DETECTION OF PHISHING WEBSITES USING MACHINE LEARNING

R. Karthikeyan[1], C. Abirami[2] ,R. Ramya[3], B. Sneha[4],U. Uma Azhagu Sudha[5]

12345*Dept. of Computer Science and Engineering, Government College of Engineering Srirangam, TamilNadu, India*

Abstract - *Phishing is a kind of worldwide spread cyber crime that uses disguised websites to trick users into downloading malware or providing personally sensitive information to attackers. With the rapid development of artificial intelligence, researchers in the cyber security field utilize machine learning algorithms to classify phishing websites. To track the phishing websites the machine learning algorithm Random Forest is used because it gives high accuracy as compared to other machine learning algorithms. The dataset is collected which contains both the malicious and legitimate url to train Machine Learning models using Random Forest Classifier and predict the websites in order to identify and prevent users from falling victim to online scams.*

1.INTRODUCTION

In today's digital world, where the internet is an integral part of our daily lives, online security has become an important concern. Phishing is a form of cybercrime, poses a significant threat to individuals, businesses, and organizations worldwide. Phishing attacks involve fraudulent attempts to obtain sensitive information such as usernames, passwords, and financial data by masquerading as a trustworthy entity in electronic communication.

The consequences of falling victim to phishing can be severe, ranging from financial loss and identity theft to compromised personal and corporate data. With the advancement of technology, phishing techniques have become increasingly sophisticated, making it more challenging to distinguish between legitimate and malicious websites.

Therefore, the ability to detect phishing websites accurately and efficiently is crucial in safeguarding against cyber threats. This detection process involves analyzing the lexical features and characteristics of websites to identify indicators of phishing or other malicious activity.

In this paper, we analyze into the methodologies and techniques employed in the detection of phishing websites. By understanding the strategies utilized by cyber criminals and leveraging innovative detection techniques, we aim to empower users and organizations to mitigate the risks posed by phishing attacks.

2. RELATED WORKS

[1] Researchers have investigated different feature sets and selection techniques to improve the performance of phishing detection systems. Studies have explored the importance of features such as URL characteristics, domain registration information, website content analysis, SSL certificate attributes, and user-related factors in distinguishing between phishing and legitimate websites. Feature selection methods, including information gain, chi-square test, and recursive feature elimination, have been employed to identify the most discriminating features for classification.[2]Some research focuses on analyzing the dynamic behavior of website interactions, including mouse movements, keystrokes, and navigation patterns, to detect phishing attempts in real-time. These approaches aim to capture subtle indicators of malicious intent that may not be apparent from static website features alone, enhancing the adaptability and responsiveness of detection systems to evolving phishing tactics.[3]Hybrid and multi-modal detection systems integrate multiple sources of information, such as website content, network traffic, user behavior, and reputation data, to improve the comprehensiveness and effectiveness of phishing detection. These approaches combine the strengths of different detection techniques and data sources to enhance detection accuracy and resilience to evasion strategies employed by cyber criminals.[4]With the growing sophistication of phishing attacks, researchers have explored adversarial machine learning techniques to enhance the robustness of detection systems against evasion attempts. Adversarial training, robust feature representation learning, and generative adversarial networks (GANs) are among the approaches investigated to defend against adversarial manipulation and stealthy phishing tactics.

By building upon and extending the insights gained from related works in the field, the proposed system aims to advance the state-of-the-art in phishing website detection, addressing key challenges and limitations to enhance cybersecurity resilience in the face of evolving threats.

