


Article

Intelligent Ensemble Learning Approach for Phishing Website Detection Based on Weighted Soft Voting

Altyeb Taha 

Department of Information Technology, Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, Jeddah 21589, Saudi Arabia; aaataha@kau.edu.sa

Abstract: The continuous development of network technologies plays a major role in increasing the utilization of these technologies in many aspects of our lives, including e-commerce, electronic banking, social media, e-health, and e-learning. In recent times, phishing websites have emerged as a major cybersecurity threat. Phishing websites are fake web pages that are created by hackers to mimic the web pages of real websites to deceive people and steal their private information, such as account usernames and passwords. Accurate detection of phishing websites is a challenging problem because it depends on several dynamic factors. Ensemble methods are considered the state-of-the-art solution for many classification tasks. Ensemble learning combines the predictions of several separate classifiers to obtain a higher performance than a single classifier. This paper proposes an intelligent ensemble learning approach for phishing website detection based on weighted soft voting to enhance the detection of phishing websites. First, a base classifier consisting of four heterogeneous machine-learning algorithms was utilized to classify the websites as phishing or legitimate websites. Second, a novel weighted soft voting method based on Kappa statistics was employed to assign greater weights of influence to stronger base learners and lower weights of influence to weaker base learners, and then integrate the results of each classifier based on the soft weighted voting to differentiate between phishing websites and legitimate websites. The experiments were conducted using the publicly available phishing website dataset from the UCI Machine Learning Repository, which consists of 4898 phishing websites and 6157 legitimate websites. The experimental results showed that the suggested intelligent approach for phishing website detection outperformed the base classifiers and soft voting method and achieved the highest accuracy of 95% and an Area Under the Curve (AUC) of 98.8%.



Citation: Taha, A. Intelligent Ensemble Learning Approach for Phishing Website Detection Based on Weighted Soft Voting. *Mathematics* **2021**, *9*, 2799. <https://doi.org/10.3390/math9212799>

Academic Editor: Abeer Alsadoon

Received: 2 October 2021

Accepted: 3 November 2021

Published: 4 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: phishing website detection; machine learning; ensemble learning

1. Introduction

Due to their flexibility, convenience, and simplicity of use, the number of web users who utilize online services, e-banking, and online shopping has increased rapidly in recent years. This massive increase in the use of online services and e-commerce has encouraged phishers and cyber attackers to create misleading and phishing websites in order to obtain financial and other sensitive information [1,2]. Online phishing sites typically utilize similar page layouts, fonts, and blocks to imitate official web pages in order to persuade web visitors to provide personal information, such as login credentials. Due to the evolution of online hacking techniques and a lack of public awareness, internet users are frequently exposed to cyber dangers, such as phishing, spam, trojans, and adware. Phishing has grown in popularity as a means of collecting users' private information, such as login details, credit card information, and social security numbers, via fraudulent websites [3]. Therefore, phishing attacks represent a serious cybersecurity problem that significantly affects commercial websites and the users of the web [4,5]. Personal information collected in this way can be used to steal money via stolen credit cards, debit cards, bank account fraud, and gaining illegal access to people's social media profiles.

Phishing attacks have already resulted in significant losses and may have a negative impact on the victim, not just financially, but also in terms of reputation and national security. In comparison to 2018 and 2019, in 2020, there was a 15% increase in the number of phishing attacks. In addition, Kaspersky Lab's anti-phishing security systems stopped over 482 million phishing threats in 2018, a twofold increase over 2017 [6]. Based on the Anti Phishing Working Group's (APWG) report (APWG 2020), the number of phishing attacks is rising continually, with 146,994 phishing websites discovered in the second quarter of 2020 [7]. In 2020, the anticipated average cost of a business breach caused by phishing attacks was 2.8 million USD. It is important to utilize anti-phishing methods to avoid such significant losses.

Several anti-phishing technologies for identifying phishing sites have been suggested and designed by various cybersecurity professionals and researchers [8–10]. One of these methods is the detection of website phishing attacks using blacklists. To determine validity, web browsers use the blacklist technique, which matches the universal resource locator (URL) with previously recorded phishing website URLs. As a result of its dependence on a database of blacklisted phishing URLs, blacklist anti-phishing systems cannot identify new phishing URLs, which is a significant drawback [11]. Concerning the dynamic of cyber-attacks, machine learning (ML)-based solutions can be utilized to validate websites in order to handle online phishing attacks depending on the website characteristics [12]. The goal is to make it easier to distinguish legitimate websites from phishing ones [13–15].

However, phishing websites are becoming increasingly capable of avoiding detection as a result of the developing nature of phishing attacks since there are ways of avoiding the existing defenses. Many machine-learning (ML) techniques for identifying phishing websites have low detection accuracy and high false-positive rates [15,16]. Ensemble learners integrate the perspectives of several learners to improve performance; they have been utilized in many applications, including critical power system applications. Ensemble learning compensates for a classifier's weakness with the strength of other classifiers, resulting in a superior performance over an individual classifier [17,18].

