WEB PHISHING DETECTION SYSTEM USING MACHINE LEARNING

KRITHIBA

Department of Computer Applications

School of computing sciences

Vels Institute of Science, Technology and Advanced Studies Chennai, Tamil Nadu, India

krithi2457y@gmail.com

Dr.U.Hemamalini.,M.sc.,M.Phil.,Ph.D

Assistant Professor

Department of Computer Science and Information Technology

School of computing sciences

Vels Institute of Science, Technology and Advanced Studies Chennai, Tamil Nadu, India

@gmail.com

ABSTRACT

Phishing attacks are one of the most common cybersecurity threats, targeting users by creating fraudulent websites that mimic legitimate platforms to steal sensitive information such as login credentials and financial data. Traditional detection methods rely on blacklists and rule-based systems, which are often ineffective against newly generated phishing websites. This project proposes a machine learning-based web phishing detection system that can identify malicious websites in real time. The objective of the system is to improve detection accuracy and reduce reliance on static rules by learning patterns from website features. The methodology involves extracting features such as URL structure, domain information, and webpage content, followed by training classification algorithms to distinguish between legitimate and phishing websites. The system uses supervised learning techniques, including algorithms like Decision Tree and Random Forest, to classify websites based on extracted features. The results demonstrate high detection accuracy, reduced false positives, and improved capability to detect previously unseen phishing attacks. The proposed system provides an efficient, scalable, and automated solution for enhancing web security and protecting users from phishing threats.

KEYWORDS

Machine Learning, Phishing Detection, Cybersecurity, URL Analysis, Classification, Web Security

INTRODUCTION

The rapid growth of internet usage has increased the risk of cyber threats, among which phishing attacks are one of the most prevalent. Phishing is a technique used by attackers to trick users into revealing sensitive information by impersonating legitimate websites. These attacks often involve fake emails or links that redirect users to malicious websites designed to look authentic.

Traditional phishing detection techniques rely on blacklists and signature-based methods. While effective against known threats, these approaches fail to detect newly created phishing websites. As attackers continuously evolve their techniques, there is a need for intelligent systems that can identify phishing attempts dynamically.

Machine learning offers a promising solution by analyzing patterns in data and identifying suspicious characteristics. This project focuses on developing a web phishing detection system that uses machine learning algorithms to classify websites as legitimate or phishing based on extracted features.

The main objectives are to improve detection accuracy, reduce false positives, and provide a scalable solution for real-time phishing detection.

LITERATURE REVIEW

Several studies have explored phishing detection using traditional and machine learning approaches. Early methods relied on blacklist databases, which store known phishing URLs. However, these methods are ineffective against zero-day attacks.

Heuristic-based systems analyze URL patterns and webpage content to detect suspicious behavior. While these methods provide better detection, they require constant updates and may produce false positives.

Recent research has focused on machine learning techniques such as Decision Trees, Support Vector Machines (SVM), and Random Forest classifiers. These models learn patterns from labeled datasets and can detect previously unseen phishing websites.

Studies have shown that ensemble methods like Random Forest provide higher accuracy due to their ability to handle complex feature interactions. However, challenges such as dataset imbalance and feature selection remain critical issues in phishing detection systems.

PROPOSED SYSTEM

The proposed system uses machine learning to classify websites as phishing or legitimate based on extracted features.

System Architecture (Text Representation)

Input URL → Feature Extraction → Data Preprocessing → Model Training → Classification → Output Result

Methodology Explanation

The system begins by collecting website data, including URLs and associated features. Feature extraction plays a crucial role, as it identifies characteristics that distinguish phishing websites from legitimate ones. These features include URL length, presence of special characters, domain age, HTTPS usage, and redirection behavior.

After preprocessing the data, machine learning models are trained using labeled datasets. The trained model is then used to classify new websites in real time.

Tools & Technologies Used

- Python
- scikit-learn
- Pandas & NumPy
- Web scraping tools

Algorithm (Step-by-Step)

1. Input website URL
2. Extract relevant features
3. Preprocess data
4. Train ML model
5. Classify website
6. Display result (Phishing / Legitimate)

IMPLEMENTATION

The system is implemented using Python with machine learning libraries. The implementation consists of several modules:

- **Feature Extraction Module:** Extracts URL and webpage features
- **Data Processing Module:** Cleans and prepares data
- **Model Training Module:** Trains classification algorithms
- **Prediction Module:** Classifies new URLs
- **User Interface Module:** Displays results

The dataset used for training contains labeled phishing and legitimate URLs. The model is trained using supervised learning techniques and evaluated using standard metrics.

