IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

An Intelligent System For Real-Time Phishing URL Detection Using Hybrid ANN Algorithm

[1]Dr.J.S.Kanchana, [2]Dinesh Balaji, [3]Tamil Ezhil, [4]Yoki Raj,

[1]Professor, [2]Student, [3]Student,[4]Student,

Department of Cyber Security,

KLN College Of Engineering , Sivagangai

Abstract Phishing attacks are one of the major cyber security threats, where attackers deceive users into revealing sensitive information through fraudulent websites. This project proposes a real-time phishing website detection system using a hybrid machine learning algorithm that combines Extreme Gradient Boosting (XGBoost) and an Artificial Neural Network (ANN). A labeled dataset containing both legitimate and phishing URLs is used to train and evaluate the proposed model. Various URL-based features, including structural properties, security indicators, and domain-related attributes, are extracted to enhance detection accuracy. The XGBoost model effectively captures complex feature interactions, while the ANN improves generalization by learning non-linear patterns. The hybrid approach integrates predictions from both models to achieve improved robustness and reduced overfitting. Experimental results demonstrate that the proposed system achieves high accuracy and reliable performance in detecting phishing websites. This approach enables early identification of phishing threats and can be integrated into web-based security solutions to enhance user safety and prevent online fraud..

.

Index Terms **- Phishing, URL, ANN, Feature extraction, Classification model**

I. INTRODUCTION

The rapid growth of internet usage and online services has increased the risk of cybersecurity threats, among which phishing attacks are one of the most common and dangerous. Phishing occurs when attackers create fake websites or messages that appear legitimate in order to trick users into revealing sensitive information such as passwords, banking details, or personal data. Traditional detection methods like blacklists and rule-based systems are often ineffective against newly created phishing websites. To address this issue, machine learning techniques are increasingly used to automatically detect phishing websites by analyzing patterns in URLs and related features. In this project, a hybrid machine learning approach combining Extreme Gradient Boosting (XGBoost) and an Artificial Neural Network (ANN) is proposed to improve detection accuracy and reliability. By extracting various URL-based features such as structural characteristics, security indicators, and domain-related attributes, the system can effectively identify phishing websites in real time. The hybrid model leverages the strengths of both algorithms to enhance performance, reduce overfitting, and provide a robust solution for protecting users from online fraud.

II. PROPOSED TECHNIQUE

The proposed system for detecting phishing URLs using machine learning and ANN algorithms consists of the following components:

1. URL Scanner: The URL scanner is responsible for collecting and scanning URLs from the user input. It extracts different structural features from the URL such as domain name, URL length, number of subdomains, presence of special characters, and suspicious keywords.
2. Feature Selection Module: This module analyzes the URL and extracts important features used for phishing detection. These features include lexical features, domain-based features, and content-based features that help in identifying malicious URLs.
3. Feature Extraction Module: This module analyzes the URL and extracts important features used for phishing detection. These features include lexical features, domain-based features, and content-based features that help in identifying malicious URLs.

4.ANN Classifier:

The selected features are fed into an **Artificial Neural Network (ANN)** model. The ANN learns patterns from the training data and classifies the URL as either **phishing or legitimate**. The neural network consists of input layers, hidden layers, and output layers that process the URL The features and produce the final prediction

.

5.User Interface: The system provides a user-friendly interface where users can enter a URL and check whether it is phishing or legitimate. The interface also displays performance metrics such as **accuracy, precision, recall, and F1-score**.

The **ANN model** learns complex patterns and relationships among the extracted features through multiple layers and backpropagation, while the **XG-Boost algorithm** improves prediction accuracy by building multiple decision trees and correcting the errors of previous trees. By combining these two techniques, the system provides an effective and reliable solution for detecting phishing URLs and improving online security

Working process of proposed system

1. URL preprocessing: The input URL is preprocessed by removing unnecessary elements such as **"https://", "http://", and "www."** in order to standardize the format of the URL.
2. Feature selection: A feature selection algorithm is applied to select the most important features that contribute to phishing detection. This helps reduce the dimensionality of the dataset and improves the performance of the ANN model..
3. ANN model training: The selected features are used to train the **Artificial Neural Network model**. During training, the ANN learns patterns from both phishing and legitimate URLs.
4. Model optimization: The ANN model is optimized by adjusting parameters such as the number of hidden layers, learning rate, and activation functions to improve the detection accuracy.
5. Model evaluation: The trained model is evaluated using a test dataset containing phishing and legitimate URLs. The performance of the system is measured using evaluation metrics include accuracy, precision, recall, and F1 score.
6. Deployment: The proposed system can be deployed in real-time applications such as web browsers, email filters, and network security systems to detect phishing URLs and alert users.

