PHISHING DETECTION IN WEBSITES

Mrs. M.AMANI

Assistant Professor

LAKKIDI SRIHITH REDDY, MANDADI PRAVEEN, PITTALA ADHITHYA, NANDIPATI SONIA

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

TKR COLLEGE OF ENGINEERING AND TECHNOLOGY (AUTONOMOUS)

(Accredited By NBA and NAAC with 'A^{+}' Grade)

Medbowli, Meerpet, Balapur (M), Hyderabad-500097

ABSTRACT

This project proposes a phishing detection system that functions as a real-time Chrome browser extension powered by Artificial Intelligence. Traditional methods relying solely on blacklists or URL structure often fail to catch cleverly disguised phishing websites. To overcome this, our approach introduces a hybrid detection model that first checks websites against a known phishing database for instant blocking. If no match is found, the system extracts both URL features and visible text content from the page. Using ensemble learning for URL analysis, the system evaluates threats with deeper context. The ensemble model generates the final result by combining evaluations from both machine learning and deep learning models. This two-layer detection significantly improves accuracy and reduces false positives by going beyond rule-based logic. The result is a faster, smarter, and more reliable tool for protecting users during everyday browsing.

Phishing is one of the most common and dangerous cyberattacks, where attackers create fake websites that mimic legitimate ones to steal sensitive information such as usernames, passwords, and banking details. With the rapid growth of online services, phishing attacks have become more sophisticated, making it difficult for users to distinguish between genuine and malicious websites. Traditional security methods such as blacklists and rule-based systems are no longer sufficient to detect newly generated phishing websites.

This project proposes a machine learning-based phishing detection system that analyzes website features and classifies

them as legitimate or phishing. The system uses various features such as URL structure, domain age, presence of HTTPS, website content, and abnormal behavior patterns. These features are extracted and fed into machine learning algorithms like Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines to train a predictive model.

The proposed system improves detection accuracy and reduces false positives compared to traditional approaches. It can detect zero-day phishing attacks by identifying suspicious patterns rather than relying solely on known phishing databases. The system can be integrated into web browsers or deployed as a web application to provide real-time protection to users.

INTRODUCTION

With the increasing use of the internet for banking, shopping, and communication, cybersecurity threats have also grown significantly. One of the most common threats is phishing, where attackers attempt to trick users into providing sensitive information by creating fake websites that closely resemble legitimate ones. These websites often look identical to trusted platforms, making it difficult for users to identify them.

Phishing attacks are typically carried out through emails, social media messages, or fake advertisements that redirect users to malicious websites. Once the user enters their personal information, attackers can misuse it for financial fraud, identity theft, or unauthorized access.

Traditional phishing detection methods rely on blacklists, which store known phishing URLs. However, attackers frequently create new websites, making blacklist-based detection ineffective against zero-day attacks. Therefore, there is a need for intelligent systems that can detect phishing websites based on their characteristics.

Machine learning provides a powerful solution by enabling systems to learn patterns from data and make predictions. By analyzing features such as URL length, presence of special characters, domain registration details, and webpage content, machine learning models can identify suspicious websites with high accuracy.

This project focuses on building a phishing detection system using machine learning techniques. The system extracts relevant features from websites and trains classification models to distinguish between phishing and legitimate sites. The goal is to provide a fast, accurate, and

scalable solution that can be used in real-time applications.

EXISTING SYSTEM

Existing phishing detection systems primarily rely on traditional techniques such as blacklist-based detection, heuristic rules, and manual analysis. Blacklist-based systems maintain a database of known phishing URLs and compare visited websites against this list. If a match is found, the website is flagged as malicious. While this method is simple and fast, it fails to detect new phishing websites that are not yet included in the database.

Heuristic-based approaches use predefined rules to identify suspicious websites. For example, URLs containing excessive special characters, mismatched domain names, or missing security certificates may be flagged as phishing. However, these rules are limited and can be easily bypassed by attackers who design more sophisticated websites.

Another approach involves visual similarity detection, where the system compares the appearance of a website with known legitimate sites. Although this method can detect cloned websites, it is computationally expensive and not suitable for real-time detection.

