WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

PhishGuard – Phishing Detection Tool

VIDHUR CHENGAPPA [1], NIIHARKA S [1], MR. SHASHIKIRAN A[2]

[1]UG Scolar, Department of Computer Applications, BMS College Of Commerce and Management, India
[2]Assistant Professor, Department of Computer Applications, BMS College Of Commerce and Management, India

A B S T R A C T :

Phishing attacks are among the most prevalent and dangerous threats in today's digital world, often tricking users into revealing sensitive information through deceptive URLs and fake websites. This project presents a Phishing URL Detection Website that leverages machine learning to accurately classify URLs as either safe or suspicious. The system is built using a full-stack architecture: the frontend is developed with HTML and JavaScript, allowing users to input potentially malicious URLs through a simple interface; the backend, powered by Node.js, processes the input and communicates with a Python-based machine learning model.

The ML model extracts features from the URL and predicts a phishing probability score. If the score is below a certain threshold , the URL is flagged as phishing and if it is higher than the threshold, it is considered safe. This approach ensures fast, real-time detection and improves user awareness against phishing threats. The project is scalable, lightweight, and serves as a practical example of how intelligent systems can enhance cybersecurity in both personal and organizational contexts.

Keywords: Phishing Detection, Cybersecurity, Machine Learning, URL Analysis, URL Classification , Behavioural Analysis, Threat Prevention, Suspicious Link Detection, Online Threat Protection.

Introduction

In today's digital age, phishing has emerged as one of the most common and dangerous cybersecurity threats, targeting individuals, businesses, and organizations worldwide. Phishing attacks typically involve fraudulent websites or URLs that mimic legitimate services to deceive users into revealing sensitive information such as passwords, banking credentials, or personal data. These attacks are often difficult to detect manually, especially for non-technical users, and traditional protection methods like blacklists or browser filters may not identify newly created phishing URLs in time.

To address this growing concern, this project presents a Phishing URL Detection Website that utilizes machine learning to predict the legitimacy of a given URL. The platform is built with a full-stack approach: the frontend uses HTML and JavaScript to provide a user-friendly interface where users can submit suspicious URLs, while the backend is powered by Node.js, which processes requests and communicates with a Python-based machine learning model. This model analyzes key features of the URL and returns a prediction score.

If the score is below 75%, the URL is flagged as suspicious or phishing; otherwise, it is marked as safe. This real-time, intelligent system offers users a quick and effective way to protect themselves against online phishing threats, enhancing both awareness and digital safety

Literature Review

1. Blacklist-Based Detection: Traditional systems rely on blacklists like Google Safe Browsing or Phishtank to block known phishing URLs, but they cannot detect new or zero-day phishing websites.
2. Heuristic-Based Approaches: These methods analyse URL characteristics such as length, use of special characters, and presence of subdomains. While helpful, they often produce false positives and are not adaptive.
3. Rule-Based Systems: Security tools use predefined rules to classify URLs, but these rigid rules can be bypassed by attackers who continuously evolve their techniques.
4. Machine Learning in Phishing Detection: Studies show that ML models (e.g., Decision Trees, Logistic Regression, SVM) trained on URL features outperform traditional methods in detecting phishing attacks.
5. Feature Engineering from URLs: Research highlights the importance of extracting meaningful features from URLs, such as the number of digits, hyphens, and the domain structure, to improve detection accuracy.
6. Real-Time Detection Models: Literature supports the need for real-time phishing detection tools, as phishing sites are often short-lived and require immediate response systems.

7. Web-Based Detection Systems: Several academic works and prototypes demonstrate the effectiveness of integrating ML models into web platforms for user-friendly, on-demand phishing protection.

Methodology

The Phishing URL Detection Website operates through a series of stages that involve collecting user input, extracting URL features, and using machine learning to classify the input as safe or phishing. These stages are outlined below:

3.1 URL Input Submission

- The system provides a simple web interface where users can enter or paste a suspicious URL for analysis.
- This input mechanism is designed for ease of use, supporting real-time detection with instant feedback upon submission.

3.2 Feature Extraction

- The system analyzes static URL features such as length, number of special characters, subdomain count, use of HTTPS, presence of IP addresses, and keyword patterns.
- These features are pre-processed and converted into numerical values suitable for machine learning input.
- This step is essential for identifying common traits associated with phishing URLs.

3.3 Model Training and Testing

- A labelled dataset containing both phishing and legitimate URLs is used to train a supervised machine learning model (e.g., Logistic Regression, Decision Tree, or XGBoost).
- The model is trained to recognize patterns in the features and distinguish between safe and malicious URLs.
- Evaluation metrics such as accuracy, precision, recall, and F1-score are used to validate the model's performance.