3. PROPOSED SYSTEM

In the Proposed system, the ensemble learning method Random Forest is used to train the model and predict the websites whether it is malicious or legitimate website. The dataset used contains more records around 6,51,191 URLs, out of which 4,28,103 are safe URLs, 96,457 are defacement URLS, 94,111 are phishing URLS and 32,520 are malware URLS. There are 26 lexical features are extracted from the urls in the dataset to build the model. The features includes check for Having ip address, Abnormal url, Google index, Count of dot, Count of www, Count of @, Count of directory, Count embedded domain, Suspicious words in URL, Short url, Count of https, Count of http, Count of %, Count of ?, Count of -, Count of =, Url length, Hostname length, Firstdirectory length, Length of top-level domains, Count of digits, Count of letters, Count the number of .com, Average length of word in hostname, ratio of digits in hostname. Not only phishing URLS are predicted and other Malicious URLs are detected. The System Architecture diagram of our proposed system is given in fig.1

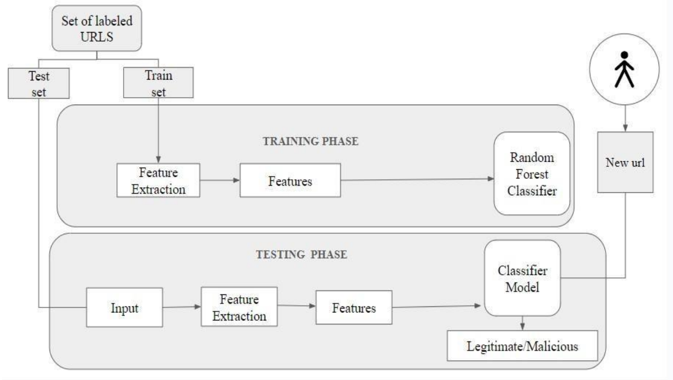


Figure-1: System Architecture Diagram

4. METHODOLOGY

4.1 Dataset Collection

The dataset is collected from the Kaggle website. It contains a total of 6,51,191 URLs, out of which 4,28,103 are safe URLs, 96,457 are defacement URLS, 94,111 are phishing URLS and 32,520 are malware URLS.

4.2 Data Preprocessing

Data preprocessing is the first and crucial step after data collection. The raw dataset obtained for phishing detection was prepared by removing redundant and irregular data. The dataset contains some irrelevant columns such as checking websites, date and unwanted type of target variable. These are rectified and make the dataset suitable for building the model.

4.3 Feature Extraction

Feature extraction is the process of transforming raw data into numerical features that preserve the information in the original data set. It's a way of identifying and selecting the most important information or characteristics from a data set, while filtering out less significant details. In this step, the features are extracted from the raw URLs, as these features will be used as the input features for training the machine learning model. The features are Having ip address, Abnormal url, Google index, Count of dot, Count of www, Count of @, Count of directory, Count embed domain, Suspicious words in URL, Short url, Count of https, Count of http, Count of %, Count of ?, Count of -, Count of =, URL length, Hostname length, First directory length, Length of top-level domains, Count of digits, Count of letters.

4.4 Exploratory data analysis

The data visualization method was employed to analyze, explore and summarize the dataset. These visualizations consist of histograms, box plots, scatter plots, catplot and count plots to uncover patterns and insights within data. The google index feature denotes if the URL is indexed in google search console or not. As a result of this step, the google index features were dropped because In this dataset, all the URLs are google indexed and have a value of 1.

4.5 Model Building and Evaluation

The model is developed using tree-based ensemble machine learning method include Random Forest. Model Evaluation involves assessing the performance of a trained model on unseen data to understand how well it generalizes to new, unseen urls. The model is evaluated based on the performance metrics include accuracy, precision, recall and F1 score.

4.6 Prediction

After the model has been trained and evaluated, it can be used to make predictions on new, unseen websites. When presented with the features of a website, the model applies the decision rules learned during training to classify the website as either phishing or legitimate. In the context of Random Forests, prediction involves aggregating the predictions of multiple decision trees to arrive at a final classification. Each tree in the forest independently classifies the input data, and the final prediction is determined by a majority vote of the individual tree predictions.

5. MACHINE LEARNING

Machine Learning (ML) is a branch of artificial intelligence (AI) where computer systems are trained to learn patterns and make decisions based on data without explicit programming instructions and accurately process large volumes of data, generating insights and predictions with minimal human intervention. ML enables organizations to streamline decision-making processes, improve productivity, and achieve better outcomes across various domains. ML includes many techniques that allow software applications to improve their performance as time progresses. It requires understanding mathematical and statistical concepts to select appropriate algorithms and training them with sufficient data to achieve accurate results. Prediction techniques leveraging machine learning involve the application of computational algorithms across various industries to anticipate future outcomes, trends, and patterns based on historical data analysis.

