



# PHISHNET: A Machine Learning-Powered Threat Intelligence System for Multi-Vector Phishing Attack Detection

<sup>1</sup>Asha R,<sup>2</sup>Bhoomika K Shetty,<sup>3</sup>Chaitra S Moger, <sup>4</sup>H G Rakshitha, <sup>5</sup>Prof. Shana Santhosh

<sup>1</sup>Final year B.E. Student, <sup>2</sup>Final year B.E. Student, <sup>3</sup>Final year B.E. Student, <sup>4</sup>Final year B.E. Student, <sup>5</sup> Assistant Professor, CSE  
1Department of Computer Science & Engineering,  
1Srinivas Institute of Technology, Mangaluru, India

**Abstract:** Phishing remains a prevalent and evolving cyber threat, designed to trick others into disclosing private information. This essay presents PHISHNET, a comprehensive threat intelligence system that leverages machine learning to detect phishing attacks across multiple vectors, including URLs, emails, and SMS messages. The system employs a rich feature set, including the extraction of 30 distinct URL characteristics, alongside natural language processing (NLP) for email and SMS content analysis. We utilize algorithms such as Gradient Boosting for URL classification and Random Forest for email and SMS spam/phishing identification. PHISHNET integrates these detection mechanisms into a unified platform, offering real-time analysis and proactive defense. Experimental results demonstrate high efficacy, with URL detection achieving 98% accuracy, email detection 97%, and SMS detection 96%. This work contributes to building resilient defenses against sophisticated phishing attacks, thereby enhancing trust in digital communication channels.

**Keywords:** Phishing Detection, Machine Learning, Threat Intelligence, URL Analysis, Email Security, SMS Security, Gradient Boosting, Random Forest, Natural Language Processing.

## I. INTRODUCTION

The proliferation of digital communication has unfortunately led to a surge in cyber threats, with phishing attacks being among the most pervasive and damaging. These attacks exploit human vulnerabilities, employing deceptive tactics through fake emails, malicious URLs, and fraudulent SMS messages to steal credentials, financial details, and personal information [9]. The increasing sophistication of these attacks often allows people to get over conventional security procedures, which makes their detection and prevention a crucial obstacle for individuals and organizations alike

Manual verification of potential phishing attempts is time-consuming, inefficient, and prone to error, especially given the volume and evolving nature of these threats. Existing solutions often focus on a single attack vector (e.g., only URL filtering or email spam detection) or lack the adaptability to counter new phishing techniques effectively.

This paper presents PHISHNET, a Phishing Detection System designed to mitigate these threats by leveraging machine learning and advanced analytics. PHISHNET offers a comprehensive solution by analyzing emails, URLs, and SMS messages in real-time. It utilizes structured analysis of URL features, natural language processing for message content, and robust machine learning models to identify phishing indicators such as mismatched domains, suspicious links, and manipulative language. The system aims to provide automated, accurate, and adaptive multi-channel protection against phishing.

## II. LITERATURE REVIEW

Numerous strategies have been put forth for detecting phishing attacks. Early methods often relied on blacklist/whitelist approaches, which are insufficient against zero-day attacks. Machine learning (ML) has become a potent instrument for more dynamic and robust detection.

Studies such as [1] and [9] explored spam SMS/email detection using Naïve Bayes and TF-IDF vectorizers, achieving high precision. For website phishing, [2] reviewed various ML methods such as SVM, Random Forest, and Ada Boosting, highlighting challenges like overfitting. Deep learning, and Generative Adversarial Networks (GANs) to bypass

detection systems, as explored in [6], underscores the need for adaptive countermeasures. [4] introduced PADSTM for text message phishing detection using KNN, Random Forest, and SVM, with Random Forest showing superior performance. [10] compared XG Boost.

Multinomial Naive Bayes, and Logistic Regression for phishing website detection, finding Logistic Regression to be highly effective.

While these works address specific aspects of phishing, PHISHNET aims to provide a holistic solution by:

1. Integrating detection across multiple vectors (URL, Email, SMS).
2. Employing a comprehensive set of 30 lexical and host-based features for URL analysis.
3. Utilizing robust classifiers like Gradient Boosting and Random Forest tailored for each vector.
4. Offering a unified platform for real-time threat intelligence.

### III. PROPOSED SYSTEM: PHISHNET

PHISHNET is designed as a multi-component system to provide real-time phishing detection capabilities.

