Phishing Website Detection Using Gan And Random Forest

Vedantika Jagtap[1], Vaishnavi Baraskar[2], VaibhaviGavali[3], BhaktiJadhav[4], Suvarna Sonavane[5]
[1]Assistant Professor, Dept of Computer Engineering
[2, 3, 4, 5]Dept of Computer Engineering
[1, 2, 3, 4, 5] SVPM's College of Engineering Malegaon(bk), Baramati

Abstract- *Cyber attackers create fake websites for various nefarious purposes, such as promoting their products, distributing malware, or stealing login credentials through a deceitful tactic known as phishing. Phishing involves impersonating legitimate entities via spoofed websites or emails to trick users into divulging sensitive information. Traditional methods for detecting these spoofed or phishing websites rely on fixed patterns, making them ineffective against newly created ones. To address this challenge, researchers are turning to machine learning and deep learning techniques.*

This paper proposes a method that utilizes a diverse range of robust features grouped into three main categories: webpage features, URL features, and HTML-based features. Initially, these features are individually employed to classify webpages. Subsequently, the paper suggests integrating all features to enhance classification accuracy.

As the Internet becomes increasingly integrated into our daily lives, it exposes us to a growing number of sophisticated security threats. Recognizing these threats, especially those that are novel and unseen before, is a critical challenge that requires immediate attention. Phishing site URLs are specifically designed to extract private information such as user identities, passwords, and financial details, underscoring the urgency of addressing this issue.The ways to recognize various network threats, specifically attacks notseen before, is a primary issue that needs to be looked into immediately. The aim of phishing site URLs is to collect the private information like user's identity, passwords and onlinemoney related exchanges.

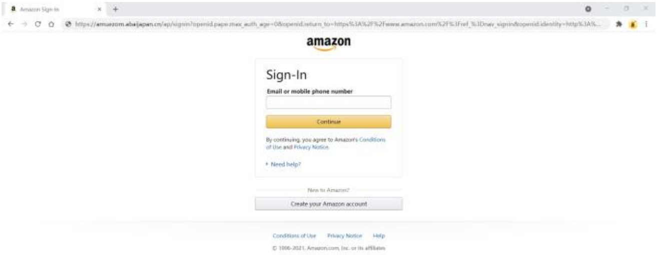
Keywords- Feature extraction; Deep learning; phishing detection; Spoofed websites.

I. INTRODUCTION

Due to the speedy growth of the Internet and its availability at low price, users have shifted from the traditional shopping to e-commerce. Attackers find their victims online by using some tricks instead of taking the risk of robbery. They use innovative techniques such as phishing to mislead the victims to make them to visit fake websites and collect their sensitive credentials. Phishing is a type of identity theft in which unsuspecting users are fooled to provide their valuable and sensitive information like credit card details, password and personal information on fraudulent/spoofed websites Phishing attack has gained serious attention from security researchers in recent years. Even for attackers, it is an area where they can show their skills in launching new deceitful webpages/websites, which look exactly similar to the popular and legal ones. The fake pages have similar graphical user interface (GUI), but different uniform resource locators (URLs) from the actual ones. Usually a careful and practiced user can easily detect these fake webpages by just looking at the URL. However, due to the busy life style, sometimes they do not care for the same and come into the trap of attackers.
Phishing attacks have become a significant concern owing to an increase in their numbers. It is one of the most widely used, effective, and destructive attacks, in which attackers try to trick users into revealing sensitive personal information, such as their passwords and credit card information. A typical phishing attack technique involves using a phishing website, where the attacker lures users to access fake websites by imitating the names and appearances of legitimate websites, such as eBay, Facebook, and Amazon.

As shown in Figure 1, it is difficult for the average person to distinguish phishing websites from normal websites because phishing websites appear similar to the websites they imitate. In many cases, users do not check the entire website URL, and, once they visit a phishing website, the attacker can access sensitive and personal information.

With the growth in the field of e-commerce, phishing attack and cybercrimes are rapidly growing. Attackers use websites, emails, and malware to conduct phishing attacks. According to the Anti-Phishing Working Group (APWG) Q4 2020 report, in 2020, there was an average of 225,759 phishing attacks per month, an increase of 220% compared to 2016 [1]. The country most affected by phishing sites is China, with 47.9% of machines infected. Phishing has become one of the biggest threats in cybersecurity. According to the FBI Internet Crime Center data records, the economic loss due to phishing crimes can reach $3.5 billion in 2019 [2].



https://amuezom.abaijapan.cn/ap/signin?openid.pape.max_auth_age=0&openid.return_to=https%3A%2F%2Fwww.amazon.com%2F%3Fref_%3Dnav_signin&openid.identity=http%3A%...