Although ensemble learning based on soft voting has many advantages, traditional soft voting may be unable to combine systems successfully if they lack an effective method for assessing confidence in their predictions. Excessively optimistic or pessimistic systems can significantly distort the results, resulting in a classifier that performs worse than the best system among those selected in the vote. Due to the changing nature of phishing attacks, effective and better methods for detecting them are required since there is no one-size-fits-all solution for phishing deletion. According to the literature review, the majority of existing machine learning methods have limitations, including a high false alarm rate, a low detection rate, and the inability of single classifiers and some hybridized methods to produce highly effective and efficient phishing website detection solutions [19–21]. Therefore, the proposed ensemble approach for phishing websites detection utilizes a weighted soft voting method to assign greater weights of influence to stronger base learners and lower weights of influence to weaker base learners. Then, it uses weighted soft voting to integrate the findings of each classifier in order to distinguish between phishing and legitimate websites. The following are the study's key contributions and their importance in terms of improving the detection of phishing websites:

- To improve the detection accuracy of phishing websites, a novel weighted soft voting method based on the k-statistic is suggested to evaluate the contributions of each classifier and assign higher influence weights to stronger classifiers and lower impact weights to weaker classifiers.
- The proposed intelligent ensemble technique for phishing websites incorporates the results of individual learners in accordance with their significance in distinguishing phishing from legitimate websites.
- Individual classifiers and soft voting ensembles were outperformed by the proposed approach, which achieved a detection accuracy of 95% for phishing websites using publicly available datasets. One of the issues is that single classifier approaches are

unable to accurately detect evolving phishing websites. This was demonstrated by the fact that most single classifier models are usually outperformed by ensemble techniques or hybridization algorithms. As a result, this study focused on developing a more effective approach (Intelligent Ensemble Learning Approach) for detecting phishing websites.

2. Related Work

In this section, we discuss some of the recent research studies on detecting phishing websites using machine learning approaches.

The researchers in [22] used a nonlinear regression approach to determine whether a website was phishing. The Harmony Search and Support Vector Machine (SVM) methods were used to operate the system. They made use of 11,055 websites and 20 features. Their proposed method had a detection accuracy of 92.80%. In [23], the researchers presented a phishing detection system based on 209 word-vector features and 17 natural language processing (NLP) features. The Random Forest, SMO, and Naive Bayes algorithms were tested, and the Random Forest method, with an accuracy rate of 89.9%, achieved the best results with a hybrid approach.

Using a c4.5 decision tree technique, MacHado and Gadge [24] introduced a method for detecting phishing URL websites. This method analyzes the sites and derives heuristic values. These variables were used to assess whether the site was phishing or not using the c4.5 decision tree method. The dataset was compiled using data from PhishTank and Google. This method is divided into two stages: pre-processing and detection. Features are retrieved using rules in the pre-processing phase, and the features and their associated values are fed to the c4.5 algorithm, which achieved an accuracy of 89.40%. Mohammad et al. [25] suggested a self-structuring neural network-based intelligent phishing detection system. The authors gathered 17 features from URLs, source code, and a third-party source in order to train the system using a neural network. The weights of the network were adjusted using the backpropagation technique. This method has the benefit of adjusting its neural network to the changing features of phishing attacks. The proposed strategy resulted in a detection accuracy of 89.40%.

Chiew et al. [26] proposed a visual similarity-based approach in which the logo of a suspected website is retrieved using machine learning and fed to Google image search to obtain the target identity. When the real domain of the suspected site does not match the domain returned by Google image search, it is considered phishing. This approach depends upon the success rate of machine learning-based logo extraction. This approach is ineffective in detecting phishing attacks on websites that lack a logo; however, the accuracy of phishing website detection is 93.4%. Aggarwal et al. [27] proposed PhishAri, a method for detecting phishing URLs in tweets. They detected phishing URLs by combining URL, WHOIS, tweets, and network-based data with a Random Forest classifier. The accuracy of phishing website detection was 92.52%. Because this approach is simply based on the URL text and not on the content of a website, it may be ineffective if the phishing URL is housed on a hacked domain.

Dedakia and Mistry [28] presented a technique for phishing detection called Content-Based Associative Classification (CBAC). By including content-based characteristics, the suggested approach expanded the Multi-Label Class Associative Classification (MCAC) algorithm. The proposed approach (CBAC) had an accuracy value of 94.29% based on the testing results. To choose optimum features for phishing website detection, Chiew et al. [1] presented a hybrid ensemble feature selection (HEFS) technique based on a unique cumulative distribution function gradient (CDF-g) method. The phishing website detection accuracy achieved by HEFS using the Random Forest was 94.6%.

Wei et al. [19] proposed a method for detecting malicious URL addresses with nearly 100% accuracy using convolutional neural networks. In contrast to earlier research that analyzed URLs, traffic statistics, or web content, they analyzed the URL text. As a result, their technique was more efficient and identified zero-day threats.

Alsariera et al. [15] presented four meta-learner models based on the extra-tree basis classifiers: AdaBoost-Extra Tree (ABET), Bagging-Extra Tree (BET), Rotation Forest-Extra Tree (RoFBET), and LogitBoost-Extra Tree (LBET). Their suggested AI-based meta-learners were fitted and assessed on phishing website datasets (currently with the most up-to-date features). Their models obtained a detection accuracy of at least 97% and a false-positive rate of less than 0.028.

Azeez et al. [20] proposed a method for recognizing phishing sites that would protect internet users from suffering from any type of phishing attack by checking the conceptual and literal consistency of the uniform resource locator (URL) and the web content. Their suggested PhishDetect technique obtained a 99.1% accuracy rate, showing that it is successful at identifying various types of phishing attacks.

The review of these relevant existing methods showed that the majority of existing machine learning methods have several limitations, including a high false alarm rate, a low detection rate, and the inability for single classifiers and some hybridized methods to produce highly effective and efficient phishing website detection solutions.

3. Proposed Intelligent Ensemble Learning Approach for Phishing Website Detection Based on Weighted Soft Voting

In this section, the proposed approach is described in detail. We believe that combining the “opinion” of various machine learning algorithms on a given task can yield better results than any individual approach. The structure of the proposed approach consists of two parts: First, the base learners consist of four heterogeneous machine-learning algorithms, which have different weaknesses and strengths, and give the final classification results based on individual decisions. Second, a weighted soft voting method is utilized to assign greater weights of influence to stronger base learners and lower weights of influence to weaker base learners, and then it combines the results from all the classifiers according to their weights to distinguish between phishing websites and legitimate websites. Figure 1 shows the proposed approach for phishing website detection.

