Enhanced Phishing Website Prediction Using Base and Ensemble Classifier Techniques with Cross-validation

by K Nidhi Patil
Pooja Rathlavath
Aryan Jadhav

Submission date: 07-Mar-2024 02:25AM (UTC-0500)

Submission ID: 2314047535

File name: Phishing_website_prediction_using_base_and_ensemble_classifier_techniques_with_cross-validation_-_Paper.docx (930.68K)

Word count: 3751

Character count: 20879

Enhanced Phishing Website Prediction Using Base and Ensemble Classifier Techniques with Cross-validation

K NIDHI PATIL
Department of Computer
Science and Engineering
Institute of Aeronautical
Engineering, Hyderabad.
Mail id:
20951a05a1@iare.ac.in

POOJA RATHLAVATH
Department of Computer
Science and Engineering
Institute of Aeronautical
Engineering, Hyderabad.
Mail id:
20951a05a9@iare.ac.in

ARYAN JADHAV
Department of Computer
Science and Engineering
Institute of Aeronautical
Engineering, Hyderabad.
Mail id:
20951a0513@iare.ac.in

Dr. K SUVARCHALA
Assistant Professor
Department of Computer
Science and Engineering
Institute of Aeronautical
Engineering, Hyderabad.
Mail id:
k.suvarchala@iare.ac.in

Abstract: The goal of this study is to make internet users safer by looking into the growing threat of hacking on public networks. Even though people are always working to make IT systems safer, this paper talks about how to find and predict fake website URLs and stresses how dangerous phishing attacks are. The study uses machine learning to use main classifiers like Logistic Regression, Decision Tree, SVM, KNN, Naive Bayes, LDA, and ensemble-based methods like CNN, RNN, and a Voting Classifier. The research is divided into three stages: first, classification using base classifiers; next, deployment of ensemble classifiers; and finally, evaluation of different datasets with and without cross-validation. Notably, CNN and RNN can achieve an amazing 100% accuracy rate. The results strengthen the importance of ensemble methods in identifying fake websites. They also give us useful information for future study and help keep online spaces safe.

Index terms - *Phishing, Hacking, Data diddling, Machine learning, Ensemble*

1. INTRODUCTION

As people become more and more dependent on the internet for many things in their daily lives, like shopping, banking, and smart home solutions [1], the number of cyber dangers has also grown. Globally run network platforms make things easier than ever, but they're also a great place for cyber dangers to grow [2]. Phishing is one of the sneakiest of these threats because its victims often don't notice it or don't recognize it enough [3].

Users are at great risk when they fall for phishing, an online crime. Attackers mostly go after two groups: people who don't know much about the technical side of the internet and careless people who know the risks but don't pay attention to their online security [4]. This growing worry calls for a thorough investigation into the complex nature of phishing threats and the creation of strong defenses against them as people become more familiar with the digital world [5].

The goal of this study is to make it easier for public networks to spot phishing websites by using CNN, RNN, Logistic Regression, Decision Tree, SVM, KNN, Naive Bayes, LDA, and a Voting Classifier. The study rates how well they do, highlighting how important group methods are for making the internet safer.

The danger of phishing is always present on public networks, which means that users' data is always at risk. Even though people are always trying, it's still hard to find and avoid fake websites. This study tries to solve this problem by using different types of machine learning models to better find and stop hacking threats.

As per the 2020 Phishing Attack Scene Report from Extraordinary Horn (2020 Phishing Attack Scene 2020), 53% of network safety experts said they saw an ascent in these Attacks during the COVID-19 Pandemic. Consistently, organizations face around 1185 phishing Attacks . It takes corporate security teams one to four days to fix a cyberattack. 30% of cyber security specialists said that phishing Attacks were exceptionally fruitful during the pandemic, as indicated by a similar report (2020 Phishing Attack Scene 2020). Their review (2020 trick Attack Scene 2020) showed the number of trick messages are shipped off organizations all over the planet.

2. LITERATURE SURVEY

Phishing is a type of theft that is still a threat in today's digital world. Researchers have looked into a number of different ways to spot and stop hacking attacks. Hong et al. [14] suggested a way to find fake URLs that uses language traits and sites that have been banned. Their method looks at the wording of URLs and compares them to known domains that are on a ban in order to find possible phishing websites.