Machine learning for phishing detection is like teaching a computer to recognize the difference between real and fake websites by showing it lots of examples and letting it learn from them.

Ensemble learning is a machine learning technique that combines multiple base models to produce a stronger predictive model. Instead of relying on the output of a single model, ensemble methods aggregate the predictions of multiple models to make more accurate and robust predictions. Ensemble learning can be applied to both classification and regression tasks, and it is particularly effective when individual base models have different strengths and weaknesses.

5.1 RANDOM FOREST ALGORITHM

Random Forest belongs to the family of ensemble learning methods. Ensemble learning is a machine learning technique that combines multiple base models to produce a stronger predictive model. Instead of relying on the output of a single model, ensemble methods aggregate the predictions of multiple models to make more accurate and robust predictions. Ensemble learning can be applied to both classification and regression tasks, and it is particularly effective when individual base models have different strengths and weaknesses.

Random Forest works by creating a number of Decision Trees during the training phase. Each tree is constructed using a random subset of the data set to measure a random subset of features in each partition. This randomness introduces variability among individual trees, reducing the risk of overfitting and improving overall prediction performance. In prediction, the algorithm aggregates the results of all trees by voting classification tasks. This collaborative decision-making process, supported by multiple trees with their insights, provides an example of stable and precise results. Random forests are widely used for classification and regression functions, which are known for their ability to handle complex data, reduce overfitting, and provide reliable forecasts in different environments. Random Forest tends to achieve high accuracy in classification tasks, even with relatively little tuning. Its ability to handle high-dimensional data and capture complex relationships makes it well-suited for a wide range of applications. By aggregating the predictions of multiple decision trees trained on random subsets of data and features, Random Forest mitigates the risk of overfitting and generalizes well to unseen data.

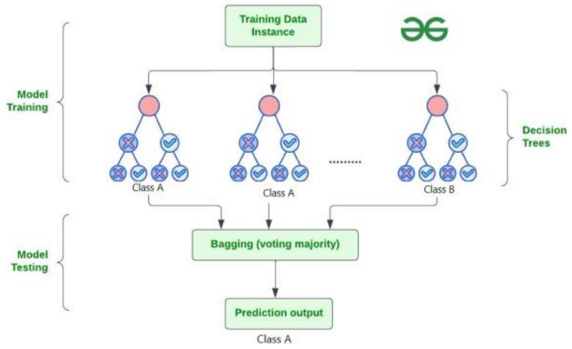


Figure-2: Algorithm

5.1.1 Ensemble of Decision Trees:

Random Forest gives the power of ensemble learning by constructing an army of Decision Trees. These trees are like individual experts, each specializing in a particular aspect of the data. Importantly, they operate independently, minimizing the risk of the model being overly influenced by the nuances of a single tree.

5.1.2 Random Feature Selection:

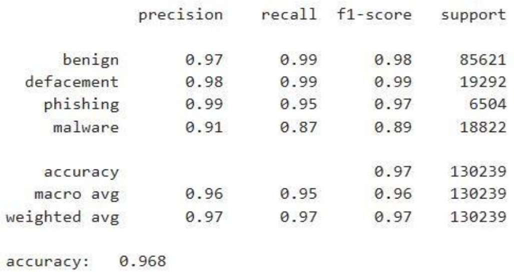
To ensure that each decision tree in the ensemble brings a unique perspective, Random Forest employs random feature selection. During the training of each tree, a random subset of features is chosen. This randomness ensures that each tree focuses on different aspects of the data, fostering a diverse set of predictors within the ensemble.

5.1.3 Bootstrap Aggregating or Bagging:

The technique of bagging is a Random Forest training strategy which involves creating multiple bootstrap samples from the original dataset, allowing instances to be sampled with replacement. This results in different subsets of data for each decision tree, introducing variability in the training process and making the model more robust.

