Indonesian Journal of Electrical Engineering and Computer Science
Vol. 28, No. 3, December 2022, pp. 1582∼1591
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v28.i3.pp1582-1591

Vote algorithm based probabilistic model for phishing website detection

Md. Sazzadul Islam Prottasha[1], Md. Zihadur Rahman[2], ABM Kabir Hossain[2], Samia Ferdous Mou[3], Md. Bulbul Ahmed[3], Md. Shamim Kaiser[4]
[1]Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh
[2]Department of Information and Communication Engineering, Bangladesh Army University of Engineering and Technology, Natore, Bangladesh
[3]Department of Electrical and Electronics Engineering, Bangladesh Army University of Engineering and Technology, Natore, Bangladesh
[4]Institute of Information Technology, Jahangirnagar University, Dhaka, Bangladesh

Article Info

Article history:

Received Jun 8, 2022
Revised Aug 17, 2022
Accepted Sep 5, 2022

Keywords:

Cyber security
Fraud detection
Machine learning
Online scam
Phishing website
Vote algorithm

ABSTRACT

Internet scams have been a major concern for everyone over the past decade. With the advancement of technology, attackers have formulated different kinds of contemporary fraudulent procedures to obtain user's sensitive information. Phishing is one of the oldest and common fraudulent attempts by which every year millions of internet users fall victim to scams resulting in losing their money. Different techniques and algorithms have been proposed by researchers in detecting phishing websites. However, the detection of phishing websites has few challenges since there are different subjective considerations and ambiguities involved in the detection process. This paper presents a two-stage probabilistic method for detecting phishing websites based on the vote algorithm. In the first stage, 29 different base classifiers have been used and their probabilistic values were calculated. In the second stage, the voting algorithm aggregated the probabilistic values of several base classifiers and the phishing websites were detected using the average of probabilities approach. The voting technique achieved an accuracy of 97.431% outperforming all of the single base classifiers in terms of accuracy.

Corresponding Author:

Md. Sazzadul Islam Prottasha
Department of Information and Communication Technology, Bangladesh University of Professionals
Mirpur Cantonment, Dhaka-1216, Bangladesh
Email: 19541026@bup.edu.bd

1. INTRODUCTION

Phishing is the deceitful utilization of electronic correspondences to mislead and exploit clients. It could be defined as a criminal mechanism to steal user's personal information like username, password, and monetary account details (credit card information). Phishing is the most popular attack among attackers since it is simple to target an individual by analyzing his behavior and preferences, which can be done simply by stalking him on social networking sites and then personalizing phishing sites/spoofed emails based on the analysis. Usually the attack starts with the victim receiving a message containing malicious software like wheel of fortune or quiz game. This type of applications are often used by attackers to lure the victim by offering them money, gift cards, free coupons or exclusive items. An individual may not understand or recognize that he is currently browsing through a phishing site and easily can fall victim to it.

As online business or e-commerce grows rapidly users become more vulnerable to phishing. According to a research study by 'Verizon' shows that 30% of phishing messages or spoofed emails get opened by the targeted individual [1]. A study by 'AVANAN' (a cloud security platform) shows 51% of phishing attacks contain malicious links. Statistics show the average financial loss for breach in confidential data is 3.86M [2]. The "2018 internet crime report" from the Internet Crime Complaint Center (IC3) indicates that $48,241,748 was reportedly lost per victim due to phishing/vishing/smishing attacks in the same year [3]. In fact, nearly 86% of all phishing was targeted on U.S. entities alone [4]. Which makes the U.S. the top-most vulnerable country for phishing. There were 26,379 victims of phishing in 2018 according to the 2018 internet crime report from the IC3. Although phishers use several kinds of techniques, most of the phishing website corresponds to some common attributes such as redirecting link, prefix or suffix, and (HTTP) token in the uniform resource locator (URL). By analyzing a total of 30 attributes, in this paper we proposed a machine learning approach using vote algorithm that aggregates multiple base classifiers to detect phishing websites