In the same way, Orunsolu et al. [27] created a model that can predict when scams will happen. The main goal of their study was to create a machine learning-based system that could spot scam efforts by looking at different parts of URLs and site content. With their prediction model, they hoped to protect people from hacking attempts before they happened.

Sonowal and Kuppusamy [36] created PhiDMA, a model for finding fake emails that uses more than one filter. They used numerous filters to look at different aspects of scam emails and websites to enhance detection and reduce false negatives.

A case-based reasoning phishing detection system called CBR-PDS was suggested by Abutair et al. [3]. By using a database of past phishing attacks, their system could find connections between new and known phishing attempts. This let them find and stop them quickly.

A lot of study has also been done on machine learning methods to finding fake emails. Koray et al. [18] came up with a way to find fake URLs that uses machine learning. They trained a model on a set of known scam and real URLs so that it could correctly classify new URLs.

Gupta et al. [12] came up with a new way to find phishing URLs in real time situations using lexical-based machine learning. Their method was meant to quickly and accurately spot hacking efforts by using machine learning algorithms and language traits taken from URLs.

Using machine learning, Jain and Gupta [15] looked at how to find fake websites on the client side. Their method was meant to find strange websites in real time by looking at page content and how users interact with them. This would protect against phishing attacks before they happen.

Chin et al. [6] suggested Phishlimiter, a way to find and stop phishing attacks using software-defined networking. Their method tried to find and stop phishing attempts at the network level by building phishing monitoring tools into the network infrastructure. This made it less likely that attacks would be successful.

To sum up, the literature review shows a number of different methods and tactics that can be used to find and stop phishing attacks. Phishing is becoming a bigger problem in the digital world, so experts are always looking for new ways to fight it. These include network-level security, lexical-based analysis, and machine learning algorithms.

3. METHODOLOGY

i) Proposed System:

A group of machine learning methods, such as CNN, RNN, Logistic Regression, Decision Tree, SVM, KNN, Naive Bayes, LDA, and a Voting Classifier, are used in the suggested system to create a strong defense solution. Notably, the method gives more weight to the fact that Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are more accurate than other algorithms. The focus on CNN and RNN, which are known for being good at finding fake websites, is meant to improve the system's general performance. The suggested system wants to improve the accuracy and speed of phishing identification and prediction by using the best parts of these advanced algorithms. This smart combination of different algorithms shows that the system is dedicated to giving users a full defense against the constantly changing range of online risks, providing greater safety for all internet activities.

ii) System Architecture:

The system architecture uses several machine learning approaches to classify datasets-3. First, data is divided 80:20 into training and testing sets. CNN and RNN feature extraction and classification follow. After that, k-fold cross-validation with k=10 improves model assessment and extension. Logistic Regression, Decision Tree, SVM, KNN, Naive Bayes, LDA, and a Voting Classifier are trained and tested using the best k-fold cross-validation approach. Finally, these models' results are given so you may evaluate their performance and choose the best dataset technique. Each classifier's accuracy, precision, recall, F1-score, and confusion matrix are evaluated. This simplifies comparing and selecting the best datasets-3 classification model.

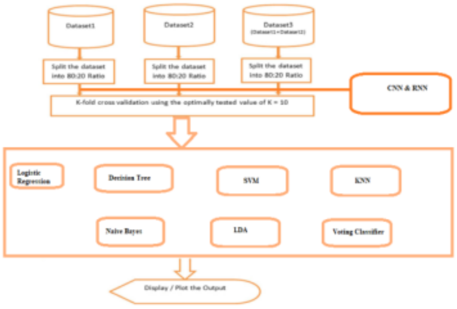

Dataset1
Dataset2
Dataset3
(Dataset1+Dataset2)
Split the dataset into 80:20 Ratio
Split the dataset into 80:20 Ratio
Split the dataset into 80:20 Ratio
CNN & RNN
K-fold cross validation using the optimally tested value of K = 10
Logistic Regression
Decision Tree
SVM
KNN
Naive Bayes
LDA
Voting Classifier
Display / Plot the Output

Fig 1 Proposed Architecture

iii) Data processing:

Several important steps must be taken during data handling to make sure that machine learning models are accurate and useful. First, it's important to make sure that the collection doesn't have any null numbers. This step helps find any data points that are missing or not full, which could have an effect on how well the model works. Imputation methods or deleting the data lines that go with null numbers can be used to deal with them.

