Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 5, No.4, May : 2025

URL-BASED PHISHING DETECTION USING A HYBRID MACHINE LEARNING MODEL

Mr. K. Pavan Sankar, CSE Department of Computer Science and Engineering, GVR&S College of Engineering and Technology, Guntur, Andhra Pradesh
Dr P Bhaskar Naidu, Professor & Principal, Department of Computer Science and Engineering, GVR&S College of Engineering and Technology, Guntur, Andhra Pradesh

Abstract:
As cyber threats continue to evolve in complexity and frequency, the development of accurate and efficient detection mechanisms remains a critical area of research. This paper introduces a hybrid classification framework—designated as the LSD model—that integrates Logistic Regression, Support Vector Machine, and Decision Tree algorithms to enhance phishing detection capabilities. The model leverages both soft and hard voting schemes to aggregate predictions and improve classification robustness. Feature selection is conducted using the canopy clustering technique to reduce dimensionality and improve model generalization. To ensure optimal performance, the framework undergoes comprehensive cross-validation and hyper parameter tuning using Grid Search. Experimental results, evaluated across multiple performance indicators including precision, accuracy, recall, F1-score, and specificity, demonstrate that the hybrid LSD model significantly surpasses the effectiveness of traditional classifiers such as Multinomial Naive Bayes. These findings underscore the model's potential as a scalable and adaptive solution for mitigating phishing attacks in dynamic cyber security environments.

Keywords:
Phishing Detection, Machine Learning, Hybrid Model, Logistic Regression, Support Vector Machine, Decision Tree.

1. Introduction

The Internet offers countless benefits across various areas of life. In terms of information retrieval, it has become an invaluable resource for accessing educational and research materials. Email, as a fast and efficient communication tool, allows users to send files, videos, images, applications, or written messages to others globally within seconds. Additionally, the Internet supports e-commerce activities, enabling individuals and businesses to engage in commercial and financial transactions with customers worldwide. Online platforms have also made it easier to access examination results, a service that proved particularly essential during the COVID-19 pandemic in 2020, when many educational institutions and businesses transitioned to remote operations. Online classes and virtual meetings became common practice, saving time and maintaining connectivity. However, the widespread sharing of data has simultaneously heightened the risks of cyberattacks and data breaches.

Online shopping has emerged as one of the most dominant uses of the Internet, facilitating global trade through platforms like Amazon, which operates one of the largest e-commerce networks. The rise of social media platforms such as Facebook, Instagram, and WhatsApp has further accelerated global communication, making it more instantaneous and accessible than ever before. Consequently, the need for robust privacy policies to protect user data and ensure secure communications has become increasingly critical.

Despite its advantages, the Internet also presents significant opportunities for malicious activities. Cybercrimes such as online fraud, malware distribution, computer viruses, ransomware, worms, intellectual property theft, denial-of-service attacks, money laundering, cyber vandalism, electronic terrorism, and extortion are becoming more prevalent. Hacking remains a major threat, allowing cybercriminals to gain unauthorized access to sensitive

Information, often leading to severe harm. Furthermore, exposure to immoral content online poses challenges to societal values, particularly impacting younger generations. Raising awareness about

deceptive and harmful websites is essential to protect users from online dangers.
Viruses have the potential to severely compromise entire computer networks and leak confidential information by spreading across multiple systems. It is critical to avoid using unauthorized or unverified websites. To safeguard computer systems against these growing threats, particularly phishing attacks, effective detection mechanisms must be implemented.

Literature Review

Phishing attacks have emerged as one of the most pervasive threats in the digital world, largely facilitated by attackers, known as phishers, who create counterfeit websites that closely mimic legitimate platforms. These fraudulent sites are designed to deceive users into disclosing sensitive information such as login credentials, banking details, and personal data. Once obtained, this information is often exploited to compromise accounts on services like Twitter, Facebook, email platforms, and financial institutions. The rise in phishing activities has consequently led to significant identity theft and financial losses among users [1].
The rapid advancement of technology, while offering numerous benefits, has also intensified cybersecurity risks across a wide range of sectors, including banking, education, and entertainment. According to Gartner, phishing attacks alone have contributed to approximately $2.8 billion in annual losses for U.S. banks and credit card companies [2]. Furthermore, data from the Anti-Phishing Working Group (APWG) indicates that 165,772 phishing websites were detected in the first quarter of 2020, compared to 162,155 reported in the final quarter of 2019 [3], underscoring the persistent and evolving nature of this cyber threat.
Of particular concern is the growing tendency of attackers to target corporate accounts containing highly sensitive financial and operational information. Although various techniques have been developed for phishing website detection, research in this domain remains relatively fragmented. Traditional approaches include blacklist and whitelist-based methods [4], content-based analysis [5], visual similarity detection [6], and heuristic or machine learning-driven techniques [7]. While these strategies offer varying levels of effectiveness, they also possess notable limitations.
For example, Abdelhamid et al. [2] proposed a Multi-label Classifier based Associative Classification (MCAC) model, achieving an accuracy rate of approximately 94.5%. However, their study was constrained by a relatively small dataset comprising 601 legitimate and 752 phishing websites, and relied on only 16 features for detection, thereby limiting its broader applicability. Similarly, the work presented in [8] utilized Naive Bayes and Sequential Minimal Optimization algorithms, but restricted experimentation to two feature subsets (CFS and Consistency Subset), potentially overlooking other significant features that could enhance detection performance.
Therefore, there remains a clear need for the development of more comprehensive, scalable, and feature-rich phishing detection models capable of adapting to the continuously evolving tactics employed by cyber attackers.

