2581-4575
International Journal For Recent Developments in Science & Technology
Crossref
A Peer Reviewed Research Journal

PHISHING DETECTION SYSTEM THROUGH HYBRID MACHINE LEARNING BASED ON URL

S.Preethi[1], Kesari Prasanna Rao[2], Kukkala Chandana[3], Errolla Bindu[4]

[1]Associate Professor, School of CSE,Malla Reddy Engineering College For Women (UGC-Autonomous), Maisammaguda,Dhulapally,Secunderabad,Telangana-500100

[234]UG Student, School of CSE,Malla Reddy Engineering College for Women, (UGC-Autonomous), Maisammaguda,Dhulapally,Secunderabad,Telangana-500100

Email: Preethisingireddy19@gmail.com

ABSTRACT

Cybercrime, especially phishing, is one of the growing concerns today. First discovered in 1996, phishing has become one of the most dangerous forms of cybercrime, using email manipulation and fraudulent websites to steal sensitive information. Although many studies have been conducted on the prevention, detection, and solutions of phishing, there is no holistic approach that has been established.This paper presents an automated phishing detection framework using machine learning algorithms. The objective is that of distinguishing phishing from benign while keeping the false positive as low as possible. Therefore, the research focuses on a static phishing detector for windows operating systems using more than 11,000 genuine and phishing website URLs that have been preprocessed on which several machine learning algorithms are applied, including: Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), Gradient Boosting Classifier (GBM), K-Nearest Neighbor (KNN), and Support Vector Classifier (SVC). In addition, a hybrid LSD model combining LR, SVC, and DT with soft and hard voting mechanisms is proposed. The LSD model applies canopy feature selection, cross-fold validation, and grid search for hyperparameter optimization. The models are tested using metrics such as accuracy, precision, recall, F1-score, and specificity. The results show that the LSD model outperforms others in accuracy and efficiency.The model, although successful, does not inherently learn to change with the changing phishing techniques as attackers constantly modify their methods. This can be overcome by retraining the system using updated data. Although this study is on static analysis, dynamic analysis may be integrated in future work to enhance detection. Overall, the proposed framework provides an effective solution for phishing detection, although periodic updates are necessary to maintain effectiveness.

Keywords-Decision Tree, Logistic Regression, Random Forest, Naive Bayes, Gradient Boosting Classifier, k-Nearest Neighbor, Support Vector Classifier, Hybrid Model.

I. INTRODUCTION

The increased dependence on the internet has presented both opportunities and challenges, especially in the realm of cybersecurity. Among the many forms of cybercrime, phishing attacks have become one of the most common and harmful threats. Phishing is an activity where attackers impersonate legitimate entities, usually through fake emails or websites, to steal sensitive information such as login credentials,

credit card numbers, and other personal data. With the widespread use of the internet for communication, financial transactions, and social interaction,phishing attacks have become a serious concern for individuals and organizations alike.

Phishing is not a new threat; it has existed since the mid-1990s. However, it has evolved significantly, becoming more sophisticated and difficult to detect. Attackers continuously adapt their methods, making it increasingly challenging for traditional security systems to keep up. Most antivirus software and detection methods fail to prevent new or modified phishing techniques. There is a growing need for more advanced and effective solutions to identify and prevent phishing attacks.

The essence of this project lies in using machine learning for an automated phishing detection framework. Thus, the solution would leverage a dataset of phishing and legitimate URLs for training and validating the machine learning models on their ability to classify sites between malicious and benign ones. Decision Trees, Random Forests, Logistic Regression, and Support Vector Machines would be applied for finding reliable, accurate methods to detect the URLs, phishing, and non-phishing ones.

Besides the traditional machine learning-based algorithms, this project extends to a hybrid model for combining multiple classifiers to build a more accurate detection engine. The hybrid approach intends to integrate Logistic Regression (LR), Support Vector Classifier (SVC), and Decision Trees (DT) to make detection more robust against phishing. Techniques such as feature selection and hyperparameter optimization are also used to enhance performance in the detection system.