#### A. System Architecture

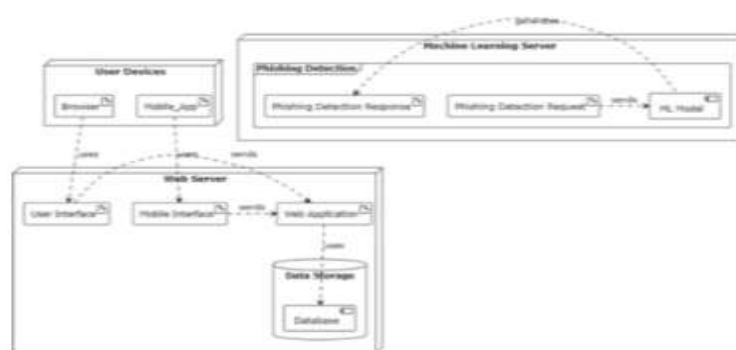


Fig.3.1: Architecture of a of Proposed System

5. **User Interface (Frontend)**: A web-based interface (HTML, CSS, JavaScript) allowing users to submit URLs, email content, or SMS messages for analysis.
6. **Web Server (Backend)**: Built using Python (e.g., Flask/Django), it responds to user inquiries, interacts with the ML models, and manages data.
7. **Machine Learning Core**: This component houses the trained ML models for URL, email, and SMS analysis. It processes input features and returns a prediction (phishing/legitimate).

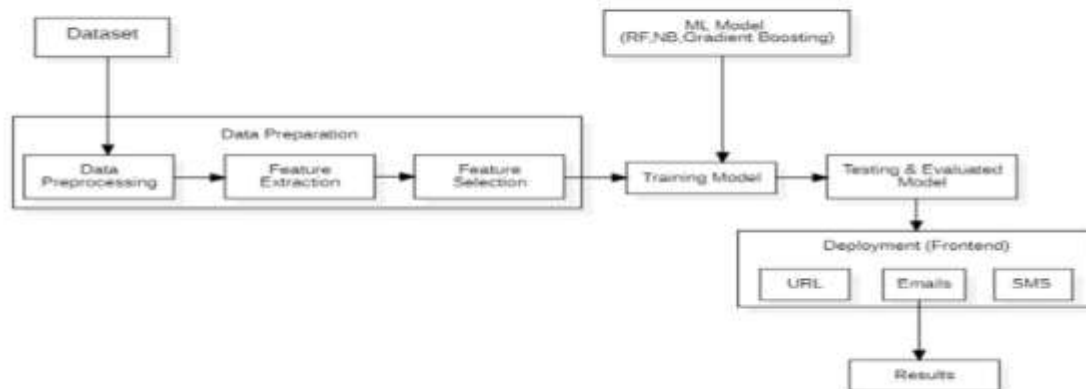


Fig.3.2: Data flow diagram for Phishing

**Data Flow:** Input data undergoes preprocessing and feature extraction. After that, these attributes are fed to the trained ML model (e.g., Gradient Boosting, Random Forest) for classification. The results are then relayed back to the user.

### Detection Modules

PHISHNET incorporates three primary detection modules:

1. **URL Phishing Detection:**

- **Feature Extraction:** The provided URL yields a comprehensive set of 30 characteristics.

- These consist of the following: Address Bar-based Features: IP Address, Long URL, Short URL, URLs with the "@" symbol, adding a prefix or suffix to the domain, sub-domains and multiple sub-domains, redirecting using "//," and HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer) are some examples.
- Features based on domains include the domain registration length, favicon, non-standard port, and the presence of the "HTTPS" token in the URL's domain section.
- The following features are based on HTML and JavaScript: Request URL, Anchor URL, Links in <Meta>, <Script>, and <Link> tags, Server Form Handler (SFH), Email submission, unusual URLs, website forwarding, customizing the status bar, turning off right-click, using pop-up windows, and IFrame redirection.
- External Features: Domain Age, DNS Record, and Website Traffic, PageRank, the quantity of links pointing to the Google Index Page, Statistical-Reports Based on Top Phishing IPs and Domains

**Classification Model:** A Gradient Boosting classifier was found to be most effective for URL classification.

2. **Email Phishing Detection:**

- **Feature Extraction:** Features are taken from emails. headers (e.g., sender information, return-path) and body content (e.g., presence of suspicious keywords, urgency, links, attachments). Natural Language Processing (NLP) techniques (e.g., TF-IDF) are used for text analysis.
- **Classification Model:** A Random Forest classifier is employed.

3. **SMS Phishing Detection (Smishing):**

- **Feature Extraction:** Similar to email, features include sender ID, presence of URLs (often shortened), suspicious keywords, and calls to urgent action. NLP techniques are applied to the message text.
- **Classification Model:** A Random Forest classifier is utilized.

### B. Technologies Used

- **Backend:** Python (Flask/Django)
- **Frontend:** HTML, CSS, JavaScript
- **Machine Learning:** Scikit-learn, Pandas, NLTK (for NLP tasks like tokenization, stop-word removal, TF-IDF).
- **Dataset:** A comprehensive dataset of authentic and phishing URLs, emails, and SMS messages was curated for training and evaluation.

## III. IMPLEMENTATION DETAILS AND TECHNOLOGIES

The PHISHNET system was implemented with distinct modules for URL, email, and SMS analysis.

### A. URL Feature Implementation Example:

1. Using IP: Checks if the URL uses an IP address instead of a domain name (e.g., ipaddress.ip\_address(URL)). Returns - 1 (phishing suspicion) if IP, 1 (legitimate) otherwise.
2. Long URL: Classifies URLs based on length: <54 chars (legitimate), 54-75 chars (suspicious), >75 chars (phishing).
3. Prefix-Suffix: Checks for hyphens within the domain name, often used to imitate authentic domains.