Manual verification is also used in some cases, where security experts analyze suspicious websites. This process is time-consuming and not scalable, especially with the increasing number of phishing attacks.

Disadvantages

- Cannot detect zero-day phishing attacks
- High dependency on predefined rules
- Requires frequent updates to blacklists
- Limited scalability and efficiency
- Higher chances of false positives and false negatives

PROPOSED SYSTEM

The proposed system uses machine learning techniques to detect phishing websites based on their features rather than relying solely on known phishing databases. This approach allows the system to identify new and unknown phishing websites effectively.

The system works in several stages. First, it collects a dataset containing both legitimate and phishing URLs. Next, feature extraction is performed to analyze

various attributes such as URL length, presence of HTTPS, domain age, number of subdomains, use of IP address instead of domain name, and webpage content features.

These features are then used to train machine learning models such as Random Forest, Decision Tree, Logistic Regression, and Support Vector Machine. Among these, Random Forest often provides higher accuracy due to its ensemble learning capability.

Once the model is trained, it can classify new websites as phishing or legitimate in real-time. The system can be deployed as a browser extension or web application, allowing users to check the safety of websites before accessing them.

The proposed system improves accuracy, reduces false positives, and provides faster detection compared to traditional methods. It also has the capability to adapt to new phishing techniques by retraining the model with updated datasets

Advantages

- Detects zero-day phishing attacks
- High accuracy using machine learning
- Real-time detection capability
- Reduced dependency on blacklists
- Scalable and adaptable system

SYSTEM ARCHITECTURE

TECHNOLOGIES USED

 Python – Main programming language for development

 Scikit-learn – Used for machine learning algorithms

 Pandas& NumPy – Data processing and numerical operations

 Matplotlib / Seaborn – Data visualization

 Flask / Django – Web application development

 HTML, CSS, JavaScript – Front-end design

 BeautifulSoup / Regex – Feature extraction from websites

 Jupyter Notebook / VS Code – Development tools

 Git& GitHub – Version control

 Kaggle Datasets – Training data for the model

APPLICATIONS

 Web browser security (detects unsafe websites)

 Email filtering (blocks phishing links in emails)

 Online banking protection

 E-commerce website security

 Corporate network protection

 Social media link scanning

 Mobile application security

 Cybersecurity tools integration

CHALLENGES & RISKS

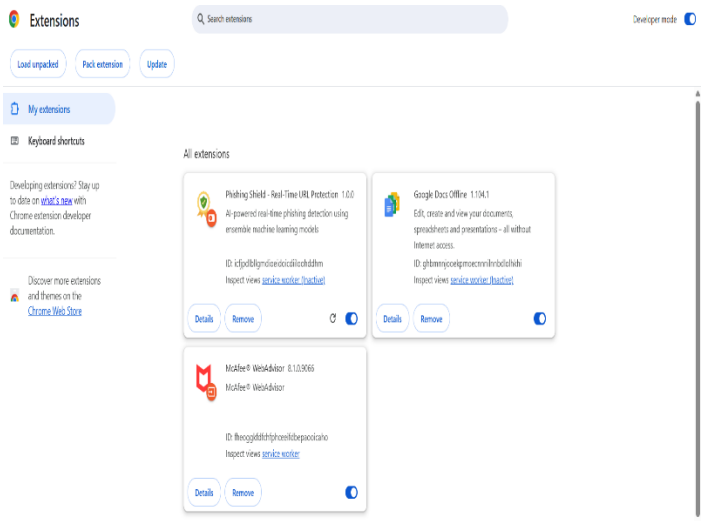
Challenges

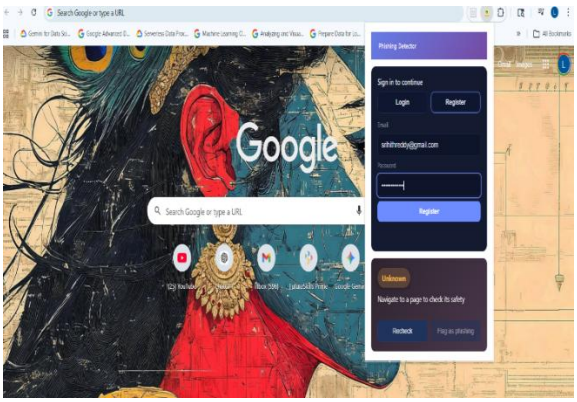
- Detecting new (zero-day) phishing websites
- Selecting important features
- Handling imbalanced datasets
- Maintaining real-time performance
- Reducing false positives/negatives