By continuing, you agree to Amazon's Conditions of Use and Privacy Notice.

© 1996-2021, Amazon.com, Inc. or its affiliates

Figure 1. Example of phishing website.

Phishing crimes are usually underreported. New phishing detection techniques have been developed to mitigate phishing attacks. A detailed review of the methodologies of various anti-phishing papers is given by Mohammad et al. [3]. Phishing website detection techniques are categorized into four types, whitelist/blacklist-based techniques, deep learning-based detection, machine learning-based detection, and heuristic-based detection techniques, as described in Figure 2.

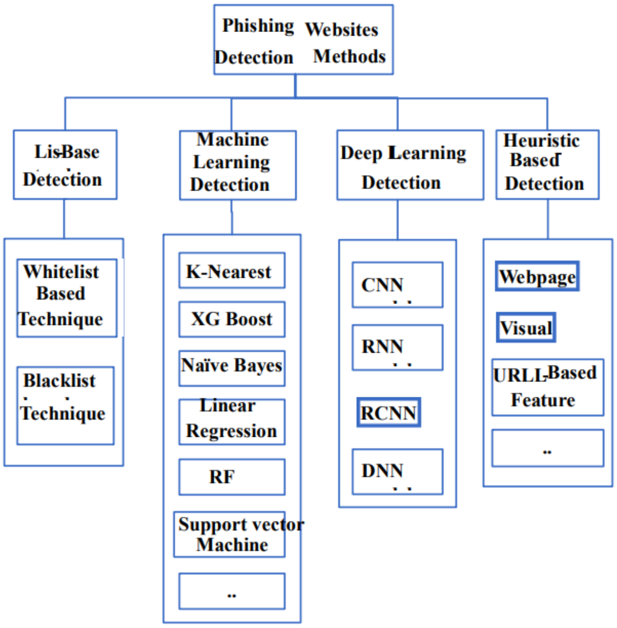


Figure 2. Category of phishing detection techniques.

It outputs phishing website detection results by aggregating the outputs of multiple classifiers using a winner-take-all strategy. The main advantages of the proposed method are listed as follows:

1.Strong generalization ability: The proposed method has strong generalization ability. The multi-level features used by the proposed method obtain better generalization ability and check-side accuracy. The low-level features in the hidden layer are common and similar for different but related distributed datasets or tasks; these are combined with the low-latitude features in the hidden layer.

2.Third-party service independence: The proposed method relies only on website URL features for detection, without extracting third-party features, such as page rank, search engine index, web traffic measurement, and domain age, which can improve the efficiency of detection and reduce the detection time.

3.Independence of cybersecurity experts: Reduced required expert function engineering, the deep neural network CNN model proposed in this paper can automatically extract URL features without the need for experts.

4.Language-independent: The approach proposed in this paper is effective for the detection of websites with content in various languages using character-level features. The main contributions of this paper are as follows.

1.This paper proposes a phishing website detection technique based on integrated learning and deep learning with fast and accurate detection of phishing websites using only URL features.
2.We built a real dataset by crawling 22,491 phishing URLs from phishtank and 24,719 legitimate URLs from Alex and conducted experiments on the dataset.
3.The phishing website detection process based on ensemble learning and deep learning is described, and the constructed dataset is extensively experimented. The results of the experiments indicate that our proposed method shows good performance in terms of accuracy and false positive rate.

The remainder of the paper is organized as follows: Section 2 introduces some problems related to phishing website detection, Section 3 introduces character embedding, CNN, RF, and the phishing website detection method proposed in this paper, Section 4 analyzes the experimental results of the proposed method, and Section 5 provides the conclusion and future scope of this work.

II. RELATEDWORK

A very effective detection of phishing website model which is focused on optimal feature selection technique and also based on neural network (OFS-NN) is proposed . In this proposed model, an index called feature validity value (FVV) has been generated to check the effects of all those features on the detection of such websites. Now, based on this newly generated index, an algorithm is developed to find from the phishing websites, the optimal features. This selected algorithm will be able to overcome the problem of over-fitting of the neural network to a great extent. These optimal features are then used to build an optimal classifier that detects phishing URLs by training the neural network.

