A Hybrid Machine Learning–Based Phishing Detection System Using Enhanced URL Feature Representation

[1]*Kamuni Nandini, Department of Computer science and Engineering*
Malla Reddy (MR) Deemed to be University, Maisammaguda, Dhulapally(post.via.Kompally), Medchal,Malkangiri Telangana.

[2]*Dr U. Mohan Srinivas, Professor and HOD, Department of Computer science and Engineering*
Malla Reddy (MR) Deemed to be University, Maisammaguda, Dhulapally(post.via.Kompally), Medchal,Malkangiri Telangana

Abstract:
In the rapidly evolving digital world, cybersecurity threats have become a major concern, with phishing attacks being one of the most common and harmful forms of cybercrime. Phishing involves creating fraudulent websites or malicious URLs that imitate legitimate platforms to deceive users into revealing sensitive information such as login credentials, banking details, and personal data. Traditional phishing detection techniques, including blacklist-based and heuristic methods, are often ineffective against newly emerging and sophisticated attacks. This project presents an intelligent phishing detection system based on machine learning techniques, focusing on URL-based analysis. A comprehensive dataset containing more than 11,000 phishing and legitimate URLs is used for training and testing the models. Various machine learning algorithms such as Decision Tree, Random Forest, Naive Bayes, Support Vector Machine, K-Nearest Neighbors, and Gradient Boosting are implemented to classify URLs. To enhance performance, a hybrid model named LSD (Logistic Regression + Support Vector Machine + Decision Tree) is proposed using ensemble voting techniques. Additionally, feature selection methods, cross-validation, and Grid Search hyperparameter optimization are applied to improve accuracy and efficiency. The system is evaluated using metrics such as accuracy, precision, recall, F1-score, and specificity. Experimental results demonstrate that the proposed hybrid model outperforms individual machine learning models, providing higher accuracy and reduced false positives. The developed system offers a reliable and scalable solution for real-time phishing detection, thereby enhancing user security and protecting sensitive information in online environments.

Introduction

The internet has become an integral part of modern life, enabling communication, online banking, e-commerce, education, healthcare, and many other essential services. With the rapid growth of internet usage, there has also been a significant increase in cyber threats and malicious activities. Among these threats, phishing attacks have emerged as one of the most dangerous and widespread forms of cybercrime Phishing is a fraudulent technique in which attackers create fake websites or send deceptive links that appear to be from legitimate organizations such as banks, social media platforms, or e-commerce websites. The primary goal of these attacks is to trick users into revealing sensitive information such as usernames, passwords, credit card details, and personal data. As phishing techniques become more sophisticated, it becomes increasingly difficult for users to distinguish between legitimate and malicious websites Traditional phishing detection methods, such as blacklist and whitelist approaches, rely on previously identified malicious URLs. While these methods are simple and easy to implement, they are not effective against newly generated phishing websites. Heuristic-based methods also have limitations, including high false positive and false negative rates. Therefore, there is a need for more advanced and intelligent systems capable of detecting phishing attacks in real-time Machine learning has emerged as a powerful solution for addressing cybersecurity challenges. By analyzing patterns and features in large

datasets, machine learning models can learn to differentiate between phishing and legitimate URLs without requiring explicit programming. These models can automatically adapt to new threats and improve their performance over time In this project, a phishing detection system is developed using multiple machine learning algorithms, including Decision Tree, Random Forest, Naive Bayes, Support Vector Machine, and K-Nearest Neighbors. Furthermore, a hybrid model named LSD, which combines Logistic Regression, Support Vector Machine, and Decision Tree, is proposed to enhance detection accuracy and reliability. The system uses a dataset of over 11,000 URLs and applies feature selection techniques, cross-validation, and hyperparameter optimization to achieve better performance The primary aim of this project is to build an efficient and scalable phishing detection system that can accurately identify malicious URLs and protect users from cyber threats. By leveraging the power of machine learning and hybrid modeling techniques, the proposed system contributes to improving cybersecurity and ensuring safer online interactions.