The final end product will be a fully functional automatic real-time detection of phishing sites with significant enhancement to protect users' personal online activities. Enhancing detection techniques of phishing attempts thus plays its part to protect and ensure security in this broader fight of securing online

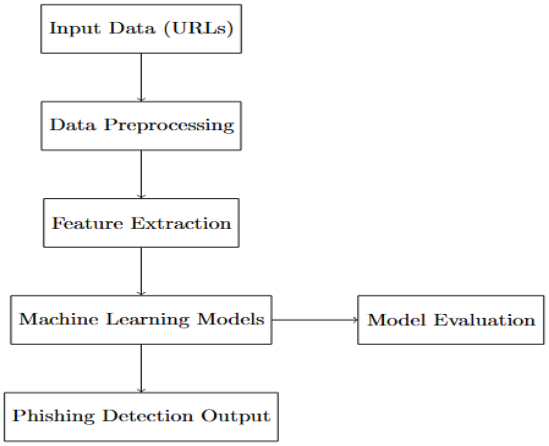

Fig 1:System architecture

II. RELATED WORK

1.R. Jenni and S. Shankar (2018) carried out a comprehensive review of different methods of phishing detection. The study underlines the various detection techniques and approaches used in identifying phishing attacks. The authors discuss a variety of methods based on machine learning, heuristics, and blacklists and how they vary in terms of efficiency, accuracy, and scope. This review stresses the need for further research in developing efficient phishing detection methods to mitigate the increasing threat of cybercrime. The study further indicates that future solutions might be required to integrate multiple methods to enhance the robustness and accuracy of phishing detection systems.

2.V. Jyothsna et al. (2020) proposed a Network Intrusion Detection System (NIDS) that integrates Hybrid Dimensionality Reduction and a Neural Network-Based Classifier. Their system detects network intrusions of the various types, including phishing attacks, with high efficiency by reducing the dimensionality of data for improvement in processing time and accuracy, plus a neural network classifier able to detect complex patterns associated with a phishing attack. This paper demonstrates how the fusion of dimensionality reduction and machine learning can create an efficient and scalable system to detect phishing within network traffic that is helpful for the betterment of overall cybersecurity.

3.R. Anusuya et al. (2023) explored the detection of DDoS attacks in SDNs using a Machine Learning approach. While this study primarily addresses DDoS attacks, it provides valuable insight on how machine learning techniques can be adapted to detect other types of cyber threats, such as phishing. Utilizing machine learning, their approach can effectively identify patterns of network anomalies that often indicate malicious activity. This research advances machine learning in network security enhancement and provides an essential foundation for its application in the detection of phishing attacks in SDNs as well.

4. P. George and P. Vinod (2018) proposed a methodology for spam email identification using composite email features. Although their paper mainly targets the issue of spam detection, the presented techniques are pretty relevant for phishing detection since both relate to malicious communication intentions. They introduced a system that extracts the features of an email related to content, headers, and other patterns relevant for spam and malicious emails. Using these composite features, their approach enhances the accuracy of spam classification and has implications for phishing email detection. Their approach underscores the importance of feature extraction in identifying malicious communications and sets the stage for applying similar techniques to phishing emails, where attackers attempt to steal sensitive information. This work demonstrates how email characteristics can be used to effectively distinguish between legitimate and phishing emails.

III. IMPLEMENTATION

The implementation of the phishing detection system begins with data collection, where a dataset containing both phishing and legitimate URLs is obtained. The dataset consists of over 11,000 URLs, with various attributes such as domain name, URL length, use of HTTPS, and the presence of suspicious keywords. This dataset is preprocessed to handle missing values, remove irrelevant features, and normalize numerical attributes.Once the data is preprocessed, it is divided into training and testing sets using an 80-20 split.Feature selection techniques like SelectKBest are used to retain the most relevant features, which help improve model performance by eliminating redundant or irrelevant information.

The core of the system involves training several machine learning algorithms, including Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), Gradient Boosting (GBM), K-Nearest

Neighbors (KNN), and Support Vector Classifier (SVC). Each model is trained on the training dataset, and predictions are made on the testing dataset. A hybrid model combining LR, SVC, and DT is also developed to improve the accuracy of the system.

Hyperparameter optimization is carried out using Grid Search to fine-tune the models for better performance. The models are evaluated based on key performance metrics such as accuracy, precision, recall, F1-score, and specificity. The best performing model is selected for deployment, ensuring high accuracy and minimal false positives.

The final system can effectively classify URLs as phishing or legitimate, providing real-time protection against phishing attacks.

IV. ALGORITHM

Below is a step-by-step outline of the algorithm used for phishing detection:

1. Data Collection

Dataset: The dataset used for this project contains over 11,000 URLs, including both phishing and legitimate URLs, gathered from various repositories. Each URL in the dataset has multiple features (attributes) representing its structure and behavior, such as domain name, URL length, presence of special characters, etc.

2. Preprocessing

1.Data Cleaning: In this step, missing or erroneous data points are handled, and the dataset is cleaned for further analysis

2.Feature Engineering: Relevant features are selected or created based on the URL's characteristics.

3.Feature Normalization: The features are normalized to ensure they are on the same scale and do not influence the machine learning models disproportionately.