A theory called Fuzzy Rough Set (FRS) was devised to a tool that finds the most appropriate features from a few standardized dataset. These features are then sent to a few classifiers for detection of phishing. To investigate the feature selection for FRS in building a generalized detection of phishing, the models by a different dataset of 14,000 website samples are trained. Feature engineering plays a vital role in finding solutions for detection of phishing websites, although the accuracy of the model greatly will be based on knowledge of the features. Though the features taken from all these various dimensions are understandable, the limitation lies in the time taken to collect these features.

This section presents a phishing website detection method based on character embedding, CNN, and RFs. The overall structure of the proposed method is shown in Figure 3. The phishing website detection method proposed in this paper consists of three main components. First, URL data is transformed into a character vector using the character embedding method. The converted URLs have the same data structure, which is beneficial for the detection of phishing websites. Second, an improved CNN network is designed, and the model is trained using the transformed URL data. After the model is trained, the URL features are extracted to obtain the features of different layers in the CNN network. Third, the features extracted from different network layers are classified in random forests separately. The classifier with the best classification result is used as the final classifier to classify the website.

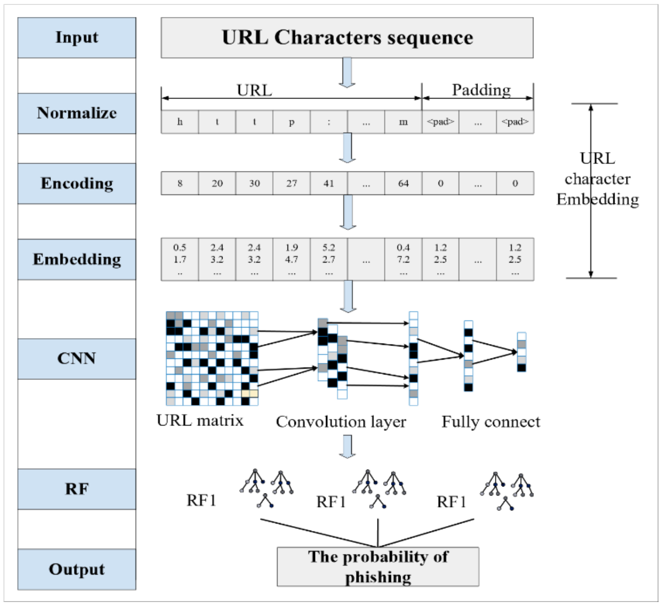


Figure 3. Framework of the proposed method.

SUMMARY OF LITERATURE SURVEY

Sr. No	Title	Technologies Used	Advantages	Disadvantages
1	A Deep Learning-Based Framework for Phishing Website Detection [1]	RNN , LSTM	deep learning models can significantly reduce the number of false positives, minimizing the inconvenience to users while maintaining high security.	Deep learning models typically require large amounts of labeled data for training. Collecting and annotating a comprehensive dataset of phishing websites can be time-consuming and costly.
2	OFS-NN: \An Effective Phishing Websites Detection Model Based on Optimal Feature Selection and Neural Network [2]	The New FVV Index Algorithm	The model can provide real-time or near real-time responses , Effective feature selection helps reduce the number of false positives,This improves the overall user experience.	Developing, training, and maintaining an effective phishing detection model can be costly, both in terms of hardware and personnel with the necessary expertise. Neural network models can be complex and challenging to design, train, and maintain.
3	Fuzzy Rough Set Feature Selection to Enhance Phishing Attack Detection [3]	Random Forest Algorithm , Multilayer perceptron ,	Fuzzy Rough Set Feature making datasets scalable for different phishing detection scenarios. , Fuzzy rough sets can handle uncertainty and noise in the data.	Overfitting can occur when feature selection is not carefully performed. , No Guarantee of Optimality.

Table1.: Summary of Literature Survey

III. METHODOLOGY

The proposed system for phishing website detection leverages the Gan model, which incorporates advanced natural language processing techniques and semantic analysis to extract meaningful features from textual content on websites. The Gan model is seamlessly integrated with the Random Forest algorithm, a robust ensemble learning method. In this execution, the Gyan model processes the textual content of websites to capture semantic nuances and linguistic patterns associated with phishing activities. The extracted features are then fed into the Random Forest algorithm, which utilizes a collection of decision trees to collectively assess the likelihood of a website being a phishing site based on the learned patterns. This integrated approach enhances the accuracy and reliability of phishing detection by combining the strengths of semantic analysis with the predictive power of Random

Forest, providing a comprehensive and effective solution for identifying and thwarting potential phishing threats.