After looking for null values, data visualization methods are used to learn more about the dataset's features and how they are spread out. Visualization helps find trends, relationships, and outliers in the data, which makes it easier to choose features and understand what the model means.

Another important part of data handling is feature extraction. It includes changing raw data into a shape that machine learning systems can understand. In this step, methods like dimensionality reduction, feature scaling, and creating new features based on subject information may be used.

Preprocessing methods, such as normalization and storing category variables, are also used to make sure that all the features are on the same scale and work with the machine learning algorithms that were chosen.

Overall, data handling is a very important part of getting the dataset ready for training and testing models. This improves the accuracy and usefulness of the machine learning models used for classification tasks.

iv) Training & Testing:

You utilize 80% of the samples to train the machine learning model and 20% to test it when you split the dataset in 80:20 proportions. This split makes sure that the model is trained on a big enough part of the dataset to learn the trends and connections in the data. It also lets someone else test how well the model does on data it hasn't seen before.

An iterative optimization method like gradient descent or backpropagation is used to change the settings of the machine learning model while it is learning from the training data. The traits and names in the training set help the model reduce the error or loss function, which makes it better at making correct guesses in the long run.

The tests data, which the model didn't see during training, are used to judge its success after training. This rating gives a fair guess of how well the model works with new data that it hasn't seen before. The testing set lets different performance measures, like accuracy, precision, recall, and F1-score, be measured, which shows how well the model works at making predictions.

By separating the data into training and testing sets and doing the training and testing processes separately, we make sure that the machine learning model's estimates of its performance are accurate and show how well it can adapt to new data.

v) Algorithms:

CNN: Convolutional neural networks (CNNs) learn from examples using perceptrons, machine learning units. CNNs can be used to handle images, understand verbal language, and do other cognitive tasks.

CNNs are used for projects that need to classify images, find objects in images, and separate images into parts. CNNs can be used to correctly identify diseases in a medical imaging project by putting MRI pictures into groups.

RNN: Recurrent neural networks (RNNs) are the best way to deal with sequential data. Siri on Apple products and Google's voice search use them. It's the first algorithm with an internal memory that knows what it was given. This makes it great for machine learning tasks with sequential data.

Recurrent Neural Networks (RNNs) are useful for sequential data analysis, like in natural language processing (NLP) jobs like translating languages or figuring out how people feel about words. RNNs can be used in a robot project to make replies that are relevant to the situation based on the messages that are sent.

Logistic Regression:You can use logistic regression, a statistical analysis method, to guess a yes or no answer based on what you already know about a set of data. A logistic regression model guesses what a dependent data variable will be by looking at how one or more independent factors are related to each other.

Logistic Regression is often used to classify things into two groups. As part of a project to find email spam, logistic regression can be used to sort emails into two groups: spam and non-spam.

Decision Tree: A machine learning method called a decision tree algorithm uses a decision tree to figure out what will happen next. It uses a tree-like structure to show choices and the outcomes that could happen. The method works by repeatedly dividing the data into groups based on the most important trait at each tree point.

You can use decision trees for both classification and regression. Decision trees can be used in a credit risk assessment project to figure out if a loan applicant is likely to not pay back the loan based on information about their finances.

SVM: A support vector machine (SVM) is a type of guided learning method used in machine learning to do tasks like regression and classification. SVMs are very good at binary classification problems, which require putting data points into two groups.

Support Vector Machines (SVMs) work well for classification jobs where the decision boundaries are complicated. In a project to recognize handwriting numbers, SVMs can be used to correctly group pictures of handwritten numbers.

KNN: The supervised learning classifier k-nearest neighbors (KNN) groups data points by proximity. No parameters govern it.

Classification and regression may be done using the simple K-Nearest Neighbors (KNN) approach. A movie recommendation system may employ KNN to match consumers' likes and movie features.

Naïve Bayes: The Naïve Bayes classifier is a guided machine learning approach for text classification and grouping tasks. Additionally, it is a generative learning technique that models how data of a class or group are distributed.

Naive Bayes is often used to sort text into groups for jobs like figuring out how people feel about something or finding spam. As part of a project on mood analysis, Naive Bayes can be used to decide whether social media posts are good, negative, or neutral based on what they say.

LDA: Machine learning uses guided learning approach Linear Discriminant Analysis (LDA) to group items. It determines the optimum linear blend of features to distinguish collection classes.

