model illustrated in Figure 11 presents an impressive picture of the model's performance in classifying websites as either legiti- mate (class 0) or phishing (class 1). For legitimate websites, the model shows a high precision of 0.96, indicating that 96% of the websites it identified as legitimate were indeed so. With a somewhat lower recall of 0.93 for this class, the model appears to have acquired 93% of all genuine websites. The F1-score for legitimate websites is 0.94, reflecting a harmonious balance between precision and recall.

In terms of identifying phishing websites, the ensemble model achieved a precision of 0.94 and an even higher recall of 0.97, indicating that it correctly identified 97% of all phishing websites in the dataset. The F1-score for phishing sites stands at 0.96, showcasing a slightly more balanced precision-recall performance compared to that for legitimate sites.

The overall accuracy of the ensemble model reaches 0.95, underscoring its ability to accurately classify the majority of the cases. The macro average and weighted average for precision, recall, and F1-score are all 0.95, which further confirms the model's consistent and balanced classification capabilities across both classes. These results demonstrate the effectiveness of combining multiple classifiers using hard voting to achieve high performance in the critical task of phishing detection.

	precision	recall	f1-score	support
0	0.96	0.93	0.94	976
1	0.94	0.97	0.96	1235
accuracy			0.95	2211
macro avg	0.95	0.95	0.95	2211
weighted avg	0.95	0.95	0.95	2211

Fig. 11. Classification report of ensemble hard voting

E. Discussion

The comparative analysis of the models indicates that our proposed ensemble majority voting approach has the edge over the individual models in terms of accuracy as illustrated in Figure I. LR, robust and straightforward, yielded an accuracy of 93.31%, which is commendable for many applications but not as high as other more complex models. GB demonstrated its predictive prowess with an accuracy of 94.93%, reflecting the strength of this technique in handling complex, non- linear relationships within the data. KNN with an accuracy of 94.35%, also showed its mettle, particularly as a non- parametric method that makes few assumptions about the form of the mapping function from inputs to outputs.

However, the ensemble hard voting model, which amalga- mates the decisions from LR, GB, and KNN, achieved the highest accuracy of 95.02%. This superior performance can be attributed to the ensemble method's ability to harness the di- verse strengths of the individual models. The ensemble model benefits from the variance reduction of GB, the simplicity and interpretability of LR, and the instance-based learning of KNN. By leveraging the majority vote to make final decisions, the ensemble model reduces the likelihood of overfitting and increases the model's robustness against the diverse tactics used in PAs.

TABLE I COMPARISON OF MODEL ACCURACIES

Model	Accuracy
Logistic Regression	93.31%
Gradient Boosting	94.93%
K-Nearest Neighbors	94.35%
Ensemble Hard Voting	95.02%

CONCLUSION

Finally, our thorough investigation has shown the tremendous promise of an ensembles hard-voted classifier for phishing detection. We have established a strong basis for model building by carefully selecting a dataset, carrying out a perceptive exploratory analysis of the data, and carefully preparing the data. With a precision level of 95.02%, our ensemble method—which incorporates the benefits of GB, KNN, and LR—performed better than any of the individual models. This is a significant step in the direction of building a safer digital environment, not just a statistical victory.

The enhanced performance of the ensemble model highlights the importance of diversity in machine learning algorithms when handling intricate issues like phishing detection. It is this diversity that allows for a more comprehensive capture of the multifaceted nature of PAs, ensuring that a wider array of tactics can be identified and neutralized. Moreover, the nuanced balance between precision and recall achieved by the ensemble model points to its effectiveness in minimizing both false positives and false negatives—a balance that is of paramount importance in practical cybersecurity applications where the stakes are high.

Our findings have important implications for the cyberse- curity community, suggesting that an ensemble hard voting approach could be employed to significantly bolster the accu- racy of phishing detection systems. This, in turn, could lead to more effective prevention of data breaches and financial fraud, thereby safeguarding the integrity of digital assets for individuals and organizations alike.

For future work, we aim to explore the integration of deep learning techniques within our ensemble framework. Deep learning models, with their ability to learn hierarchical representations and complex abstractions from data, could potentially enhance the feature extraction process and improve the ensemble's predictive performance. Additionally, we plan to expand the ensemble with more advanced and sophisticated classifiers. This could include the incorporation of models with state-of-the-art performance on related tasks, such as Gradient Boosting Machines (GBMs) with optimized hyperparameters, or exploring the use of meta-learners that can learn the optimal combination of model predictions. We also anticipate utilizing AutoML frameworks to systematically search for the best ensemble configurations and ML pipelines.

REFERENCES

[1]. Z. Alkhalil, C. Hewage, L. Nawaf, and I. Khan, "Phishing Attacks: A Recent Comprehensive Study and a New Anatomy," Front Comput Sci, vol. 3, p. 563060, Mar. 2021.

[2]. E. Dzuba, "Introducing Cloudflare's 2023 phishing threats report," Cloudflare Blog, Oct. 2023.