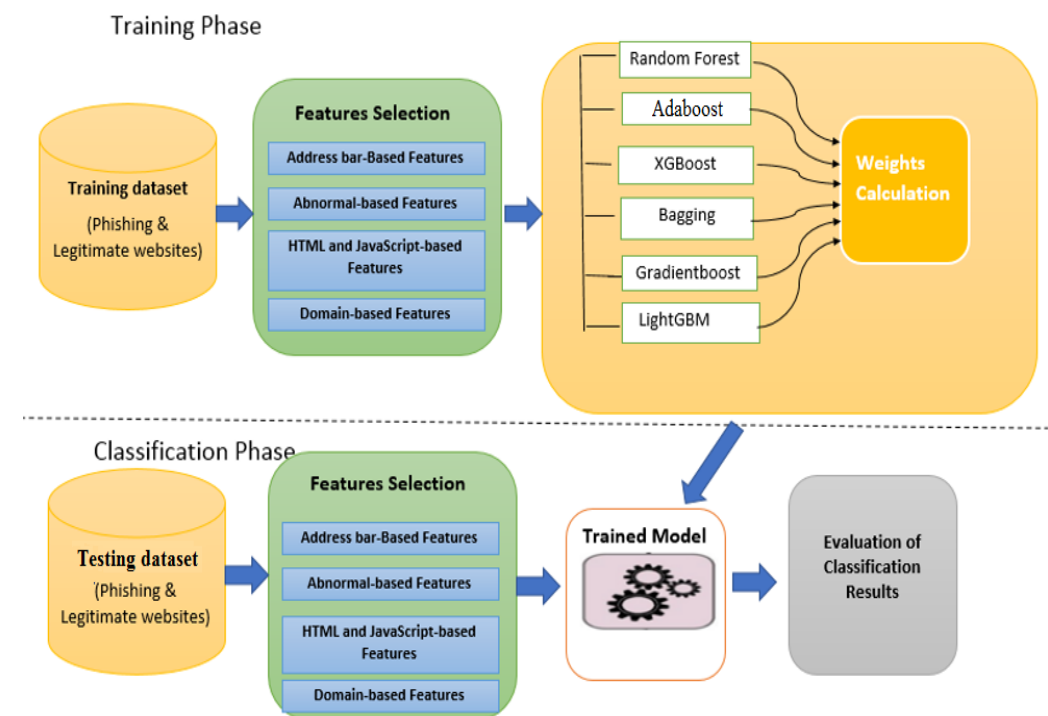


Figure 1. The proposed approach for phishing website detection.

3.1. Dataset and Data Preprocessing

In this paper, we conducted experiments using the publicly available phishing website dataset from the UCI Machine Learning Repository [29] to assess the efficacy of the pro-

posed Intelligent Ensemble learning technique for improving phishing website detection. We used this dataset because it has been widely and successfully used for phishing website detection [1,22,23,25]. There are 4898 phishing websites and 6157 legitimate websites in this phishing website dataset, with a total of 11,055 total websites. Table 1 summarizes the key attributes of the phishing website datasets utilized in the experiments and evaluation.

Table 1. The fundamental attributes of the dataset of phishing websites utilized in the experiments.

Attributes	Description
Website Features	The address bar-based type has 12 features, the abnormality-based type has 6 features, the HTML and JavaScript types have 5 features, and the domain-based type has 7 features.
Number of features	30
Classes	Phishing or legitimate website
Number of classes	2
Number of websites	11,055
Number of phishing websites	4898
Percentage of phishing websites	44%
Number of legitimate websites	6157
Percentage of legitimate websites	56%

The Correlation-based Feature Selection (CFS) is utilized to select the most significant features to differentiate between phishing and legitimate websites. CFS determines the value of a subset of features by taking into account each feature's unique predictive power, as well as the degree of redundancy between them. The CFS algorithm's primary component is a heuristic for determining the utility or merit of a subset of attributes, as specified in Equation (4). This heuristic demonstrates the utility of individual features in predicting the class label, as well as their degree of intercorrelation [30].

$$Merit_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad (1)$$

where Merits is the heuristic "merit" of an attribute subset S, including k attributes, $\overline{r_{cf}}$ is the average attribute class correlation, and $\overline{r_{ff}}$ is the average attribute-attribute correlation. The heuristic's objective is to eliminate unnecessary and duplicated attributes that are ineffective predictors of the class. Figure 2 shows the average merits and ranks for all of the features.

3.2. Weighted Soft Voting Based on k Statistics

Weighted soft computing methods can be defined at the classifier, class, or instance level [31,32]. Soft voting is more effective than hard voting because it prioritizes extremely confident votes and incorporates each classifier's significance into the final decision. The Kappa statistic is widely employed in many classification tasks [33–35]. It evaluates a classifier's competency by comparing successful predictions to the statistical distribution of the data classes; therefore, correcting any agreements caused by statistical chance.

We aimed to create an ensemble learning model based on weighted soft voting to detect phishing websites. To accomplish this, we first trained the various learning techniques in the base learners using the training dataset and generated the confusion matrix for each classifier to evaluate its performance. As seen in Table 2, the horizontal direction corresponds to the predicted label, while the vertical direction corresponds to the real label. The diagonal line represents the number of websites that were successfully classified.