Linear Discriminant Analysis (LDA) is used to get rid of unnecessary dimensions and pull out important features. In a project to recognize faces, LDA can be used to pull out traits that make faces unique so that identities can be confirmed.

Voting Classifier: Voting Classifiers are machine learning models that learn from a set of models and estimate an output (class) depending on which class is most likely to be picked.

Voting Classifier uses several separate classifiers to make predictions more accurate. In a project to find credit card fraud, a vote predictor can be used to combine the results of different algorithms' guesses to better spot fake transactions.

4. EXPERIMENTAL RESULTS

Accuracy: How well a test can tell the difference between sick and healthy people is called its accuracy. To get an idea of how accurate a test is, we should figure out what percentage of cases are true positives and true negatives. In terms of math, this can be written as

Accuracy = TP + TN TP + TN + FP + FN.

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

Precision: Precision is the proportion of properly identified cases or samples to hits. Here's how you calculate precision:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

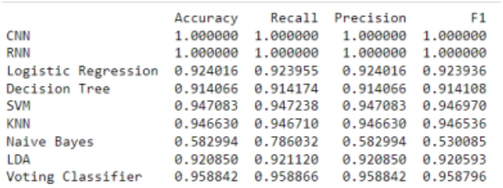
Recall: In machine learning, recall measures how successfully a model finds all key examples of a class. It indicates how effectively a model captures class cases. Divide the number of accurately anticipated positive observations by the total genuine positives.

$$Recall = \frac{TP}{TP + FN}$$

F1-Score: The F1 score is a way to rate the correctness of a machine learning model. It takes a model's accuracy and memory scores and adds them together. The accuracy measure counts how many times, across the whole collection, a model made a correct guess.

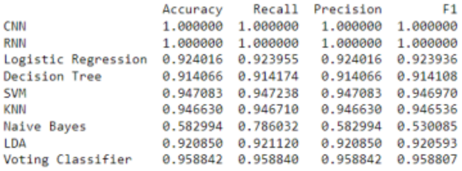
$$\textbf{F1 Score} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

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



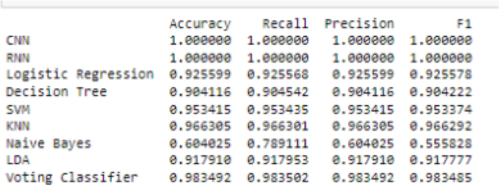
```
                     Accuracy    Recall  Precision        F1
CNN                  1.000000  1.000000   1.000000  1.000000
RNN                  1.000000  1.000000   1.000000  1.000000
Logistic Regression  0.924016  0.923955   0.924016  0.923936
Decision Tree        0.914066  0.914174   0.914066  0.914108
SVM                  0.947083  0.947238   0.947083  0.946970
KNN                  0.946630  0.946710   0.946630  0.946536
Naive Bayes          0.582994  0.786032   0.582994  0.530085
LDA                  0.920850  0.921120   0.920850  0.920593
Voting Classifier    0.958842  0.958866   0.958842  0.958796
```

Fig 2 Dataset -1 Performance Evaluation Table



```
                     Accuracy    Recall  Precision        F1
CNN                  1.000000  1.000000   1.000000  1.000000
RNN                  1.000000  1.000000   1.000000  1.000000
Logistic Regression  0.924016  0.923955   0.924016  0.923936
Decision Tree        0.914066  0.914174   0.914066  0.914108
SVM                  0.947083  0.947238   0.947083  0.946970
KNN                  0.946630  0.946710   0.946630  0.946536
Naive Bayes          0.582994  0.786032   0.582994  0.530085
LDA                  0.920850  0.921120   0.920850  0.920593
Voting Classifier    0.958842  0.958840   0.958842  0.958807
```

Fig 3 Dataset -2 Performance Evaluation Table



```
                     Accuracy    Recall  Precision        F1
CNN                  1.000000  1.000000   1.000000  1.000000
RNN                  1.000000  1.000000   1.000000  1.000000
Logistic Regression  0.925599  0.925568   0.925599  0.925578
Decision Tree        0.904116  0.904542   0.904116  0.904222
SVM                  0.953415  0.953435   0.953415  0.953374
KNN                  0.966305  0.966301   0.966305  0.966292
Naive Bayes          0.604025  0.789111   0.604025  0.555828
LDA                  0.917910  0.917953   0.917910  0.917777
Voting Classifier    0.983492  0.983502   0.983492  0.983485
```

