Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.4, May : 2024

ENHANCING CYBERSECURITY: MACHINE LEARNINGAPPROACHES TO PHISHING DETECTION

V.V.R Manoj *Assistant Professor, Dept. of CSE Andhra Loyola Institute of Engineering and Technology*,Vijayawada, Andhra Pradesh, India manoj.professorcse@gmail.com
Srija Talamarla *Computer Science and Engineering Andhra Loyola Institute of Engineering and Technology*,Vijayawada, Andhra Pradesh, India talamarlasrija@gmail.com
Navya Sri Oleti *Computer Science and Engineering Andhra Loyola Institute of Engineering and Technology*,Vijayawada, Andhra Pradesh, India onavyasri@gmail.com
Venu Madhavi Goli *Computer Science and Engineering Andhra Loyola Institute of Engineering and Technology*,Vijayawada, Andhra Pradesh, India venumadhavigoli7@gmail.com

Abstract—
Phishing attacks are becoming more common in the digital era, presenting significant threats to security also placing confidential information at risk globally. Phishing is a type of cyberattack that involves misleading users into visiting fake websites that appear legitimate, aiming to steal the user's personal data and financial account credentials. Phishers create URLs for websites that resemble legitimate ones but can be identified by keen observation. Our project explores the application of machine learning techniques to develop an effective phishing website detection system, focusing on the automated analysis of website URLs. By leveraging only URL data, our method eliminates the need to visit potentially malicious sites, reducing the risk of exposure to harmful content through making use of the metadata characteristics of URLs, including domain information and lexical features. The Random Forest algorithm, known for its resistance to overfitting, plays a pivotal role in our model, providing a robust framework for classification based on extracted features.
Index Terms—Phishing, Machine Learning, Random Forest Algorithm, URL, Feature extraction.

INTRODUCTION

Phishing is one of the most common forms of cyberattacks and continues to be of concern to individuals, companies, and even governments. Phishing is a cybercrime involving Social engineering techniques which includes deceiving consumers into believing that they are accessing information from legitimate sources, resulting in the disclosure of sensitive information like personal financial information, usernames, and passwords [14]. Email phishing, SMS phishing (smishing), voice phishing (vishing), and fraudulent websites are some of the several types of phishing attacks. [1] These attacks can have severe repercussions including money loss, identity theft to compromised networks and data breaches. Given the pervasive nature of phishing threats, effective detection mechanisms are essential to safeguarding individuals, businesses, and critical infrastructure. [2]

Our study aims to investigate the field of identifying counterfeit websites among different types of phishing attacks. For detecting these fake websites, traditional methods like blacklisting known malicious websites, email filters, heuristic rule-based approaches [5] [7] struggle to keep pace with the evolving tactics employed by cybercriminals, who continuously refine their techniques to evade detection. Machine learning approaches capable of identifying complex patterns can be employed in order to improve conventional phishing website detection systems. Machine learning, a subset of artificial intelligence that enables systems to learn from data and improve over time without explicit programming, holds promise for enhancing cybersecurity defences. By analysing vast datasets containing examples of legitimate and fraudulent communications, machine learning algorithms can identify minute patterns and indicators indicative of phishing attempts, enabling automated detection with high accuracy and efficiency. Supervised learning algorithms, such as decision trees, random forests, support vector machines (SVM), and neural networks, [1] can be trained on labeled datasets containing examples of phishing and legitimate websites without the need to open the website. Through iterative training and validation, these algorithms learn to distinguish between genuine and fraudulent

messages, thereby improving their ability to detect phishing attempts in real-time.

In summary, the convergence of machine learning and cybersecurity represents a paradigm shift in the fight against phishing and other cyber threats [3]. By harnessing the power of data-driven algorithms and predictive analytics, organizations can strengthen their defences, safeguarding sensitive information and preserving the integrity of digital ecosystems.