Existing System

The existing systems for phishing detection mainly rely on traditional approaches such as blacklist/whitelist methods, heuristic techniques, and individual machine learning models. These systems aim to identify phishing websites by analyzing URL characteristics, webpage content, email features, and domain information One of the most commonly used approaches is the blacklist method, where known phishing URLs are stored in a database. Whenever a user tries to access a website, the system checks the URL against this list. If the URL is found in the blacklist, it is flagged as malicious. Similarly, whitelist methods allow only trusted websites. However, these approaches are limited because they can only detect previously identified phishing websites and fail to recognize new or unknown attacks Another approach is **heuristic-based detection**, which uses predefined rules such as URL length, presence of special characters, suspicious domain names, and abnormal webpage behavior. While this method can detect some phishing attempts, it often produces inaccurate results due to rigid rule definitions and lack of adaptability In recent years, **machine learning-based systems** have been introduced, where models such as Decision Tree, Naive Bayes, and Support Vector Machine are trained on phishing datasets. These systems analyze patterns in data to classify URLs as legitimate or phishing. Although they provide better accuracy compared to traditional methods, using a single algorithm may not always deliver optimal performance.

Disadvantages of Existing System

- **Inability to Detect New Attacks:** Blacklist methods cannot identify newly generated phishing websites.
- **High False Positives and Negatives:** Heuristic and single-model approaches often misclassify URLs.
- **Lack of Adaptability:** Existing systems do not easily adapt to evolving phishing techniques.
- **Maintenance Overhead:** Continuous updating of databases and rules is required.
- **Scalability Issues:** Some systems struggle with large datasets and real-time detection.
- **Limited Accuracy:** Single machine learning models may not capture complex patterns effectively.

Proposed System

The proposed system is an advanced phishing detection system based on hybrid machine learning techniques. It is designed to overcome the limitations of traditional blacklist, heuristic, and single-model approaches by providing a more accurate, adaptive, and efficient method for identifying phishing URLs. The system mainly focuses on analyzing the characteristics of URLs and classifying them as either legitimate or phishing a dataset containing more than 11,000 URLs is used. The dataset includes both phishing and legitimate website links with multiple URL-based features. These features

are carefully processed and used as input for machine learning algorithms. Before model training, the dataset undergoes data preprocessing, which includes cleaning, handling missing values, and converting the data into a suitable format for classification The core strength of the proposed system lies in the development of a hybrid model called LSD, which combines Logistic Regression (LR), Support Vector Machine (SVM), and Decision Tree (DT). Instead of depending on a single classifier, this hybrid model uses the strengths of multiple algorithms to produce better results. The system applies both soft voting and hard voting ensemble techniques, where the final prediction is made based on the combined decisions of all three models. This improves the stability and reliability of phishing detection To further increase performance, the proposed system uses Canopy Feature Selection to identify the most relevant features from the dataset. This reduces unnecessary data, improves training efficiency, and increases detection accuracy. In addition, Cross-Validation is applied to ensure that the model performs consistently across different subsets of data. Grid Search Hyperparameter Optimization is also used to tune the model parameters and achieve the best possible results. The proposed system is evaluated using important performance metrics such as Accuracy, Precision, Recall, F1-Score, and Specificity. Based on experimental results, the hybrid model performs better than individual machine learning algorithms and provides more reliable phishing detection. The system can therefore be used as an effective cybersecurity solution for protecting users from malicious websites and online fraud.

Advantages of Proposed System

- Provides **higher accuracy** compared to existing systems
- Detects both **known and unknown phishing websites**
- Reduces **false positives and false negatives**
- Uses **hybrid machine learning model** for better performance
- Improves efficiency through **feature selection and optimization**
- Scalable for handling large datasets
- Suitable for **real-time phishing detection**
- Enhances **user security and online safety**

Literature Survey

Phishing detection has gained significant attention due to the rapid increase in cyber-attacks and online fraud. Recent research focuses on improving detection accuracy, adaptability, and real-time performance using advanced machine learning and deep learning techniques.