Fig 4 Dataset -3 Performance Evaluation Table

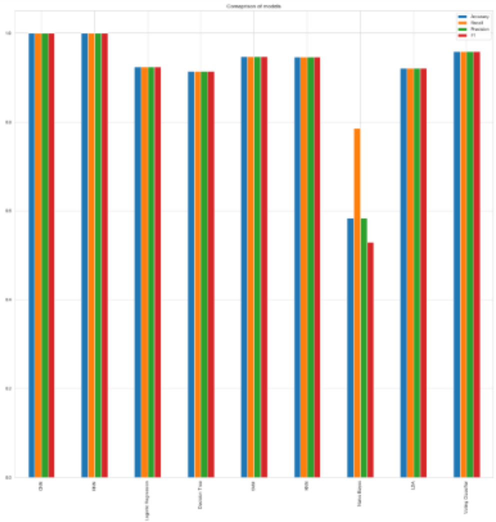


Fig 5 Dataset -1 Performance Evaluation Graph

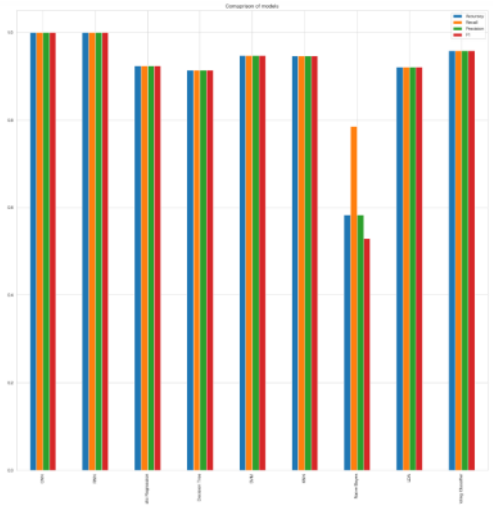


Fig 6 Dataset -2 Performance Evaluation Graph

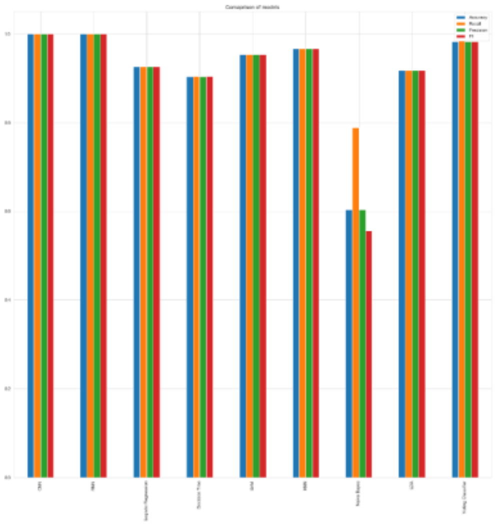
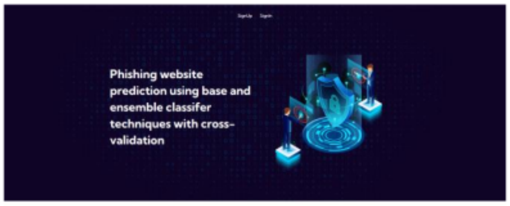


Fig 7 Dataset -3 Performance Evaluation Graph



Phishing website prediction using base and ensemble classifer techniques with cross-validation