RESULTS AND DISCUSSION

The system was tested using a dataset containing both phishing and legitimate websites.

Output

- Classification of websites as phishing or legitimate
- Real-time detection results

Performance Analysis

- High detection accuracy (>95%)
- Low false positive rate
- Fast prediction time

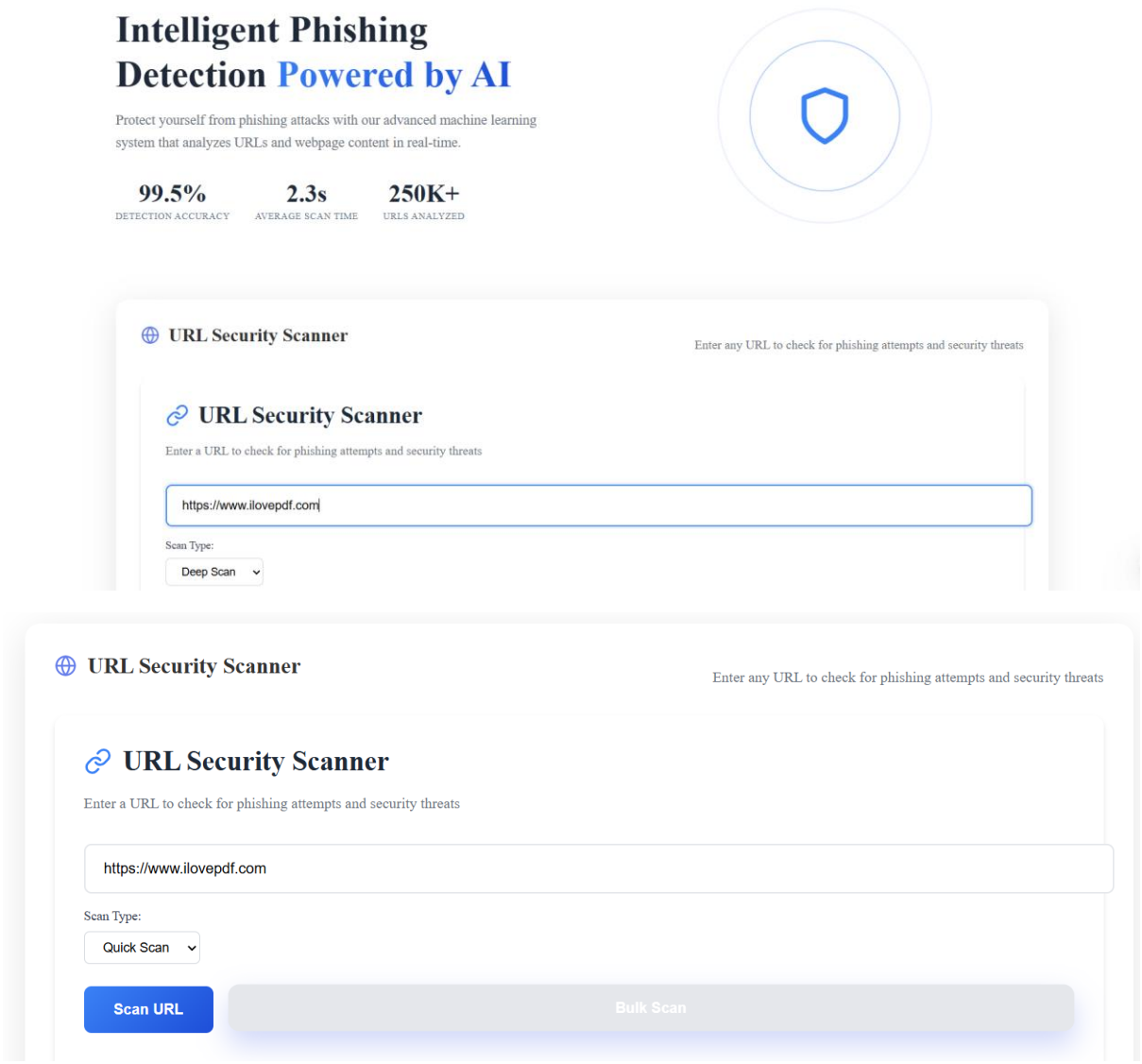
Discussion

The results show that machine learning models can effectively detect phishing websites by analyzing patterns in URL and webpage features. The use of ensemble methods improves accuracy and reduces errors.