Different researcher has presented various methodologies for detecting phishing websites. We took cues from prior research. Jain and Gupta [5] provided visual similarities-based techniques to identify phishing websites from analyzing different feature sets. They analyzed different URL features, hypertext markup language (HTML) tags, cascading style sheet (CSS), and images to distinguish a phishing website from a legitimate website. The work also analyzed different phishing methods and their exploitation. Ali [6] used wrapper-based feature selection technique in combination with machine learning classifiers to detect phishing websites. This work demonstrated that wrapper-based feature selection improved the overall accuracy of the classifiers. The research was conducted using 7 different machine learning classifiers. Among them, the random forest classifier achieved the best accuracy of 97.1%. However, the wrapper-based feature selection technique may require more time and can consume extra computational overhead with some classifiers. Yang *et al.* [7] proposes a multidimensional feature-driven phishing detection technique using deep learning methods. In the first step, they extracted the character sequence features of the URL and later they combined the URL statistical features, webpage text features, and the classification result into multidimensional features thus identifying a phishing website. They achieved an accuracy of 98.99% while conducting the research on random URLs from the internet. The work by Karabatak and Mustafa [8] uses different classifiers on reduced dataset to detect phishing website. After taking the dataset [9] instead of using 30 attributes they reduced the dataset to 24-27 attributes using various feature selection algorithms. They achieved the highest accuracy of 97.58% using Lazy KStar classifier on a reduced dataset of 26 attributes. However, there are no comparison provided based on the time required to perform the classification on the reduced dataset. The work by Pan and Ding [10] uses the SVM technique to detect phishing web-page. Taking keyword, request URL, server form handler, the main body of a web page they tried to detect whether or not the web page is a legitimate site. Using the support vector machine (SVM) approach they achieved 84% of success rate. James *et al.* [11] uses various machine learning classifiers to detect phishing websites by analyzing the URL. They collected websites URL from Alexa, Dmoz and PhishTank. After analyzing the lexical feature of the URL's and using 90% test data split they achieved a maximum accuracy of 93.78% using the J48 decision tree algorithm.

Mhaske-Dhamdhere and Vanjale [12] proposes K-means algorithm to detect phishing emails. By taking 160 emails, they used K-means algorithm to distinguish between phishing emails and legitimate emails in real time. The work by Wardman and Warner[13] proposes an automatic phishing website detection technique using the message-digest algorithm. After downloading all the files from a phishing URL and using the MD5 database provided by the digital PhishNet (DPN), they matched the MD5 checksum with the URLs homepage. Using this technique they have been able to identify 30% of phishing websites by matching only the main HTML MD5. Mohammad [14] proposed a rule-based phishing website detection method where they imposed rules on the data set attributes that can define phishing website. They studied the minimum set of features that can be utilized to detect phishing websites. At the initial phase their proposed method achieved an average error rate of 5.76%. Later using a reduced feature sets they achieved an accuracy of 95,25%. Several studies [15]-[17] have suggested that URLs are the key attribute to easily detect phishing websites. Kumar *et al.* [18] proposes a hybrid methodology of SVM combined with probabilistic neural network model to identify phishing emails. Identification of malicious JavaScript-based code has been discussed [19]. Following a thorough examination of these works, we used multiple feature sets in our dataset, which includes 30 attributes and aggregated various algorithms using the voting technique to effectively identify phishing websites with high precision.

2. DATA PREPARATION

We collected the phishing website dataset from the UCI machine learning repository [9]. The dataset contains 11,055 instances of 30 different attributes. Among the 11,055 instances, 4,898 instances are phishing websites and 6,157 instances are legitimate websites. We used the feature selection [20] method among the attributes and grouped them according to their similarities. Table 1 shows the feature groups created from the phishing website dataset attributes. The Feature groups summarizes the key attributes that help in identifying the phishing website. Each attribute represents phishing characteristics in a unique way. Further details on these feature groups can be found in the work by Mohammad *et al.* [14].

Table 1. Feature groups of phishing website dataset attributes

Feature group	Attributes
URL based features	1. Having IP address 2. URL length 3. Shortinig service 4. Having at symbol 5. Double slash redirecting 6. Prefix suffix 7. Having sub domain 8. SSLfinal state 9. Domain registration length 10. Favicon 11.Port 12.HTTPS token 13.Request URL
JavaScript based features	14.Redirect 15.On mouseover 16.RightClick 17.popUpWidnow 18.Iframe
Anomaly based features	19.URL of anchor 20.Links in tags 21.SFH 22.Submitting to email 23.Abnormal URL 24.Links pointing to page
Statistics based features	25.Age of domain 26.DNSRecord 27.Web traffic 28.Page rank 29.Google index 30.Statistical report

2.1. URL based feature

URL's can provide a lot of information regarding a webpage. We take into account 13 attributes in the URL-based feature that indicates a phishing website. The features include having IP address instead of URL, long URL lengths that can potentially have hidden links inside it, URL shortening services like "Bitly" or "Tiny URL", URL having @ symbol that will potentially submit the information into an email, redirecting using double slash "//", having prefix-suffix in any URL, having no secure sockets layer (SSL) final state, Short domain registration link, using an uncommon port, having any subdomain, having HTTPS token in the URL and having any request URL strongly indicates that the website is unauthorized.

2.2. Anomaly based feature

In anomaly-based features, we take into account 6 attributes that indicates a phishing website. The features include URL of anchors connected to a different domain, having links in tags, server form handler is either empty or "about:blank", submitting information to email, abnormal URL where host name is absent in the URL and having links pointing to a page strongly indicates that the website is unauthorized.