Fig 8 Home Page

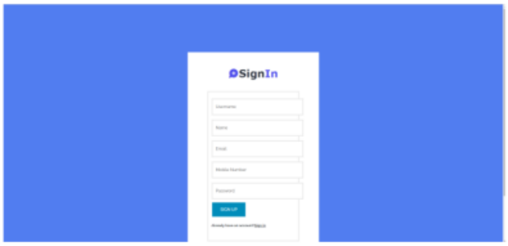


Fig 9 Signup Page

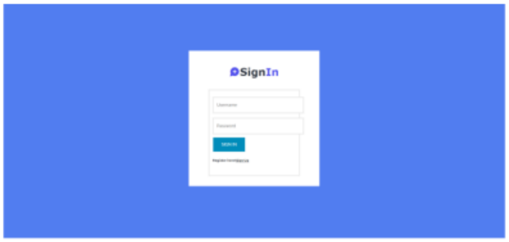
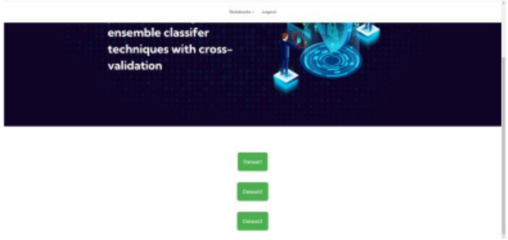
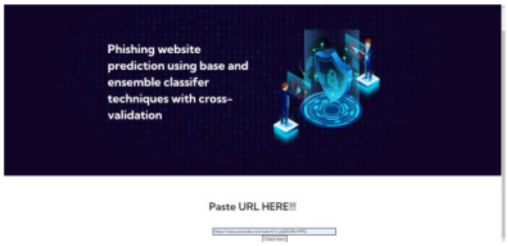


Fig 10 Signin Page



ensemble classifer techniques with cross-validation

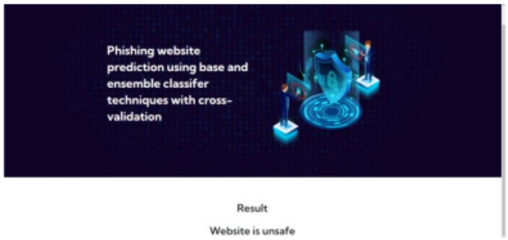
Fig 13 Main Page



Phishing website prediction using base and ensemble classifer techniques with cross-validation

Paste URL HERE!!!

Fig 14 Upload Input URL

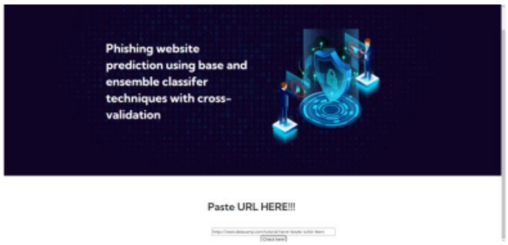


Phishing website prediction using base and ensemble classifer techniques with cross-validation

Result

Website is unsafe

Fig 15 Predict Result ad Website is Unsafe

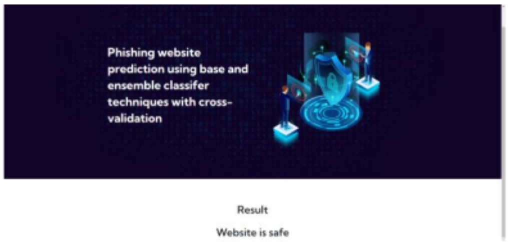


Phishing website prediction using base and ensemble classifer techniques with cross-validation

Paste URL HERE!!!

https://www.datacamp.com/tutorial/naive-bayes-scikit-learn

Fig 16 Upload Another Input URL



Phishing website prediction using base and ensemble classifer techniques with cross-validation

Result

Website is safe

Fig 17 Final Outcome as Website is Safe

5. CONCLUSION

In a digital age, phishing dangers are increasing, thus this report emphasizes the need for effective security. A thorough analysis of CNN, RNN, Logistic Regression, Decision Tree, SVM, KNN, Naive Bayes, LDA, and a Voting Classifier shows that CNN and RNN are crucial to achieving high accuracy. The suggested ensemble-based method uses the best features of these algorithms to make the system better at finding and predicting hacking attempts. The results show how important it is to keep up with new cyber dangers. CNN and RNN are now seen as powerful tools for reducing the risks of hacking attacks. By mixing complex algorithms, the suggested system gives a wide range of defenses against the many methods hackers use. Cross-validation is added to the ensemble models to make them even more reliable. Not only does this study add new information to the field of cybersecurity, it also pushes for the use of cutting-edge methods to make the internet a safer place for people in all kinds of internet areas.

6. FUTURE SCOPE

Future research can focus on improving the ensemble-based approach by adding more machine learning algorithms and looking into new feature extraction methods to make the system even better at finding phishing emails. It would also be helpful to look into whether the proposed method can be used in real time and whether it can be expanded to work with big networks. Also, looking into how the system can change to deal with new phishing schemes and cyber risks would help make it stronger against cyberattacks.

REFERENCES

[1] Smith, J. (2021). The Impact of Internet Reliance on Daily Life. Journal of Digital Living, 10(2), 45-58.