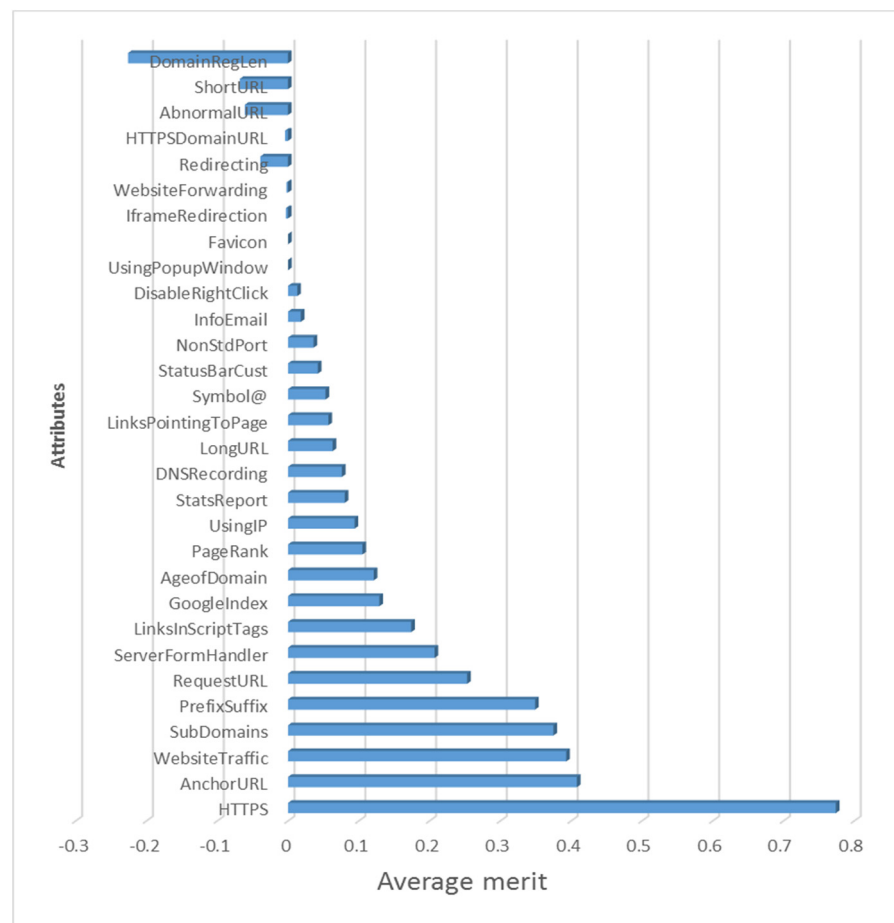


Figure 2. Average merits and ranks for website features.

Table 2. Confusion matrix.

		Predicted Class	
		Legitimate	Phishing
Actual Class	Legitimate	True Negative (TN)	False Positive (FP)
	Phishing	False Negative (FN)	True Positive (TP)

Numerous class-specific measures can be derived from the confusion matrix, such as:

True Positive (TP): The number of phishing websites that the classifier categorized as phishing websites. False Positive (FP): The number of legitimate websites that the classifier incorrectly categorized as phishing websites. False Negative (FN): The number of phishing websites that the classifier classified as legitimate. True Negative (TN): The number of legitimate websites that the classifier categorized as legitimate.

The Kappa test is a method used to assess consistency in statistics and measures the consistency of two judgments [36]. Because Kappa statistics aid classifiers in achieving superior prediction performance in binary classification issues [37], we used Kappa statistics to assign weights to the individual classifiers in the proposed approach. The Kappa coefficient was calculated from the confusion matrix as follows:

$$K = \frac{accuracy - expected\ accuracy}{1 - expected\ accuracy} \tag{2}$$

where the formula of accuracy was calculated as follows:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

and the formula of expected accuracy is given by:

$$expected\ accuracy = \left(\frac{TP + FP}{N} \cdot \frac{TP + FN}{N} \right) + \left(\frac{TN + FP}{N} \cdot \frac{TN + FN}{N} \right) \tag{4}$$

where N is the number of samples in the dataset.

The suggested method determined the final classification results based on the outcomes of the individual classifiers and the appropriate weights provided by the Kappa statistics. For a dataset sample x , the output of the ensemble classifier can be expressed by:

$$H(x) = arg_c\ max\ \sum_{t=1}^T w_t h_t^c(x) \tag{5}$$

where $h_t^c(x)$ denotes the probability that the classifier h_t classifies sample x to class c . Although the predictions provided by the classifiers are usually not very precise, soft voting still presents a slightly enhanced performance compared to hard voting [38]. Therefore, we utilized the soft voting strategy in this research.

3.3. Performance Evaluation Measures

Given the True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) counts, the accuracy, precision, recall, and f1-score were calculated to evaluate the proposed approach for phishing website detection, as shown in Table 3.

Table 3. The classification measures used to evaluate the proposed intelligent ensemble learning approach for phishing website detection.

Classification Measure	Formula
Accuracy	$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$
Precision	$Precision = \frac{TP}{TP + FP}$
Recall	$Recall = \frac{TP}{TP + FN}$
F1-score	$F1-score = \frac{2TP}{2TP + FP + FN}$

4. Results and Discussion

In this section, we assessed the performance of the proposed approach by using the information collected from the experiments and comparing it with four basic classifiers. The basic classifiers were the Decision Tree, Naïve Bays, Random Forest, Gradient Boosting, and Logistic Regression. The proposed approach and the basic classifiers were trained and tested using five feature subsets (A–E), which were used to select the top 5%, 10%, 15%, 20%, and 100% of the features of the ordered ranking, respectively. The average run time (in seconds) of the proposed approach for classifying phishing websites was 25 s, calculated using a PC with an Intel(R) Core (TM) i7-8550U CPU @ 1.80 GHz 1.99 GHz and 8 GB RAM. The proposed approach was evaluated using the Accuracy, Precision, Recall, F-score, and Kappa statistics measures. Table 4 shows the results of these performance evaluation metrics.

As shown in Table 5, the proposed approach obtained the highest accuracy score of 95% when the top 20% of ranking features were used, demonstrating the ability of the proposed approach to distinguish legitimate from phishing websites. Additionally, the proposed approach earned an accuracy score of 95%, which represents the proportion of accurately categorized websites to all websites. Moreover, the proposed approach achieved a recall score of 95%, demonstrating its ability to accurately categorize 95% of websites with a low number of false positives. Furthermore, the proposed technique achieved an F1-score of 95%. The F1-score quantifies the recall-precision trade-off. As a result, it takes into account both FPs and FNs.