GAN Algorithm:

A Generative Adversarial Network (GAN) is a deep learning architecture that consists of two neural networks competing against each other in a zero-sum game framework. The goal of GANs is to generate new, synthetic data that resembles some known data distribution. Generative Adversarial Networks (GANs) are a powerful class of neural networks that are used for unsupervised learning. It was developed and introduced by Ian J. Goodfellow in 2014. GANs are basically made up of a system of two competing neural network models which compete with each other and are able to analyze, capture and copy the variations within a dataset. It has been noticed most of the mainstream neural nets can be easily fooled into misclassifying things by adding only a small amount of noise into the original data. Surprisingly, the model after adding noise has higher confidence in the wrong prediction than when it predicted correctly. The reason for such an adversary is that most machine learning models learn from a limited amount of data, which is a huge drawback, as it is prone to overfitting. Also, the mapping between the input and the output is almost linear. Although, it may seem that the boundaries of separation between the various classes are linear, but in reality, they are composed of linearities, and even a small change in a point in the feature space might lead to the misclassification of data.

GAN Model :

Generative Adversarial Networks (GANs) can be broken down into three parts:

Generative: To learn a generative model, which describes how data is generated in terms of a probabilistic model.

Adversarial: The training of a model is done in an adversarial setting.

Networks: Use deep neural networks as artificial intelligence (AI) algorithms for training purposes.

In GANs, there is a Generator and a Discriminator. The Generator generates fake samples of data(be it an image, audio, etc.) and tries to fool the Discriminator. The Discriminator, on the other hand, tries to distinguish between the real and fake samples. The Generator and the Discriminator are both Neural Networks and they both run in competition with each other in the training phase. The steps are repeated several times and in this, the Generator and Discriminator get better and better in their respective jobs after each repetition. The work can be visualized by the diagram given below:

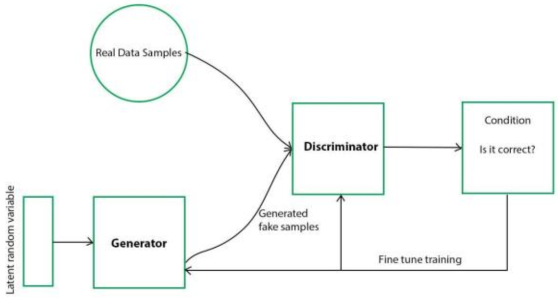


Figure 4: GAN Architecture

Phishing website detection using a Gan model integrated with Random Forest involves several steps. Below are the steps executed in sequence:

1. Data Collection:

- Gather a dataset that includes features relevant to website characteristics and potential phishing indicators.
- Ensure the dataset has labeled examples distinguishing between phishing and legitimate websites.

2. Data Preprocessing:

- Clean the dataset by handling missing values, removing irrelevant features, and converting categorical variables into a suitable format.
- Normalize or scale numerical features to ensure uniformity.

3. Feature Engineering:

- Extract meaningful features from the dataset.
- Consider features such as URL structure, domain information, SSL/TLS certificates, and content characteristics.

4. Train Gan Model:

- Utilize the Gan model (assuming it's a pre-trained model for semantic analysis or context understanding).
- Fine-tune or train the Gan model on a dataset that captures the semantic context of websites.

5. Extract Features from Gan Model:

- Apply the trained Gan model to the dataset to extract relevant semantic features for each website.

6. Combine Features:

- Combine the semantic features obtained from the Gan model with the original features from the dataset.

7. Train Random Forest Model:

- Split the dataset into training and testing sets.
- Train a Random Forest classifier using the combined features.

8. Model Evaluation:

- Evaluate the performance of the Random Forest model using metrics like accuracy, precision, recall, and F1 score on the testing set.

9. Integration:

- Integrate the Gan model and the trained Random Forest model into a cohesive system.
- Ensure seamless communication between the models for feature extraction and phishing detection.

10. Deployment:

- Deploy the integrated model in a suitable environment for real-time or batch processing.

11. Monitoring and Updates:

- Implement a monitoring system to track the model's performance over time.
- Periodically update the Gan model and retrain the Random Forest model to adapt to evolving phishing techniques.