3.4 Real-Time URL Classification

- Once the model is trained, it is deployed in the backend to analyse new URLs submitted through the frontend.
- When a URL is received, features are extracted and passed to the model, which returns a prediction score.
- If the score is below 0.75, the URL is classified as suspicious or phishing; if 0.75 or higher, it is classified as safe.

3.5 Loading of URL Dataset

- A CSV file containing labelled examples of phishing and legitimate URLs is used during training and evaluation.
- The dataset includes a variety of phishing patterns to improve the model's generalization and robustness.

3.6 Python Libraries Used

The following Python libraries were utilized during development and training of the machine learning model:

- **Pandas** – for dataset loading and preprocessing
- **NumPy** – for numerical operations
- **Scikit-learn** – for model training, testing, and evaluation
- **Matplotlib** – for visualization of performance metrics
- **Seaborn** – for feature correlation and data distribution plotting
- **Joblib** – for saving and loading the trained model efficiently

Results

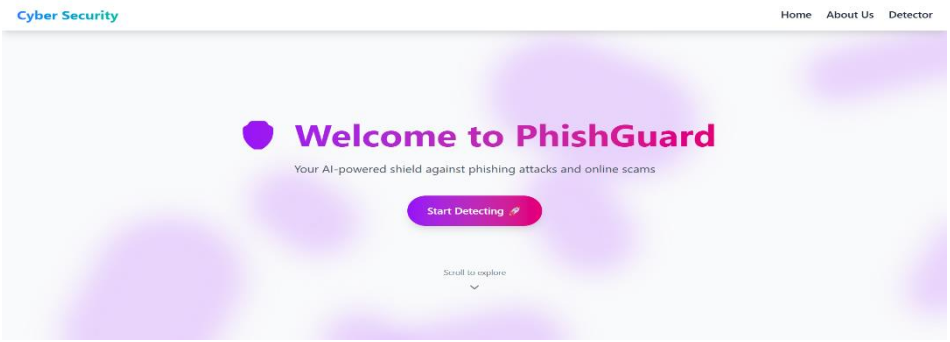

Figure 1

This is the home page of the web application. It displays a welcome message and provides navigation to the different features of the page

Figure 2

This is our **About** page, where users can test their knowledge of phishing and learn about common phishing indicators, as well as the appropriate actions to take if they encounter a phishing attempt. Our quiz includes questions on topics such as the common features of phishing emails, what to do when receiving an unexpected link from a colleague, and how to respond if a website unexpectedly asks for your password, among others.

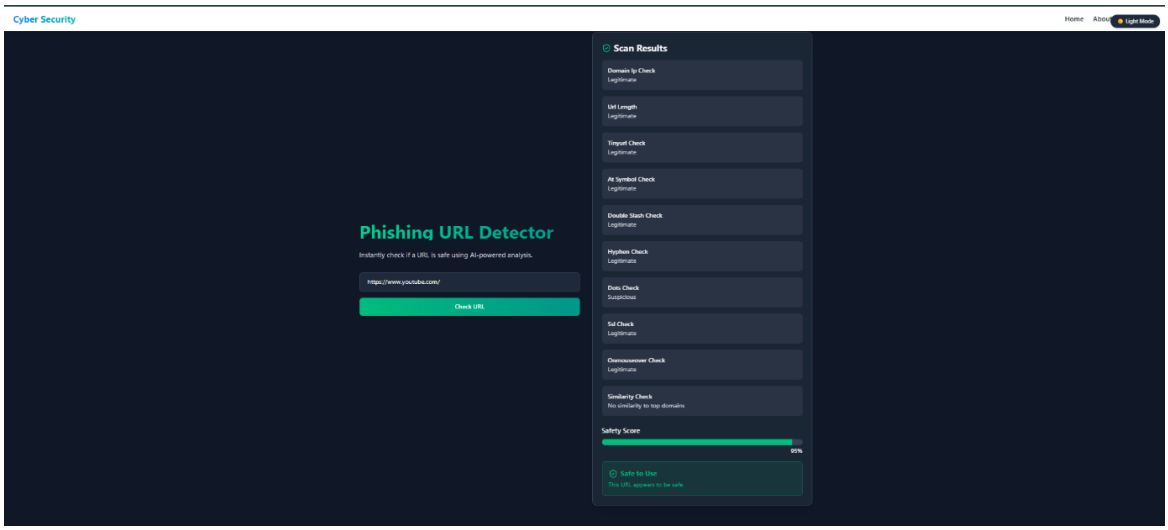

Figure 3

This is the **Detector** page, where users can enter a suspected URL to check whether it is a phishing attempt or not. In this case, it shows that the link is safe to use.

