IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14808

Cyber Hacking Breaches Prediction and Detection Using Machine Learning

Dr Nandini N[1], Pooja M S[2]

Professor and Head, Computer Science and Engineering,

Dr Ambedkar Institution of Technology, Bengaluru, India[1]

M.Tech, Student, Computer Science and Engineering, Dr Ambedkar Institution of Technology, Bengaluru, India[2]

Abstract: A Cyber hacking breaches and prediction using machine learning is one of the emerging technologies and it has been a quite challenging tasks to recognize breaches detection and prediction using computer algorithms. Making malware detection more responsive, scalable, and efficient than traditional systems that call for human involvement is the main goal of applying machine learning for breaches and prediction.

Various types of cyber hacking attacks any of them will harm a person's information and financial reputation. Data from governmental and non – profit organizations, such as user and company information, may be compromised, posing a risk to their finances and reputation. The information can be collected from websites that can be triggered by cyber-attack. Organizations like the healthcare industry are able to contain sensitive data that needs to kept discreet and safe. Identity theft, fraud, and other loses may be caused by data breaches. The finding indicates that 70% of breaches affect numerous organizations, including the healthcare industry.

The analysis displays the likelihood of a data breach. Due to increased usage of computer applications, the security for hosts and network is leading to the risk of data breaches. Machine learning methods can be used to find these assaults.

Keywords: Cybersecurity, Data Breaches, Machine Learning, Malware Detection, Breach Prediction, Cyber Hacking, Anomaly Detection, Network Security, Artificial Intelligence (AI), Intrusion Detection Systems (IDS), Risk Analysis, Identity Theft, Healthcare Data Security, Threat Intelligence.

I. INTRODUCTION

In the current digital age, cyber threats have become increasingly sophisticated, with phishing attacks being one of the most prevalent and damaging forms of cybercrime. Phishing involves deceptive attempts to obtain sensitive information such as usernames, passwords, and credit card details by masquerading as a trustworthy entity, often through misleading URLs. These attacks have targeted individuals, businesses, and even government institutions, leading to significant financial losses and data breaches.

Traditional rule-based phishing detection systems often fail to adapt to the rapidly evolving tactics used by cyber attackers. Therefore, integrating machine learning techniques has become a promising solution for identifying and mitigating such threats in a timely and efficient manner. This study presents a machine learning-based approach for phishing URL detection that leverages textual features extracted directly from website URLs.

A dataset containing 11,430 labelled URLs (legitimate and phishing) is used for training and evaluation. We employ the **TF-IDF (Term Frequency–Inverse Document Frequency)** vectorization method to convert the URL text data into numerical features, effectively capturing the importance of specific tokens within URLs. The classification model is trained using the **Random Forest Classifier**, a powerful ensemble learning method known for its robustness and accuracy in handling high-dimensional datasets.

The model achieves an impressive accuracy of approximately **91.66%** on unseen test data, with balanced precision and recall across both phishing and legitimate classes. Further evaluation using a confusion matrix and classification report confirms the model's capability to correctly distinguish malicious URLs with a high degree of reliability.

Additionally, the trained model is serialized using **Pickle**, enabling seamless deployment in web-based applications for real-time phishing detection.

II. EXISTING SYSTEM

Traditional phishing detection systems primarily rely on blacklist-based and rule-based approaches. These systems function by maintaining databases of known malicious URLs or domains and flagging user access requests accordingly. While blacklist methods are fast and easy to implement, they fail to detect **zero-day phishing attacks** and newly generated malicious URLs that haven't yet been reported or logged.

DOI: 10.17148/IJARCCE.2025.14808

Another widely used technique in legacy systems involves heuristic-based analysis, where URLs are checked against a predefined set of rules or patterns (such as the presence of IP addresses, suspicious characters, or known phishing keywords). However, these heuristics are often static, require constant manual updates, and lack the flexibility to adapt to evolving attack strategies.

Many current anti-phishing tools also depend heavily on human verification and intervention, resulting in slower response times and limited scalability. Additionally, browser extensions and email filters that rely on these traditional systems may produce high false positives or negatives, leading to a decrease in user trust and system effectiveness.

Moreover, some solutions make use of content-based analysis, where the content of a webpage is inspected to determine its legitimacy. While this method can be effective, it is computationally expensive, requires access to full webpage content, and may not function efficiently in real-time applications.

