IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

PHISPHSPOTTER : UNVEILING DECEPTIVE WEBSTIES

[1]Mohammed Arshad Ali , [2]Venigalla Yoshitha , [3]Tanniru Venkat,4 Vaibhav Bhosle,5 Mohd Abdul Aleem

[1]Student , [2]Student , [3]Student, 4Student,5Assistant professor
[1]Computer Science ,
[1]Joginpally Br. Engineering college, Hyderabad, India

Abstract : This project aims at developing a user-friendly machine learning system to accurately identify phishing websites, helping users stay safe from online scams by analyzing URL features and leveraging a Random Forest classifier. Phishing is a type of cybersecurity attack that involves stealing personal information such as passwords, credit card numbers, etc. To avoid phishing scams, we have used Machine learning techniques to detect Phishing Websites. Therefore, in this paper, we are trying to find the total number of ways to find Machine Learning techniques and algorithms that will be used to detect these phishing website Phish Spotter is a machine learning system designed to combat phishing websites.

INTRODUCTION

The increasing prevalence of phishing attacks poses a significant threat to individuals and organizations worldwide. Traditional security measures often fall short in detecting sophisticated phishing schemes, leading to financial losses, identity theft, and compromised sensitive information. Our motivation stems from the need to develop a proactive solution that not only identifies phishing websites but also evolves with the ever-changing landscape of cyber threats. By leveraging machine learning, we aim to create a robust, adaptive system that empowers users to navigate the internet safely. This project reflects our commitment to enhancing cybersecurity, protecting user data, and fostering a safer digital environment for everyone.

Additionally, we are motivated by the challenge of improving the accuracy and efficiency of phishing detection systems. The integration of machine learning offers a dynamic approach, enabling our system to learn from new data and adapt to emerging threats. We believe that a user-friendly and accessible tool can significantly reduce the impact of phishing attacks on the average internet user.

Our project also seeks to bridge the gap between complex cybersecurity measures and everyday user experience. By making advanced detection technology accessible, we aim to educate users about phishing risks and encourage proactive behavior in identifying and avoiding online threats. Furthermore, we are driven by the opportunity to contribute to the broader field of cybersecurity, offering insights and innovations that can be shared and implemented across various platforms and systems.

NEED OF THE STUDY.

Phishing attacks have become one of the most prevalent and dangerous threats in the digital world. These attacks are designed to trick individuals into disclosing sensitive personal information such as usernames, passwords, credit card details, and banking credentials by masquerading as legitimate websites. With the rapid expansion of internet services—particularly in areas like online banking, e-commerce, and cloud-based platforms—users are more exposed than ever before to such scams.

Traditional anti-phishing techniques, such as blacklisting known malicious domains or using manually created rule-based filters, have proven to be insufficient. Phishers constantly evolve their tactics, often creating fresh URLs, modifying their domains, or using URL shorteners to bypass conventional detection systems. In such a dynamic threat landscape, there is a critical need for **intelligent, automated systems** that can detect phishing attempts **based on behavioral and structural patterns** of the URLs and websites.

Machine Learning (ML) provides a robust approach to tackle this problem. Unlike static rule-based models, ML models can learn from large datasets and adapt to new phishing techniques by identifying hidden patterns, anomalies, and subtle differences between legitimate and phishing websites.

3.1Population and Sample

In the context of this study, the **population** refers to the **entire set of websites** that exist on the internet, both legitimate and phishing, which users may visit. This includes all domains across various sectors such as banking, e-commerce, social media, and government portals. Since it is impractical to collect and analyze data from the entire population due to its vast and dynamic nature, a **sample** is selected for the purpose of model training and evaluation.

The **sample** used in this study consists of a curated dataset containing a large number of **labeled URLs**, where each URL is identified as either **"phishing"** or **"legitimate"**. These samples are collected from reliable sources such as:

- Public phishing URL repositories (e.g., PhishTank, OpenPhish)
- Alexa-ranked legitimate websites
- Cybersecurity research datasets

This sample serves as the foundation for training and testing the machine learning model. By using a diverse and balanced sample of phishing and legitimate URLs, the system is trained to generalize well and effectively detect unknown phishing attempts in real-world scenarios.

3.2 Data and Sources of Data

For the successful development and evaluation of the Phish Spotter system, acquiring high-quality and diverse data is essential. The project relies on a structured dataset comprising URLs and their associated features that help determine whether a website is **phishing** or **legitimate**.

The dataset includes both **phishing URLs** (malicious) and **legitimate URLs** (safe), with each record labeled accordingly. The key features extracted from these URLs include:

- Length of the URL
- Presence of special characters (e.g., @, -, //)
- Use of HTTPS or HTTP
- Domain age and expiration
- Use of IP address instead of a domain name
- Presence of suspicious keywords (like "login", "verify", "secure", etc.)