3. Feature Selection

Canopy Feature Selection: This technique is applied to reduce the dimensionality of the feature space by selecting the most relevant and important features. It helps improve the model's performance and reduces overfitting.

4.Model Evaluation

Cross-fold Validation: To ensure the model's robustness, a k-fold cross-validation is used to evaluate the performance on different subsets of the dataset.

Performance Metrics: The models are evaluated using several metrics

6. Prediction

Once the models are trained and validated, the system can predict whether a given URL is phishing or legitimate by applying the trained models to unseen URLs (from the test set).

7. Final Decision

Combining Predictions: The predictions from different models (in case of the hybrid model) are combined using voting techniques (either soft or hard) to finalize the detection decision.

Alert Generation: If the system detects a phishing URL, it generates an alert for the user or blocks access to the phishing site.

2581-4575
International Journal For Recent Developments in Science & Technology
Crossref
A Peer Reviewed Research Journal

V. RESULTS

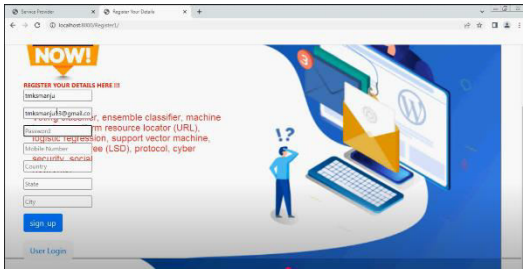

Fig 1 :Register

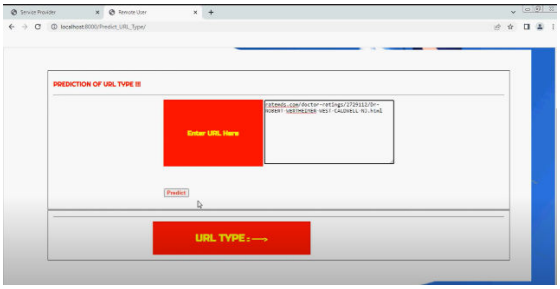

Fig 2 : Predction of URL Type

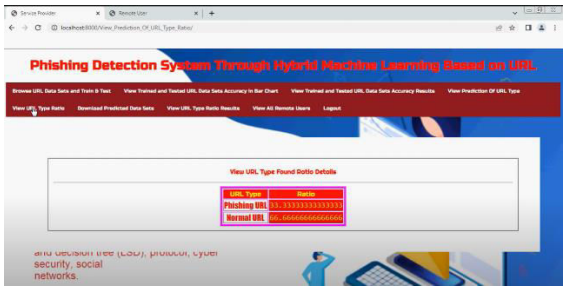

Fig 3 :View URL Type Found Ratio Details

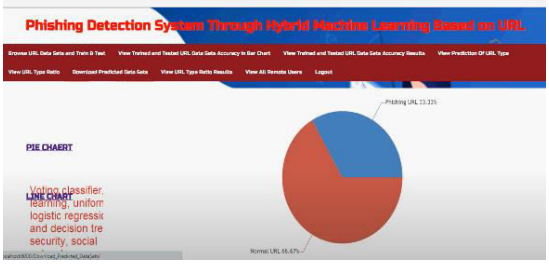

Fig 4 : Pie Chart

VI. CONCLUSION

In summary, this project provides a machine learning-based framework for detecting phishing attacks. Phishing attacks are some of the most common and destructive cybercrimes in today's digital world. Phishing attacks are evolving to be more sophisticated, thereby making traditional security measures ineffective. This project addresses the issue by using machine learning techniques to develop an automated system that can distinguish between phishing and legitimate websites with high accuracy.

The proposed system makes use of a wide variety of machine learning models, including Decision Trees, Random Forests, Logistic Regression, Naive Bayes, Gradient Boosting, K-Neighbors Classifiers, and Support Vector Machines. By using a hybrid model combining Logistic Regression (LR), Support Vector Classifier (SVC), and Decision Trees (DT), the framework enhances the accuracy and robustness of the detection. Techniques like feature selection, cross-fold validation, and hyperparameter optimization further ensure the reliability and efficiency of the system.

The evaluation results indicate that the proposed hybrid model has outperformed the traditional machine learning approaches by various key performance metrics, including precision, accuracy, recall, F1-score, and specificity. It means that combining different classifiers will increase the detection capability for phishing attacks and provides a better defense mechanism.

But limitations persist on the proposed approach. Since the system is based on static analysis, there may not be effective detection of

new or dynamically evolving phishing attacks. This issue can, however, be mitigated with periodic updates on the training dataset. For an even better detection system incorporating dynamic analysis, the system may be further enhanced in accuracy and responsiveness to phishing.