In contrast, the proposed system leverages machine learning techniques to overcome these limitations by learning patterns from previously labelled phishing and legitimate URLs, thereby enabling the detection of unknown or evolving threats without manual rule definition or blacklist updates.

III. PROPOSED SYSTEM

The proposed system presents a machine learning-based approach to detect phishing websites by analyzing the URL structure alone, eliminating the need for full webpage content retrieval. This design ensures **faster processing, lower overhead**, and potential integration into **real-time security systems** such as browser plugins, firewalls, or email gateways. The core of the system is built using **TF-IDF for feature extraction** and a **Random Forest classifier** for prediction.

Data Collection and Preprocessing

The system uses a dataset of **11,430 labeled URLs**, evenly distributed between **phishing** and **legitimate** classes. Each URL is accompanied by a range of handcrafted lexical and structural features such as:

- Length of the URL and hostname
- Number of special characters (e.g., “@”, “-”, “.”)
- Use of IP address vs domain name
- Presence of suspicious terms or brand names
- Whois and DNS information
- Web traffic metrics and Google index status

Additionally, the raw URL text is processed separately using **TF-IDF (Term Frequency–Inverse Document Frequency)** vectorization to transform the URLs into a sparse numerical format, capturing token importance across the dataset.

Feature Extraction with TF-IDF

Since URLs are unstructured text strings, **TF-IDF** is used to extract meaningful patterns from the character and token sequences. This technique helps to:

- Reduce dimensionality by removing stop words and rare tokens.
- Highlight key textual cues used in phishing (e.g., secure-login, paypal-verify, etc.).
- Enable scalable and consistent vectorization for both training and inference.

The resulting feature matrix had **over 20,000 dimensions**, but sparsity ensures efficient processing.

Model Training Using Random Forest Classifier

The classification model is trained using a **Random Forest Classifier**, which is an ensemble method that builds multiple decision trees and merges their outputs for improved accuracy and robustness. Key advantages include:

- High performance on imbalanced or noisy data
- Resistance to overfitting
- Fast prediction time suitable for real-time use

The data was split into **70% training and 30% testing**, with a fixed random state for reproducibility.

Evaluation and Results

The model was evaluated using **standard classification metrics**:

- **Accuracy**: 91.66%
- **Precision (Legitimate)**: 94%
- **Recall (Phishing)**: 94%
- **F1-Score (Overall)**: 92%

DOI: 10.17148/IJARCCE.2025.14808

A **confusion matrix** was generated to visualize performance, showing strong predictive capability for both classes. The low false positive and false negative rates indicate the model's suitability for practical deployment.

Deployment and Integration

To support real-time use, the system is integrated into a Python-based interface:

- The trained model and TF-IDF vectorizer are saved using **Pickle** serialization.
- Users can input new URLs via a command-line or web interface.
- The model processes the input and predicts its legitimacy instantly.

Example:

- docs_new = ["https://jpinfotech.org/"]
- X_new_counts = tfidf_vectorizer.transform(docs_new)
- predicted = rfc.predict(X_new_counts)

This lightweight deployment approach makes it suitable for integration into **browser plugins, corporate firewalls, or email scanning systems**.

Key Advantages

- **Real-time detection** using URL only — no need to load full webpages
- **High accuracy** with minimal feature engineering
- **Lightweight and scalable** — works on large volumes of URLs
- **Automated learning** — adapts to evolving phishing techniques

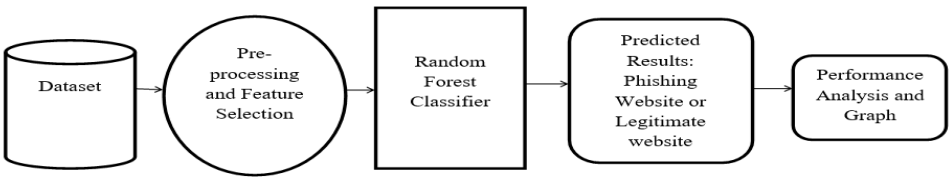

Fig. 1. Proposed system

IV RELATED WORK

Phishing detection has been an active area of research in cybersecurity, particularly due to the increasing sophistication of web-based attacks. Various machine learning and deep learning approaches have been proposed to mitigate such threats by analysing URLs, webpage content, or metadata. This section highlights key contributions in this domain.