The remaining sections in this paper are organized as follows: the next section talks about the literature survey. Section 3 describes our proposed system. It discusses ourmethodology and its details. Also, we discuss the classification algorithm used for the classification between phishing and legitimate URLs. The outcomes produced by the model are covered in Section 4. Finally, Section 5 includes conclusion.

LITERATURE SURVEY

The issue of cybersecurity has become increasingly critical in today's digital age, with phishing attacks posing a significant threat to individuals, businesses, and organizations worldwide. Phishing attacks involve fraudulent attempts to obtain sensitive information, such as passwords, financial details, and personal data, by masquerading as trustworthy entities. To combat this pervasive threat, researchers and practitioners have turned to machine learning (ML) approaches for phishing detection. ML techniques offer a data-driven approach to identifying and mitigating phishing threats, leveraging algorithms to analyze patterns and characteristics indicative of fraudulentbehavior. This literature survey explores existing research in the field of ML-based phishing detection, focusing on key studies and methodologies employed to enhance cybersecurity, particularly in the context of developing countries like India.

Christopher N. Gutierrez et al. (2020) explored the detection of new forms of phishing attacks, emphasizing the importance of continuous learning and adaptation in cybersecurity systems. Their study highlighted the role of ML algorithmsin identifying emerging threats and improving the resilience of phishing detection systems. Overall, the literature survey highlights the growing interest in ML-based approaches to phishing detection and their potential to enhance cybersecurity in developing countries like India. By leveraging advanced algorithms and techniques, researchers and practitioners can develop more effective and scalable solutions to combat the evolving threat landscape posed by phishing attacks.

PROPOSED SYSTEM

The proposed system aims to enhance cybersecurity measures, specifically focusing on the detection of phishing attacks. Through the utilisation of machine learning techniques, particularly the Random Forest algorithm, this system presents a novel strategy for countering phishing scams. Fundamentally, the system utilises a vast dataset of phishing and legitimate websites, sourced from diverse regions and tailored to reflect the unique characteristics of developing countries' online ecosystems. The Random Forest model is trained onthis dataset, which helps it identify complex patterns and features that point to phishing attempts. By incorporating region-specific data, the system enhances its adaptability and efficacy in identifying phishing threats relevant to the target demographic. The Random Forest algorithm, renowned for its robustness and ability to handle complex datasets, forms the backbone of the phishing detection mechanism. Through an ensemble learning approach, the model aggregates the insightsof numerous decision trees, each trained on different subsetsof the dataset. This ensemble strategy not only enhances the system's predictive accuracy but also mitigates the risk of over-fitting, ensuring reliable performance across diverse phishing scenarios. Furthermore, the system incorporates dynamic feature extraction techniques to capture evolving phishing tactics and strategies. By continuously analyzing emerging threats and extracting relevant features from website content and user interactions, the system adapts its detection capabilities in real-time, effectively staying ahead of evolving cyber threats.

Methodology

The methodology for enhancing phishing detection involves several key steps. Firstly, data collection and preprocessingare conducted, where a diverse dataset containing examplesof phishing and legitimate communications is gathered from various sources. This dataset is then preprocessed to extract relevant features, ensuring data quality and consistency. Next, feature engineering is carried out to transform raw data into features that capture the underlying patterns of phishing behaviour. Subsequently, the Random Forest algorithm is trainedon the preprocessed dataset. During the training phase, the algorithms learn to distinguish between phishing and legitimate communications based on the extracted features. Once trained,the performance of the machine learning models is evaluated using metrics such as accuracy, precision, recall, and F1-score.The model is then fine-tuned and optimized through parametertuning and cross-validation to enhance their effectiveness in detecting phishing attempts. Finally, the trained models are deployed into production environments where they can be used to analyze incoming communications in real-time and classify them as either phishing or legitimate.