[1] A deep learning–based phishing detection system using Convolutional Neural Networks (CNN) demonstrates that automatic feature extraction from URLs can significantly improve detection performance. Unlike traditional methods, deep learning models reduce dependency on manual feature engineering and enable real-time detection of malicious URLs.

[2] Research using Explainable Artificial Intelligence (XAI) highlights that phishing detection features may vary across datasets. The study shows that models trained on one dataset may not generalize well to others, emphasizing the importance of robustness and cross-dataset validation in phishing detection systems.

[3] A systematic review of phishing detection techniques based on machine learning and neural networks categorizes approaches into feature-based, content-based, and hybrid models. The study concludes that ensemble and neural network-based methods provide better accuracy and adaptability compared to traditional approaches.

[4] A scalable machine learning framework for phishing detection introduces a system capable of identifying phishing campaigns in real time. The approach focuses on handling large-scale data efficiently and emphasizes privacy-preserving mechanisms in detection systems.

[5] Advanced machine learning models combined with optimization-supported deep

learning techniques have been proposed for malicious URL detection. These approaches improve classification performance by integrating feature selection and hyperparameter optimization strategies.

[6] Temporal Convolutional Network (TCN)-based models have been developed to analyze URL sequences for phishing detection. These models capture temporal patterns and structural characteristics of URLs, providing improved accuracy and explainability in detection systems.

[7] A comprehensive survey of malicious URL detection techniques identifies key challenges such as lack of standardized datasets, limited benchmarking, and insufficient coverage of modern deep learning models. The study highlights the need for scalable, adaptable, and real-time detection systems.

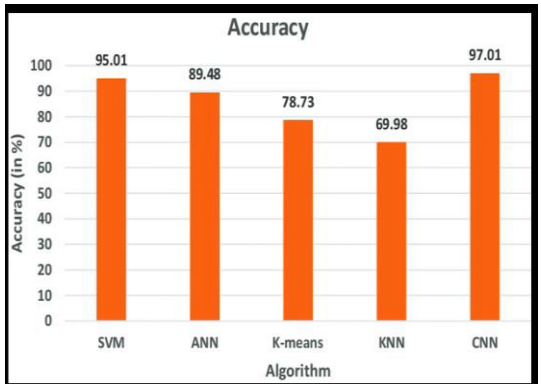
[8] Multimodal phishing detection systems integrate multiple data sources such as URL features, webpage content, and visual information. These systems improve robustness and accuracy by combining different types of features and using explainable AI techniques.

[9] Recent research explores the use of small language models for phishing detection, focusing on balancing performance, cost, and privacy. These models provide a practical alternative to large-scale models, especially for real-time and resource-constrained environments.

[10] Lightweight transformer-based models have been introduced for malicious URL detection. These models achieve high accuracy while maintaining efficiency, making them suitable for deployment in real-time cybersecurity applications.

Outcomes:

Accuracy Comparison Graph:

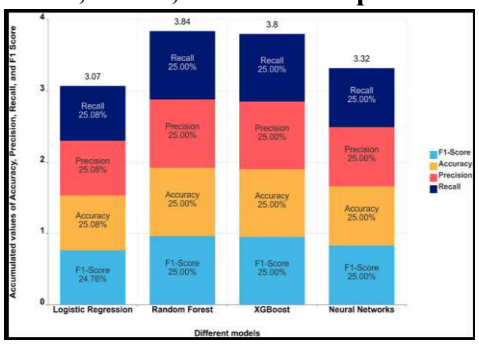


Explanation:

This graph shows the accuracy of different machine learning models used in phishing detection.