Risks

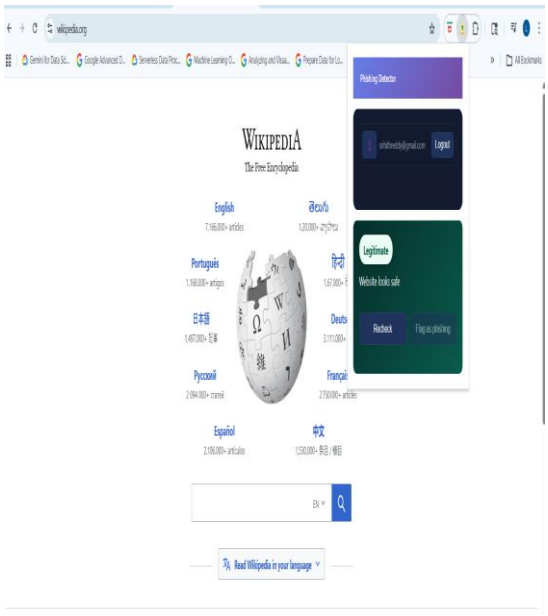
- Failure to detect phishing attacks
- Privacy issues with user data
- Model overfitting
- Adversarial attacks by hackers
- Need for regular updates

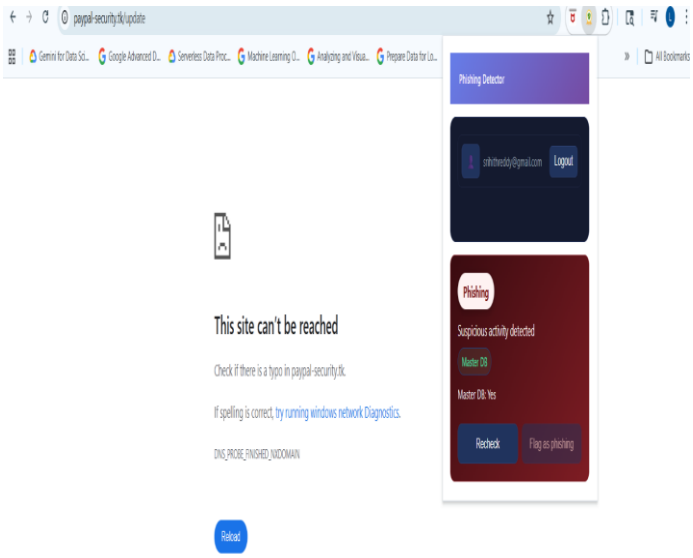
RESULTS

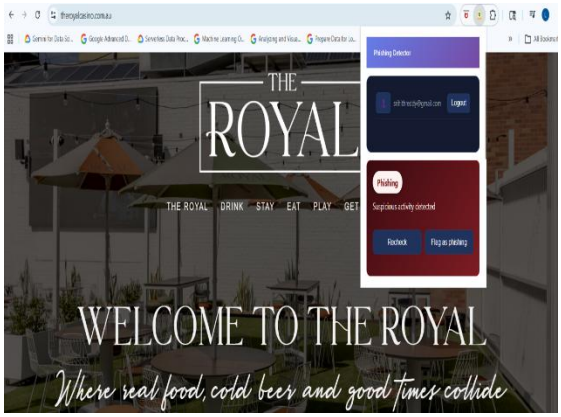


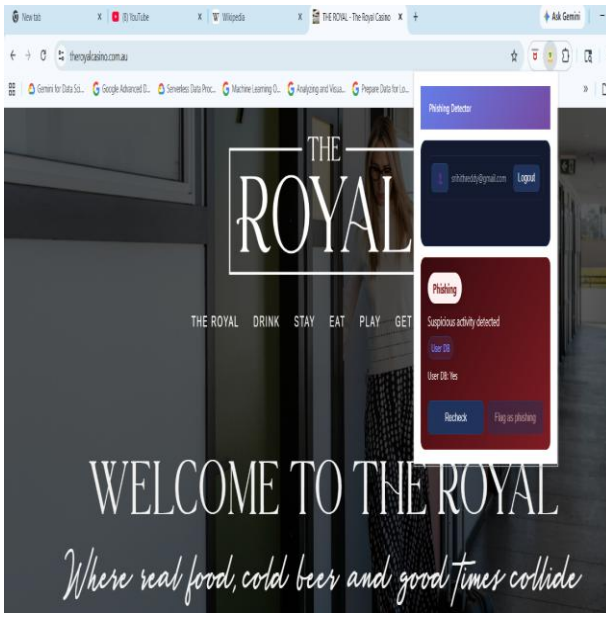


International Journal of Engineering Research and Science & Technology









CONCLUSION

Phishing attacks continue to pose a serious threat to online users and organizations. Traditional detection methods are no longer sufficient due to the increasing sophistication of phishing techniques. This project presents a machine learning-based approach to detect phishing websites effectively.

By analyzing various features of websites and using classification algorithms, the system can accurately identify phishing websites, including zero-day attacks. The proposed system offers improved accuracy, scalability, and real-time detection capabilities compared to existing methods.

The implementation demonstrates that machine learning can significantly enhance cybersecurity measures. The system can be further improved by incorporating deep learning models, real-time threat intelligence, and browser integration.

FUTURE ENHANCEMENTS

- Use Deep Learning (CNN, LSTM)
- Real-time browser extension
- Integration with threat intelligence APIs
- AI-based URL content scanning
- Mobile application for phishing detection

BIBILIOGRAPHY

[1]F.Salahdine,Z.E.Mrabet,andN.Kaabouch,‘‘Phishingattacksdetectiona machine learning-based approach,’’ in Proc. IEEE 12th Annu. Ubiquitous Comput., Electron. Mobile Commun. Conf. (UEMCON), New York, NY, USA, Dec. 2021, pp. 250–255.