Algorithm - Random Forest

Several characteristics of Random Forest are as follows:

- Ensemble Technique : Random Forest is an ensemble learning technique that utilizes multiple decision treesto make predictions. Each decision tree in the forest is trained independently on a subset of the training data, using a technique called Bootstrap Aggregation (bagging). Bagging involves randomly sampling the training data with replacement to create multiple subsets, ensuring diversity among the trees.
- Multiple Decision Trees : Random Forest consists of a collection of decision trees, where each tree is trainedon a different subset of the data. This approach helps to reduce over-fitting and improve generalization by incorporating diverse perspectives from the individual trees.
- Combining Predictions : Once the decision trees are trained, Random Forest combines their predictions through a process called averaging (for regression tasks) or voting (for classification tasks). In regression tasks, the final output is the average of the predictions madeby each tree. In classification tasks, the final output is determined by a majority vote among the trees.
- Flexibility and Robustness : Random Forest is knownfor its flexibility and robustness. It can handle both regression and classification tasks, making it suitable for various types of data. Additionally, Random Forest is less sensitive to noisy data and outliers compared to individual decision trees, making it a reliable choice for real-world applications.
- Feature Importance : Random Forest provides a measure of feature importance, indicating the relative significance of different features in making predictions. This information can be valuable for understanding the underlying patterns in the data and identifying key indicators of phishing behavior.

RESULTS AND ANALYSIS

The class diagram encapsulates the essential entities and relationships involved in the implementation of machine learning techniques for phishing detection. At its core lies the ”Phishing Detection System” class, representing the overarching system responsible for identifying and mitigating phishing threats. This class serves as a container for various components, including the ”Random Forest Classifier” class, which embodies the specific machine learning algorithm employed for detection. The ”Random Forest Classifier” class encapsulates methods and attributes pertinent to the Random Forest algorithm, such as training data, decision trees, and feature importance scores. Additionally, the diagram includes classes representing input data sources, such as ”Emails” and ”Websites,” which serve as the primary sources of informationfor phishing detection. These classes are connected to the ”Phishing Detection System” via association relationships, indicating their dependency on the system for analysis and classification. Furthermore, auxiliary classes like ”Feature Extractor” and ”Data Preprocessor” are included to illustrate the preprocessing steps involved in

extracting relevant features from input data before feeding them into the Random Forest classifier. Associations between these classes demonstrate the flow of data and operations within the system, highlighting the interconnectedness of its components. Overall, the class diagram provides a comprehensive visual representation of the architecture and functionality of the machine learning-based phishing detection system, facilitating a deeper understanding of its inner workings and relationships.

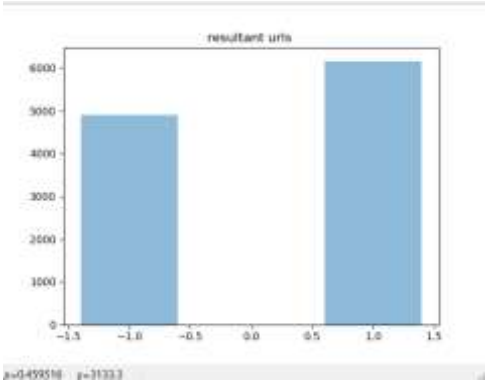

Fig. 1. Class diagram.