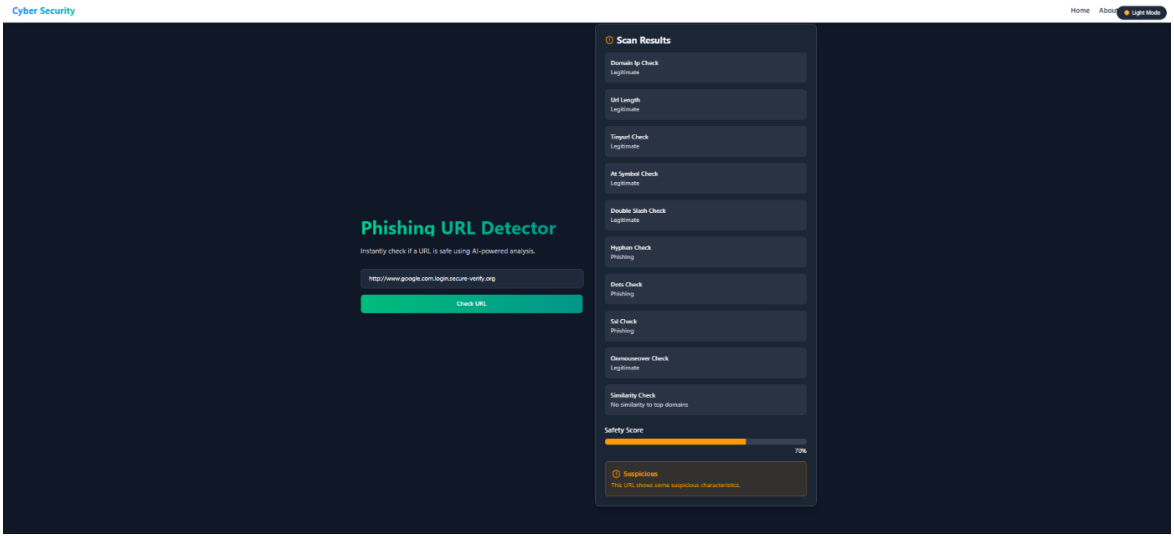

Figure 4

At present, the system has completed its analysis of the uploaded URL and identified indicators consistent with the features of a phishing attempt.

Conclusion

The Phishing URL Detection Website provides a simple yet effective solution for identifying potentially harmful URLs using machine learning. By combining a user-friendly interface with intelligent backend analysis, the system allows users to check the safety of any URL in real time. The model analyzes key features of the URL and classifies it based on a defined risk threshold. This project not only enhances online safety but also raises user awareness about phishing threats. With future improvements, such as real-time threat integration and browser support, the tool can become a powerful asset in everyday cybersecurity practices.

Acknowledgements

I would like to express my sincere gratitude to my mentors, faculty members, and peers for their valuable guidance, support, and encouragement throughout the development of this ransomware detection tool. I also extend my appreciation to the open-source community for providing essential tools, datasets, and frameworks that made this work possible.

REFERENCES

1. Aburrous, M., Hossain, M. A., Dahal, K., & Thabtah, F. (2010). Intelligent phishing detection system for e-banking using fuzzy data mining. *Expert Systems with Applications*, 37(12), 7913–7921. **https://doi.org/10.1016/j.eswa.2010.04.044**
2. Basnet, R., Sung, A. H., & Liu, Q. (2012). Learning to detect phishing URLs. *International Journal of Research in Computer Science*, 2(3), 47–56.
3. IEEE Std 1619™-2007 (2008). IEEE Standard for Cryptographic Protection of Data on Block-Oriented Storage Devices. *IEEE Computer Society*.
4. Marchal, S., Saari, K., Singh, N., & Asokan, N. (2016). Know your phish: Novel techniques for detecting phishing sites and their targets. ***Proceedings of the IEEE 36th International Conference on Distributed Computing Systems***, 323–333. https://doi.org/10.1109/ICDCS.2016.45
5. Mohammad, R. M., Thabtah, F., & McCluskey, L. (2014). Predicting phishing websites based on self-structuring neural network. ***Neural Computing and Applications***, 25(2), 443–458. https://doi.org/10.1007/s00521-013-1490-z
6. Xiang, G., Hong, J., Rose, C., & Cranor, L. (2011). Cantina+: A feature-rich machine learning framework for detecting phishing websites. ***ACM Transactions on Information and System Security (TISSEC)***, 14(2), 1–28. https://doi.org/10.1145/2019599.2019606
7. Jain, A. K., & Gupta, B. B. (2018). Phishing detection: Analysis of visual similarity-based approaches. ***Security and Privacy***, 1(1), e8. https://doi.org/10.1002/spy2.8
8. Verma, R., & Das, A. (2017). What's in a URL: Fast feature extraction and classification of phishing URLs. ***Proceedings of the 3rd ACM Workshop on Artificial Intelligence and Security***, 55–66. https://doi.org/10.1145/3128572.3140453
9. Sahoo, D., Liu, C., & Hoi, S. C. (2017). Malicious URL detection using machine learning: A survey. ***arXiv preprint arXiv:1701.07179***. **https://arxiv.org/abs/1701.07179**