Table 4. Comparison of the performance of the proposed approach and other machine learning algorithms.

Features Subset	Algorithm	Accuracy	Precision	Recall	F1-Score	Kappa Statistics
A	AdaBoost	0.91	0.91	0.91	0.91	0.87
	Decision Tree	0.91	0.91	0.91	0.91	0.87
	Random Forest	0.91	0.91	0.91	0.91	0.85
	Gradient Boosting	0.91	0.91	0.91	0.91	0.85
	Soft Voting	0.91	0.91	0.90	0.91	0.85
	The proposed approach	0.91	0.91	0.91	0.91	0.85
B	AdaBoost	0.91	0.91	0.91	0.91	0.87
	Decision Tree	0.92	0.92	0.92	0.92	0.89
	Random Forest	0.91	0.92	0.91	0.91	0.88
	Gradient Boosting	0.91	0.91	0.91	0.91	0.91
	Soft Voting	0.92	0.92	0.92	0.92	0.88
	The proposed approach	0.92	0.92	0.92	0.92	0.88
C	AdaBoost	0.91	0.91	0.91	0.91	0.86
	Decision Tree	0.93	0.93	0.93	0.93	0.93
	Random Forest	0.92	0.92	0.92	0.92	0.90
	Gradient Boosting	0.91	0.91	0.91	0.91	0.91
	Soft Voting	0.94	0.94	0.94	0.94	0.93
	The proposed approach	0.94	0.94	0.94	0.94	0.93
D	AdaBoost	0.92	0.92	0.92	0.92	0.88
	Decision Tree	0.94	0.94	0.94	0.94	0.94
	Random Forest	0.93	0.93	0.93	0.93	0.91
	Gradient Boosting	0.91	0.91	0.91	0.91	0.91
	Soft Voting	0.94	0.94	0.94	0.94	0.94
	The proposed approach	0.95	0.95	0.95	0.95	0.95
E	AdaBoost	0.92	0.92	0.92	0.92	0.89
	Decision Tree	0.92	0.92	0.92	0.92	0.89
	Random Forest	0.91	0.91	0.91	0.91	0.92
	Gradient Boosting	0.91	0.91	0.91	0.91	0.86
	Soft Voting	0.93	0.93	0.93	0.93	0.92
	The proposed approach	0.93	0.93	0.93	0.93	0.92

Table 5. Comparison between the proposed approach and prior research.

Approach	Classification Accuracy
Chiew et al. [1]	94.6%
Babagoli et al. [22]	92.80%
Buber et al. [23]	89.9%
MacHado and Gadge [24]	89.40%
Mohammad et al. [25]	92.18%
Chiew et al. [26]	93.4%
Aggarwal et al. [27]	92.52%
Dedakia and Mistry [28]	94.29%
Mao et al. [21]	93%
The proposed approach	95%

Among these accuracy measures, Kappa statistics were used to test the proposed approach since they demonstrate that classifiers perform well in a binary classification issue. Kappa statistics neglect classifications based on chance. A high Kappa statistic indicates that instances are not randomly assigned to classes. Figure 2 illustrates the Kappa statistics for the proposed method and various machine learning algorithms. Kappa statistics were utilized to determine the inter-rater reliability or agreement between predicted and real website occurrences. The Kappa statistics values ranged from 0 to 1. A Kappa statistic value of less than 0.4 indicates an extremely low similarity; a value between 0.4 and 0.55 indicates an acceptable level of similarity; a value between 0.55 and 0.70 indicates a good level of similarity; a value between 0.70 and 0.85 indicates an extremely high level of

similarity; and a value greater than 0.85 indicates a perfect match between predicted and actual website instances.

We can see in Figure 3 that the proposed approach outperformed the other classifiers and achieved the highest Kappa statistic of 0.95 using the top 20% ranked features, which indicated perfect matching between the predicted and actual website instances.

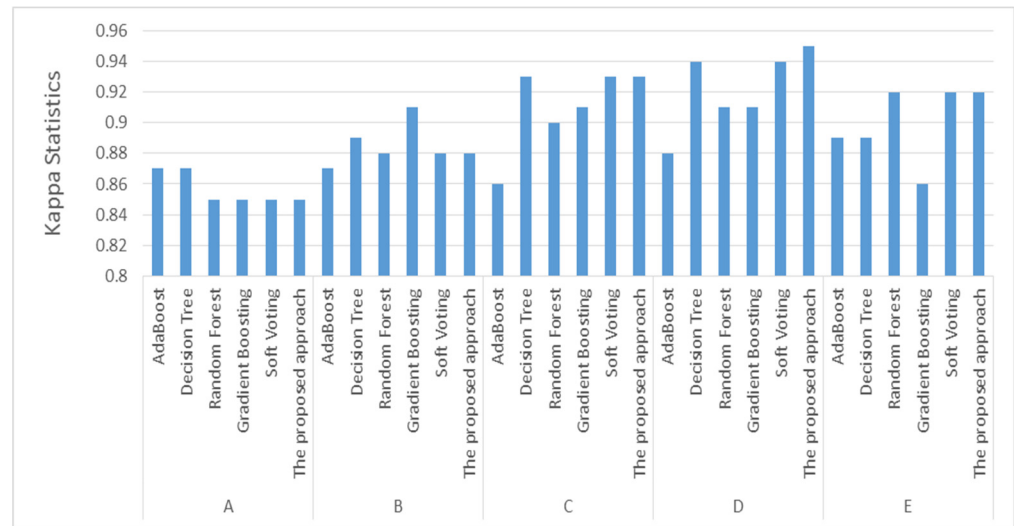


Figure 3. Kappa statistics for the proposed approach and other machine learning algorithms.