The construction of a confusion matrix serves as a pivotal tool in evaluating the effectiveness and performance of machine learning algorithms. This matrix is a comprehensive tabular representation that systematically organizes the results of the classification process, categorizing predictions made by the model against the actual outcomes. With rows representing the actual classes (e.g., legitimate websites, phishing websites) and columns indicating the predicted classes, the confusion matrix provides insights into the algorithm's ability to correctly classify in- stances and identify potential areas of improvement. In the context of phishing detection, the confusion matrix delineates true positives (correctly classified phishing websites), true negatives (correctly classified legitimate websites), false positives (legitimate websites misclassified as phishing), and false negatives (phishing websites misclassified as legitimate). Through this granular breakdown, cybersecurity practitioners can assess the algorithm's precision, recall, accuracy, and F1 score, crucial metrics in gauging its efficacy in identifying and mitigating phishing threats. By scrutinizing the confusion matrix, stakeholders can identify patterns of misclassification, discerning whether the model tends to overpredict or under predict certain classes, thus informing refinements to the algorithm's training data, feature selection, or hyperparameters. Additionally, the confusion matrix facilitates the identification of potential biases or imbalances in the dataset, allowing for corrective measures to ensure equitable performance across all classes. Ultimately, the deployment of a well-constructed con- fusion matrix enables cybersecurity professionals to iteratively enhance the robustness and reliability of machine learning- based phishing detection systems, fortifying defenses against evolving cyber threats in developing country settings.

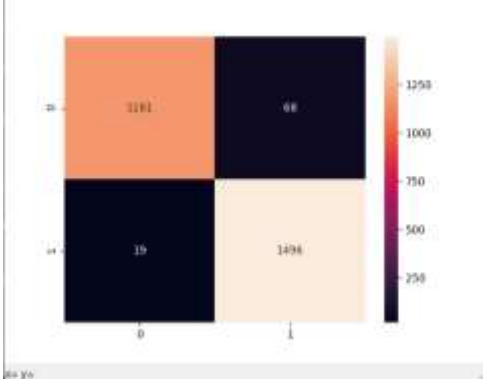

Fig. 2. Confusion matrix.

In this, the concepts of variable importance and relative importance play pivotal roles in understanding and optimizing the effectiveness of the implemented machine learning model. Variable

importance refers to the contribution of individual features or variables in the model towards its predictive accuracy and performance. In the case of phishing detection, variables could encompass a range of factors such as email sender details, URL characteristics, content analysis, and metadata attributes. Through techniques like Random Forest, variable importance can be quantified, providing insights into which features have the most significant influence on distinguishing between legitimate and phishing emails. On the other hand, relative importance extends beyond the absolute impact of variables, considering their significance concerning each other within the model. This nuanced perspective enables a deeper understanding of the interplay between different features and their collective effect on the model's decision-making process. In the context of phishing detection, relative importance elucidates not only which individual features are crucial but also how they interact and complement each other in identifying malicious content. By analyzing both variable and relative importance, practitioners can refine feature selection, optimize model performance, and develop more robust cybersecurity solutions tailored to the unique challenges and contexts of developing countries. This comprehensive approach empowers cybersecurity professionals to enhance detection capabilities, mitigate risks, and safeguard digital assets against evolving phishing threats, ultimately contributing to a more secure and resilient cyber landscape.

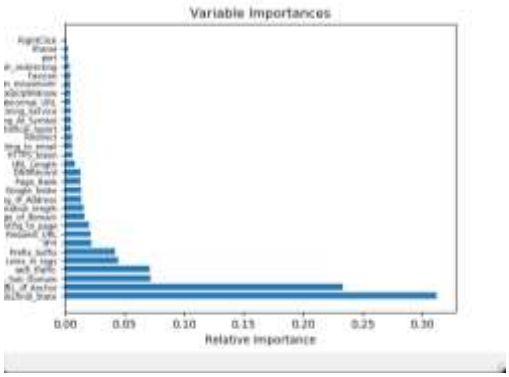

Fig. 3. Variable importance vs Relative importance.

The achievement of a 97% accuracy rate for the random forest algorithm marks a significant advancement in bolstering cybersecurity measures, particularly in regions with developing infrastructure and limited resources. This remarkable accuracy underscores the efficacy of machine learning methodologies, specifically random forest, in combating the pervasive threat of phishing attacks. By leveraging a diverse ensemble of decision trees, the random forest algorithm demonstrates unparalleled robustness in discerning fraudulent from legitimate communications, thereby fortifying the defenses of vulnerable networks and users.

Fig. 4. Accuracy for random forest algorithm.