5.1.4 Decision Making and Voting:

When it comes to making predictions, each decision tree in the Random Forest casts its vote. For classification tasks, the final prediction is determined by the mode most frequent prediction across all the trees. This internal voting mechanism ensures a balanced and collective decision-making process.



```
              precision    recall  f1-score   support

      benign       0.97      0.99      0.98     85621
 defacement       0.98      0.99      0.99     19292
    phishing       0.99      0.95      0.97      6504
     malware       0.91      0.87      0.89     18822

    accuracy                           0.97    130239
   macro avg       0.96      0.95      0.96    130239
weighted avg       0.97      0.97      0.97    130239

accuracy:   0.968
```

Figure-3: Accuracy of Random Forest Algorithm

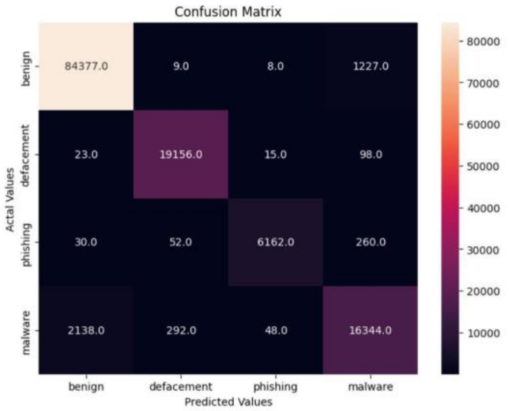


Figure-4: Confusion Matrix

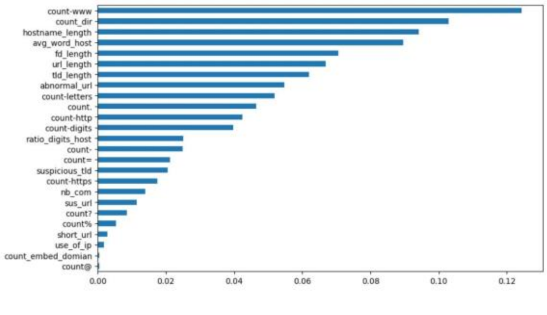


Figure-5: Feature Importance given by Random Forset

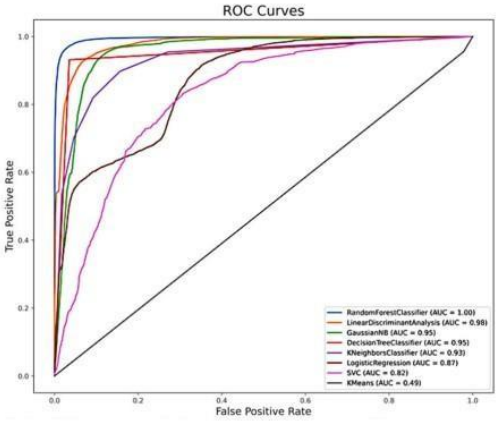


Figure-6: ROC Curve of Comparing Eight other Machine Learning Algorithm

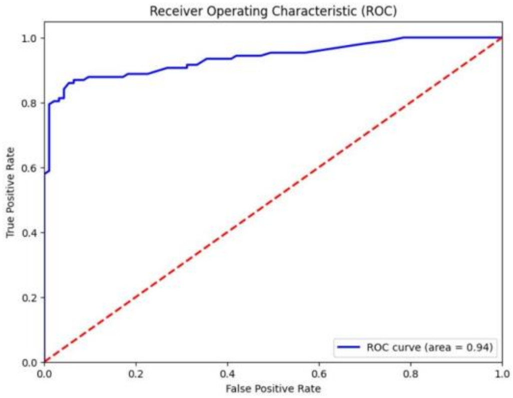


Figure-7: ROC Curve of Random Forest

6. CONCLUSIONS

We presented a comprehensive approach for detecting phishing websites using machine learning techniques, with a focus on the Random Forest algorithm. Through extensive experimentation and analysis, we have demonstrated the effectiveness of our proposed methodology in accurately identifying phishing websites and mitigating potential risks to users' online security. Our study involved the collection of a diverse dataset comprising both legitimate and phishing websites, followed by feature extraction and model training using the Random Forest algorithm. We evaluated the performance of our detection model using standard evaluation metrics and conducted cross-validation to ensure its robustness and generalization capability. The results of our experiments indicate that our Random Forestbased approach achieves high levels of accuracy, precision, recall, and F1-score in distinguishing between legitimate and phishing websites. Furthermore, comparative analysis with other machine learning algorithms highlights the superiority of our proposed methodology in terms of detection performance and efficiency. Overall, our research contributes to the advancement of web security by offering a practical and effective solution for detecting phishing websites using machine learning techniques. By leveraging the power of the Random Forest algorithm and feature engineering strategies, we have developed a robust detection system capable of effectively identifying and mitigating phishing threats in realworld scenarios.