2.3. JavaScript based feature

JavaScript is basically a scripting language used on the client-side of a website to make. Developers use JavaScript for making an interactive and animated web page. When a user sends some request in JavaScript enabled page, the script is sent to the browser to process the request. The attackers use these features to deceive the users by adding JavaScript on the phishing web page and making it look authentic. In JavaScript based feature we take account into 5 attributes that indicates a phishing website. The features include web page redirecting, using on mouseover to hide any link, right click disabled, showing pop up window and Iframe redirecting indicates that the website is unauthorized.

2.4. Statistics based feature

In statistics based features, we take account into 6 different attributes to detect a phishing website. These attributes mainly corresponds to statistical analysis. The features include the age of domain is less than 6 months, having no DNS record, less web traffic, page rank is lower, low google index score and lack of a statistical report suggests that the website is unauthorized.

3. PROPOSED METHOD

We employed a two-stage probabilistic model in our proposed model to detect phishing websites more accurately by minimizing the variance error. In the first stage, we calculated the probabilistic values given by the individual base classifiers for each output class. In the second stage, we took the probabilistic values given by each base classifier and used the voting algorithm to aggregate them. In the vote algorithm, we combine multiple base classifiers and using the output probabilities of different base classifiers we make the decision.

Different kinds of voting techniques are available, such as majority voting, average of probabilities, product of probabilities, median, minimum probabilities, maximum probabilities [21]. Vote algorithm can be used in any kind of class such as binary, nominal, date class, and numeric class. In this study, we employed the average of probabilities voting algorithm on our binary class phishing website dataset. In the average of probabilities, the algorithm checks the probabilities of every individual base classifier and averages the net probability. Considering each of the base classifier's output probabilities independent of each other, then averaging the probabilities helps in reducing the variance error that could be caused by a single base classifier. After computing the net average probability, the class label is assigned to the class having the maximum probability. Since there are only two class labels in our dataset hence, the voting algorithm simply calculated the probabilities of every single base classifier in the first stage, then averaged the probabilities in the second stage and predicted the class label. Figure 1 shows the flow diagram of our proposed method.

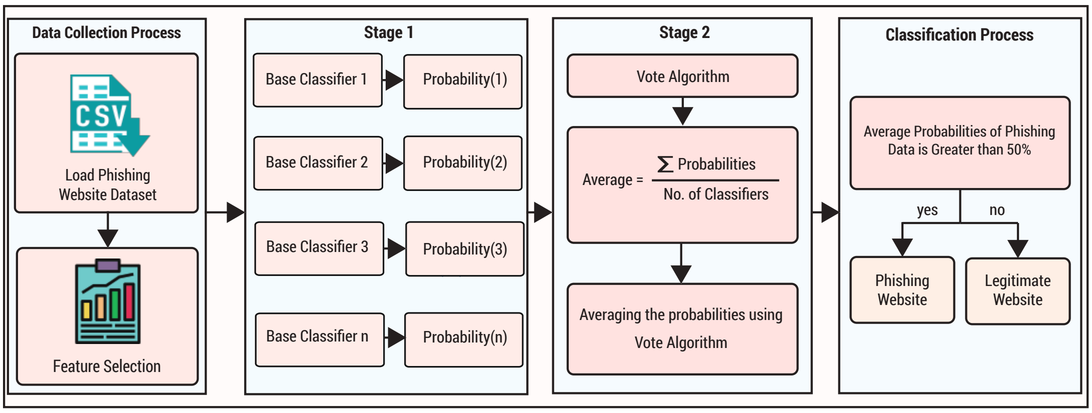


Figure 1. A method of detecting phishing website using voting algorithm

We combined multiple base classifiers using vote algorithm. And for each classifier we got a probabilistic value for our class label phishing website. The (1) shows the sum of the probabilities given by all the base classifiers when the class label is -1 in our dataset which is phishing website.

$$\sum P_{phishing} = P_{phishing}(1) + P_{phishing}(2) + ... + P_{phishing}(n) \quad (1)$$

Then the algorithm average the probabilities by dividing it with the number of classifiers used. The (2) shows the average probability for phishing calculation. Here ***n*** denotes the number of classifiers used.

$$Avg(P_{phishing}) = \frac{\sum P_{phishing}}{n} \quad (2)$$

We compare the average probability value with 50% because we have binary class labels in our dataset. When the average probability of a phishing website is greater than 0.5, we define the class label as phishing website. Conversely, it is the same for the legitimate website.

$$\textbf{if} Avg(P_{phishing}) > 0.5, \textbf{then} class - label = phishing$$

$$\text{Mean variance error} = \frac{\sum \sigma^2}{n} \quad (3)$$

Now assuming that errors of the base classifiers are independent of each other then for given n individual observations $P_1, P_2, P_3, ..., P_n$ each having variance σ^2, the mean variance error is given by (3). Here the mean-variance error of the voting algorithm can be smaller than the variance error of any single base classifier. Thus in several cases, the voting algorithm reduces the variance error of the individual base classifiers resulting in overall better accuracy.