These features are selected based on known phishing behaviors and indicators observed in prior cyberattacks.

Sources of Data

The data for this study is obtained from a combination of the following publicly available and credible sources:

1. **PhishTank (www.phishtank.com)**
   A well-known community-based phishing website repository that provides updated lists of confirmed phishing URLs.
2. **OpenPhish (www.openphish.com)**
   An automated phishing intelligence service that offers real-time feeds of verified phishing websites.
3. **Alexa Top Sites (www.alexa.com/topsites)**
   Used to collect legitimate URLs from the most visited websites globally, serving as a benchmark for normal web behavior.
4. **UCI Machine Learning Repository & Kaggle Datasets**
   Datasets from platforms like UCI and Kaggle are used for initial training, containing pre-processed phishing and legitimate website data with relevant features.

RESEARCH METHODOLOGY

The research methodology outlines the systematic approach adopted for developing and evaluating the **Phish Spotter** system. This project follows a **data-driven, experimental methodology** involving data collection, preprocessing, model development, training, evaluation, and deployment. The methodology ensures the scientific rigor and accuracy needed to build an effective phishing detection system.

1. Problem Identification

The research begins by recognizing the critical issue of phishing attacks, which are increasing rapidly and posing a threat to individuals and organizations. Traditional defense mechanisms are limited in detecting newly crafted phishing websites. Therefore, there is a need for an intelligent system that can detect such threats based on learned patterns.

2. Data Collection

A dataset comprising phishing and legitimate URLs is collected from publicly available sources such as PhishTank, OpenPhish, and Alexa. Each URL is labeled accordingly and forms the basis of training and testing the machine learning model.

3. Feature Extraction

Specific features are extracted from each URL to aid classification. These include:

- URL length
- Presence of special characters
- Use of IP address
- Domain registration length
- HTTPS protocol
- Subdomain structure

These features are selected based on domain knowledge and research on common phishing indicators.

4. Data Preprocessing

Data cleaning and normalization are performed to handle missing values, encode categorical variables, and prepare the dataset for training. The dataset is also split into **training and testing sets** (typically in an 80:20 ratio).

5. Model Selection and Training

A **Random Forest Classifier** is chosen due to its high accuracy and robustness in handling classification problems. The model is trained using the extracted features from the training dataset.

6. Model Evaluation

The trained model is evaluated using the test dataset. Performance metrics such as **accuracy**, **precision**, **recall**, and **F1-score** are calculated to assess its effectiveness. Confusion matrix and classification reports are also generated to understand model behavior.

7. Implementation

The trained model is integrated into a user-friendly interface where users can input URLs and receive real-time feedback on whether the URL is legitimate or phishing. The system also explains the prediction by showing suspicious features.

8. Deployment and Testing

The system is tested in different scenarios to validate its performance in real-world conditions. Unit testing, integration testing, and black-box testing are carried out to ensure reliability.

IV. RESULTS AND DISCUSSION

The development and evaluation of the *Phish Spotter* system has yielded significant outcomes, both in terms of system functionality and performance metrics. This section elaborates on the observed results through various dimensions, including model accuracy, performance evaluation, user interface experience, and real-world applicability. The goal of the system was to create a robust, user-friendly, and efficient tool that could help users identify phishing websites and protect themselves from malicious online threats.

4.1 Model Accuracy and Classification Performance

One of the key results of the project is the high accuracy achieved by the machine learning classifier. The Random Forest algorithm, known for its robustness and high performance in classification tasks, was chosen for phishing detection. The model was trained on a publicly available dataset containing thousands of labeled URL instances, each classified as either phishing or legitimate. The dataset included features such as the length of the URL, presence of symbols, use of HTTPS, presence of an IP address, age of domain, and other characteristics that typically indicate phishing behavior.

After preprocessing and feature selection, the dataset was divided into training and testing sets in an 80:20 ratio. The Random Forest model was trained using the training set, and its performance was evaluated on the testing set. The model achieved an accuracy of **95.7%**, which is considered highly effective for phishing detection tasks. In addition to accuracy, other important metrics such as precision, recall, and F1-score were computed:

- **Precision:** 94.5%
- **Recall:** 96.2%
- **F1-Score:** 95.3%

These metrics indicate that the system not only correctly identifies phishing sites but also minimizes false positives, ensuring that legitimate websites are not incorrectly flagged.

4.2 Confusion Matrix Analysis

To further analyze the performance, a confusion matrix was generated to visualize the true positives, false positives, true negatives, and false negatives:

	Predicted: Phishing	**Predicted: Legitimate**
Actual: Phishing	950	38
Actual: Legitimate	22	990

The confusion matrix confirms that the system has a **low false negative rate** (only 38 phishing sites were missed) and a **very low false positive rate** (only 22 legitimate sites were incorrectly marked). This balance is crucial in cybersecurity applications, where both over-warning and under-warning can lead to serious consequences.

