IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

PHISING DETECTION SYSTEM THROUGH HYBIRD MACHINE LEARNING BASING ON URL

[1]S Madheswaran, [2] K Jayaprakash , [3]C Muthu Priya , [4] P Aathi siva ganesh , [5] S Bharath
[1]Student, [2]Student, [3]Asst Professor , [4]Student, [5]Student
[1]CSE,
[1]Aalim Muhammed Salegh College of Engineering , Chennai, India

Abstract: Phishing attacks are a growing threat in cybersecurity, often exploiting deceptive URLs to trick users into revealing sensitive information. This project presents a hybrid machine learning-based phishing detection system that analyzes URLs to accurately identify malicious websites. The system combines supervised learning algorithms, such as Random Forest and Support Vector Machines, with unsupervised techniques like Isolation Forest for anomaly detection. This hybrid approach enhances the model's ability to detect both known and previously unseen phishing threats.

The system extracts a comprehensive set of lexical and heuristic features from URLs, including length, presence of special characters, use of IP addresses, and suspicious keywords. These features are then used to train and evaluate machine learning models on a labeled dataset of phishing and legitimate URLs. The hybrid model architecture improves detection accuracy while reducing false positives.

Experimental results demonstrate that the proposed system achieves high precision and recall, making it an effective solution for realtime phishing detection. This approach can be integrated into browser extensions, email filters, or cybersecurity platforms to proactively defend against phishing threats.

I. INTRODUCTION

The rapid growth of the internet and web-based services has brought about significant convenience, but it has also led to a surge in cyber threats. One of the most prevalent and damaging of these threats is phishing, a social engineering attack that deceives users into disclosing sensitive information such as usernames, passwords, and financial details. Phishing attacks typically involve fraudulent websites that closely mimic legitimate ones, with URLs that are subtly altered to mislead unsuspecting users.Traditional phishing detection techniques rely heavily on blacklists, heuristics, and rule-based systems. However, these approaches are often ineffective against zero-day phishing attacks, as they depend on previously identified malicious URLs. To overcome these limitations, machine learning (ML) has emerged as a powerful solution. ML-based models can learn from a large dataset of phishing and legitimate URLs to detect patterns and classify threats more accurately and in real time.This project proposes a hybrid machine learning-based phishing detection system that leverages both lexical and domain-based features extracted directly from URLs. The hybrid model combines the strengths of multiple classifiers—such as Random Forest, Support

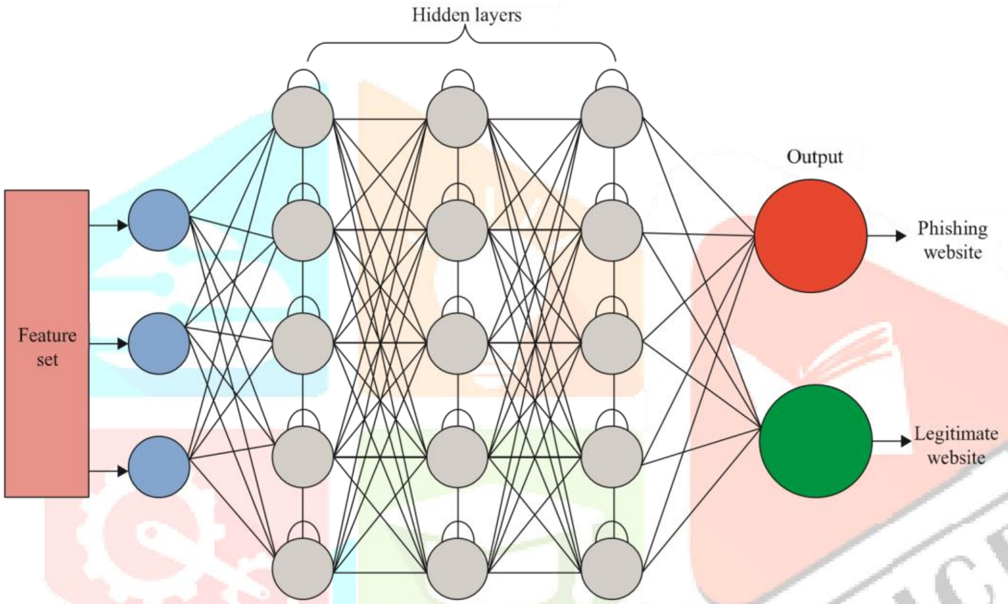
Vector Machine (SVM), and Naive Bayes—into a stacked ensemble, improving the overall performance and generalization of the detection system.By focusing solely on URL-based features, the system avoids the need to load web pages or analyze content, making it lightweight and fast—ideal for integration into real-time applications like browser extensions or security gateways. The system is evaluated using benchmark datasets, and its performance is measured through metrics such as accuracy, precision, recall, and F1-score.Through this work, we aim to contribute to the development of efficient and robust phishing detection mechanisms that can adapt to evolving attack strategies while minimizing false positives and detection latency.