4. RESULTS ANALYSIS

4.1. Classification performance

We employed the machine learning tools weka [22] and rapidminer [23] for the classification of the phishing websites. The experiment was carried out on a system with a GeForce GTX 1060 graphics card and 16 GB of RAM. 29 different base classifier with a ten-fold cross-validation was used to evaluate the performance of each classifier on raw data. The classification accuracy of our experiment is shown in Figure 2.

From Figure 2, we observe that random forest [24] achieved the highest accuracy among the single base classifiers in the first stage. Random committee [25], Lazy KStar [26] and IBK (k nearest neighbor) [27] all of them achieved an accuracy of more than 97%. So we discard all of the classifiers having less than 97% accuracy and considered the base classifiers that achieved more than 97% accuracy in the second stage. Along with classification accuracy, we have taken account of the receiver operating characteristic (ROC) and the time complexity of the base classifiers. In the second stage, we compared the accuracy, ROC and time complexity of the base classifiers and combined the base classifiers into different combinations using the voting algorithm to calculate the net probability for the binary class. Table 2 shows the confusion matrix of phishing website classification. From the confusion matrix we can observe true positive rate, true negative rate, false positive rate, false negative rate and accuracy of the classifier and hence the accuracy is calculated using the formula $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}(\%)$

Considering time constraint we observed that, random committee performed best with a time of 1.57 seconds while completing 10 fold cross-validation. Random forest and IBK performed very similar while having a time complexity of 10.54 seconds and 9.60 seconds respectively. The Lazy kStar took maximum time of 348.67 seconds on our machine for 10 fold cross-validation while predicting the phishing websites, which is inconvenient for a large dataset. Therefore, we excluded the Lazy KStar from voting technique. The result analysis of vote algorithm on pre-selected classifiers is shown in Table 3.

Based on the results reported in Table 3, we can clearly observe that the vote algorithm with every combination outperformed every other single base classifier in terms of accuracy. Firstly, we considered 3 base classifiers random forest, random committee, IBK and combined them using the vote algorithm. This combination achieved the maximum accuracy of 97.431% with a time of 21.71 seconds. Later we considered 2 base classifiers with different combinations and compared the accuracy. A combination of random committee and IBK achieved an accuracy of 97.359% with a time of 10.17 seconds. And a combination of random forest and IBK achieved an accuracy of 97.332% with a time of 20.14 seconds. Among the single base classifiers, random forest achieved the highest accuracy on 10-fold cross validation.

