PHISHING DETECTION SYSTEM THROUGH HYBRID MACHINE LEARNING BASED ON URL

[1]Masanam Veda Kumar Raja,[2]Dr.S.Swathi Rao

[1]Student, Department of CSE, Ellenki College of Engineering and Technology (UGC Autonomous).

[2]Professor,Department of CSE, Ellenki College of Engineering and Technology (UGC Autonomous).

ABSTRACT

Phishing attacks have become one of the most dangerous forms of cybercrime since they were first introduced in 1996. These attacks often use deceptive emails and fake websites to trick individuals into revealing sensitive information. Despite numerous studies addressing methods for detecting and preventing phishing, a comprehensive solution to fully combat these attacks remains elusive. This study focuses on utilizing machine learning to tackle phishing threats, specifically through analyzing URLs. It leverages a phishing URL-based dataset, containing over 11,000 websites, with both phishing and legitimate URL attributes. These URLs are processed and transformed into vector form for machine learning applications. Several machine learning algorithms have been tested and implemented to identify phishing URLs and provide enhanced protection. These include well-established models like Decision Tree (DT), Linear Regression (LR), Random Forest (RF), Naive Bayes (NB), Gradient Boosting Classifier (GBM), K-Neighbors Classifier (KNN), and Support Vector Classifier (SVC). In addition, the study proposes a hybrid model called LSD, combining Logistic Regression, Support Vector Machine, and Decision Tree (LR+SVC+DT). This hybrid model uses both soft and hard voting mechanisms to improve the system's accuracy and efficiency in detecting phishing URLs. To further refine the results, the study applies feature selection using the Canopy technique, along with cross-fold validation and Grid Search for hyperparameter optimization. Several evaluation metrics, including precision, accuracy, recall, F1-score, and specificity, are used to assess the performance of the models. Comparative analysis shows that the proposed LSD hybrid model outperforms the other tested algorithms, offering the best overall results in terms of accuracy and efficiency in phishing detection.

Keywords : Phishing Detection, Cybersecurity, Machine Learning, URL-Based Analysis, Hybrid Model, Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, Gradient Boosting, K-Nearest Neighbors, Naive Bayes, Feature Selection, Hyperparameter Optimization, Voting Mechanism, Fraud Prevention.

I.INTRODUCTION

The internet has revolutionized various aspects of human life, serving as a global network of interconnected computers that allows people to access information, communicate, and conduct business. It has become a central hub for activities ranging from education and banking to social media and e-commerce. In particular, the internet facilitates rapid communication, online shopping, and virtual meetings, all of which have become increasingly important, especially in the context of the COVID-19 pandemic. However, with the rise in data sharing and online interactions, there has also been an increase in cybercrime, which poses significant threats to internet users' privacy and security. [1] Phishing attacks, in which criminals trick individuals into revealing sensitive information, have emerged as one of the most prevalent and dangerous cybercrimes. Phishing, which involves the creation of fake websites and misleading communications, is a growing threat in the digital world. Attackers often create deceptive websites that appear legitimate, sending links to these sites in the hopes of tricking users into disclosing personal information such as passwords, social security numbers, or credit card details. [2] These attacks can lead to serious consequences, including financial loss and identity theft. Despite efforts to detect and mitigate phishing, many people still fail to recognize the risks posed by suspicious websites, often because they do not scrutinize website URLs carefully. This lack of awareness makes individuals vulnerable to phishing scams, which continue to increase in both frequency and sophistication. [3] The study presented focuses on improving phishing detection through the analysis of URLs using machine learning techniques. URLs are a critical component of phishing attacks, as attackers often use misleading or malicious URLs to direct victims to fraudulent websites. The research aims to develop a robust system to detect phishing URLs based on a large dataset of over 11,000 phishing-related URL attributes. By analyzing these attributes, the system seeks to classify URLs as either legitimate or malicious, thus preventing users from falling victim to phishing scams. [4] The study leverages several machine learning models, including decision tree (DT), linear regression (LR), random forest (RF), naive Bayes (NB), gradient boosting machine (GBM), K-neighbors classifier (KNN), and support vector classifier (SVC), as well as a hybrid model that combines logistic regression, support vector machine, and decision tree (LR+SVC+DT). [5] To further improve the performance of the model, the study employs advanced techniques such as cross-fold validation, grid search parameter optimization, and the canopy feature selection method. These approaches help fine-tune the model to increase its accuracy in identifying phishing URLs. The hybrid model, known as the LSD model, utilizes both soft and hard voting mechanisms to enhance prediction results. [6] The effectiveness of the proposed system is evaluated using various performance metrics, including accuracy, precision, recall, specificity, and F1-score. These metrics are essential for assessing the model's ability to correctly identify phishing URLs while minimizing false positives and false negatives. [7] The research builds upon previous efforts in phishing detection, contributing a more accurate and efficient system to protect users from online threats. Phishing remains a serious concern, and the development of better detection systems is crucial to reducing the risks associated with cybercrime. [8] The results of this study demonstrate the potential of machine learning to improve phishing detection, offering a promising solution to the growing problem of internet security. The study's proposed methodology offers a novel approach to phishing URL detection, combining various machine learning models to achieve superior results compared to previous systems. [9] In conclusion, this study emphasizes the importance of addressing phishing threats to ensure the safety and privacy of internet users. By leveraging machine learning and sophisticated techniques like hybrid models and feature selection, the proposed system offers a more accurate and efficient solution to phishing detection. [10] As phishing attacks continue to evolve, it is crucial to develop robust systems that can adapt to new challenges and effectively safeguard users from online scams. The paper is organized into sections that detail the background and related work, the materials and methods used in the study, the experimental results, and the conclusions drawn from the research, offering valuable insights for future advancements in the field of cybersecurity. [11]