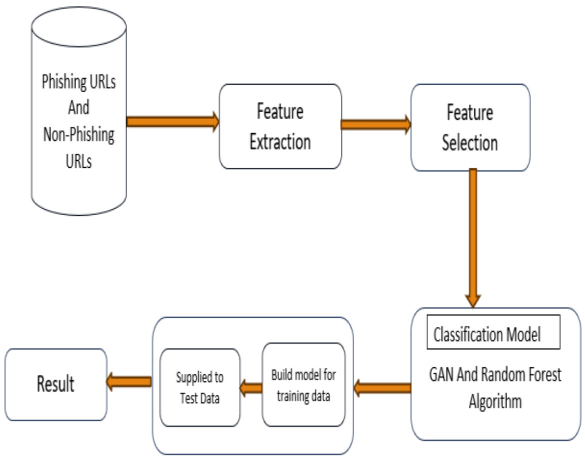

Phishing URLs And Non-Phishing URLs
Feature Extraction
Feature Selection
Classification Model
GAN And Random Forest Algorithm
Build model for training data
Supplied to Test Data
Result

Figure5*:Architecture of deep learning-based framework for detecting phishing URLs.*

The classification of phishing websites can be achieved using multi-level features to improve the accuracy and generalization ability of the classification algorithm. In this paper, multi-level URL features are extracted from the improved CNN network, as shown in Figure 6, and URL features are extracted using the pooling layer, L1 layer, and L3 layer. The aforementioned features are return (F1,F2,F3).

In order to extract more comprehensive URL feature information, high latitude features, mid-latitude features, and low latitude features are extracted separately. In addition, ensemble learning has a great impact on the performance improvement of the model and is widely used. So, three RFs are used to classify these features.

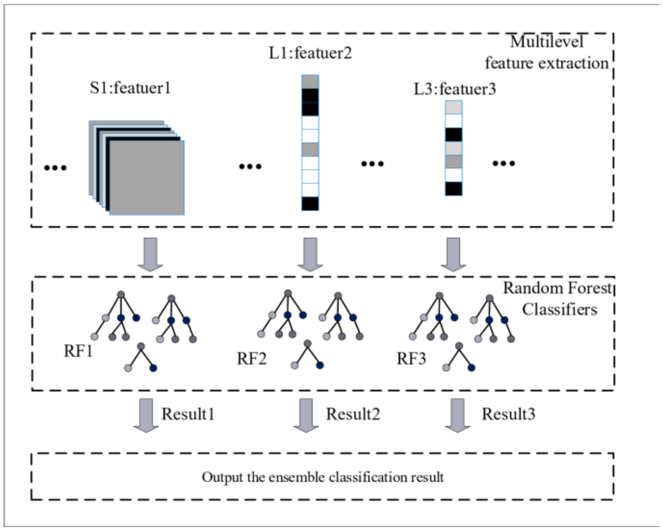
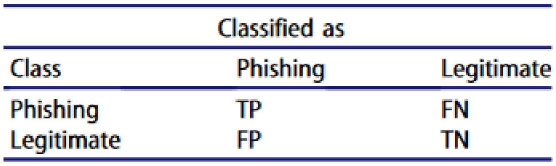

Multilevel feature extraction
S1:featuer1
L1:featuer2
L3:featuer3
Random Forest Classifiers
RF1
RF2
RF3
Result1
Result2
Result3
Output the ensemble classification result

Figure 6. Ensemble classifiers.

For different RF classifiers, features are extracted from different CNN network layers and used as training data for the RF classifiers. The results of each classifier are output after the training of the three RF classifiers is completed. The best RF classification result is used as the final classification result of phishing websites. Using this classification strategy of combining multiple classifiers can improve the accuracy and increase the generalization ability of the phishing website detection model. Using the max voting strategy, the output results are consistent between all ensemble classifiers and the base classifier. The best classification results are obtained in different layers.

Evaluation Parameters Used

For validating the proposed system, a k-fold cross validation has been used in the experiments conducted as it is well-accepted and the standard method to estimate likely predictions over unseen data. In our experiments, we have used 5-fold cross validation to increase the chance of learning all relevant information in the training set. In 5-fold cross validation, the original sample is randomly portioned into 5 sub-samples of equal size. Out of 5 sub-samples, 4 are used for training and remaining 1 is retained as the validation data for testing purpose. The process is then repeated 5 times with each of the 5 sub-samples used exactly once as the validation data. The results of these tests are averaged to obtain a measure of algorithm's performance over the entire dataset. The evaluation parameters used for comparing various classifiers are False Positive Rate (FPR), False Negative Rate (FNR), precision, F-Measure, percentage of overall Accuracy and Matthews Correlation Coefficient (MCC) (Gandotra, Bansal, and Sofat 2017). These are computed on the basis of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN)



	Classified as	
Class	Phishing	Legitimate
Phishing	TP	FN
Legitimate	FP	TN

Figure7:Classification of websites in Confusion matrix.