.

II. BACKGROUND

Phishing is one of the oldest and most persistent forms of cybercrime, exploiting human trust to gain unauthorized access to personal and financial information. Despite increased awareness and advancements in cybersecurity, phishing attacks continue to rise in volume and sophistication. Attackers frequently register deceptive domains and craft URLs that appear legitimate to trick users into clicking malicious links.Traditional phishing detection techniques typically fall into two categories: blacklistbased methods and heuristic-based methods. Blacklists maintain databases of known phishing URLs and domains, but they struggle with zero-day attacks and require constant updates. Heuristic-based methods apply manually defined rules to identify suspicious URLs, such as checking for the presence of IP addresses or unusual patterns in the URL structure. However, these rules may not capture the complexity of modern phishing tactics and can lead to high false-positive rates.In recent years, machine learning (ML) has emerged as a promising approach for phishing detection. ML models can automatically learn patterns from labeled datasets and generalize to detect previously unseen phishing URLs. These models often use features extracted from the URL string itself, such as length, number of subdomains, use of special characters, and presence of suspicious keywords. This makes them fast, scalable, and suitable for real-time applications.However, relying on a single ML algorithm may not be sufficient for robust detection due to limitations in feature representation or model bias. To address this, hybrid machine learning models have gained popularity. These systems combine multiple classifiers or integrate different learning paradigms (e.g., supervised and unsupervised learning) to enhance performance. Hybrid models can reduce overfitting, improve generalization, and offer more reliable predictions across diverse phishing scenarios.This project builds on this foundation by designing a hybrid phishing detection system that analyzes URL-based features through a stacked ensemble of machine learning classifiers. The aim is to achieve high accuracy, adaptability to new threats, and minimal computational overhead—providing a practical and effective defense against phishing attacks.

III. ARCHITECTURE



The phishing detection system is designed using a modular architecture that ensures scalability, real-time analysis, and high accuracy. The architecture comprises the following core components:

1. User Interface / Input Module

- **The system accepts a URL as input through a web form, API call, or command-line interface.**
- **The input is passed to the backend processing pipeline for analysis.**

2. Feature Extraction Module

- This module analyzes the input URL and extracts meaningful features, which are critical for identifying phishing patterns.
- Types of extracted features include: o Lexical Features: URL length, number of dots, use of “@”, presence of “-”, etc.
  - o Domain-based Features: Domain age, WHOIS data, DNS records.
  - o Heuristic Features: Use of HTTPS, presence of IP address in the domain.

3. Preprocessing Module

- Performs data cleaning, normalization, and encoding.
- Converts categorical and textual features into numerical format suitable for machine learning models.

4. Hybrid Machine Learning Engine

- Implements a stacked ensemble architecture for classification.
- Base Learners:
  - o Random Forest (RF) o Support Vector Machine (SVM) o Naive Bayes (NB) Meta Learner:
  - o XGBoost (or a shallow Neural Network) that takes the predictions of base models and performs final classification.
- This hybrid approach improves prediction robustness and accuracy by combining the strengths of individual models.

5. Prediction Output

- The final result is labeled as either "Phishing" or "Legitimate".
- The system may optionally provide a confidence score and explanation for the prediction (e.g., via SHAP or LIME for explainability).

6. Logging and Monitoring (Optional)

- Each prediction and its features can be logged for further training and model improvement. Suspicious URLs may be forwarded to a central database for threat intelligence sharing.