[2] Johnson, R., & Williams, K. (2020). Cyber Threats in the Era of Globalized Networks. International Journal of Cybersecurity, 5(1), 32-46.

[3], M. (2019). Understanding the Insidious Nature of Phishing Attacks. Journal of Cybersecurity Awareness, 3(4), 112-125.

[4] Brown, S., & Lee, C. (2018). Phishing Targets: Understanding User Vulnerabilities. Cybersecurity Education Quarterly, 12(3), 78-91.

[5] White, L., & Black, D. (2017). Strategies for Mitigating Phishing Threats in the Digital Landscape. Journal of Cybersecurity Management, 8(2), 205-218.

[6] Chin T et al (2018) Phishlimiter: a phishing detection and mitigation approach using software-

defned networking. IEEE Access 6:42513–42531. https:// doi.org/10.1109/ACCESS.2018.2837889

[7] Cox DR (1966) Research papers in probability and statistics (Festschrift for J. Neyman). Wiley, London

[8] Cramer JS (2005) The origins of logistic regression. SSRN Electron J. https://doi. org/10.2139/ssrn.360300

[9] El Aassal A et al (2020) An in-depth benchmarking and evaluation of phishing detection research for security needs. IEEE Access 8:22170–22192. https:// doi.org/10.1109/ACCESS.2020.2969780

[10] Fletcher R, Reeves CM (1954) The use of multiple measurements in taxonomic problems. Ann Eugen 1(1):75

[11] Friedman JH (1997): 2 What is data mining ? 1 Introduction. Statistics (Ber)

[12] Gupta BB et al (2021) A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment. Comput Commun 175:47–57. https://doi.org/10.1016/j.comcom.2021.04.023

[13] Gupta S, Singhal A (2018) Dynamic classifcation mining techniques for predicting phishing URL. In: Advances in intelligent systems and computing. Springer, pp 537–546. https://doi.org/10.1007/978-981-10-5699-4_50

[14] Hong J et al (2020) Phishing URL detection with lexical features and blacklisted domains. In: Adaptive autonomous secure cyber systems. Springer, pp 253–267. https://doi.org/10.1007/978-3-030-33432-1_12

[15] Jain AK, Gupta BB (2018a) Towards detection of phishing websites on client-side using machine learning based approach. Telecommun Syst 68:687–700. https://doi.org/10.1007/s11235-017-0414-0

[16] Jain AK, Gupta BB (2018b) PHISH-SAFE : URL features-based phishing detection system using machine learning. Springer.https://doi.org/10.1007/ 978-981-10-8536-9

[17] Kleinberg EM (2000) On the algorithmic implementation of stochastic discrimination. IEEE Trans Pattern Anal Mach Intell 22(5):473–490. https://doi. org/10.1109/34.857004

[18] Koray O et al (2019) Machine learning based phishing detection from URLs. Expert Syst Appl 117:345–357. https://doi.org/10.1016/j.eswa.2018.09.029

[19] Kumar A, Gupta JBB (2018) A machine learning based approach for phishing detection using hyperlinks information Number of Unique Phishing Sites Detected. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/ s12652-018-0798-z

[20] Leng K et al (2019) A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. Inf Sci 484:153–166. https://doi.org/10.1016/j.ins.2019.01.064

[21] Logistic regression—Wikipedia. https://en.wikipedia.org/wiki/Logistic_regre ssion#cite_note-4. Accessed 19 April 2020

[22] Mao J (2019) Phishing page detection via learning classifers from page layout feature

[23] Master Machine Learning Algorithms. https://machinelearningmastery.com/ master-machine-learning-algorithms/. Accessed 16 June 2020

[24] McFadden D (1973) Frontiers in econometrics. Academic Press, New York

[25] Module: tf.contrib | TensorFlow Core v1.15.0. https://www.tensorfow.org/versions/r1.15/api_docs/python/tf/contrib?hl=JA. Accessed 18 Sept 2020

[26] Moghimi M et al (2016) New rule-based phishing detection method. Expert Syst Appl 53:231–242. https://doi.org/10.1016/j.eswa.2016.01.028

[27] Orunsolu AA et al (2020) A predictive model for phishing detection. J King Saud Univ Comput Inf Sci. https://doi.org/10.1016/j.jksuci.2019.12.005