To illustrate the assessments of the proposed approach, the Area Under the Curve (AUC) and recall-precision curves are displayed in Figures 3 and 4. The AUC is a useful and meaningful measure of overall performance [39]. A high AUC value indicates a higher categorization ability. As seen in Figure 4, the proposed method achieved an AUC value of 98.8%, indicating that it is capable of distinguishing between legitimate and phishing websites.

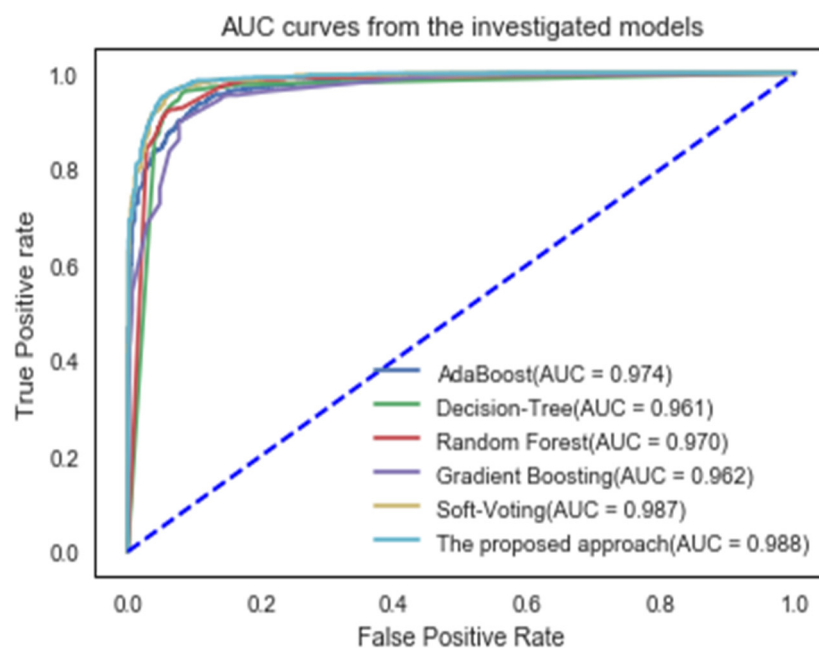


Figure 4. AUC curves of the suggested classifier and some machine learning techniques.

Generally, the precision–recall curve is used to compare classification models on the basis of their precision and recall. The precision–recall curve is a graph in which the *y*-axis represents the precision percentage, and the *x*-axis represents the recall percentage. The

precision–recall curve provides a comprehensive view of the classifiers’ performance [40]. The proposed approach’s accuracy and recall curves are compared to those of existing machine learning methods in Figure 5.

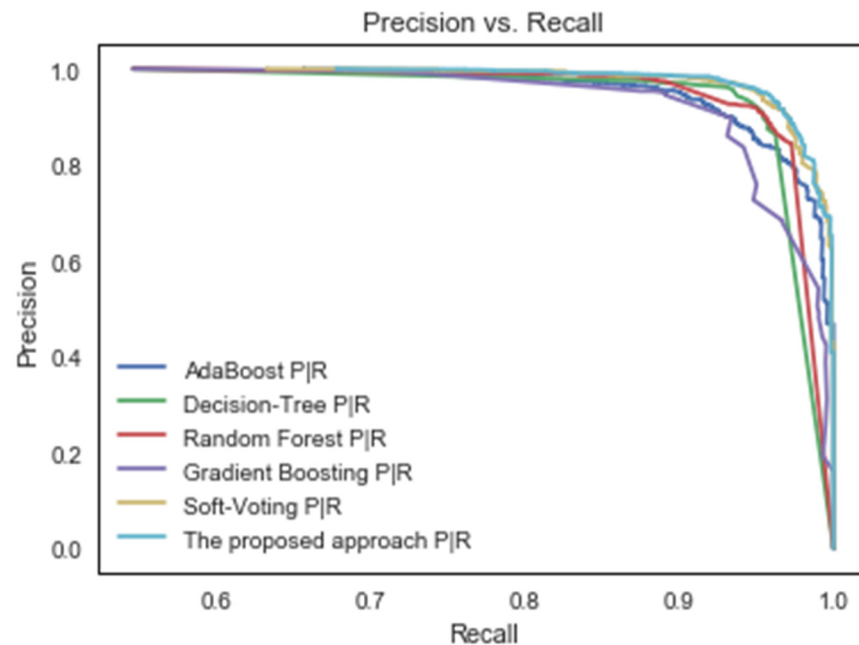


Figure 5. Precision–recall curves of the suggested classifier and some machine learning techniques.

In addition to the basic measures listed above, the statistical significance of the proposed approach and other methods was measured. Figure 6 presents the critical difference diagram, where the models with statistically similar values of performance are connected to one another.

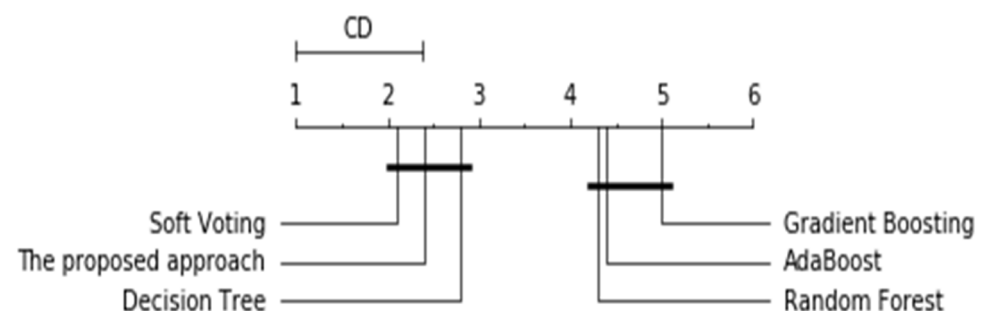


Figure 6. Critical difference diagram for the Nemenyi test.