III. CHALLENGES

There are several challenges that need to be addressed in the proposed system for detecting phishing URLs using machine learning and ANN algorithms. Some of these challenges are:

1. Dataset bias: The quality and quantity of the dataset can significantly affect the accuracy of the classification model. The dataset should be representative of the real-world scenario and contain a balanced distribution of phishing and legitimate URLs.
2. Feature engineering: The selection of relevant features can be challenging, and some features may not be effective in identifying phishing URL's in different contexts .The feature selection algorithm need to be carefully designed to ensure that the most relevant features are selected.
3. Adversarial attacks: Phishing attackers can use advanced techniques, such as URL obfuscation and evasion techniques, to evade detection by the phishing detection system. The proposed system needs to be robust to adversarial attacks and able to detect advanced phishing attacks.
4. Real-time processing: The proposed system needs to process URLs in real-time and provide instant feedback on their legitimacy. This requires efficient algorithms and hardware resources to handle large

volumes of requests.

5. Generalizability: The proposed system needs to be able to generalize to different contexts and be effective in detecting phishing URLs in different languages and domains. The system should be adaptable to different environments and easily customizable to different user needs.

Addressing these challenges will be critical to the success of the proposed system and improving the security of online services.

IV. METHODOLOGIES

1. Universe of the study: The universe of the study is all the URLs that are available on the internet. The study focuses on identifying phishing URLs among them using machine learning and ANN+XG-BOOST algorithms.
2. Sample of the study: The study uses a dataset of 3,000 URLs, with 1,500 phishing URLs and 1,500 legitimate URLs. The URLs are randomly selected from various sources on the internet to ensure that the dataset is representative of the real world scenario.
3. Data and sources of data: The data used in the study includes structural and content features of URLs, such as domain name, path length, presence of special characters, and the number of subdomains. The dataset is collected from various sources on the internet, including public phishing databases and legitimate websites.

ANN+XG-BOOST

Artificial Neural Networks (ANN) and XG-Boost are powerful machine learning techniques used for classification tasks such as phishing URL detection. ANN is inspired by the human brain and consists of interconnected neurons organized into input, hidden, and output layers that learn complex patterns from the input data through training and backpropagation.

In the proposed system, extracted URL features such as URL length, domain information, number of subdomains, and presence of suspicious characters are provided as inputs to the ANN model to learn patterns that distinguish phishing URLs from legitimate ones.

XG-Boost (Extreme Gradient Boosting) is an advanced ensemble learning algorithm based on decision trees that improves prediction performance by combining multiple weak learners into a strong model using gradient boosting techniques.

It sequentially builds trees where each new tree corrects the errors of the previous one, resulting in improved accuracy and robustness.

By combining ANN and XG-Boost, the proposed system benefits from the deep learning capability of ANN to capture complex feature relationships and the high efficiency and optimization ability of XG-Boost, leading to better phishing URL detection performance and improved classification accuracy/

V. FUTURE SCOPE

There are several future directions that can be explored to further improve and extend the proposed system for phishing URL detection using machine learning and ANN + XG-BOOST.