[2]M. Baykara and Z. Z. Gürel, ‘‘Detection of phishing attacks,’’ in Proc. 6th Int. Symp. Digit. Forensic Secur. (ISDFS), Antalya, Turkey, Mar. 2018, pp. 1–5.

[3]K. L. Chiew, K. S. C. Yong, and C. L. Tan, ‘‘A survey of phishing attacks: Their types, vectors and technical approaches,’’ Expert Syst. Appl., vol. 106, pp. 1–20, Sep. 2018.

[4]F. Yahya, ‘‘Detection of phishing websites using machine learning approaches,’’ in Proc. Int. Conf. Data Sci. Appl. (ICoDSA), Bandung, Indonesia, 2021, pp. 40–47.

[5]A. Alswailem, B. Alabdullah, N. Alrumayh, and A. Alsedrani, ‘‘Detecting phishing websites using machine learning,’’ in Proc. 2nd Int. Conf. Comput. Appl. Inf. Secur. (ICCAIS), Riyadh, Saudi Arabia, May 2019, pp. 1–6.

[6]H. Zuhair, A. Selamat, and M. Salleh, ‘‘Feature selection for phishing detection: A review of research,’’ Int. J. Intell. Syst. Technol. Appl., vol. 15, no. 2, p. 147, 2016.

[7]S.DangwalandA.-N.Moldovan,‘‘Featureselection formachinelearning based phishing websites detection,’’ in Proc. Int. Conf. Cyber Situational Awareness, Data Anal. Assessment (CyberSA), Dublin, Ireland,

Jun. 2021, pp. 1–6.

[8]P. Chinnasamy, N. Kumaresan, R. Selvaraj, S. Dhanasekaran, K. Ramprathap, and S. Boddu, ‘‘An efficient phishing attack detection using machine learning algorithms,’’ in Proc. Int. Conf. Advancements Smart, Secure Intell. Comput. (ASSIC), Bhubaneswar, India, Nov. 2022, pp. 1–6.

[9] learnopencv. An Example of FNN With One Hidden Layer. Accessed: May 5, 2024. [Online]. Available: https://learnopencv.com/understandingfeed forward-neural-networks/