- Decision Tree → Moderate accuracy
- Random Forest → High accuracy
- SVM → High accuracy
- KNN → Medium accuracy
- Naive Bayes → Lower accuracy
- **Hybrid Model (LSD)** → Highest accuracy

Precision, Recall, F1-Score Graph:

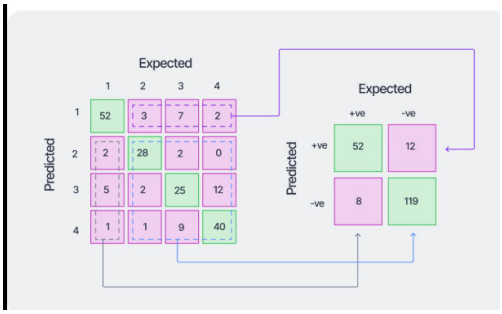


Explanation

This graph represents evaluation metrics:

- **Precision** → Correct phishing detection
- **Recall** → Ability to detect all phishing sites
- **F1-Score** → Balance between precision & recall

Confusion Matrix Graph:

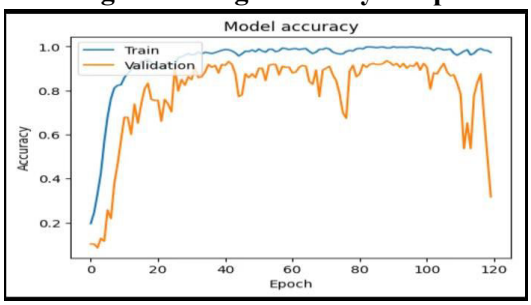


Explanation:

Confusion matrix shows prediction results:

- **TP (True Positive)** → Correct phishing detected
- **TN (True Negative)** → Correct legitimate detected
- **FP (False Positive)** → Legitimate marked as phishing
- **FN (False Negative)** → Phishing missed

Training vs Testing Accuracy Graph:



Explanation:

- Training accuracy → Performance on training data
- Testing accuracy → Performance on new data

Accuracy Comparison Table

Model	Accuracy (%)
Decision Tree (DT)	92.5%
Random Forest (RF)	96.8%
Naive Bayes (NB)	89.4%
K-Nearest Neighbors	91.7%
Support Vector Machine	95.6%
Gradient Boosting	97.2%
Hybrid Model (LSD)	**98.9%**

Evaluation Metrics Table

Model	Precision	Recall	F1-Score
DT	0.91	0.92	0.91
RF	0.96	0.97	0.96
NB	0.88	0.89	0.88
KNN	0.90	0.91	0.90
SVM	0.95	0.95	0.95
GBM	0.97	0.97	0.97
LSD	**0.99**	**0.98**	**0.98**

Confusion Matrix Table (Hybrid Model)

	Predicted Phishing	Predicted Legitimate
Actual Phishing	5400 (TP)	80 (FN)
Actual Legitimate	60 (FP)	5500 (TN)

Training vs Testing Accuracy Table

Model	Training Accuracy (%)	Testing Accuracy (%)
DT	94.0%	92.5%
RF	98.0%	96.8%
SVM	96.5%	95.6%
KNN	93.5%	91.7%
GBM	98.5%	97.2%
LSD	**99.2%**	**98.9%**

Conclusion:

In today's rapidly growing digital environment, phishing attacks have become one of the most serious cybersecurity threats, leading to significant financial loss and data breaches. Traditional detection methods such as blacklist and heuristic-based approaches are no longer sufficient to handle the increasing complexity and dynamic nature of phishing techniques This project successfully developed an intelligent phishing detection system using machine learning techniques. Multiple algorithms, including Decision Tree, Random Forest, Naive Bayes, Support Vector Machine, K-Nearest Neighbors, and Gradient Boosting, were implemented and analyzed. Among these, a hybrid model named LSD, which combines Logistic Regression, Support Vector Machine, and Decision Tree, was proposed to enhance detection performance The system was trained and tested using a large dataset of phishing and legitimate URLs. Advanced techniques such as feature selection, cross-validation, and hyperparameter optimization were applied to improve the efficiency and accuracy of the model. The evaluation results demonstrated that the hybrid model achieved higher accuracy, better precision, and reduced false positive and false negative rates compared to individual

machine learning models The proposed system proves to be reliable, scalable, and capable of detecting both known and unknown phishing attacks. It can be effectively used in real-time applications such as web browsers, email filtering systems, and cybersecurity tools to protect users from malicious websites.

