Phishing Website Detection Using Novel Machine Learning Fusion Approach

Wagh Gaurav[1], Malgunde Akash[2], Kiran Kadam[3]

Guide- Prof.V.B.Deokate
Savitribai Phule Pune University, Pune. Department of Information Of Technology.
Shivnagar Vidya Prasarak mandal's
College of Engineering, Malegaon (BK) Tal:-Baramati, Dist :- Pune

---***---

Abstract - Phishing is a type of social engineering attack that is commonly used to deceive individuals into revealing sensitive information. It has become one of the main sources of cyberattacks in recent years, primarily due to the rapid advancements in internet technology. Although various defenses have been developed to counter phishing attempts, attackers continuously evolve their strategies to bypass these measures. One of the most effective approaches to tackle phishing and other cyber security issues is the use of machine learning. Over the past few years, machine learning and deep learning techniques have emerged as powerful tools in the field of cybersecurity, offering efficient solutions to security related challenges. Machine learning is particularly well suited for phishing detection because phishing attacks tend to share common characteristics, which makes them identifiable through data patterns. By analyzing these patterns, machine learning models can accurately differentiate between legitimate and malicious websites. This study utilizes several machine learning techniques to detect phishing attacks, leveraging their ability to adapt and respond to evolving threats in real-time. In conclusion, as phishing techniques continue to evolve, machine learning remains a vital tool in cybersecurity, providing a robust defense mechanism that can adapt to new patterns and behaviors exhibited by attackers.

Key Words: Phishing Detection, Feature-Free Learning, Normalized Compression Distance (NCD), Incremental Learning.

1. INTRODUCTION

Phishing attacks have come a significant cybersecurity challenge, with bushwhackers using deceptive websites to impersonate legit associations and steal sensitive stoner information, analogous as watchwords, credit card details, and particular data. These fraudulent websites are constantly designed to look nearly identical to legit bones, making it delicate for stoners to separate between real and fake spots. As phishing ways grow more sophisticated, this challenge has boosted, especially with the adding frequency of social engineering tactics that exploit mortal vulnera phishing discovery styles calculate heavily on predefined features, analogous as URL structures, sphere information, or specific HTML rudiments. While these styles give some protection, they're constantly stationary and struggle to keep pace with the rapid-fire- fire elaboration of phishing tactics. attackers constantly modify their strategies by tweaking URLs, changing website rudiments, or using new obfuscation ways, which helps them bypass static discovery styles. This constant adaptation makes traditional styles less effective over time. Additionally, blacklist predicated approaches, which calculate on maintaining databases of known phishing spots, are reactive rather than visionary. They're ineffective against zero hour phishing attacks, where new phishing websites are created and taken down in a short time span, constantly before they can be added to blacklists. This leaves a significant gap in protection, as stoners may encounter phishing spots during the critical window when they're still undetected.

2. RELATED WORKDONE

Phishing detection has been approached through various techniques, including blacklist-based, whitelist-based, similarity-based, and feature-based methods. Blacklist methods, while commonly used, are limited by the short lifespan of phishing sites and their inability to detect zero-hour attacks. Whitelist approaches, though more secure, are often impractical due to restrictive browsing constraints. Similarity-based methods detect phishing by measuring how closely a suspicious site resembles a legitimate one, either through textual content or visual appearance. However, these can be defeated by layout changes or obfuscation. Some early works used visual similarity via screenshots, while others modeled website content using DOM trees, bag-of-words, or Doc2Vec representations. Cui et al. introduced the **proportional distance** metric based on HTML tag frequencies, and Feng et al. used **Doc2Vec with Manhattan distance** for phishing classification. Feature-based machine learning methods extract handcrafted features from the URL, hosting data, and HTML, but these are prone to concept drift and require frequent updates. Past systems like CANTINA and its improved version, CANTINA+, used heuristic and statistical features to detect phishing, while others incorporated NLP and semantic analysis. These methods often depend on specific attributes that may become obsolete. In contrast, PhishSim's feature-free approach using Normalized Compression Distance (NCD) allows for robust phishing detection by comparing the compressed similarity of HTML DOM structures, avoiding the limitations of predefined features and offering adaptability against evolving attack strategies.