7. REFERENCES

[1] Andrew J. Park, Ruhi Naaz Qadri, and Herbert H. Tsang developed a "Phishing website detection framework through web scraping and data mining", in 2017.

[2] Chen, Kai, et al. "PhishDetect: Real-Time Phishing Detection Using Machine Learning and URL Analysis." IEEE Transactions on Dependable and Secure Computing in 2021.

[3] K. L. Chiew, C. L. Tan, K. Wong, K. S. C. Yong, andW. K. Tiong, ‘‘A new hybrid ensemble feature selection framework for machine learning-based phishing detection system in May 2019.

[4] C. Emilin Shyni, Anesh D. Sundar, and G.S. Edwin Ebby proposed "Phishing detection in websites using parse tree validation", in 2018.

[5] Gao, Xin, et al. "Enhanced Phishing Website Detection Using Machine Learning and Feature Engineering." IEEE Access 8 in 2020.

[6] Gupta, Ananya, and Rajesh Kumar. "Phishing Detection using Ensemble Learning and Feature Fusion." International Journal of Computer Applications 10.7 in 2020.

[7] Jhen-Hao Li and Sheng-De Hang proposed "PhishBox-An approach for phishing validation and detection", in 2018.

[8] Li, Wei, et al. "Phishing Website Detection Using Feature Fusion and Deep Learning." Proceedings of the ACM Conference on Computer and Communications Security (CCS). 2020.

[9] U. Ozker and O. K. Sahingoz, ‘‘Content based phishing detection with machine learning,’’ in Proc.Int. Conf. Electr. Eng. (ICEE), Sep. 2020.

[10] Patel, Priya, and Arjun Sharma. "Web Phishing Detection using Machine Learning and Natural Language Processing Techniques." Proceedings of the IEEE International Conference on Cybersecurity and Privacy (ICCSP). 2021.

[11] Patel, Sanjay, and Ritu Gupta. "A Novel Approach to Phishing Website Detection Using Neural Networks and Behavioral Analysis." International Journal of Information Security in 2021.

[12] Peng Yang, Peng Zeng, and Guangzhen Zhao introduced "Phishing website detection method based on multidimensional features driven by deep learning", in 2019.

[13] V. Shahrivari, M. M. Darabi,and M. Izadi, ‘‘Phishing detection using machine learning techniques in 2020.

[14] Singh, Rahul, and Priya Sharma. "Detecting Phishing Websites Using Machine Learning and Feature Engineering." Journal of Cybersecurity Research in 2020.

[15] S. Wang, S. Khan, C. Xu, S. Nazir, and A. Hafeez, ‘‘Deep Learning-based efficient model development for phishing detection using random forest and BLSTM classifiers,’’ Complexity,Sep. 2020.

[16] Wang, Xiaoyong, et al. "Deep Learning for Phishing Website Detection." IEEE Transactions on Information Forensics and Security in 2019.

[17] Yongjie Huang, Qiping Yang, Jinghul Qin, and Wu Shao Wen developed a method for "Phishing URL detection via CNN and attentionbased hierarchical RNN", in 2019.