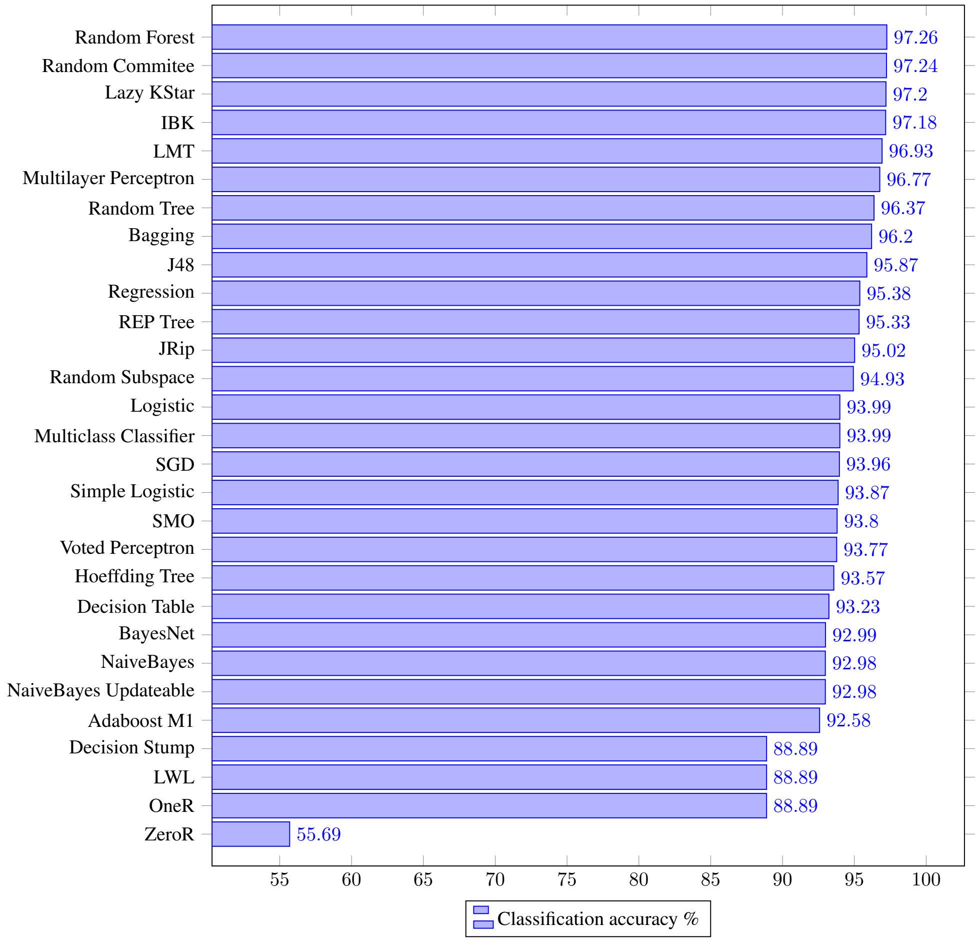


Figure 2. Performance of different classifiers on raw phishing website dataset

Table 2. Confusion matrix of phishing website classification

	Predicted phishing	Predicted Legitimate
Actual phishing	True positive (TP)	False negative (FN)
Actual Legitimate	False positive (FP)	True negative (TN)

Table 3. Classification accuracy, confusion matrix, ROC and time needed for pre-selected classifiers (results only for raw sample dataset, sorted by accuracy in descending order)

Classifier	Accuracy (%)	Precision	Recall	ROC	Time (Sec)
Vote (random forest + IBK + random vommittee)	**97.431%**	**0.974**	**0.974**	0.996	21.71 s
Vote (random vommittee + IBK)	97.359%	0.974	0.974	0.993	10.17 s
Vote (random forest + IBK)	97.332%	0.973	0.973	0.996	20.14 s
Random forest	97.259%	0.973	0.973	0.996	10.54 s
Rando committee	97.241%	0.972	0.972	0.992	**1.57 s**
Lazy KStar	97.196%	0.972	0.972	**0.997**	348.67 s
IBK	97.178%	0.972	0.972	0.989	9.60 s

The confusion matrix of vote algorithm along with other classifiers is shown in Figure 3 respectively shown in Figures 3(a)-3(f). By comparing the confusion matrix of random forest in Figure 3(d) to the matrix 3(a)-3(c) of the vote method, we can observe that the vote algorithm reduced the number of false positive and false negative occurrences, resulting in a lower error rate. The same thing happened with the random committee and IBK.

Considering the Area under the ROC curve(AUC) covered by the classifiers, the vote algorithm performs considerably well in AUC along with other classifiers. Figure 4 demonstrates the RUC curve of different algorithms. The ROC curve for the voting method is nearly a perfect curve, covering an area of 0.996 in the AUC. The ROC curve for the ZeroR method is the lowest, covering a 0.902-square-meter region.

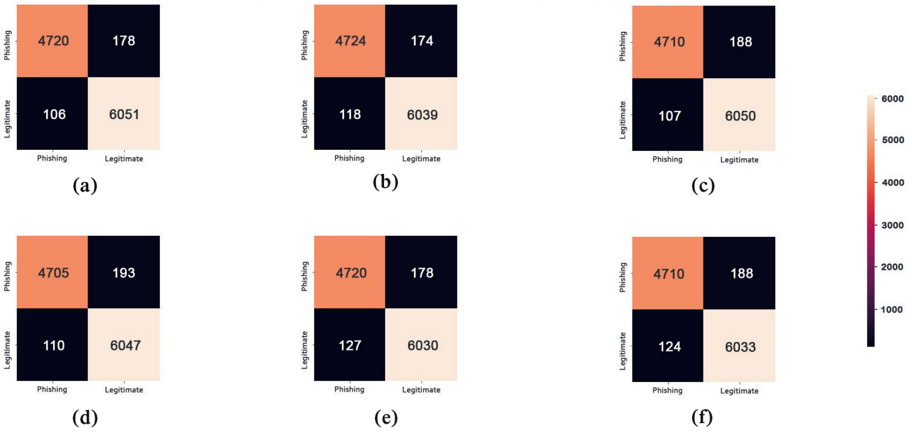


Figure 3. Confusion matrix of different classifiers (a) vote (random forest+IBK+random committee), (b) vote (random committee+IBK), (c) vote (random forest+IBK), (d) random forest, (e) random vommittee, and (f) IBK