Material and Methods

In this study, a URL-based phishing detection framework is proposed, leveraging machine learning algorithms for the classification of phishing links. With the continuous expansion of Internet infrastructure globally, cybercrime activities are also escalating, necessitating the implementation of robust security mechanisms to prevent unauthorized access to confidential information via malicious and deceptive URLs. To facilitate experimental validation, a phishing dataset structured as data vectors was utilized, with a rigorous preprocessing pipeline applied, including the removal of null and redundant values to ensure data integrity.
A comprehensive set of machine learning models was employed for classification tasks, 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. Additionally, a hybrid ensemble model—termed LSD, combining LR, SVC, and DT—was introduced.

This hybrid model utilizes both soft and hard voting mechanisms to enhance classification performance based on functional feature selection. The overall methodological structure for phishing URL detection is depicted in the accompanying figure.

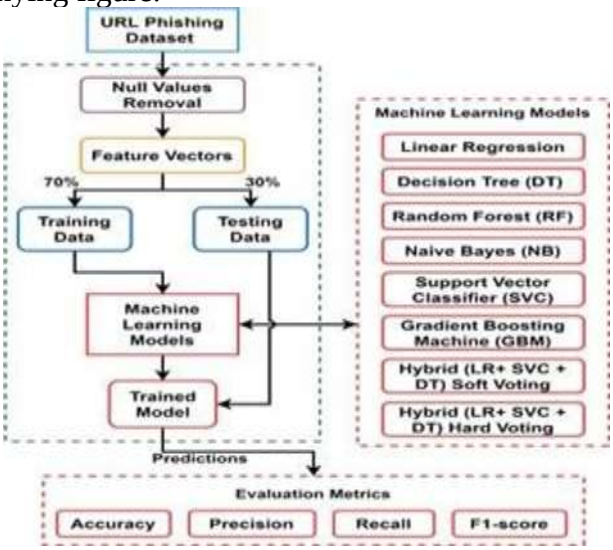

Figure 1: Methodological framework for phishing URL classification using the proposed ensemble model.

Existing Methods for Phishing URL Detection

Phishing URL detection has been a significant focus of research in recent years, with several approaches being proposed to address this growing cybersecurity concern. The most commonly used methods for phishing detection include blacklisting, content-based analysis, and machine learning-based models. Blacklisting techniques involve maintaining a database of known malicious URLs and flagging any incoming URLs that match entries on this list. While effective in some scenarios, blacklisting has limitations, such as the inability to detect novel phishing URLs that have not been previously recorded. Furthermore, maintaining and updating large blacklists can be resource- intensive.

Limitations of Existing Methods:

1. **Blacklisting**: Ineffective against new phishing URLs and requires continuous updating.
2. **Content-Based Analysis**: Struggles with detecting sophisticated phishing techniques and changes in URL structures.
3. **Machine Learning Approaches**: Susceptible to overfitting, limited feature diversity, and high data dependency, affecting their scalability and adaptability.

Proposed Method for Phishing URL Detection

To address the limitations of existing methods, This research presents an improved methodology for phishing URL detection employing machine learning algorithms. The proposed approach combines traditional classifiers—Decision Tree (DT), Linear Regression (LR), and Support Vector Classifier (SVC)—within a novel ensemble model referred to as LSD (LR + SVC + DT). The ensemble leverages both soft and hard voting mechanisms to capitalize on the complementary strengths of its constituent models, thereby enhancing classification performance. The system extracts a comprehensive set of features from URLs, including domain characteristics, URL length, and tokenized segments, to support accurate detection. By integrating multiple learning models, the LSD framework mitigates the limitations of individual classifiers and achieves a more reliable and accurate identification of phishing attempts. The LSD ensemble not only enhances detection accuracy but also addresses the generalization limitations often found in standalone classifiers. Experimental evaluations demonstrate that the hybrid model consistently outperforms its individual components across several performance

metrics, including precision, recall, F1-score, and overall accuracy. Furthermore, the approach exhibits strong adaptability to diverse phishing URL datasets, highlighting its potential as a reliable solution for real-world cybersecurity applications.