- **FPR:** It is the rate of incorrectly identified legitimate webpages.

$$FPR = \frac{FP}{FP + TN} \qquad (3)$$

- **FNR:** It is the rate of incorrectly identified phishing webpages.

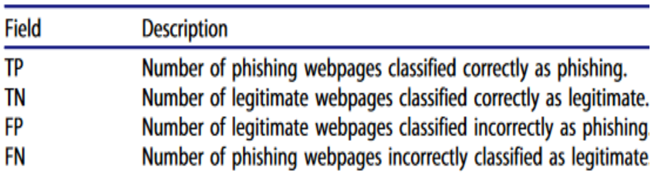
$$FNR = \frac{FN}{TP + FN} \qquad (4)$$

- **Precision**: It measures the exactness of a model. It is the probability for a true result to be classified correctly.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (5)$$

- **F-Measure**: It is the harmonic mean of precision and recall. It lies between 0 and 1, and provides a simple way to compare classifiers.

$$\text{F-Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 * \text{TP}}{2 * \text{TP} + \text{FN} + \text{FP}} \qquad (6)$$



Field	Description
TP	Number of phishing webpages classified correctly as phishing.
TN	Number of legitimate webpages classified correctly as legitimate.
FP	Number of legitimate webpages classified incorrectly as phishing
FN	Number of phishing webpages incorrectly classified as legitimate

Figure8: Fields of Confusion Matrix.

IV. RESULT AND ANALYSIS

This section will introduce the details of the model and analyze the experimental results. To evaluate the effectiveness of the proposed method in phishing detection, two phishing datasets were analyzed and studied separately.

The extracted URL features are fed into the three RF classifiers separately, and the training error curves are shown in Figure 9. It is important to note that the training errors of all three RF classifiers are close to zero, which indicates further that the feature mapping maps in the implicit layer also contain the significant information that contributes to the phishing websites detection results.

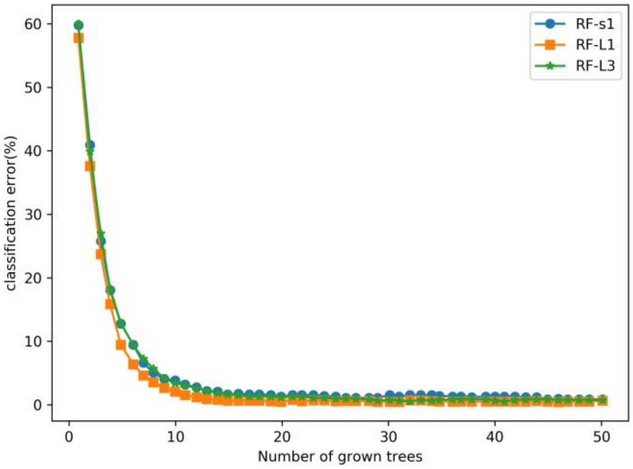
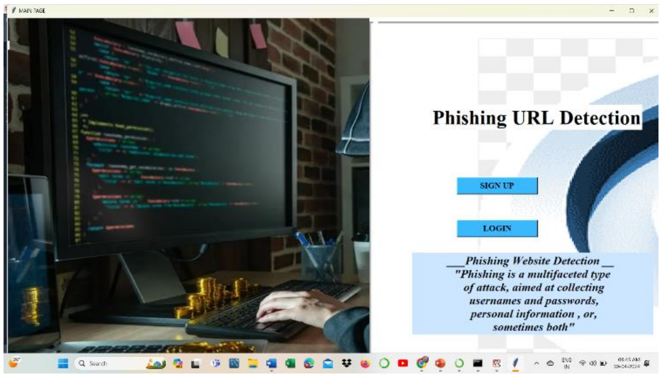
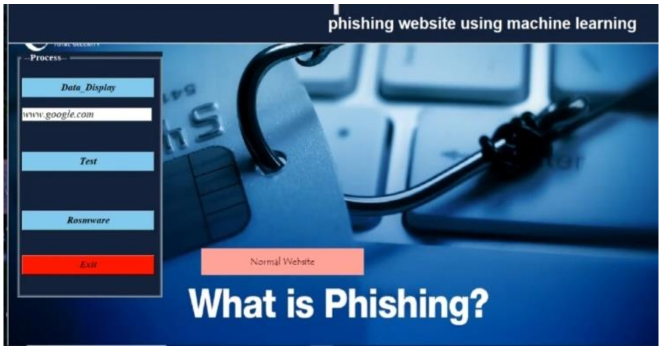