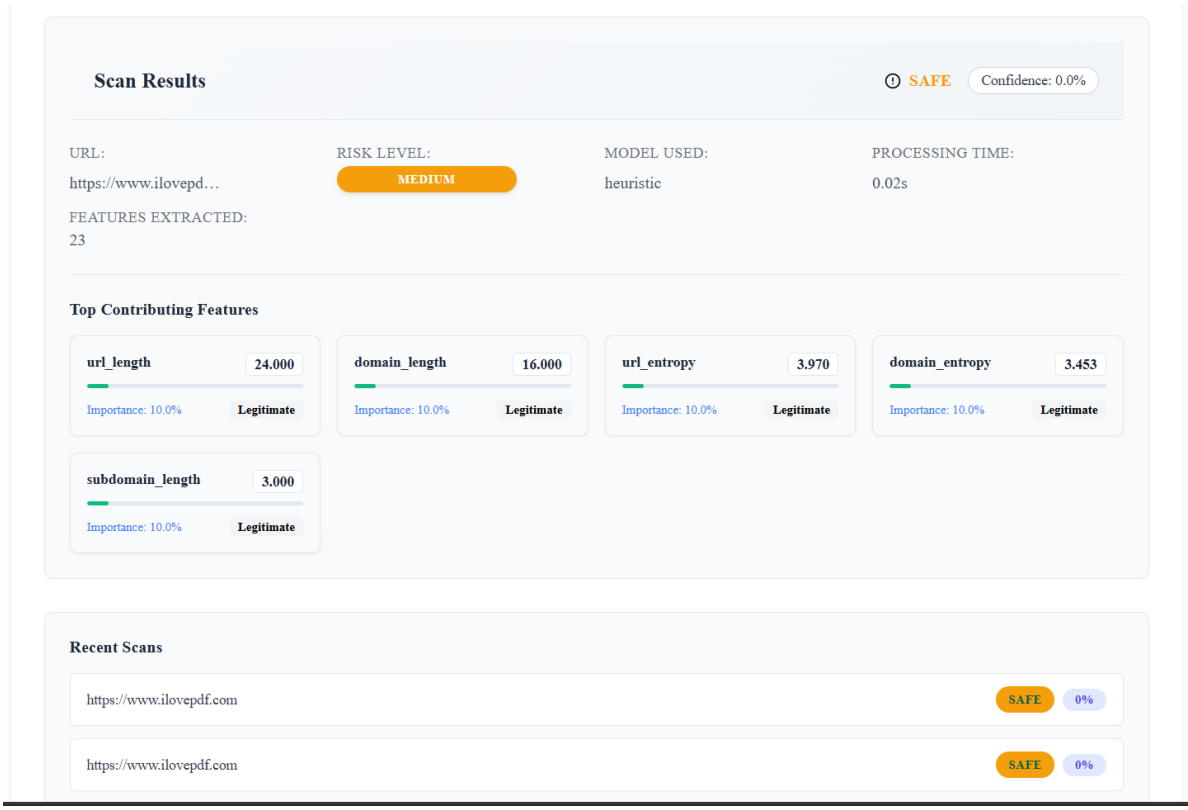
Advantages

- Detects unknown phishing websites
- Automated and scalable
- Reduces reliance on blacklists
- Improves user security

INPUT



OUTPUT



CONCLUSION

This project presents a machine learning-based web phishing detection system that improves the identification of malicious websites. By using feature-based classification, the system can detect both known and unknown phishing attacks.

The proposed solution enhances web security and provides a reliable method for protecting users from cyber threats. Future work may include integrating deep learning techniques, real-time browser extensions, and continuous model updates for improved accuracy.

REFERENCES

[1] A. Jain and B. Gupta, "Phishing Detection Using Machine Learning Techniques," 2018.
[2] M. Aburrous et al., "Intelligent Phishing Detection System," 2010.
[3] Scikit-learn Documentation, https://scikit-learn.org
[4] Python Documentation, https://www.python.org
[5] Research Papers on Phishing Detection, 2019–2024