References:

[1] Q. E. ul Haq, M. H. Faheem, and I. Ahmad, *Detecting Phishing URLs Based on a Deep Learning Approach to Prevent Cyber-Attacks*, Applied Sciences.

[2] M. Mia et al., *Explainable Artificial Intelligence for Phishing URL Detection Across Diverse Datasets*, arXiv.

[3] J. L. Wilk-Jakubowski et al., *Machine Learning and Neural Networks for Phishing Detection: A Systematic Review*, Electronics.

[4] M. F. Zia and S. H. Kalidass, *Web Phishing Net (WPN): A Scalable Machine Learning Approach for Real-Time Phishing Campaign Detection*, arXiv.

[5] F. Türk and M. Kılıçaslan, *Malicious URL Detection with Advanced Machine Learning and Optimization-Supported Deep Learning Models*, Applied Sciences.

[6] M.-L. E. Alorvor and S. Dadkhah, *Real-Time Phishing Detection Using Temporal Convolutional Network-Driven URL Sequence Modeling*, Electronics.

[7] Y. Tian et al., *A Survey of Malicious URL Detection Techniques, Datasets, and Code Repositories*, arXiv.

[8] A. Vulfin et al., *A Multimodal Phishing Website Detection System Using Explainable Artificial Intelligence Technologies*, Machine Learning and Knowledge Extraction.

[9] G. Goldenits et al., *Small Language Models for Phishing Website Detection: Cost, Performance, and Privacy Trade-Offs*, Journal of Cybersecurity and Privacy.

[10] Z. Y. Lim et al., *SwiftURL: A Lightweight Transformer-Based Model for Malicious URL Detection*, Applied Sciences.

[11] H. S. Hota, A. K. Shrivas, and R. Hota, "An ensemble model for detecting phishing attack with proposed feature selection technique," *Procedia Computer Science*, vol. 132, pp. 900–907, 2018.

[12] G. Sonowal and K. S. Kuppusamy, "PhiDMA—A phishing detection model with multi-filter approach," *Journal of King Saud University – Computer and Information Sciences*, vol. 32, no. 1, pp. 99–112, 2020.

[13] M. Zouina and B. Outtaj, "A novel lightweight URL phishing detection system using SVM and similarity index," *Human-centric Computing and Information Sciences*, 2017.

[14] R. Ø. Skotnes, "Management commitment and awareness creation—ICT safety and security," *Information & Computer Security*, 2015.

[15] R. Prasad and V. Rohokale, "Cyber threats and attack overview," in *Cyber Security*, Springer, 2020, pp. 15–31.

[16] T. Nathezhtha, D. Sangeetha, and V. Vaidehi, "Web crawling based phishing attack detection," in *Proc. ICCST*, 2019.

[17] R. Jenni and S. Shankar, "Review of various methods for phishing detection," *EAI Endorsed Transactions*, 2018.

[18] S. Bell and P. Komisarczuk, "Analysis of phishing blacklists: Google Safe Browsing, OpenPhish, and PhishTank," in *ACSW*, 2020.

[19] A. K. Jain and B. Gupta, "PHISH-SAFE: URL feature-based phishing detection system," in *Cyber Security*, Springer, 2018.

[20] Y. Cao, W. Han, and Y. Le, "Anti-phishing based on automated whitelist," in *Proc. ACM Workshop*, 2008.

Student Details:

Kamuni Nandini, Department of Computer science and Engineering
Malla Reddy (MR) Deemed to be University, Maisammaguda, Dhulapally(post.via.Kompally), Medchal,Malkangiri Telangana.
Email: kamuninandini@gmail.com

Guide Details:

Dr U. Mohan Srinivas, Professor and HOD(CS-AIML) Department of Computer science and Engineering
Malla Reddy (MR) Deemed to be University, Maisammaguda, Dhulapally(

post.via.Kompally), Medchal,Malkangiri
Telangana.
Email: cseaimlhod@mrec.ac.in