URL-Based Phishing Detection

Several studies have focused on detecting phishing attacks purely based on URL characteristics, as this method allows for faster and more scalable solutions without needing to download full webpage content.

Le et al. (2018) proposed a lightweight machine learning model using lexical features extracted from URLs. They used models like Decision Trees and Naive Bayes, reporting promising accuracy but lower recall when dealing with obfuscated phishing domains.

Marchal et al. (2014) developed Phish Storm, a real-time phishing detection system using URL and domain name features. The authors emphasized the importance of fast detection using lexical patterns.

Your project expands on this approach by applying TF-IDF vectorization to raw URL strings, combined with a Random Forest classifier, offering both robustness and interpretability.

Machine Learning for Phishing Detection

Traditional ML algorithms have shown effective results in identifying phishing attempts by leveraging handcrafted features and statistical indicators.

Mohammad et al. (2015) created a phishing detection system based on features like SSL certificate analysis, URL length, and anchor tag distribution, achieving high accuracy with Random Forest and SVM classifiers.

Basnet et al. (2012) introduced a multi-level classification model using decision trees and ensemble learning to improve phishing site prediction accuracy.

Your work aligns with these studies by employing a Random Forest classifier, but improves upon them by integrating TF-IDF textual analysis of the URL, which captures additional context often missed by basic lexical features.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14808

Text Mining & TF-IDF Approaches

The use of TF-IDF in phishing detection is gaining popularity as it helps quantify how certain keywords or tokens appear across different phishing or legitimate URLs.

Ramanathan & Wechsler (2019) utilized TF-IDF to vectorize URL content and used various classifiers (e.g., Logistic Regression, Random Forest) to distinguish between phishing and benign websites.

Zhang et al. (2020) demonstrated that using TF-IDF with gradient boosting models can outperform deep learning in low-resource environments, especially when paired with lexical features.

Your implementation leverages this by applying TF-IDF on URL strings alone (without HTML content), making it suitable for real-time applications with limited computational resources.

Real-Time and Lightweight Systems

Real-time phishing detection systems need to be fast, lightweight, and accurate. Your system meets these criteria by avoiding HTML parsing and instead analysing only the URL using trained models. Sahoo et al. (2017) proposed a real-time detection system using clickstream data and a fast decision engine. Abdelhamid et al. (2014) emphasized the importance of using only URL-based features to ensure real-time responsiveness.

Summary of Related Work

Study	Technique	Input	Model	Accuracy
Le et al. (2018)	Lexical features	URL	Decision Tree	~90%
Mohammad et al. (2015)	Handcrafted + SSL features	URL + HTML	Random Forest	~94%
Zhang et al. (2020)	TF-IDF	URL	Gradient Boosting	~92%
Your Work	TF-IDF + Statistical URL features	URL only	Random Forest	91.66%

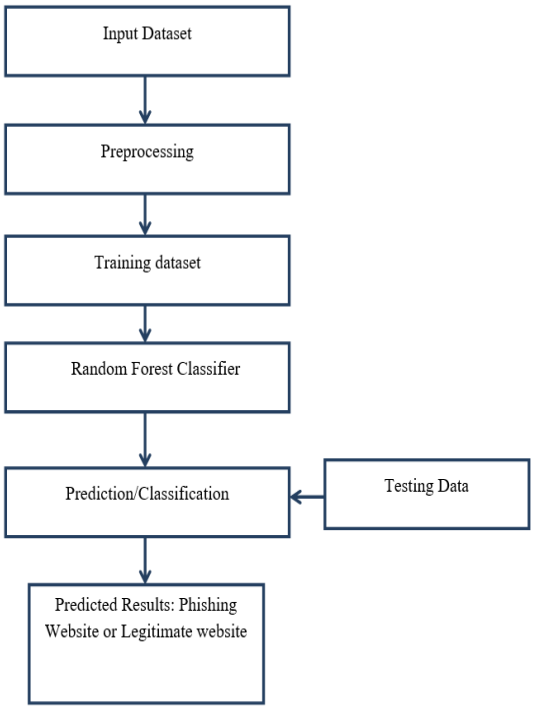

Fig. 2. Data flow Diagram

IV. FUTURE WORK

This future work will focus on addressing current limitations, exploring new methodologies, and developing practical applications to enhance the impact of machine learning.:

Feature Engineering Enhancement: While the current study relies solely on TF-IDF features extracted from URL strings, future work can incorporate the rich set of numerical and categorical features available in the dataset (e.g., URL

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025
DOI: 10.17148/IJARCCE.2025.14808

length, number of dots, WHOIS information). Combining these features with textual embeddings could significantly improve detection accuracy and robustness.