Figure 9. The training error curves of 3 RF classifiers.



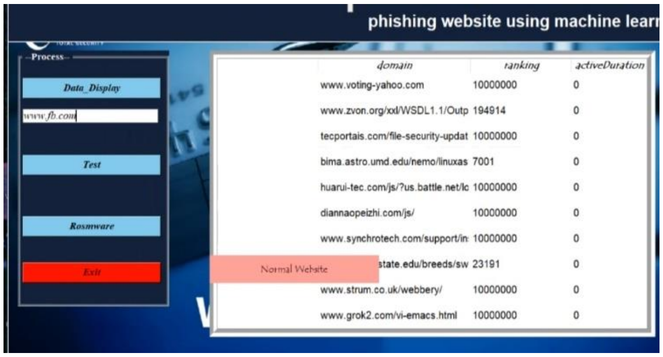
Phishing URL Detection

___Phishing Website Detection __
"Phishing is a multifaceted type of attack, aimed at collecting usernames and passwords, personal information , or, sometimes both"

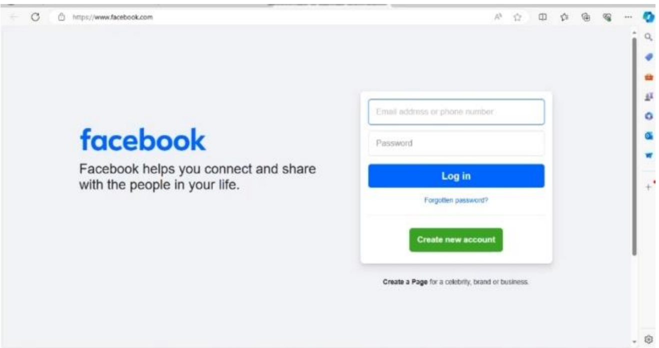


www.google.com

Normal Website



domain	ranking	activeDuration
www.voting-yahoo.com	10000000	0
www.zvon.org/xxl/WSDL1.1/Outp	194914	0
tecportais.com/file-security-updat	10000000	0
bima.astro.umd.edu/nemo/linuxas	7001	0
huarui-tec.com/js/?us.battle.net/lc	10000000	0
diannaopeizhi.com/js/	10000000	0
www.synchrotech.com/support/in	10000000	0
state.edu/breeds/sw	23191	0
www.strum.co.uk/webbery/	10000000	0
www.grok2.com/vi-emacs.html	10000000	0



Facebook helps you connect and share with the people in your life.

Create a Page for a celebrity, brand or business.

V. CONCLUSION AND FUTURE WORK

The paper presented a machine learning based model for detecting phishing websites. The proposed model makes use of a set of features which are categorized into different categories, i.e., Page, URL and HTML based features. The performance of each category of features is evaluated andcompared with the proposed approach which uses integration of all features belonging to three categories. The experimental results show that the features under URL based category are most effective in classifying the webpages. The performance results also report a significant improvement when an integration of features is considered. Nowadays, the attackers are getting more sophisticated by using advanced and novel phishing techniques. Consequently, it is pertinent to develop more robust and effective anti-phishing approaches to handle complex phishing attacks. In future, we intend to focus on the extended feature set and deep learning algorithms for improving the classification accuracy of detecting phishing webpages at large scale.

In this paper, we proposed a multi-level feature phishing website classification method based on character embedding CNN and RF. The main features of this model is as follows.

1
)Character embedding of URLs is performed to convert URLs into normalized matrices, containing much important phishing website classification information in the URL characters. This information helps classify phishing websites. URLs are transformed into uniform signals by the character embedding technique, more suitable for CNN networks' input.

2)Automatic phishing web feature extractor using CNN. The CNN model is pre-trained using the converted URL data to optimize and improve the CNN model parameters. The pre-trained model can extract multi-level features from the URL data. The extracted multi-level features contain sensitive information that can classify phishing websites and provide knowledge for phishing website classification.