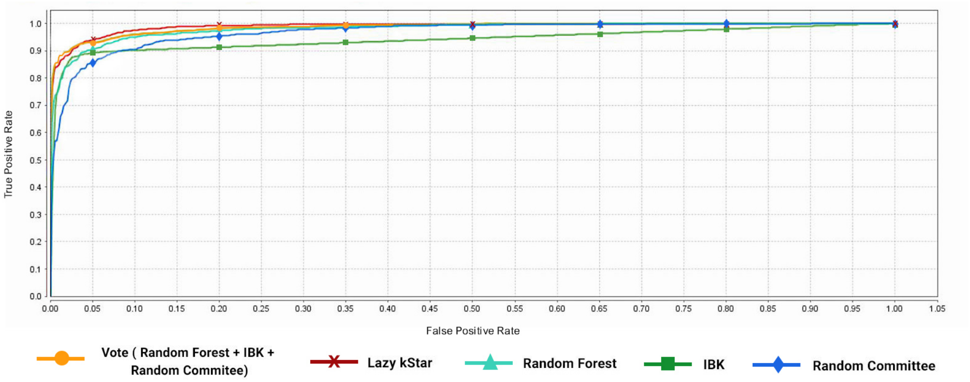


Figure 4. ROC curve of different algorithms on phishing dataset

4.2. Discussion and findings

After analyzing the overall results, we have acquired some interesting findings in our study. The findings are as follows:

- In multiple instances, the vote algorithm reduced the False Positive and False Negative instances resulting in higher accuracy than the single base classifiers. However, the voting technique required more time to perform the classification task than the single base classifiers.
- The Lazy KStar achieved the maximum ROC while it also took considerably long time to perform the classification task. Hence, there is obviously a trade-off between the time and the ROC of the base classifiers.
- The Lazy KStar took the minimum time to perform the classification task yet provided a similar accuracy level to the voting algorithm. Hence, the Lazy KStar should be preferred for a faster classification process over the voting algorithm.
- In case of time constraint is not a concern, the vote algorithm should be preferred for the classification task, since it will result in higher overall accuracy.

Various authors have used different approaches towards phishing website detection. A statistical comparison between different phishing detection techniques along with our proposed model is shown in Table 4. From table 4, in terms of accuracy and time complexity, the vote algorithm provided much better accuracy than the wrapper-based machine learning technique proposed by Ali [6] on the same dataset. Also, without reducing the parameters, the vote algorithm achieved a similar level accuracy to the result reported by Karabatak and Mustafa [8]. Comparing the accuracy, Precision, Recall, ROC and time complexity, we conclude that the vote algorithm reduced the variance error of different single base classifiers and performed better in identifying phishing websites accurately.

Table 4. Comparison between existing phishing detection approaches with our proposed technique

Author	Approach	Dataset used	Accuracy
Ali [6]	Wrapper based feature selection approach	UCI machine learning repository phishing dataset	97.1%
Yang *et al.* [7]	Deep learning based multidimensional feature driven approach	Random Url's from the internet	98.99%
Karabatak and Mustafa [8]	Reduced feature selection based approach	UCI machine learning repository phishing dataset	97.58%
Pan and Ding [10]	DOM object anomalies based anti-phishing approach	Random Url's from the internet	84%
James *et al.* [11]	Lexical feature based approach	Url's from Alexa, DMOZ, and PhishTank	93.78%
Mohammad *et al.* [14]	Intelligent rule-based approach	Url's from PhishTank and Millersmiles	95.25%
Proposed model	Vote algorithm based approach	UCI machine learning repository phishing dataset	97.431%

5. CONCLUSION

In the age of the internet, cyber security is a major concern for everyone. Phishing is a prevalent type of cyber attack that everyone should be aware of in order to stay safe. In this study, a two-stage probabilistic model based on vote algorithm has been proposed for detecting phishing websites. Firstly, we performed classification using 29 different base classifiers on phishing website dataset taken from the UCI machine learning repository. Based on the results of 29 base classifiers, we selected four base classifiers having more than 97% accuracy. By analyzing the confusion matrix, ROC area and time required to complete 10 fold cross-validation on selected classifiers, we discarded the Lazy KStar algorithm due to its time constraints. We aggregated the other three base classifiers using our proposed vote algorithm.

The classification results indicate that the voting method minimizes false positive and false negative instances of single base classifiers for any combination of base classifiers, thus reducing the error rate. Combining three base classifiers, vote algorithm achieved a maximum accuracy of 97.431% outperforming all single base classifiers in terms of accuracy. However, the voting technique takes longer than single base classifiers to perform classification. Our experiment was employed on raw data without any filter or data segmentation. The accuracy can further be increased by using filters or data segmentation on raw data. In the future, we plan to integrate our proposed vote algorithm based phishing detection algorithm into a browser extension that will detect any phishing website or phishing links in real-time.

REFERENCES

[1] Verizon, "Verizon data breach investigations report," 2018. Accessed: Jul. 3, 2020. [Online]. available: https://www.verizon.com/business/resources/reports/2018-data-breach-digest.pdf.