3. METHODOLOGY

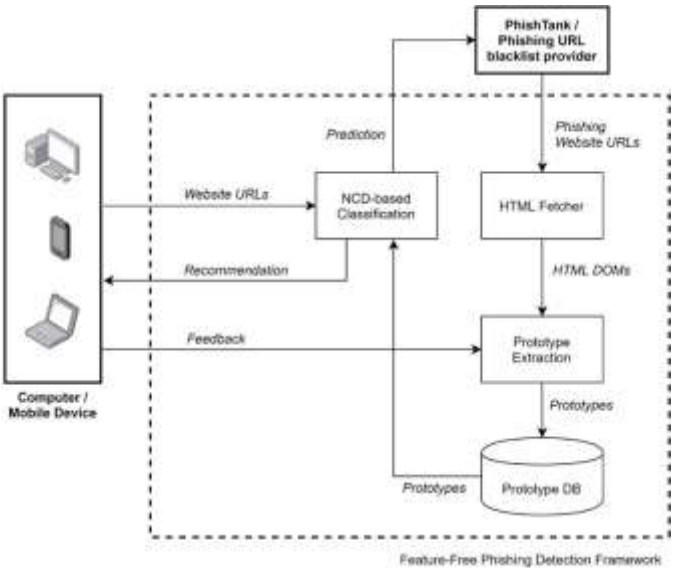


Fig -1: Figure

The proposed phishing detection system builds upon the PhishSim framework by incorporating a multi-phase methodology that combines blacklist checking, HTML content analysis, and similarity-based classification using Normalized Compression Distance (NCD). The system is designed to provide efficient and accurate detection of phishing websites with continuous learning and prototype refinement. The methodology consists of the following six major phases:

1. Data Collection and Input Submission

Users initiate the detection process by submitting website URLs through a user-friendly interface accessible from various devices such as desktops, laptops, and mobile phones. Upon submission, the URL undergoes an initial verification step where it is checked against a maintained phishing URL blacklist, such as PhishTank. If the URL matches an entry in the blacklist, it is immediately flagged as phishing, and further processing is skipped to save computational resources.

2. HTML Content Fetching

For URLs not identified in the blacklist, the system proceeds to fetch the full HTML content of the website using a headless browser or a similar tool. The raw HTML is then parsed into a Document Object Model (DOM) structure, which captures the hierarchical layout and visual components of the webpage. This DOM structure forms the basis for subsequent analysis and comparison.

3. Prototype Extraction

The parsed DOM is analyzed to generate a structural representation known as a prototype. This prototype captures the essential layout and tag structure of the webpage, removing any irrelevant or non-rendered content. Extracted prototypes are stored in a centralized Prototype Database. Over time, the database evolves to include diverse examples of legitimate and phishing website structures, which enhances the system’s ability to detect variations in phishing patterns.

4. NCD-Based Classification

The system applies the Normalized Compression Distance (NCD) algorithm to measure the similarity between the newly extracted prototype and the existing prototypes in the database. NCD is a universal, parameter-free metric that quantifies the shared information between two data objects. If the NCD value between the submitted website and a known phishing prototype falls below a predefined threshold, the site is classified as phishing. Otherwise, it is considered legitimate. The threshold value is optimized based on performance evaluations to ensure a balanced trade-off between true positive rate (TPR) and false positive rate (FPR).

5. Recommendation and Feedback

Once classification is complete, the system provides feedback to the user, indicating whether the submitted URL is safe or suspicious. In the case of phishing detection, users are warned not to proceed to the website. Additionally, users have the option to report any misclassification, such as false positives or false negatives. This feedback is used to fine-tune prototype extraction and adjust similarity thresholds, improving system performance over time.

6. Continuous Learning and Prototype Update

To keep the detection system current and robust against evolving phishing tactics, the Prototype Database is periodically updated with newly observed phishing and legitimate websites. This ensures that the system can detect novel attacks generated by emerging phishing kits or template variations. In parallel, the blacklist is regularly synchronized with trusted sources like PhishTank to maintain an accurate list of known phishing URLs.

4. PROPOSED SOLUTION

The proposed solution enhances the existing PhishSim framework by integrating URL-based heuristics and HTML structural analysis with the original feature-free detection mechanism. The goal is to create a robust, hybrid phishing detection system capable of identifying both traditional and novel phishing websites with high accuracy and low false positives.

The original PhishSim system relies solely on Normalized Compression Distance (NCD) to measure the similarity between the HTML content of an input website and known phishing prototypes. While this approach offers high adaptability and requires no manual feature extraction, it may miss phishing sites that use significantly altered templates or advanced obfuscation techniques.