CONCLUSION

The conclusion of the research paper emphasizes the effectiveness of using the Random Forest classifier for URL phishing detection. The study highlights the advantages of Random Forest, noting its power as a classifier and its ability to avoid overfitting the data when parameters are carefully tuned and selected. This makes it an appropriate choice for analyzing URL Phishing datasets to determine if a URL is phishing or not. Furthermore, the paper suggests future research directions, proposing the expansion of the model to include more classification algorithms. This would facilitate a comparative analysis of several supervised learning algorithms, potentially enhancing the system's capability to identify phishing URLs accurately. This conclusion underlines the potential for continuous improvement and adaptation in the field of cybersecurity, specifically in combating URL phishing threats through machine learning techniques.

REFERENCES

[1] Zuochao Dou, Issa Khalil, Abdallah Khreishah, Ala Al-Fuqaha, Mohsen Guizani, "Systematization of Knowledge (SoK): A Systematic Review of Software-Based Web Phishing Detection", IEEE Communications Surveys & Tutorials, 2017.

[2] M. Khonji, Y. Iraqi and A. Jones, "Phishing Detection: A Literature Survey," in IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp. 2091-2121, Fourth Quarter 2013. .

[3] R. B. Basnet and A. H. Sung, "Mining Web to Detect Phishing URLs," 2012 11th International Conference on Machine Learning and Applications, Boca Raton, FL, USA, 2012, pp. 568-573.

[4] C. L. Tan, K. L. Chiew and S. N. Sze, "Phishing website detection using URL-assisted brand name weighting system," 2014 International Sym- posium on Intelligent Signal Processing and Communication Systems (ISPACS), Kuching, Malaysia, 2014, pp. 054-059.

[5] M. Aydin and N. Baykal, "Feature extraction and classification phishing websites based on URL," 2015 IEEE Conference on Communications and Network Security (CNS), Florence, Italy, 2015, pp. 769-770.

[6] Luong Anh Tuan, Nguyen Ba Lam, To Huu Khuong Nguyen, Minh Hoang Nguyen, "A Novel Approach for Phishing Detection Using URL-Based Heuristic", 2014 International Conference on Computing Management and Telecommunications (ComManTel), IEEE, 2014.

[7] S. Carolin Jeeva, Elijah Blessing Rajsingh, "Intelligent phishing URL detection using association rule mining", Human-centric Computing and Information Sciences, vol. 6, no. 1, pp. 10, 2016.

[8] Marco Cova, Christopher Kruegel, Giovanni Vigna, "Detection and analysis of drive-by-download attacks and malicious JavaScript code", Proceedings of the 19th International Conference on World Wide Web, pp. 281-290, 2010.

[9] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Predicting phishing websites based on self-structuring neural network", Neural Computing and Applications, vol. 25, no. 2, pp. 443-458, 2013-B.

[10] Hiba Zuhair, Ali Selamat, Mazleena Salleh, "Feature selection for phish- ing detection: a review of research", International Journal of Intelligent Systems Technologies and Applications, Vol. 15, No. 2, 2016.

[11] Fadi Thabtah, Rami M. Mohammad, Lee McCluskey, "A Dynamic Self- Structuring Neural Network Model to Combat Phishing", 2016 International Joint Conference on Neural Networks (IJCNN), 2016.

[12] Mayuri A. and Tech M., "Phishing detection based on visual similarity", International Journal of Scientific and Engineering Research (IJSER), Vol. 3, No. 3, March, pp.1–5.

[13] Huang H., Tan J., and Liu L., "Countermeasure techniques for deceptive phishing attack", International Conference on New Trends in Informa- tion and Service Science (NISS'09), 30 June–02 July 2009, China, pp.636–641.

[14] APWG Phishing Trends Reports, Anti Phishing Working Group.

[15] UCI Machine Learning Repository, https://archive.ics.uci.edu/ml.