[3]. T. Bilot, N. E. Madhoun, K. A. Agha, and A. Zouaoui, "A survey on malware detection with graph representation learning," arXiv preprint arXiv:2303.16004, 2023.

[4]. R. Ahmad, I. Alsmadi, W. Alhamdani, and L. Tawalbeh, "Zero-day attack detection: a systematic literature review," Artificial Intelligence Review, pp. 1–79, 2023.

[5]. O. H. Abdulganiyu, T. Ait Tchakoucht, and Y. K. Saheed, "A systematic literature review for network intrusion detection system (IDS)," Int J Inf Secur, vol. 22, pp. 1125–1162, Oct. 2023.

[6]. K. Thakur, M. L. Ali, M. A. Obaidat, and A. Kamruzzaman, "A Systematic Review on Deep-Learning-Based Phishing Email Detection," Electronics, vol. 12, p. 4545, Nov. 2023.

[7]. G. Xiang, B. A. Pendleton, J. Hong, and C. P. Rose, "A Hierarchical Adaptive Probabilistic Approach for Zero Hour Phish Detection," in Computer Security – ESORICS 2010, pp. 268–285, Berlin, Germany: Springer, 2010.

[8]. E.-S. Apostol and C.-O. Truicaˇ, "Efficient Machine Learning Ensemble Methods for Detecting Gravitational Wave Glitches in LIGO Time Series," arXiv, Nov. 2023.

[9]. Z. Li, K. Ren, Y. Yang, X. Jiang, Y. Yang, and D. Li, "Towards Inference Efficient Deep Ensemble Learning," arXiv, Jan. 2023.

[10]. A. Dziedzic, C. A. Choquette-Choo, N. Dullerud, V. M. Suriyakumar, A. S. Shamsabadi, M. A. Kaleem, S. Jha, N. Papernot, and X. Wang, "Private Multi-Winner Voting for Machine Learning," arXiv, Nov. 2022.

[11]. A. Rahman and S. Tasnim, "Ensemble classifiers and their applications: a review," arXiv preprint arXiv:1404.4088, 2014.

[12]. M. N. Alam, D. Sarma, F. F. Lima, I. Saha, R.-E. Ulfath, and S. Hossain, "Phishing Attacks Detection using Machine Learning Approach," in 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), pp. 20–22, IEEE.

[13]. B. Espinoza, J. Simba, W. Fuertes, E. Benavides, R. Andrade, and T. Toulkeridis, "Phishing attack detection: A solution based on the typical machine learning modeling cycle," in 2019 International Confer- ence on Computational Science and Computational Intelligence (CSCI), pp. 202–207, IEEE, 2019.

[14]. S. Hossain, D. Sarma, and R. J. Chakma, "Machine learning-based phishing attack detection," International Journal of Advanced Computer Science and Applications, vol. 11, no. 9, 2020.

[15]. N. F. Abedin, R. Bawm, T. Sarwar, M. Saifuddin, M. A. Rahman, and S. Hossain, "Phishing attack detection using machine learning classification techniques," in 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), pp. 1125–1130, IEEE, 2020.

[16]. A. Basit, M. Zafar, A. R. Javed, and Z. Jalil, "A novel ensemble machine learning method to detect phishing attack," in 2020 IEEE 23rd International Multitopic Conference (INMIC), pp. 1–5, IEEE, 2020.

[17]. I. Saha, D. Sarma, R. J. Chakma, M. N. Alam, A. Sultana, and S. Hossain, "Phishing attacks detection using deep learning approach," in 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), pp. 1180–1185, IEEE, 2020.

[18]. F. Salahdine, Z. El Mrabet, and N. Kaabouch, "Phishing attacks detection a machine learning-based approach," in 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), pp. 0250–0255, IEEE, 2021.

[19]. A. K. Jain and B. B. Gupta, "A machine learning based approach for phishing detection using hyperlinks information," J Ambient Intell Hum Comput, vol. 10, pp. 2015–2028, May 2019.

[20]. O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," Expert Syst Appl, vol. 117, pp. 345–357, Mar. 2019.

[21]. V. Shahrivari, M. M. Darabi, and M. Izadi, "Phishing Detection Using Machine Learning Techniques," arXiv, Sept. 2020.

[22]. "Phishing website Detector," Nov. 2023. [Online; accessed 7. Nov. 2023].

[23]. X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, “A survey on ensemble learning,” Front Comput Sci, vol. 14, pp. 241–258, Apr. 2020.

[24]. O. Kramer and O. Kramer, “K-nearest neighbors,” Dimensionality re- duction with unsupervised nearest neighbors, pp. 13–23, 2013.

[25]. A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial,” Frontiers in neurorobotics, vol. 7, p. 21, 2013.

[26]. R. E. Wright, “Logistic regression.,” 1995.

[27]. I. M. De Diego, A. R. Redondo, R. R. Ferna´ndez, J. Navarro, and J. M. Moguerza, “General Performance Score for classification problems,” Appl Intell, vol. 52, pp. 12049–12063, Aug. 2022.