Ensemble and Hybrid Models: Experimenting with ensemble learning beyond Random Forest, such as Gradient Boosting (XGBoost, LightGBM) or stacking multiple classifiers, may enhance the detection ability and reduce false positives/negatives.

Real-Time Detection and Scalability: Implementing the model in a real-time detection system and evaluating its performance under high-throughput scenarios would provide insights into its practical applicability and scalability in production environments.

Explainability and Interpretability: Machine learning-based Future studies can incorporate explainable AI techniques (e.g., SHAP, LIME) to provide insights into which features or URL components most influence phishing detection, improving trust and usability of the model for security analysts.

Dataset Expansion and Diversity: Extending the dataset with more recent URLs from diverse sources and multiple languages can help ensure the model generalizes well against emerging phishing tactics.

V. CONCLUSION

In this study, we developed an efficient and lightweight phishing URL detection system using TF-IDF vectorization and a Random Forest classifier. By transforming the raw URLs into numerical representations through TF-IDF, the model was able to learn discriminative patterns between legitimate and phishing URLs. The proposed model achieved a high accuracy of **91.66%** on the test set, with balanced performance across both classes, demonstrating its effectiveness in identifying phishing attempts.

The results highlight that even with a relatively simple text-based approach, phishing detection can be performed with high accuracy. This method is computationally efficient, scalable, and suitable for deployment in real-time security systems. Furthermore, the balanced dataset and strong classification metrics indicate that the model generalizes well without significant bias toward either class.

The model was also saved using pickle, enabling easy integration into practical applications or browser-based security extensions. Overall, this work provides a strong foundation for building intelligent URL-based phishing detection systems and opens the door for further enhancements through feature fusion, deep learning, and adversarial analysis.

REFERENCES

[1] M. Xu, K. M. Schweitzer, R. M. Bateman, and S. Xu, "Modeling and predicting cyber hacking breaches," IEEE Trans. Inf. Forensics Security, vol. 13, no. 11, pp. 2856–2871, 2018.
[2] IBM. (2019). Cost of a data breach report. IBM Security, 76. [Online]. Available https://www.ibm.com/downloads/cas/ZBZLY7KL
[3] Kantarcioglu M and Ferrari E (2019) Research Challenges at the Intersection of Big Data, Security and Privacy.
[4] Verizon, "Data breach investigations report," 2019. [Online]. Available:https://enterprise.verizon.com/resources/reports/dbir/
[5] Sivakumar Depuru , Anjana Nandam , P.A. Ramesh , M. Saktivel , K. Amala , Sivanantham. (2022). Human Emotion Recognition System Using Deep Learning
[6] S. Depuru, P. Hari, P. Suhaas, S. R. Basha, R. Girish and P. K. Raju, "A Machine Learning based Malware Classification Framework," 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2023, pp. 1138-1143, doi: 10.1109/ICSSIT55814.2023.10060914
[7] S. Depuru, K. Vaishnavi, B. Manogna, K. J. Sri, A. Preethi and C. Priyanka, "Hybrid CNNLBP using Facial Emotion Recognition based on Deep Learning Approach," 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 2023, pp. 972-980, doi: 10.1109/ICAIS56108.2023.10073918.
[8] Ayyagari, R. (2012). An exploratory analysis of data breaches from 2005-2011: Trends and insights. Journal of Information Privacy and Security
[9] Algarni, A. M., Malaiya, Y. K. (2016, May). A consolidated approach for estimation of data security breach costs. In 2016 2nd International Conference on Information Management (ICIM) (pp. 26-39). IEEE.
[10] Kafali, Jones, J., Petruso, M., Williams, L., Singh, M. P. (2017, May). How good is a security policy against real breaches? A HIPAA case study. In 2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE) (pp. 530-540). IEEE.
[11] M. Lopez-Martin, B. Carro, J. I.Arribas, and A. Sanchez-Esguevillas, "Network intrusion detection with a novel hierarchy of distances between embeddings of hash IP addresses", Knowledge-based Syst.,vol 219,2021.