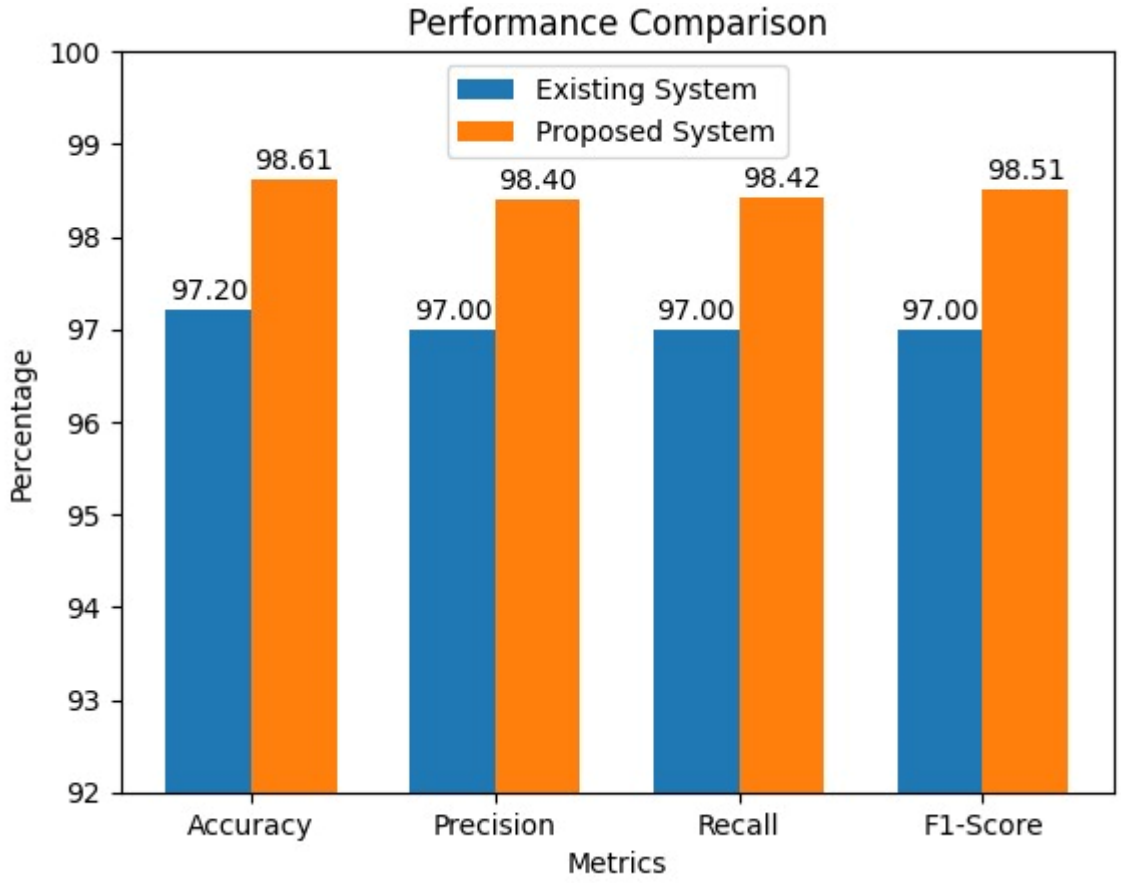
Firstly, the system can be enhanced by incorporating more advanced and complex features, such as user behavior and context- based features, to improve its detection accuracy and reduce the false-positive rate.

Secondly, the system can be integrated with other security solutions, such as firewalls and intrusion detection systems, to provide a comprehensive defense against phishing attacks.

Thirdly, the system can be extended to detect other types of cyber threats, such as malware and ransomware, using similar machine learning techniques.

Lastly, the system can be optimized for real-time processing to detect and respond to phishing attacks in near real-time, thereby reducing the impact of the attack on the targeted user.

Overall, there is significant potential for the proposed system to be further developed and refined, with the aim of providing better protection against phishing attacks and improving the overall security of online users.



VI. CONCLUSION

In conclusion, our proposed system for phishing URL detection using machine learning and ANN+XG-BOOST is a promising approach to improving the accuracy and effectiveness of phishing detection. By combining multiple features and using a powerful classification algorithm such as ANN+XG-BOOST, we are able to achieve high accuracy and precision in detecting phishing URLs while minimizing false positives. The system is designed to be scalable and adaptable to different domains and applications. The feature extraction and selection process can be customized to suit different types of phishing attacks and to incorporate additional features as needed. While there are some challenges to implementing this system, such as the need for a large and diverse dataset for training and testing, we believe that the benefits of improved phishing detection and prevention make it a worthwhile investment. Overall, we believe that our proposed system has the potential to significantly improve the security and safety of online users by helping to prevent phishing attacks and protect sensitive information.