II.LITERATURE SURVEY

N. Z. Harun, N. Jaffar, and P. S. J. Kassim (2020) explore the physical attributes essential for preserving the social sustainability of traditional Malay settlements, focusing on the role of communal spaces, architectural styles, and environmental features in fostering social interaction and community cohesion. These elements are key to maintaining cultural identity and promoting sustainable living within these communities. D. M. Divakaran and A. Oest (2022) present a comprehensive review on the use of machine learning and deep learning techniques for phishing detection, examining various models built on different types of data and their respective advantages and challenges. The review also discusses multiple deployment options aimed at improving the effectiveness of cybersecurity measures against phishing attacks. A. Akanchha (2020) investigates the potential of using SSL certificates in combination with machine learning classifiers to detect phishing domains, highlighting how this approach can improve the identification of fraudulent websites and enhance online security. H. Shahriar and S. Nimmagadda (2020) delve into the detection of network intrusions within TCP/IP packets through machine learning techniques, proposing various algorithms to identify malicious activities in network traffic, which contributes to advancing cybersecurity measures. J. Kline, E. Oakes, and P. Barford (2019) provide an analysis of the World Wide Web's structure and dynamics using URL-based metrics, offering insights into web traffic patterns and content organization, which help in understanding the functioning of the internet. A. K. Murthy and Suresha (2015) focus on classifying XML URLs based on their semantic structure orientation for web mining applications, introducing methods to categorize URLs and extract valuable information from web data. Finally, A. A. Ubing, S. Kamilia, A. Abdullah, N. Jhanjhi, and M. Supramaniam (2019) enhance phishing website detection by integrating feature selection and ensemble learning techniques, demonstrating that a combination of models and relevant feature selection can significantly improve phishing site identification, thereby strengthening cybersecurity defenses.

III.PROPOSED METHODOLOGY

The proposed methodology for detecting phishing URLs is centered around a hybrid machine learning approach, aiming to effectively prevent cybercrime and protect privacy. The study uses a comprehensive dataset containing over 11,000 attributes related to phishing URLs, which helps in the classification of these URLs based on their distinguishing features. Various machine learning models are applied, including Decision Tree (DT), Linear Regression (LR), Naive Bayes (NB), Random Forest (RF), Gradient Boosting Machine (GBM), Support Vector Classifier (SVC), and K-Neighbors Classifier (KNN). A key feature of the proposed approach is the development of a hybrid model, combining LR, SVC, and DT using both soft and hard voting mechanisms, referred to as LSD. This hybrid model aims to enhance the accuracy of phishing detection. To further improve the performance of the hybrid model, cross-fold validation combined with a grid search parameter optimization technique is employed. This is enhanced by using the canopy feature selection technique, which helps in selecting the most relevant features for phishing detection. The methodology will be rigorously evaluated using several performance metrics, including accuracy, precision, recall, specificity, and F1-score, to assess the effectiveness and reliability of the proposed system in real-world applications.

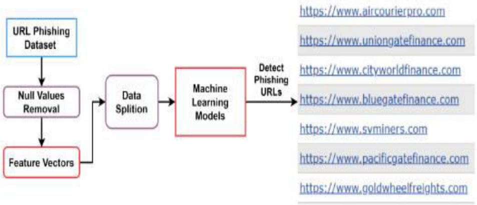

URL Phishing Dataset
Null Values Removal
Feature Vectors
Data Splition
Machine Learning Models
Detect Phishing URLs
https://www.aircourierpro.com
https://www.uniongatefinance.com
https://www.cityworldfinance.com
https://www.bluegatefinance.com
https://www.svminers.com
https://www.pacificgatefinance.com
https://www.goldwheelfreights.com