3)Using multiple RF classifiers and a winner-take-all strategy improves the model's accuracy and generalization. Extracting multi-level features for low latitude can be used to classify phishing websites. The RF classifier is trained using the extracted features of each layer, outputting the results of each RF, and, finally, choosing the one with the best results, improving the classification results.

4)The proposed method in this paper is validated by the dataset from PhishTank and Alex. A 99.35% correct classification rate of phishing websites was obtained on the dataset. Experiments were conducted on the test set and training set, and the experimental results proved that the proposed method has good generalization ability and is useful in practical applications.

Although the proposed method in this paper has achieved some good results, there are still some shortcomings. The main disadvantage is that it takes longer to train. However, the trained model is better than the others in terms of accuracy of phishing website detection. Another disadvantage is that the model cannot determine whether the URL is active or not, so it is necessary to test whether the URL is active or not before detection to ensure the effectiveness of detection. In addition, some attackers use URLs that are not imitations of other websites, and such URLs will not be detected. The next step of our work aims to use new techniques to automatically extract other features for detecting phishing sites, such as web code features, web text features, and web icon features

IV. ACKNOWLEDGMENT

It gives us great pleasure in presenting the Report on " Phishing Website Detection Using GAN and Random forest ". We would like to express our special thanks to our project guide V. G .Jagtap.I am also greateful to Dr. Y. D. Sinkar Head of Computer Engineering Department SVPMCOE for his indepensablesupport , suggestions and which also helped us to know about somany new things we are really thankful to them.we would like to thanks our Principle Dr. S. M. Mukane , College of Engineering , Malegoan(BK) who gave us the golden opportunity to work on this wonderful project.

REFERENCES

[1] L. Tang and Q. H. Mahmoud, ‘‘A survey of machine learning-based solutions for phishing website detection,’’ Mach. Learn. Knowl. Extraction, vol. 3, no. 3, pp. 672–694, Aug. 2021, doi: 10.3390/make3030034.

[2] S. Marchal, J. Francois, R. State, and T. Engel, ‘‘PhishStorm: Detecting phishing with streaming analytics,’’ IEEE Trans. Netw. Service Manage., vol. 11, no. 4, pp. 458–471, Dec. 2014.

[3] (Jun. 2021). Phishing Activity Trends Report 1st Quarter 2021. APWG. Accessed: Oct. 20, 2021. [Online]. Available: https://docs.apwg. org/reports/apwg_trends_report_q1_2021.pdf.

[4] (2020). 2020 Internet Crime Report. [Online]. Available:https://www.ic3.gov/Media/PDF/AnnualReport /2020_IC3Report.pdf

[5] R. M. Mohammad, F. Thabtah, and L. McCluskey, ‘‘Predicting phishing websites based on self-structuring neural network,’’ Neural Comput. Appl., vol. 25, no. 2, pp. 443–458, Nov. 2013, doi: 10.1007/ s00521-013-1490-z.

[6] M. A. El-Rashidy, ‘‘A smart model for web phishing detection based on new proposed feature selection technique,’’ Menoufia J. Electron. Eng. Res., vol. 30, no. 1, pp. 97–104, Jan. 2021, doi: 10.21608/ mjeer.2021.146286.

[7] B. B. Gupta, K. Yadav, I. Razzak, K. Psannis, A. Castiglione, and X. Chang, ‘‘A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment,’’ Comput. Commun., vol. 175, pp. 47–57, Jul. 2021, doi: 10.1016/j.comcom.2021.04.023.

[8] E. Gandotra and D. Gupta, ‘‘Improving spoofed website detection using machine learning,’’ Cybern. Syst., vol. 52, no. 2, pp. 169–190, Oct. 2020, doi: 10.1080/01969722.2020.1826659.

[9] W. Wang, F. Zhang, X. Luo, and S. Zhang, ‘‘PDRCNN: Precise phishing detection with recurrent convolutional neural networks,’’ Secur. Commun. Netw., vol. 2019, pp. 1–15, Oct. 2019, doi: 10.1155/2019/ 2595794.

[10] M. Sabahno and F. Safara, ‘‘ISHO: Improved spotted hyena optimization algorithm for phishing website detection,’’ Multimedia Tools Appl., Mar. 2021, doi: 10.1007/s11042-021-10678-6.