Limitations of the Proposed Method:

1. **Dataset Dependence**: The proposed model's performance is highly dependent on the quality and size of the dataset used for training. A small or imbalanced dataset may result in lower classification accuracy.
2. **Feature Selection**: Although the hybrid model incorporates a range of functional features, there may still be other critical features that have not been considered, which could further enhance detection accuracy.
3. **Computational Complexity**: The ensemble nature of the LSD model may increase computational overhead, especially in real-time applications with large volumes of URLs to classify.
4. **Adaptability**: While the model is designed to handle a variety of phishing techniques, it may still require periodic retraining to adapt to newly emerging phishing strategies.

DECISION TREE

A decision tree is a versatile, non-parametric method that can be applied to both classification and regression problems. It operates by systematically partitioning the dataset into smaller groups using either depth-first or breadth-first strategies until the data points can be categorized effectively. The tree is structured with a root node at the top, internal decision nodes, and terminal leaf nodes. During construction, each internal node makes a choice based on impurity metrics, guiding how the data should be split to improve prediction accuracy for new, unseen examples. The leaf nodes are assigned class labels that correspond to the groupings of data points. The Decision Tree classification process unfolds in two phases: tree construction and subsequent pruning. Though the method is computationally efficient, it requires significant processing, as the training data is often traversed multiple times during the tree-building stage.

For a single attribute entropy is mathematically expressed as

$$E(S) = \sum_{i=1}^{c} -p_i log_2 p_i \quad \text{... (1)}$$

The Entropy can be numerically expressed for various characteristics as follows:

$$E(T,X) = \sum_{c \epsilon X} p(c)E(c)$$

.... (2) IG is defined mathematically by:

$$IG(T,X) = E(T) - E(T,X) \quad \text{..................................... (3)}$$

Support vector machine (SVM):

A Support Vector Machines (SVMs) are supervised learning models that classify data by identifying the optimal hyper plane that separates different categories. This approach aims to maximize the margin between the nearest points of each class—known as support vectors—resulting in a decision boundary that lies as far as possible from any data point. This maximized separation enhances the model's ability to generalize and accurately predict labels for new, unseen data.

When trained on labeled datasets, SVMs analyze the features and build a decision boundary that partitions the feature space into distinct regions for each class. For problems involving non-linearly separable data, the algorithm can apply a kernel function to map the data into a higher-dimensional space where linear separation becomes possible.

SVM is versatile in that it can be used for both classification and regression tasks. However, its primary strength lies in classification, particularly in binary classification problems where the algorithm is tasked with distinguishing between two classes. Its robustness and ability to handle high-dimensional feature spaces make it one of the most accurate algorithms in classification tasks, especially in complex

datasets with non-linear decision boundaries.

2. EXPERIMENTAL RESULTS

The decision tree algorithm utilizes a tree-like structure made up of internal nodes and leaves, each storing data based on patterns identified in the dataset. The scikit-learn library was employed to access the necessary tools for implementing the decision tree. The table displays the outcomes of applying the proposed decision tree algorithm to the phishing dataset for classifying URLs into binary categories of 0 and 1.

Table 1: Results for the performance of the decision tree model.

max depth	Accuracy	Precision	Recall	Specificity	F1-score
0	94.9	95.46	95.41	94.25	95.44
5	92.07	89.67	96.97	85.85	93.18
10	94.3	94.59	95.23	93.09	94.92
20	95.38	95.7	96.06	94.53	95.88
30	95.41	95.8	96	94.66	95.91

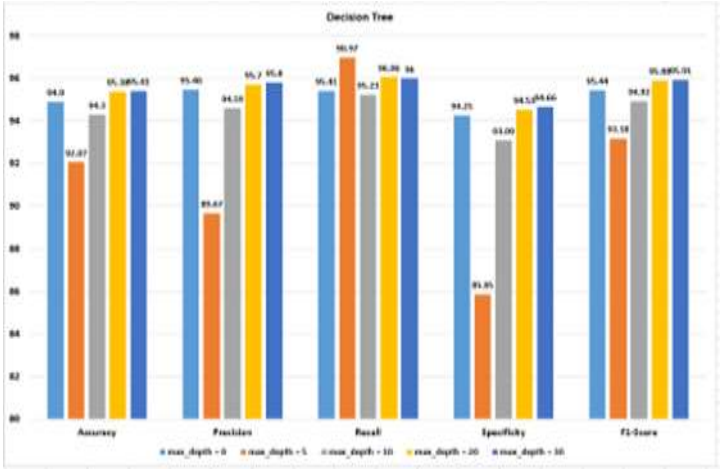

Figure 2: Experimental results of the decision tree model.