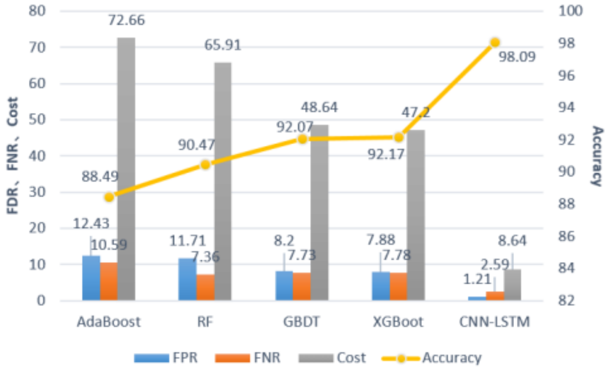
Fig1 : system architecture

IV.WORKING METHODOLOGY

The working methodology of the Phishing Detection System integrates machine learning techniques with Django-based web application functionalities to effectively identify phishing URLs. The process begins with the training phase, where various machine learning models such as Naive Bayes, SVM, Logistic Regression, Decision Tree, Gradient Boosting, and Random Forest classifiers are applied to a dataset of URLs. This dataset contains URLs labeled as either benign or phishing, and the URLs are processed using a CountVectorizer to convert them into a numerical format suitable for model training. The data is split into training and testing sets, with each model being trained on the training data. Once the models are trained, they make predictions on the test data, and their performance is evaluated using metrics like accuracy score, classification report, and confusion matrix. The results are stored in the database for further analysis and comparison. In the prediction phase, users interact with the web interface by submitting URLs for classification. The system processes these URLs using the same vectorization techniques and applies the trained models to determine whether the URL is phishing or benign. To enhance prediction robustness, a voting classifier is used to combine the predictions from multiple models. The result is displayed on the user interface, informing users whether the URL is categorized as a "Phishing URL" or "Normal URL." Additionally, the system offers data management features, allowing the administrator to view trends in phishing URL detection, compare the accuracy of various models, and track the ratio of phishing to normal URLs. Users can also download the predicted results in an Excel file, and the system keeps track of performance metrics such as detection accuracy for each classifier. The entire system is implemented using Django, where various views are created to manage operations like user login, viewing predictions, registering users, and downloading data. These operations are interconnected through Django models that interface with the database to store user data, predictions, and model performance metrics. The service provider and remote users have distinct roles within the system: the service provider is responsible for training models and analyzing results, while remote users can submit URLs for prediction.

V.IMPLIMENTATION

The implementation of the Phishing Detection System combines machine learning techniques with a Django-based web application to effectively identify phishing URLs. The process begins with setting up the environment by installing necessary libraries such as Django, Scikit-learn, Pandas, and NumPy. A dataset containing URLs labeled as phishing or benign is collected and preprocessed. The URLs are converted into numerical features using a CountVectorizer, which transforms the text data into a format suitable for machine learning models. The dataset is then split into training and testing sets, and multiple models, including Naive Bayes, Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), and Gradient Boosting, are trained on the data. After training, the models are evaluated using metrics such as accuracy score and classification reports.To improve the prediction accuracy, a Voting Classifier is used, which combines predictions from multiple models using a majority voting mechanism. This classifier is then integrated into a Django web application, where users can submit URLs for classification through a web interface. The application processes the submitted URLs and predicts whether they are phishing or benign based on the trained models. The predictions are displayed on the user interface, where users are informed of the result. Additionally, the system allows administrators to view and analyze trends, model performance, and phishing URL detection statistics. It also includes a feature for downloading prediction results in an Excel file. The Django application is structured with views for different operations like logging in, viewing predictions, registering users, and downloading data. Django models are used to interact with the database, storing user data, predictions, and performance metrics. The system is designed with distinct roles for service providers and remote users, where the service provider is responsible for training the models and analyzing the results, while remote users can predict the type of submitted URLs. Finally, the system is deployed on cloud platforms like AWS or Heroku for production use, making it accessible for users to detect phishing URLs securely and efficiently.



VI.CONCLUSION

In conclusion, the rise of the Internet has ushered in a new era of rapid growth, but it has also brought with it a significant increase in cybercrimes. Phishing, particularly through the use of malicious URLs, has become one of the most common methods for cyber attackers to breach privacy and infiltrate networks. These phishing URLs often disguise themselves as legitimate ones, making them highly effective tools for intrusion. In response to this growing threat, this study proposes a machine learning-based phishing detection system to identify and mitigate risks posed by phishing URLs. By using a comprehensive dataset with over 11,000 URLs and 32 attributes, extracted from Kaggle, the study explores various machine learning models such as Decision Tree, Random Forest, Support Vector Machine (SVM), Gradient Boosting, K-Neighbor Classifier, Naive Bayes, and hybrid models (LR+SVC+DT) with both soft and hard voting mechanisms. Advanced techniques like canopy feature selection, cross-fold validation, and Grid Search for hyperparameter optimization were employed to further improve model performance. The results of the experiments demonstrate that the proposed approach achieves effective and efficient phishing detection, surpassing traditional methods. Moving forward, future phishing detection systems could benefit from integrating list-based machine learning models with existing frameworks, allowing for even more accurate and timely identification of phishing URLs, ultimately helping to enhance the security of online environments and protecting users from malicious attacks.

VII.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, 2020, pp. 225–238.

[2] D. M. Divakaran and A. Oest, ‘‘Phishing detection leveraging machine learning and deep learning: A review,’’ 2022, *arXiv:2205.07411*.

[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. 10222/78875, 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, 2020, pp. 231–247.

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

[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, ''Phish Ari: Automatic real time phishing detection on Twitter,'' in *Proc. eCrime Res. Summit*, Oct. 2012, pp. 1–12.

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

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

[11] H. 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, p. 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, 2020, pp. 15–31.