[2] S. Shepard, "The average cost of a data breach," Security today. https://securitytoday.com/articles/2018/07/17/the-average-cost-of-a-data-breach.aspx (accessed: Jul. 3, 2020).

[3] Federal Bureau of Investigation, "2018 internet crime report," 2019. Accessed: Jul. 15, 2020. [Online]. Available: http://www.fbi.gov/news/stories/ic3-releases-2018-internet-crime-report-042219

[4] Phishlabs, "2018 phishing trends intelligence report," 2019. Accessed: Jun. 11, 2020. [Online]. Available: http://info.phishlabs.com/hubfs/2018%20PTI%20Report/PhishLabs%20Trend%20Report_2018-digital.pdf

[5] A. K. Jain and B. B. Gupta, "Phishing detection: analysis of visual similarity based approaches," *Security and Communication Networks*, vol. 2017, p. 5421046, 2017, doi: 10.1155/2017/5421046.

[6] W. Ali, "Phishing website detection based on supervised machine learning with wrapper features selection," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 9, pp. 72-78, 2017, doi: 10.14569/IJACSA.2017.080910.

[7] P. Yang, G. Zhao, and P. Zeng, "Phishing website detection based on multidimensional features driven by deep learning," *IEEE Access*, vol. 7, pp. 15196-15209, 2019, doi: 10.1109/ACCESS.2019.2892066.

[8] M. Karabatak and T. Mustafa, “Performance comparison of classifiers on reduced phishing website dataset,” in *2018 6th International Symposium on Digital Forensic and Security (ISDFS), Antalya*, 2018, pp. 1-5, doi: 10.1109/ISDFS.2018.8355357.

[9] “Phishing websites data set,” UCI Machine Learning Repository, 2020. Accessed: Jun. 11, 2020. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/phishing+websites .

[10] Y. Pan and X. Ding, “Anomaly based web phishing page detection,” in *2006 22nd Annual Computer Security Applications Conference (ACSAC’06)*, 2006, pp. 381-392, doi: 10.1109/ACSAC.2006.13.

[11] J. James, L. Sandhya, and C. Thomas, “Detection of phishing URLs using machine learning techniques,” *2013 International Conference on Control Communication and Computing (ICCC)*, 2013, pp. 304-309, doi: 10.1109/ICCC.2013.6731669.

[12] V. Mhaske-Dhamdhere and S. Vanjale, “A novel approach for phishing emails real time classification using k-means algorithm,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 6, pp. 5326-5332, 2018, doi: 10.11591/ijece.v8i6.pp5326-5332.

[13] B. Wardman and G. Warner, “Automating phishing website identification through deep MD5 matching,” in *2008 eCrime Researchers Summit, Atlanta, GA*, 2008, pp. 1-7, doi: 10.1109/ECRIME.2008.4696972.

[14] R. M. Mohammad, F. Thabtah, and L. McCluskey, “Intelligent rule-based phishing websites classification,” *IET Information Security*, vol. 8, no. 3, pp. 153-160, 2014, doi: 10.1049/iet-ifs.2013.0202.

[15] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, “Machine learning based phishing detection from URLs,” *Expert Systems with Applications*, vol. 117, pp. 345-357, 2019, doi: 10.1016/j.eswa.2018.09.029.

[16] O. V. Lee *et al.*, ”A malicious URLs detection system using optimization and machine learning classifiers,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 17, no. 3, pp. 1210-1214, 2019, doi: 10.11591/ijeecs.v17.i3.pp1210-1214.

[17] J. A. Jupin, T. Sutikno, M. A. Ismail, M. S. Mohamad, S. Kasim, and D. Stiawan, “Review of the machine learning methods in the classification of phishing attack,” *Bulletin of Electrical Engineering and Informatics*, vol. 8, no. 4, pp. 1545–1555, Dec. 2019, doi: 10.11591/eei.v8i4.1344.

[18] A. Kumar, J. M. Chatterjee, and V. G. Díaz, “A novel hybrid approach of SVM combined with NLP and probabilistic neural network for email phishing,” *International Journal of Electrical and Computer Engineering*, vol. 10, no. 1, pp. 486-493, 2020, doi: 10.11591/ijece.v10i1.pp486-493.

[19] Y. Fang, C. Huang, Y. Su, and Y. Qiu, “Detecting malicious JavaScript code based on semantic analysis,” *Computers Security*, vol. 93, p. 101764, Jun. 2020, doi: 10.1016/j.cose.2020.101764.

[20] S. Shabudin, N. Samsiah, K. Akram, and M. Aliff, “Feature selection for phishing website classification,” *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, pp. 587-595, 2020, doi: 10.14569/IJACSA.2020.0110477.

[21] B. Parhami, “Voting algorithms,” *IEEE Transactions on Reliability*, vol. 43, no. 4, pp. 617–629, 1994, doi: 10.1109/24.370218.