VII. ACKNOWLEDGMENT

We would like to express our gratitude to our principle - Dr. A.V. Ram Prasad and HOD - Dr. J.S. Kanchana and our guide Prof. Dr. J.S. Kanchana who were a continual source of inspiration., for being of great support and guiding us through the research. Their extensive knowledge, experience and expertise enabled us to successfully complete this project. This effort would not have been possible without their help and supervision. This initiative would not be successful without the contribution of everyone. We were always there to encourage each other and that kept us together until the end.

REFERENCES

[1] A. Mughaid, S. AlZu'bi, A. Hnaif, S. Taamneh, A. Alnajjar, and E. A. Elsoud, ''An intelligent cyber security phishing detection system using deep learning techniques,'' *Cluster Comput.*, vol. 25, no. 6, pp. 3819–3828, Dec. 2022.

[2] E. Kocyigit, M. Korkmaz, O. K. Sahingoz, and B. Diri, ''Enhanced feature selection using genetic algorithm for machine-learning-based phishing URL detection,'' *Appl. Sci.*, vol. 14, no. 14, p. 6081, Jul. 2024.

[3] N. T. Singh, J. Rani, P. Sharma, A. Mishra, A. Yadav, and S. Raj, ''An innovative URL-based system approach with ML based prevention,'' in *Proc. IEEE Int. Conf. Comput., Power Commun. Technol. (IC2PCT)*, Feb. 2024, pp. 1498–1503.

[4] PhishLabs. (2024). *Quarterly Threat Trends and Intelligence Report*. [Online]. Available:

https://www.phishlabs.com/resources/reports/

[5] A. Raza, K. Munir, M. S. Almutairi, and R. Sehar, ‘‘Novel class probability features for optimizing network attack detection with machine learning,’’ *IEEE Access*, vol. 11, pp. 98685–98694, 2023.

[6] A. Raza, F. Rustam, B. Mallampati, P. Gali, and I. Ashraf, ‘‘Preventing crimes through gunshots recognition using novel feature engineering and meta-learning approach,’’ *IEEE Access*, vol. 11, pp. 103115–103131, 2023.

[7] A. Raza, K. Munir, M. Almutairi, and R. Sehar, ‘‘Novel transfer learning based deep features for diagnosis of down syndrome in children using facial images,’’ *IEEE Access*, vol. 12, pp. 16386–16396, 2024.

[8] Z. Huma and A. Mustafa, ‘‘Multi-modal data fusion techniques for improved cybersecurity threat detection and prediction,’’ *Aitoz Multidisciplinary Rev.*, vol. 3, no. 1, pp. 40–53, 2024.

[9] F. S. Alsubaei, A. A. Almazroi, and N. Ayub, ‘‘Enhancing phishing detection: A novel hybrid deep learning framework for cybercrime forensics,’’ *IEEE Access*, vol. 12, pp. 8373–8389, 2024.

[10] R. Goenka, M. Chawla, and N. Tiwari, ‘‘A comprehensive survey of phishing: Mediums, intended targets, attack and defence techniques and a novel taxonomy,’’ *Int. J. Inf. Secur.*, vol. 23, no. 2, pp. 819–848, Apr. 2024.

[11] A. Newaz, F. S. Haq, and N. Ahmed, ‘‘A sophisticated framework for the accurate detection of phishing websites,’’ 2024, *arXiv:2403.09735*.

[12] K. T. Smith, L. M. Smith, M. Burger, and E. S. Boyle, ‘‘Cyber terrorism cases and stock market valuation effects,’’ *Inf. Comput. Secur.*, vol. 31, no. 4, pp. 385–403, Oct. 2023.

[13] B. B. Gupta and M. Quamara, ‘‘An overview of Internet of Things (IoT): Architectural aspects, challenges, and protocols,’’ *Concurrency* Comput, *Pract. Exper.*, vol. 32, no. 21, Nov. 2020, Art. no. e4946.