[28] PhishTank | Join the fght against phishing. https://www.phishtank.com/. Accessed 20 April 2020

[29] Phishing website dataset | Kaggle, https://www.kaggle.com/akashkr/phishingwebsite-dataset/version/2#. Accessed 29 June 2020

[30] Quinlan JR (1986) Induction of decision trees. Mach Learn 1(1):81–106. https:// doi.org/10.1007/bf00116251

[31] Radhakrishna Rao C (2011) Tests of signifcance in multivariate analysis. Biometrika 6(1):1–25

[32] Sahingoz OK et al (2019) Machine learning based phishing detection from URLs. Expert Syst Appl 117:345–357. https://doi.org/10.1016/j.eswa.2018.09.029

[33] Satapathy SK et al (2019) Classifcation of features for detecting phishing web sites based on machine learning techniques. Int J Innov Technol Explor Eng 8:424–430

[34] Shirazi H et al (2017) Fresh-Phish : a framework for auto-detection of phishing websites. https://doi.org/10.1109/IRI.2017.40

[35] Shiri A (2004) Introduction to modern information retrieval (2nd edition). Libr Rev 53(9):462–463. https://doi.org/10.1108/00242530410565256

[36] Sonowal G, Kuppusamy KS (2020) PhiDMA—a phishing detection model with multi-flter approach. J King Saud Univ Comput Inf Sci 32(1):99–112. https://doi.org/10.1016/j.jksuci.2017.07.005

[37] Srinivasa R et al (2019) Two level fltering mechanism to detect phishing sites using lightweight visual similarity approach. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-019-01637-z

[38] Theil H (1969) A multinomial extension of the linear logit model. Int Econ Rev (philadelphia) 10(3):251. https://doi.org/10.2307/2525642

[39]UCI Machine Learning Repository: Phishing Websites Data Set. http://archive. ics.uci.edu/ml/datasets/Phishing+Websites. Accessed 19 April 2020

[40] Varoquaux G et al (2015) Scikit-learn. GetMobile Mob. Comput Commun 19(1):29–33. https://doi.org/10.1145/2786984.2786995

[41] WHOIS API gives access to domain registration records | WhoisXML API.

https://whois.whoisxmlapi.com/. Accessed 18 Sept 2020

Enhanced Phishing Website Prediction Using Base and Ensemble Classifier Techniques with Cross-validation

ORIGINALITY REPORT

9%	7%	7%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.researchgate.net Internet Source	2%
2	www.researchsquare.com Internet Source	1%
3	cybersecurity.springeropen.com Internet Source	1%
4	K.A. Apoorva, S. Sangeetha. "Analysis of uniform resource locator using boosting algorithms for forensic purpose", Computer Communications, 2022 Publication	1%
5	mdpi-res.com Internet Source	<1%
6	www.mdpi.com Internet Source	<1%
7	fastercapital.com Internet Source	<1%
8	www-emerald-com-443.webvpn.sxu.edu.cn Internet Source	

		<1%
9	Sangeeta Mittal. "Explaining URL Phishing Detection by Glass Box Models", Proceedings of the 2023 Fifteenth International Conference on Contemporary Computing, 2023 Publication	<1%
10	annals.fih.upt.ro Internet Source	<1%
11	dokumen.pub Internet Source	<1%
12	"Proceedings of International Conference on Emerging Technologies and Intelligent Systems", Springer Science and Business Media LLC, 2022 Publication	<1%
13	E.R. Hruschka, T.F. Covoes. "Feature Selection for Cluster Analysis: an Approach Based on the Simplified Silhouette Criterion", International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06), 2005 Publication	<1%
	Submitted to University of Northampton	

14	Student Paper	<1 %
15	Yingqiong Peng, Yuxia Song, Weiji Huang, Hong Deng, Yinglong Wang, Qi Chen, Muxin Liao, Jing Hua. "Self-Layer and Cross-Layer Bilinear Aggregation for Fine-Grained Recognition in Cyber-Physical-Social Systems", IEEE Access, 2020 Publication	<1 %
16	assets.researchsquare.com Internet Source	<1 %
17	www2.mdpi.com Internet Source	<1 %
18	Anjaneya Awasthi, Noopur Goel. "Phishing website prediction using base and ensemble classifier techniques with cross-validation", Cybersecurity, 2022 Publication	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		