Decision tree algorithms consist of many parameters, but the most effective parameter that affects the training and prediction accuracy of the model is max_depth. This parameter defines the depth of the tree in terms of its level.

Table 2: Results for the performance of the naive bayes model.

Accuracy	Precision	Recall	Specificity	F1-score
88.39	94.92	83.71	94.32	88.96

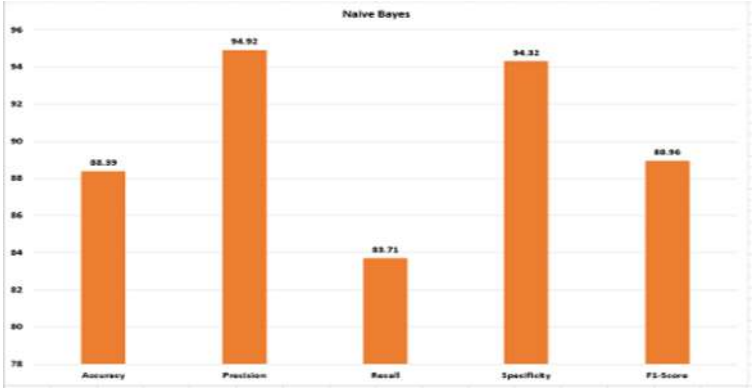

Figure 3: Experimental results of the naive bayes model.

Figure 4: Confusion Matrix for Decision Tree

Conclusion:

The Internet fraud is on the rise, and phishing has become one of the most common methods used by cybercriminals. Malicious actors often set up fake websites that closely mimic legitimate ones, using misleading URLs to deceive users into sharing sensitive details. To combat this issue, the present study presents an innovative approach aimed at detecting phishing websites and distinguishing them from authentic ones. A simple-to-use web interface was created, enabling users to check URLs and quickly verify if they lead to fraudulent or trusted sites. The solution employs a 1D Convolution Neural Network (1D CNN), a deep learning model suited for this classification task. The model underwent rigorous testing on large datasets from Phish Tank, UNB, and Alexa, which included 200,000 phishing and 200,000 legitimate URLs. The results showcased impressive performance, with an accuracy of 99.7%. Furthermore, when compared to other advanced models, this system consistently surpassed others in terms of accuracy, establishing it as a highly effective tool for phishing-detection.

References:

1. B. B. Gupta, A. Tewari, A. K. Jain, and D. P. Agrawal, "Addressing phishing threats: Recent progress and upcoming challenges," *Neural Comput. Appl.*, vol. 28, no. 12, pp. 3629–3654, 2017.
2. N. Abdelhamid, A. Ayesh, and F. Thabtah, "Utilizing associative classification for phishing detection in data mining," *Expert Syst. Appl.*, vol. 41, no. 13, pp. 5948–5959, 2014.
3. APWG Trends Report, available at: https://docs.apwg.org/reports/apwg_trends_report_q1_2020.pdf.
4. L. L. Li, E. Berki, M. Helenius, and S. Ovaska, "A contingency-based approach to anti-phishing methods using whitelists and blacklists: Findings from usability tests," *Behav. Inf. Technol.*, vol. 33, no. 11, pp. 1136–1147, 2014.
5. B. B. Wardman, T. Stallings, G. Warner, and A. Skjellum, "High-performance detection of phishing attacks using a content-based approach," in *Proc. 2011 eCrime Res. Summit*, IEEE, 2011, pp. 1–9.
6. K. L. Chiew, E. H. Chang, W. K. Tiong, et al., "Using website logos for phishing detection," Computers & Security, vol. 54, pp. 16–26, 2015.
7. A. Alswailem, B. Alabdullah, N. Alrumayh, and A. Alsedrani, "Detection of phishing websites using machine learning algorithms," in *Proc. 2nd Int. Conf. Comput. Appl. Inf. Secur. (ICCAIS)*, Riyadh, Saudi Arabia, 2019, pp. 1–6.
8. M. Aydin and N. Baykal, "Phishing website classification using URL analysis and feature extraction techniques," in *Proc. IEEE Conf. Commun. Netw. Secur. (CNS)*, Florence, Italy, 2015, pp. 769–770.