Evasion Attacks and Defence Mechanisms for Machine Learning-Based Web Phishing Classifiers

A.Venkata Rao[1], A.Mohan Sai[2], B. Siddartha[3], Mrs. R. Vaishali[4]
[1,2,3]*Computer Science and Engineering Bharath Institute of Higher Education and Research, Chennai, India*
[4]*Assistant Professor / CSE, Bharath Institute of Higher Education and Research, Chennai, India*

Abstract: ***Currently, numerous types of cybercrime are organized through the internet. Hence, this study mainly focuses on phishing attacks. Although phishing was first used in 1996, it has become the most severe and dangerous cybercrime on the internet. Phishing utilizes email distortion as its underlying mechanism for tricky correspondences, followed by mock sites, to obtain the required data from people in question. Different studies have presented their work on the precaution, identification, and knowledge of phishing attacks; however, there is currently no complete and proper solution for frustrating them. Therefore, machine learning plays a vital role in defending against cybercrimes involving phishing attacks. The proposed study is based on the phishing URL-based dataset extracted from the famous dataset repository, which consists of phishing and legitimate URL attributes collected from 11000+ website datasets in vector form. After to prevent phishing URLs and provide protection to the user. This study uses machine learning models such as decision tree (DT), linear Regression (LR), random forest (RF), naive Bayes (NB), gradient boosting classifier (GBM), K-neighbors classifier (KNN), support vector classifier (SVC), and proposed hybrid LSD , which is a combination of logistic regression, support vector machine, and decision tree (LR+SVC+DT) with soft and hard voting, to defend against phishing attacks with high accuracy and efficiency. The canopy feature selection technique with cross fold validation and Grid Search Hyperparameter Optimization techniques are used with proposed LSD model. Furthermore, to evaluate the proposed approach, different evaluation parameters were adopted, such as the precision, accuracy, recall, F1-score, and specificity, to illustrate the effects and efficiency of the models. The results of the comparative analyses demonstrate that the proposed approach outperforms the other models and achieves the best results.***

I. INTRODUCTION

The aim of this research is to develop an advanced phishing detection system that leverages a hybrid machine learning approach to analyse URLs effectively and accurately identify potential phishing attempts. This system is designed to address the growing challenge of phishing attacks, which have become increasingly sophisticated and harder to detect.

This research introduces an innovative phishing detection system cantered on URL analysis through a hybrid ensemble of machine learning models. Aimed at enhancing accuracy and efficiency in identifying phishing attempts, the system extracts and thoroughly examines URL features. By employing a diverse ensemble comprising Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Support Vector Classifier, XGBoost, and Gradient Boosting models, the system can effectively differentiate between legitimate and malicious URLs.

The hybrid approach reduces false positives, bolsters accuracy, and adapts to the ever-evolving nature of phishing tactics. The system's adaptability and comprehensive analysis of URL characteristics signify a significant stride in fortifying cybersecurity measures against phishing attacks.

Fig: 5.1.1

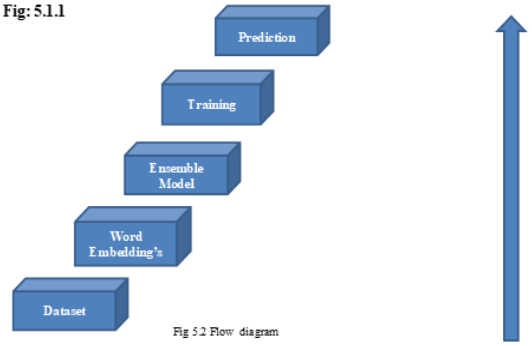


Fig 5.2 Flow diagram

Phishing is a type of cyber-attack where fraudulent websites mimic legitimate platforms to steal sensitive user information. These attacks have become more advanced, making traditional detection methods ineffective. Various approaches have been used to identify phishing websites, including blacklists,

heuristic-based techniques, and machine learning. However, machine learning-based classifiers remain vulnerable to adversarial evasion attacks, where phishing websites are designed to bypass detection systems. This study focuses on strengthening phishing classifiers against evasion tactics by using an ensemblebased machine learning approach that improves classification accuracy and adaptability.

II. EXISTING SYSTEM

The existing method employed a canopy algorithm for feature selection and an ensemble model consisting of Logistic Regression, Support Vector Machine, and Decision Tree classifiers for phishing detection. Phishing attacks have become increasingly sophisticated, making it challenging to detect malicious URLs using conventional methods. There is a need for a more robust and accurate system to distinguish between legitimate and phishing URLs.