Many other features (30 in total as mentioned in the report, e.g., shortUrl, Symbol@, SubDomains, DomainRegLen, Favicon, RequestURL, AnchorURL, AgeofDomain, PageRank, GoogleIndex) were implemented to provide a robust feature vector.

### B. Model Training and Evaluation

The system was trained and tested using a curated dataset. The performance was assessed in light of accuracy.

- **URL Detection:** The Gradient Boosting The model's accuracy was 98%. The system effectively distinguishes between valid and malicious URLs, presenting a safety percentage to the user.
- **Email Detection:** The Random Forest classifier obtained a 97% accuracy rate in identifying phishing emails. The system allows users to input email body content or connect their Gmail account for recent message analysis.
- **SMS Detection:** The Random Forest classifier achieved an accuracy of 96% for SMS phishing detection. The system analyzes SMS content for suspicious language, false claims, or deceptive links.

#### IV. RESULTS AND DISCUSSION

The PHISHNET system demonstrated high efficacy across its three core detection modules. For URL phishing, a Gradient Boosting classifier leveraging 30 distinct features achieved an excellent accuracy of 98%, effectively distinguishing malicious links and providing clear user feedback. Email phishing detection, utilizing a Random Forest classifier with NLP-derived features from email content, yielded 97% accuracy. Similarly, SMS phishing (smishing) detection, also employing Random Forest, achieved a robust 96% accuracy in successfully identifying fraudulent text messages. These figures validate the chosen machine learning approaches and the comprehensive feature engineering strategy employed for each specific attack vector.

The strong performance across modules underscores the system's robust design. Gradient Boosting excelled with the rich, structured URL feature set, while Random Forest classifiers proved adept at discerning patterns in text-heavy email and SMS content through NLP. The high accuracy across all modules, despite the unique complexities inherent to each attack vector, confirms the practical viability of PHISHNET. The integration of these diverse detection capabilities into a unified platform offers a comprehensive and user-friendly defense mechanism, validating the system's approach to multi-vector phishing threat mitigation.

#### V. CONCLUSION

In conclusion, PHISHNET successfully demonstrates a comprehensive, machine learning-driven approach for multi-vector phishing detection across URLs, emails, and SMS messages. Through robust feature engineering, including 30 distinct URL characteristics, and the application of effective classifiers like Gradient Boosting (98% accuracy for URLs) and Random Forest (97% for emails, 96% for SMS), the system achieves significant efficacy in identifying and mitigating diverse phishing threats. This research provides a practical and adaptive framework, with future work focused on advanced AI models, enhanced feature engineering, and adaptive learning to build an even more robust defense against changing online dangers and enhance digital trust.

#### REFERENCES

- [1] V. Dharani, D. Hegde, and M. Mohana, "Spam SMS (or) Email Detection and Classification using Machine Learning," Electronics and Telecommunication Engineering, RV College of Engineering®, Bengaluru, Karnataka, India.
- [2] A. Odeh, I. Keshta, and E. Abdelfattah, "Machine Learning Techniques for Detection of Website Phishing: A Review for Promises and Challenges," Princess Sumaya University for Technology, Amman, Jordan; AlMaarefa University, Riyadh, Saudi Arabia; Ramapo College of New Jersey, Mahwah, USA.
- [3] D. Patel and D. Patel, "PolyDDoSChain - Collaborative Volumetric Distributed Denial of Service Attack Detection and Prevention using Blockchain Technology," Department of Computer Science, Gujarat Vidyapith, Ahmedabad, India.
- [4] S. Uplenchwar, S. Deshpande, V. Sawant, S. Kelkar, and P. Surve, "Phishing Attack Detection on Text Messages Using Machine Learning Techniques," Computer Engineering Department, Cummins College of Engineering for Women, Pune, India.
- [5] R. Salama, F. Al-Turjman, S. Bhatla, and S.P. Yadav, "Social Engineering Attack Types and Prevention Techniques - A Survey," AI and Robotics Institute, Near East University, Turkey; Graphic Era Hill University, Dehradun, India; G.L. Bajaj Institute of Technology and Management, Greater Noida, India.
- [6] A. AlEroud and G. Karabatis, "Bypassing Detection of URL-based Phishing Attacks Using Generative Adversarial Deep Neural Networks," [Affiliation details not provided in the source document].
- [7] A. Mandadi, S. Boppana, V. Ravella, and R. Kavitha, "Phishing Website Detection Using Machine Learning," [Affiliation details not provided in the source document].
- [8] P. Saraswat and M.S. Solanki, "Phishing Detection in E-mails using Machine Learning," [Affiliation details not provided in the source document].
- [9] V. Dharani, D. Hegde, and Mohana, "Spam SMS & Email Detection and Classification using Machine Learning," [Affiliation details not provided in the source document].
- [10] G.K. Kamalam, P. Suresh, R. Nivash, A. Ramya, and G. Raviprasath, "Detection of Phishing Websites Using Machine Learning," [Affiliation details not provided in the source document].