Figure 6 shows the results of the statistical comparison of all of the approaches against one another by their mean ranks. Soft voting, the proposed approach, and the Decision Tree achieved higher ranks: 2.1, 2.4, and 2.8, respectively.

Additionally, the suggested approach was compared to prior research in terms of performance, using the same dataset and other datasets. Table 5 shows the comparison between the proposed approach and other study findings in terms of accuracy. The proposed approach attained the greatest accuracy score of 95% and outperformed the prior research using the same dataset [1,22,23,25] and the research using other datasets [24,26–28].

One of the limitations of this research is that it uses URL information solely as a feature for classifying phishing websites. The datasets employed were also limited. In future studies, more advanced features will be used to distinguish phishing from legal websites. Additionally, large datasets will be employed, as a larger dataset would result in

more reliable results. Further, in a future study, we will explore more advanced ensemble learning approaches for detecting phishing websites.

5. Conclusions

The continued development of network technologies has contributed significantly to their increased use in several aspects of our lives, including e-commerce, electronic banking, social media, e-health, and e-learning. With financial organizations continuing to suffer significant financial losses and the increasing difficulty of identifying phishing websites, it is critical to developing more effective methods for their detection. This paper proposes a two-stage intelligent ensemble learning technique for phishing website detection. First, a base classifier is formed as a collection of four heterogeneous machine-learning algorithms, each of which has unique strengths and weaknesses that influence the final classification result. Second, a weighted soft voting method based on the Kappa statistic is used to dynamically weigh the base classifiers, assigning greater influence weights to stronger base classifiers and decreasing influence weights to weaker base classifiers, and then combining the results of each classifier based on the weighted soft voting to differentiate between phishing and legitimate websites. The experimental findings suggest that the proposed approach outperformed existing machine learning algorithms in terms of accuracy, precision, and AUC. The findings of this work are expected to influence the future direction of research in phishing website prediction, as researchers have not paid much attention to classifier performance weighting. Additionally, this research demonstrates the validity and utility of assigning higher influence weights to stronger base learners and lower influence weights to weaker base learners in order to effectively detect phishing websites and contribute to increasing customer confidence in online commerce and business.

Funding: This Project was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, under grant no. (G: 205-830-1442). The authors, therefore, acknowledge with thanks DSR for technical and financial support.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This Project was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, under grant no. (G-205-830-1442). The authors, therefore, acknowledge with thanks DSR for technical and financial support.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Chiew, K.L.; Tan, C.L.; Wong, K.K.; Yong, S.C.; Tiong, W.K. A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. *Inf. Sci.* **2019**, *484*, 153–166. [CrossRef]
2. Sahingoz, O.K.; Buber, E.; Demir, O.; Diri, B. Machine learning based phishing detection from URLs. *Expert Syst. Appl.* **2019**, *117*, 345–357. [CrossRef]
3. Jain, A.K.; Gupta, B.B. A survey of phishing attack techniques, defence mechanisms and open research challenges. *Enterp. Inf. Syst.* **2021**, 1–39. [CrossRef]
4. Soon, G.K.; Chiang, L.C.; On, C.K.; Rusli, N.M.; Fun, T.S. Comparison of ensemble simple feedforward neural network and deep learning neural network on phishing detection. In *Computational Science and Technology*; Springer: Singapore, 2020; pp. 595–604.
5. Wei, B.; Hamad, R.A.; Yang, L.; He, X.; Wang, H.; Gao, B.; Woo, W.L. A deep-learning-driven light-weight phishing detection sensor. *Sensors* **2019**, *19*, 4258. [CrossRef] [PubMed]
6. Priya, S.; Selvakumar, S.; Velusamy, R.L. Evidential theoretic deep radial and probabilistic neural ensemble approach for detecting phishing attacks. *J. Ambient. Intell. Hum. Comput.* **2021**, 1–25. [CrossRef]
7. APWG. Anti Phishing Working Group Report. 2020. Available online: https://docs.apwg.org/reports/apwg_trends_report_q2_2020.pdf (accessed on 7 August 2021).
8. Yang, P.; Zhao, G.; Zeng, P. Phishing website detection based on multidimensional features driven by deep learning. *IEEE Access* **2019**, *7*, 15196–15209. [CrossRef]