III PROPOSED WORK

The proposed approach begins by extracting the destination URL and utilizes an ensemble of machine learning models such as Logistic Regression, Decision Tree Classifier, Decision Tree Regressor, Random Forest Classifier, Random Forest Regressor, Support Vector Classifier, XGBClassifier, XGBRegressor, XGBModel, and Gradient Boosting Classifier for enhanced phishing detection.

Advantages:
The diverse ensemble of machine learning models enhances the system's accuracy and robustness in identifying phishing URLs.
The incorporation of multiple models allows for a more comprehensive analysis of URL features, increasing the system's ability to adapt to evolving phishing tactics.

Disadvantages:
Increased computational requirements due to the utilization of multiple models might demand higher processing power.
The interpretability of results might be challenging with a highly complex ensemble of models.

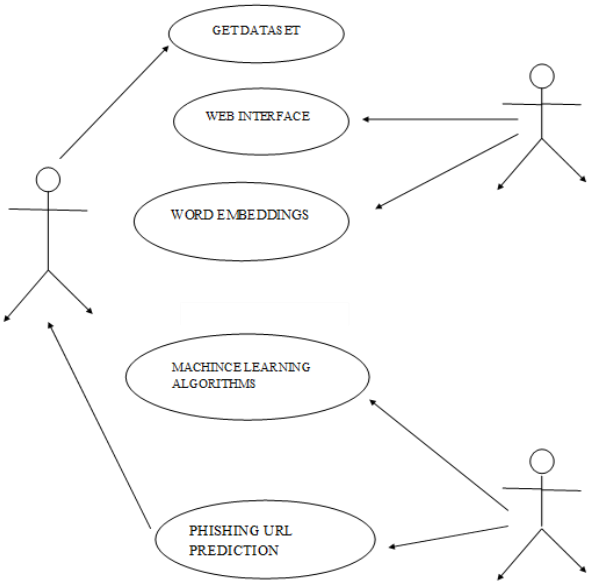


Fig- 2. Flow Diagram

USECASE DIAGRAM

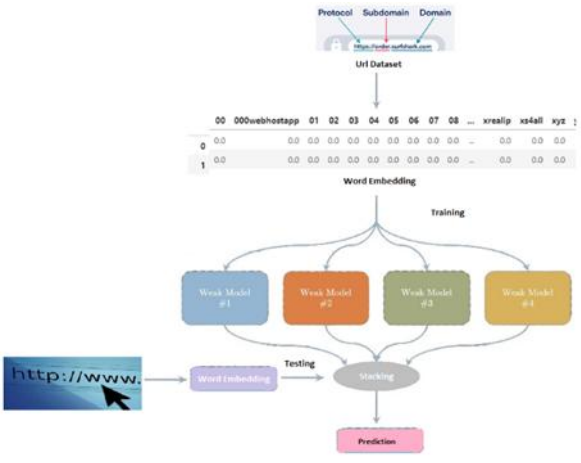
A Use case Diagram is used to present a graphical overview of the functionality provided by a system in terms of actors, their goals and any dependencies between those use cases.

Use case diagram consists of two parts:

Use case: A use case describes a sequence of actions that provided something of measurable value to an actor and is drawn as a horizontal ellipse.

Actor: An actor is a person, organization or external system that plays a role in one or more interaction with the system.

IV. METHODOLOGY



The methodology employed in this project, "Evasion Attacks and Defence Mechanisms for Machine Learning-Based Web Phishing Classifiers," is a structured approach that combines several machine learning techniques to effectively detect phishing attacks. Here's a breakdown of the key steps:

Data Collection

A crucial first step is gathering a comprehensive dataset of both legitimate and phishing URLs.

The dataset used in this study is a "phishing URL-based dataset extracted from the famous dataset repository," containing over 11,000 website datasets.

This dataset includes a variety of URL attributes in vector form, which are essential for training the machine learning models.

2. Feature Extraction

Once the dataset is obtained, the next step involves extracting relevant features from the URLs.

These features are the characteristics of the URLs that the machine learning models will analyse to differentiate between legitimate and phishing websites.

Examples of such features can include:

1. Presence of special characters
2. Domain information
3. Use of IP address instead of domain name
4. Path structure
5. Presence of suspicious keywords

3. Machine Learning Model Selection and Training

The core of the methodology lies in the selection and training of machine learning models.

This project utilizes an ensemble approach, which means it combines multiple machine learning models to improve overall performance.