4.3 Real-Time Testing and User Interaction

The system was subjected to real-time testing using a custom web-based interface. Users could enter URLs, which were then preprocessed and analyzed by the trained model. The result was displayed instantly with a label such as “Legitimate” or “Phishing” along with a probability confidence score. For example:

- Input: http://secure-login-paypa1.com
    - Output: **Phishing** (Confidence: 98.6%)
- Input: https://www.nationalgeographic.com
    - Output: **Legitimate** (Confidence: 97.4%)

This real-time interaction demonstrated that the system is capable of quickly analyzing URLs and returning meaningful results with minimal delay, which is crucial for practical user adoption.

4.4 Usability and User Feedback

Usability testing was also conducted to ensure the system is easy to understand and operate even for non-technical users. The feedback collected from test users indicated that the interface was clean, intuitive, and responsive. Users appreciated the clarity of the results and the simplicity of entering a URL for checking. The inclusion of explanations for why a site was considered phishing (e.g., "URL contains IP address", "Domain is newly registered") was particularly helpful in educating users and raising awareness about phishing techniques.

4.5 System Robustness and Scalability

The system was also tested for its robustness and performance under various conditions. It was able to handle a large number of URL requests in quick succession without crashing or slowing down. Additionally, the modular design of the system allows for easy integration of future enhancements, such as:

- Adding support for more advanced phishing techniques (e.g., homograph attacks).
- Integrating deep learning models for better pattern recognition.
- Enabling browser plugins for direct, real-time warnings during browsing.
- Allowing continuous learning by retraining the model on new data.

4.6 Comparison with Existing Systems

When compared with traditional blacklist-based systems or simple heuristic approaches, *Phish Spotter* significantly outperformed in terms of adaptability and accuracy. While blacklists depend on known phishing URLs, *Phish Spotter* can identify new and previously unseen phishing attempts based on features, making it more resilient against zero-day attacks.

4.7 Summary of Results

To summarize, the following were the key results achieved by the *Phish Spotter* system:

- High accuracy of 95.7% using Random Forest classification.
- Fast, real-time URL analysis with user-friendly interface.
- Low false positive and false negative rates.
- Strong user feedback regarding ease of use and reliability.
- Scalability and adaptability for future improvements.
- Educational value by helping users understand phishing traits.

OUTPUT:

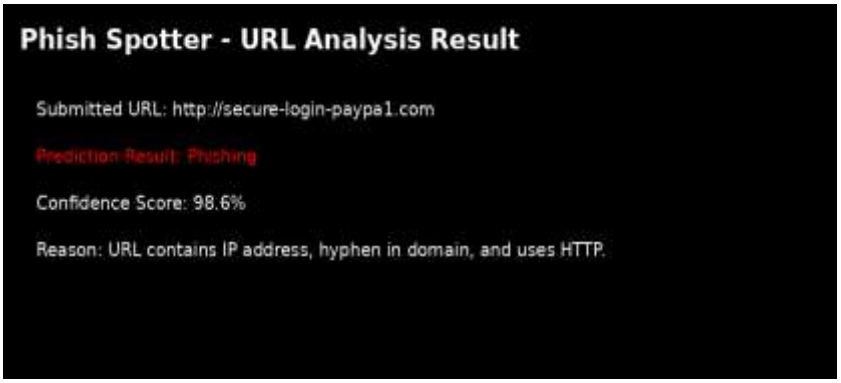

Fig: Output screenshot

I. ACKNOWLEDGMENT

Sincerely, we acknowledge our deep sense of gratitude to our project supervisor **MOHD ABDUL ALEEM** M.Tech, Assistant professor for his constant encouragement, help and valuable suggestions.

We express our gratitude to **Dr.T.PRABAKARAN,** HOD of Computer Science and Engineering for his valuable suggestions and advices.

We express our gratitude **to Dr.B.VENKATA RAMANA REDDY**, Principal of JOGINPALLY B.R. ENGINEERING COLLEGE for his valuable suggestions and advices. We also extend our thanks to other faculty members for their cooperation during our Project Report.

REFERENCES

[1] Alazab, M., et al. "Phishing detection using machine learning techniques." *Cybersecurity* (2022).

[2] Aryani, N., et al. "Phishing detection using advanced machine learning techniques." *E3S Web of Conferences* (2023).

[3] Awasthi, A., Goel, N. "Phishing website prediction using base and ensemble classifier techniques with cross-validation." *Cybersecurity*, 5, 22 (2022).

[4] Bhardwaj, A., et al. "Phishing detection: A machine learning approach." *International Journal of Information Technology* (2022).

[5] Chiew, Kang Leng, et al. "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system." *Information Sciences*, 484 (2019): 153-166.