[22] E. Frank, M. A. Hall, and I. H. Witten, *The WEKA workbench. online appendix for “data mining: practical machine learning tools and techniques*,” 4th ed. Morgan Kaufmann, 2016.

[23] I.Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler, “Yale: Rapid prototyping for complex data mining tasks,” in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*,2006, pp. 935-940, doi: 10.1145/1150402.1150531.

[24] A. Liaw and M. Wiener, “Classification and Regression by RandomForest,” *R News*, vol. 2, pp. 18-22, 2002.

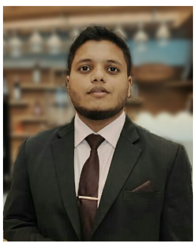
[25] T. G. Dietterich , “Ensemble methods in machine learning,” *International workshop on multiple classifier systems*,2000, pp. 1-15, doi: 10.1007/3-540-45014-9_1

[26] J. G. Cleary and L. E. Trigg, “K*: an instance-based learner using an entropic distance measure,” in *Machine Learning Proceedings 1995*, Elsevier, 1995, pp. 108–114.

[27] D. Aha, D. Kibler, and M. Albert, “Instance-based learning algorithms,” *Machine Learning*, vol. 6, no. 1, pp. 37–66, Jan. 1991, doi: 10.1007/BF00153759.

BIOGRAPHIES OF AUTHORS

Md. Sazzadul Islam Prottasha obtained his Bachelor and Master’s Degree in Information and Communication Engineering (ICE) from Bangladesh University of Professionals (BUP) in 2018 and 2021 respectively. Currently he is working on multiple research project related to image processing, deep learning and computer vision. His research interests include Machine Learning, Computer Vision, Image Processing, Big Data Analytics, Artificial Intelligence, and IOT. He can be contacted at email: 19541026@bup.edu.bd.



Md. Zihadur Rahman is working as a lecturer at the Department of Information and Communication Engineering in Bangladesh Army University of Engineering and Technology (BAUET), Qadirabad Cantonment, Natore. He received his B.Sc. degree in department of Information and Communication Technology from Bangladesh University of Professionals (BUP), Mirpur, Bangladesh, in 2018. Currently he is pursuing M.Sc. degree in Computer Software Engineering from Rajshahi University of Engineering and Technology (RUET), Rajshahi, Bangladesh. His research interest include Machine learning, Data Mining, Network security and Bio Informatics. He can be contacted at email: zihadurrahman47@gmail.com.

ABM Kabir Hossain is working as a lecturer at the Department of Information and Communication Engineering in Bangladesh Army University of Engineering and Technology (BAUET), Qadirabad Cantonment, Natore. He received his B.Sc. degree in department of Information and Communication Technology from Bangladesh University of Professionals (BUP), Mirpur, Bangladesh, in 2018. His researches are in fields of machine learning, Signal Processing, Deep learning and Cloud Computing. He can be contacted at email: kabir111192@gmail.com.

Samia Ferdous Mou is working as a lecturer at the Department of Electrical and Electronic Engineering in Bangladesh Army University of Engineering and Technology (BAUET), Qadirabad Cantonment, Natore. She received her B.Sc. degree in Electrical and Electronic Engineering from Bangladesh Army University of Engineering and Technology (BAUET), Natore, Bangladesh, in 2019. Currently she is pursuing M.Sc. degree in Electrical and Electronic Engineering from Rajshahi University of Engineering and Technology (RUET), Rajshahi, Bangladesh. Her research interest include Control system, Renewable Energy and Optical Fiber Communication System. She can be contacted at email: samia.mou.bauet@gmail.com.

Md. Bulbul Ahmed Bhadon is a graduate student at Rajshahi University of Engineering Technology and working as a lecturer at the Department of Electrical and Electronic Engineering in Bangladesh Army University of Engineering and Technology (BAUET), Qadirabad Cantonment, Natore. He obtained Bachelor degree in Electrical Electronics Engineering from Rajshahi University of Engineering Technology in 2017. His research interest includes machine learning, image processing microgrid. He can be contacted at email: badhoneee12@gmail.com.

Md. Shamim Kaiser is working as a Professor at the Institute of Information Technology, Jahangirnagar University, Dhaka He received the bachelor's and master's degrees in applied physics, electronics and communication engineering from the University of Dhaka, Bangladesh, in 2002 and 2004, respectively, and the Ph.D. degree in telecommunication engineering from the Asian Institute of Technology (AIT), Pathumthani, Thailand, in 2010. He has authored more than 100 papers in different peer-reviewed journals and conferences. His current research interests include data analytics, machine learning, cognitive radio networks,big data, cyber security, and renewable energy. He can be contacted at email: mskaiser@juniv.edu.