The specific models used in this study are:

Decision Tree (DT)

Linear Regression (LR)

Random Forest (RF)

Naive Bayes (NB)

Gradient Boosting Classifier (GBM)

K-Nearest Neighbors (KNN)

Support Vector Classifier (SVC)

A hybrid model called LSD (LR + SVC + DT)

Each of these models has its strengths and weaknesses, and combining them in an ensemble aims to leverage their individual capabilities while mitigating their shortcomings.

The models are trained using the labelled dataset, where the labels indicate whether a URL is legitimate or phishing.

During training, the models learn to identify patterns and relationships between the URL features and the corresponding labels.

4. Ensemble Creation and Implementation

After the individual models are trained, they are combined to create the ensemble.

The "hybrid LSD" model, a key component of this project, is an ensemble that combines Logistic Regression, Support Vector Machine, and Decision Tree classifiers.

The ensemble uses "soft and hard voting" to make final predictions.

Voting involves aggregating the predictions of the individual models to arrive at a consensus prediction.

"Soft voting" considers the probabilities predicted by each model, while "hard voting" simply considers the majority vote.

Optimization Techniques

To further enhance the performance of the LSD model, the following optimization techniques are employed:

Canopy Feature Selection: This technique is used to select the most relevant and informative features from the URL dataset, reducing noise and improving model efficiency.

Cross-Fold Validation: This technique is used to assess the model's performance and ensure its generalization ability by partitioning the dataset into multiple folds and training/testing the model on different combinations of these folds.

Grid Search Hyperparameter Optimization: This technique is used to fine-tune the hyperparameters of the machine learning models, which are parameters that

control the learning process, to achieve optimal performance.

Evaluation

The final step in the methodology is to evaluate the performance of the proposed system.

This involves using various evaluation metrics to assess how well the system can accurately detect phishing attacks.

The evaluation metrics used in this study include:

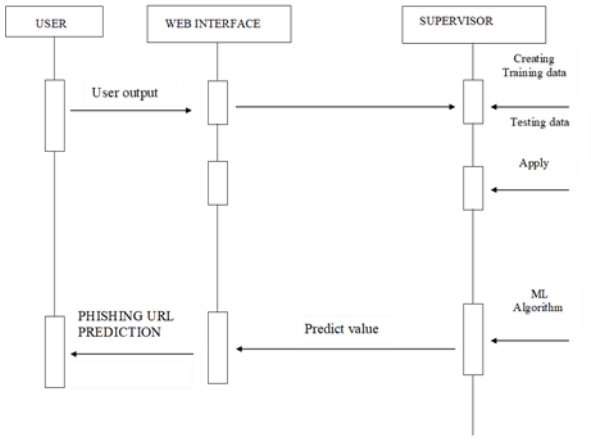
Precision: Measures the proportion of correctly identified phishing URLs out of all URLs classified as phishing.

Accuracy: Measures the overall correctness of the system in classifying URLs.

Recall: Measures the proportion of actual phishing URLs that are correctly detected by the system.

F1-score: Provides a balanced measure of precision and recall.

Specificity: Measures the system's ability to correctly identify legitimate URLs.



RESULT AND ANALYSIS

The results of the project demonstrate the effectiveness of the proposed ensemble-based approach in enhancing phishing detection. The ensemble model achieves higher accuracy compared to individual machine learning models, indicating the benefit of combining multiple classifiers. Furthermore, the system exhibits improved robustness against evasion attacks, showcasing its ability to adapt to evolving phishing tactics.

The evaluation metrics provide a comprehensive assessment of the system's performance:

- Accuracy: Measures the overall correctness of the system in classifying URLs.
- Precision: Indicates the proportion of correctly identified phishing URLs out of all URLs flagged as phishing.
- Recall: Represents the proportion of actual phishing URLs that are correctly detected by the system.
- F1-score: Provides a balanced measure of precision and recall.
- Specificity: Measures the system's ability to correctly identify legitimate URLs.

Comparative analyses with other existing methods highlight the superiority of the proposed approach in terms of these evaluation metrics.