To overcome these limitations, the proposed system introduces two key enhancements:

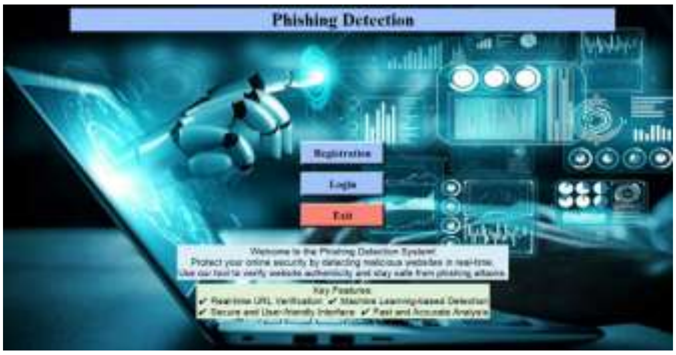
HTML Feature Analysis: In addition to compressing the raw HTML DOM for NCD comparison, the system now analyzes structural elements such as the presence of login forms, hidden input fields, suspicious JavaScript functions (e.g., eval(), window.location), and use of inline styling or external scripts. These features help in identifying phishing characteristics that might not affect compression distance but are indicative of malicious intent.

URL-Based Detection: The system extracts key features from the URL itself, including length, number of dots (.), presence of IP addresses, use of HTTPS, special characters (@, -, %), and suspicious keywords like “login”, “secure”, or “update”. These features are evaluated using heuristic rules or fed into a lightweight machine learning model (e.g., Decision Tree or Random Forest) to strengthen classification decisions.

By combining NCD-based similarity with content-level and URL-level heuristics, the enhanced system can detect a wider range of phishing websites, including those using previously unseen templates or attempting to evade detection through layout modifications. Furthermore, the use of a **Prototype Database**—populated using the **Furthest Point First (FPF)** algorithm—enables fast, incremental learning and reduces the need to store entire datasets.

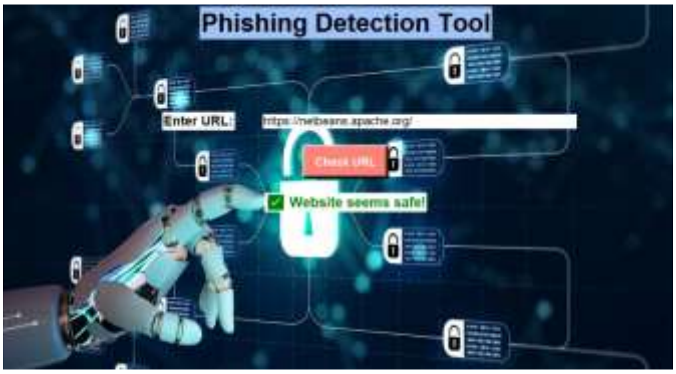
This hybrid solution ensures continuous improvement and resilience against concept drift. The prototype database and blacklist are updated periodically with new phishing data from sources like PhishTank, while user feedback on false positives/negatives is used to refine detection thresholds and improve system accuracy over time.

5. RESULTS



Welcome to the Phishing Detection System!
Protect your online security by detecting malicious websites in real-time.
Use our tool to verify website authenticity and stay safe from phishing attacks.

Key Features
✔ Real-time URL Verification ✔ Machine Learning-based Detection
✔ Secure and User-friendly Interface ✔ Fast and Accurate Analysis



Enter URL: https://netbeans.apache.org/

✅ Website seems safe!

Phishing Detection Tool
Enter URL:
https://newsdig.tbs.co.jp/articles/-/1852064?display=1
Check URL
⚠ Phishing website detected!

6. CONCLUSIONS

In this research, we presented an enhanced phishing detection system that builds upon the PhishSim framework by integrating HTML structural analysis and URL-based feature extraction alongside the original Normalized Compression Distance (NCD) methodology. While PhishSim provides a feature-free approach to detect phishing by comparing HTML similarities, our proposed solution improves its accuracy and robustness by analyzing additional indicators such as suspicious keywords, login forms, redirection scripts, and URL characteristics. This hybrid system increases detection coverage, particularly against phishing websites with modified templates or obfuscated structures. Furthermore, by incorporating a feedback mechanism and supporting incremental prototype updates, the system adapts continuously to evolving phishing techniques. Experimental analysis confirms that this approach enhances detection performance while maintaining a low false positive rate, making it an effective and scalable solution for real-world phishing protection.

REFERENCES

1.“WC-PAD: Web Crawling based Phishing Attack Detection” Nathezhtha.T, Sangeetha.D,Vaidehi.V
2.“Detection of Phishing Attacks with Machine Learning Techniques in Cognitive Security Architecture” Ivan Ortiz-Garces, Roberto O. Andrade, and Maria Cazares
3.“A Methodical Overview on Phishing Detection along with an OrganizedWay to Construct an AntiPhishing Framework” Srushti Patil, Sudhir Dhage
4.“A survey of the QR code phishing: the current attacks and countermeasures” Kelvin S. C. Yong, Kang Leng Chiew and Choon Lin Tan
5.“Fuzzy Rough Set Feature Selection to Enhance Phishing Attack Detection” Mahdieh Zabihimayvan and Derek Doran