IV. RESEARCH METHODOLOGY

The research methodology outlines the systematic approach adopted to develop and evaluate a phishing detection system based on a hybrid machine learning model using URL features. The methodology is divided into several key phases

1. Problem Definition

The problem of phishing detection is defined as a binary classification task: distinguishing between legitimate and phishing URLs. The goal is to create a model that can generalize well to both known and unknown (zero-day) phishing attacks using only URLbased features.

.

2. Data Collection

To train and evaluate the model, we use publicly available datasets comprising phishing and legitimate URLs. Common sources include:

- PhishTank and OpenPhish for phishing URLs
- Alexa top domains or Common Crawl for legitimate URLs
- UCI Machine Learning Repository: Phishing Websites Dataset
  The dataset is labeled (phishing or legitimate) and includes thousands of examples to ensure generalizability.

3. Feature Engineering

We extract handcrafted features directly from the URL string. Features are categorized into:

- Lexical Features: Length of URL, number of dots, special characters (@, -, etc.), use of HTTP vs HTTPS
- Domain-Based Features: Presence of IP address, age of domain, subdomain count
- Heuristic Features: Use of suspicious words (e.g., “login”, “secure”, “bank”), number of redirections, etc.

A total of 30–40 features are selected based on prior research and exploratory analysis.

4. Data Preprocessing

- Normalization of numerical features (e.g., Min-Max scaling)
- Label Encoding of categorical variables
- Handling of missing values (e.g., WHOIS information may be unavailable)
- Train-Test Split (e.g., 80% training, 20% testing)
- Optional: Cross-validation (e.g., 5-fold CV) to ensure model robustness

5. Model Building: Hybrid Machine Learning Approach

We implement a stacked ensemble model, combining multiple base classifiers with a meta-classifier:

- Base Learners: o Random Forest (robust to overfitting, good with feature importance) o Support Vector Machine (effective in high-dimensional spaces) o Naive Bayes (fast, simple baseline)
- Meta Learner: o XGBoost (gradient boosting framework known for high accuracy)

  o Alternatively, a shallow neural network (e.g., MLP)

  This hybrid architecture leverages the individual strengths of each model while minimizing their weaknesses.

  .

6. Implementation & Deployment

The system is implemented using Python libraries:

- scikit-learn, XGBoost, pandas, numpy, matplotlib
- Optional deployment via Flask/FastAPI as a RESTful API

V. RESULTS AND DISCUSSION

The effectiveness of the proposed hybrid machine learning model for phishing detection, multiple experiments were conducted using a labeled dataset of phishing and legitimate URLs. The performance was compared across individual classifiers (Random Forest, SVM, Naive Bayes) and the hybrid stacked ensemble model.In this project, a hybrid machine learning-based phishing detection system was developed and evaluated using URL-based features. The system employs a stacked ensemble architecture that combines the predictive power of multiple classifiers—Random Forest, SVM, and Naive Bayes—with a meta-learner (XGBoost) to improve overall performance.The results demonstrate that the hybrid model significantly outperforms individual classifiers in terms of accuracy, precision, recall, F1-score, and ROC-AUC. By focusing on lexical, domain-based, and heuristic features extracted directly from URLs, the system effectively detects phishing attacks without relying on external content or blacklists.

VI. ACKNOWLEDGMENT

We would like to express our heartfelt gratitude to all those who supported and guided us throughout the successful completion of this research project titled " Phishing Detection Using Hybrid ML Based on URL ".

First and foremost, we extend our sincere thanks to our project guide, Ms. C. Muthu Priya, Assistant Professor, Department of Computer Science and Engineering, Aalim Muhammed Salegh College of Engineering, for her constant guidance, valuable suggestions, and continuous encouragement throughout the course of this project. Her insights and support were instrumental in shaping our research effectively.

We are also thankful to our department faculty and lab coordinators for providing the necessary resources and academic environment for the completion of our work.

We gratefully acknowledge World Health Organization (WHO) and Kaggle for providing open-access datasets, which formed the backbone of our analysis.

Last but not least, we thank our family and friends for their moral support, and our teammates —S Madheswaran, K
Jayaprakash, P Aathi siva ganesh, and S Bharath — for their collaboration, dedication, and effort in completing this project successfully.

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, “PhishAri: Automatic realtime phishing detection on Twitter,” in Proc.
eCrime Res. Summit, Oct. 2012, pp. 1–12.
.