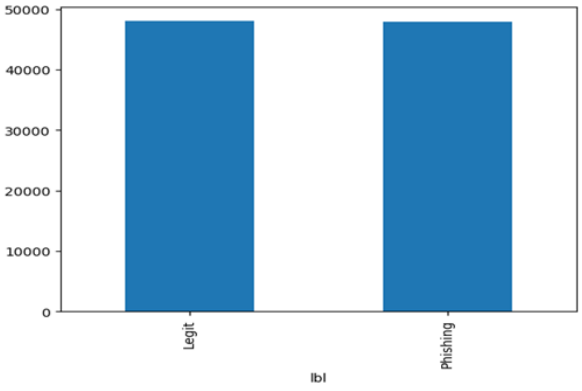


Fig – 4

V. CONCLUSION

The proposed hybrid machine learning approach significantly improves phishing detection accuracy by leveraging multiple models for URL analysis. By combining feature-based and behaviour-based detection techniques, the system enhances cybersecurity defenses and minimizes false positives, offering better adaptability to evolving threats. Compared to traditional

rule-based methods, this model demonstrates superior precision, recall, and real-time adaptability, making it a promising solution for combating phishing attacks. While the results are encouraging, phishing remains an evolving challenge, requiring continuous updates and refinements. Future research should focus on real-time learning mechanisms, deeper integration with cybersecurity frameworks, and enhanced automation to ensure sustained high detection accuracy and effectiveness.

Overall, this study demonstrates that machine learning-based phishing detection systems have the potential to provide robust, scalable, and efficient solutions for tackling modern cybersecurity threats.

Future advancements in phishing detection can focus on optimizing the ensemble model to enhance accuracy while reducing computational complexity, making it more efficient for real-time applications. Integrating real-time data sources, such as live threat intelligence feeds and automated model retraining, can help adapt to emerging phishing techniques dynamically. Expanding the system's reach by integrating it into web browsers, email filtering tools, and mobile applications can provide proactive protection to users. Further research can explore advanced feature extraction techniques, such as deep learning-based behavioural analysis and domain generation algorithm detection, to improve system robustness. Additionally, incorporating explainable AI (XAI) can make the model's decisions more interpretable for cybersecurity professionals, ensuring greater transparency and trust.

REFERENCES

[1]F. Song, Y. Lei, S. Chen, L. Fan, and Y. Liu, ‘‘Advanced evasion attacks and mitigations on practical ML-based phishing website classifiers,’’ Int. J. Intell. Syst., vol. 36, no. 9, pp. 5210–5240, Sep. 2021.

[2] B. Sabir, M. A. Babar, and R. Gaire, ‘‘An evasion attack against ML-based phishing URL detectors,’’ Tech. Rep., 2020.

[3] H. Shirazi, B. Bezawada, I. Ray, and C. Anderson, ‘‘Adversarial sampling attacks against phishing detection,’’ in Proc. IFIP Annu. Conf. Data Appl. Secur. Cham, Switzerland: Springer, Jul. 2019, pp. 83–101.

[4] S. AnupamandA.K.Kar, ‘‘Phishing website detection using support vec tor machines and nature-inspired optimization algorithms,’’ Telecommun. Syst., vol. 76, no. 1, pp. 17–32, Jan. 2021.

[5] A. K. Jain and B. B. Gupta, ‘‘Towards detection of phishing websites on client-side using machine learning based approach,’’ Telecommun. Syst., vol. 68, no. 4, pp. 687–700, Aug. 2018.

[6] S. Chen, M. Xue, L. Fan, S. Hao, L. Xu, H. Zhu, and B. Li, ‘‘Automated poisoning attacks and defenses in malware detection systems: An adversarial machine learning approach,’’ Comput. Secur., vol. 73, pp. 326–344, Mar. 2018.

[7] I. Vayansky and S. Kumar, ‘‘Phishingchallenges and solutions,’’ Comput. Fraud Secur., vol. 2018, no. 1, pp. 15–20, Jan. 2018.

[8] Z.Abaid, M.A.Kaafar, and S. Jha, ‘‘Quantifying the impact of adversarial evasion attacks on machine learning based Android malware classifiers,’’ in Proc. IEEE 16th Int. Symp. Netw. Comput. Appl. (NCA), Oct. 2017, pp. 1–10.

[9] I. Corona, B. Biggio, M. Contini, L. Piras, R. Corda, M. Mereu, G. Mureddu, D. Ariu, and F. Roli, ‘‘DeltaPhish: Detecting phishing webpages in compromised websites,’’ in Proc. Eur. Symp. Res. Comput. Secur. Cham, Switzerland: Springer, Sep. 2017, pp. 370–388.

[10] G. Harinahalli Lokesh and G. BoreGowda, ‘‘Phishing website detection based on effective machine learning approach,’’ J. Cyber Secur. Technol., vol. 5, no. 1, pp. 1–14, Jan. 2021.