9. Zamir, A.; Khan, H.U.; Iqbal, T.; Yousaf, N.; Aslam, F.; Anjum, A.; Hamdani, M. Phishing web site detection using diverse machine learning algorithms. *Electron. Libr.* **2020**, *38*, 65–80. [[CrossRef](#)]
10. Zhu, E.; Ju, Y.; Chen, Z.; Liu, F.; Fang, X. DTOF-ANN: An artificial neural network phishing detection model based on decision tree and optimal features. *Appl. Soft Comput.* **2020**, *95*, 106505. [[CrossRef](#)]
11. Gupta, B.B.; Arachchilage, N.A.; Psannis, K.E. Defending against phishing attacks: Taxonomy of methods, current issues and future directions. *Telecommun. Syst.* **2018**, *67*, 247–267. [[CrossRef](#)]
12. Harinahalli Lokesh, G.; BoreGowda, G. Phishing website detection based on effective machine learning approach. *J. Cyber Secur. Technol.* **2021**, *5*, 1–14. [[CrossRef](#)]
13. Altaher, A. Phishing websites classification using hybrid svm and knn approach. *Int. J. Adv. Comput. Sci. Appl.* **2017**, *8*, 90–95. [[CrossRef](#)]
14. He, Q.; Meng, X.; Qu, R.; Xi, R. Machine Learning-Based Detection for Cyber Security Attacks on Connected and Autonomous Vehicles. *J. Math.* **2020**, *8*, 1311. [[CrossRef](#)]
15. Alsariera, Y.A.; Adeyemo, V.E.; Balogun, A.O.; Alazzawi, A.K. Ai meta-learners and extra-trees algorithm for the detection of phishing websites. *IEEE Access* **2020**, *8*, 142532–142542. [[CrossRef](#)]
16. Chandra, Y.; Jana, A. Improvement in Phishing Websites Detection Using Meta Classifiers. In Proceedings of the 2019 6th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 13–15 March 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 637–641.
17. Agarwal, A.; Dixit, A. Fake news detection: An ensemble learning approach. In Proceedings of the 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 13–15 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1178–1183.
18. Granik, M.; Mesyura, V.; Yarovy, A. Determining fake statements made by public figures by means of artificial intelligence. In Proceedings of the 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), Lviv, Ukraine, 11–14 September 2018; IEEE: Piscataway, NJ, USA, 2018; Volume 1, pp. 424–427.
19. Wei, W.; Ke, Q.; Nowak, J.; Korytkowski, M.; Scherer, R.; Woźniak, M. Accurate and fast URL phishing detector: A convolutional neural network approach. *Comput. Netw.* **2020**, *178*, 107275. [[CrossRef](#)]
20. Azeez, N.A.; Salaudeen, B.B.; Misra, S.; Damaševičius, R.; Maskeliūnas, R. Identifying phishing attacks in communication networks using URL consistency features. *Int. J. Electron. Secur. Digit. Forensics* **2020**, *12*, 200–213. [[CrossRef](#)]
21. Mao, J.; Bian, J.; Tian, W.; Zhu, S.; Wei, T.; Li, A.; Liang, Z. Phishing page detection via learning classifiers from page layout feature. *EURASIP J. Wirel. Commun. Netw.* **2019**, *1*, 43. [[CrossRef](#)]
22. Babagoli, M.; Aghababa, M.P.; Solouk, V. Heuristic nonlinear regression strategy for detecting phishing websites. *Soft Comput.* **2019**, *23*, 4315–4327. [[CrossRef](#)]
23. Buber, E.; Diri, B.; Sahingoz, O.K. Detecting phishing attacks from URL by using NLP techniques. In Proceedings of the 2017 International conference on computer science and Engineering (UBMK), Antalya, Turkey, 5–8 October 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 337–342.
24. Machado, L.; Gadge, J. Phishing sites detection based on C4.5 decision tree algorithm. In Proceedings of the 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, India, 17–18 August 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–5.
25. Mohammad, R.M.; Thabtah, F.; McCluskey, L. Predicting phishing websites based on self-structuring neural network. *Neural. Comput. Appl.* **2014**, *25*, 443–458. [[CrossRef](#)]
26. Chiew, K.L.; Chang, E.H.; Tiong, W.K. Utilisation of website logo for phishing detection. *Comput. Secur.* **2015**, *54*, 16–26. [[CrossRef](#)]
27. Aggarwal, A.; Rajadesingan, A.; Kumaraguru, P. PhishAri: Automatic realtime phishing detection on twitter. In Proceedings of the 2012 eCrime Researchers Summit, Las Croabas, PR, USA, 23–24 October 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 1–12.
28. Dedakia, M.; Mistry, K. Phishing detection using content based associative classification data mining. *J. Eng. Comput. Appl. Sci.* **2015**, *4*, 209–214.
29. Dua, D.; Graff, C. *UCI Machine Learning Repository*; School of Information and Computer Science, University of California: Irvine, CA, USA, 2015. Available online: <https://archive.ics.uci.edu/ml/datasets/Phishing+Websites> (accessed on 10 June 2021).
30. Hall, M.A. Correlation-based feature selection for machine learning. Ph.D. Thesis, The University of Waikato, Hamilton, New Zealand, April 1999.
31. Barandela, R.; Sánchez, J.S.; Garcia, V.; Rangel, E. Strategies for learning in class imbalance problems. *Pattern Recognit.* **2003**, *36*, 849–851. [[CrossRef](#)]
32. Shukla, S.; Yadav, R.N. Unweighted class specific soft voting based ensemble of extreme learning machine and its variant. *Int. J. Comput. Sci. Inf. Secur.* **2015**, *13*, 59.
33. Ferri, C.; Hernández-Orallo, J.; Modrou, R. An experimental comparison of performance measures for classification. *Pattern Recognit. Lett.* **2009**, *30*, 27–38. [[CrossRef](#)]
34. Jeni, L.A.; Cohn, J.F.; De La Torre, F. Facing imbalanced data—recommendations for the use of performance metrics. In Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva, Switzerland, 2–5 September 2013; pp. 245–251.
35. Brzeziński, D.; Stefanowski, J.; Susmaga, R.; Szczęch, I. Visual-based analysis of classification measures and their properties for class imbalanced problems. *Inf. Sci.* **2018**, *462*, 242–261. [[CrossRef](#)]

36. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [[CrossRef](#)]
37. Ben-David, A.; Frank, E. Accuracy of machine learning models versus ‘hand crafted’ expert systems A credit scoring case study. *Expert Syst. Appl.* **2009**, *36*, 5264–5271. [[CrossRef](#)]
38. Zhou, Z.H. *Ensemble Methods: Foundations and Algorithms*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2012.
39. Dal Pozzolo, A.; Caelen, O.; Le Borgne, Y.A.; Waterschoot, S.; Bontempi, G. Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Syst. Appl.* **2014**, *41*, 4915–4928. [[CrossRef](#)]
40. Davis, J.; Goadrich, M. The relationship between precision-recall and ROC curves. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; Association for Computing Machinery: New York, NY, USA, 2006; pp. 233–240.