Overall, this project contributes to the ongoing efforts to combat phishing attacks by providing a machine learning-based solution that improves online security and user protection. With the evolution of phishing threats, the integration of such advanced detection methods, like the hybrid model proposed here, will be critical to safeguard personal and organizational data in the increasingly interconnected digital world.

REFERENCES

[1]N. Z. Harun, N. Jaffar and P. S. J. Kassim, "Physical attributes significant in preserving the social sustainability of the traditional malay settlement" in Reframing the Vernacular: Politics Semiotics and Representation, Springer, pp. 225-238, 2020.

[2]D. M. Divakaran and A. Oest, Phishing detection leveraging machine learning and deep learning: A review, 2022.

[3]A. Akanchha, "Exploring a robust machine learning classifier for detecting phishing domains using SSL certificates", Fac. Comput. Sci. Dalhousie Univ. Halifax NS Canada Tech. Rep., 2020

[4]H. Shahriar and S. Nimmagadda, "Network intrusion detection for TCP/IP packets with machine learning techniques" in Machine Intelligence and Big Data Analytics for Cybersecurity Applications, Cham, Switzerland:Springer, pp. 231-247, 2020.

[5]J. Kline, E. Oakes and P. Barford, "A URL-based analysis of WWW structure and dynamics", Proc. Netw. Traffic Meas. Anal. Conf. (TMA), pp. 800, Jun. 2019.

[6]A. K. Murthy and Suresha, "XML URL classification based on their semantic structure orientation for web mining applications", Proc. Comput. Sci., vol. 46, pp. 143-150, Jan. 2015.

[7]A. A. Ubing, S. Kamilia, A. Abdullah, N. Jhanjhi and M. Supramaniam, "Phishing website detection: An improved accuracy through feature selection and ensemble learning", Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 1, pp. 252-257, 2019.

[8]A. Aggarwal, A. Rajadesingan and P. Kumaraguru, "PhishAri: Automatic realtime phishing detection on Twitter", Proc. eCrime Res. Summit, pp. 1-12, Oct. 2012.

[9]S. N. Foley, D. Gollmann and E. Snekkenes, Computer Security-ESORICS 2017, Oslo, Norway:Springer, vol. 10492, Sep. 2017.

[10]P. George and P. Vinod, "Composite email features for spam identification" in Cyber Security, Singapore:Springer, pp. 281-289, 2018.

[11]S. Hota, A. K. Shrivas and R. Hota, "An ensemble model for detecting phishing attack with proposed remove-replace feature selection technique", Proc. Comput. Sci., vol. 132, pp. 900-907, Jan. 2018.

[12]G. Sonowal and K. S. Kuppusamy, "PhiDMA—A phishing detection model with multi-filter approach", J. King Saud Univ. Comput. Inf. Sci., vol. 32, no. 1, pp. 99-112, Jan. 2020.

[13]M. Zouina and B. Outtaj, "A novel lightweight URL phishing detection system using SVM and similarity index", Hum.-Centric Comput. Inf. Sci., vol. 7, no. 1, pp. 17, Jun. 2017.

[14]R. Ø. Skotnes, "Management commitment and awareness creation—ICT safety and security in electric power supply network companies", Inf. Comput. Secur., vol. 23, no. 3, pp. 302-316, Jul. 2015.

[15]R. Prasad and V. Rohokale, "Cyber threats and attack overview" in Cyber Security: The Lifeline of Information and Communication Technology, Cham, Switzerland:Springer, pp. 15-31, 2020.

[16]T. Nathezhtha, D. Sangeetha and V. Vaidehi, "WC-PAD: Web crawling based phishing attack detection", Proc. Int. Carnahan Conf. Secur. Technol. (ICCST), pp. 1-6, Oct. 2019.

[17]R. Jenni and S. Shankar, "Review of various methods for phishing detection", EAI Endorsed Trans. Energy Web, vol. 5, no. 20, Sep. 2018.

[18]V. Jyothsna, A.N. Sreedhar, D. Mukesh and A. Ragini, "A Network Intrusion Detection System with Hybrid Dimensionality Reduction and Neural Network Based Classifier" in ICT Systems and Sustainability. Advances in Intelligent Systems and Computing, Singapore:Springer, vol. 1077, 2020.

[19]R. Anusuya, M. Ramkumar Prabhu, Ch. Prathima and J. R. Arun Kumar, "Detection of TCP UDP and ICMP DDOS attacks in SDN Using Machine Learning approach", Journal of FisheriesSciences, vol. 10, no. 4S, 2023.