Phishing Website Detection Using Machine Learning

Neeraj Kumar (2022UCA1887),
Abhishek Jethuri (2022UCA1911)
Branch- Computer Science and
Engineering (Artificial Intelligence)
Section-2
Netaji Subhas University of Technology
New Delhi, India

***Abstract*—Phishing attacks have become one of the most common cyber threats, tricking users into revealing sensitive information by mimicking legitimate websites. This project aims to develop a machine learning-based system to detect phishing websites automatically.**

Keywords—phishing, cybersecurity, machine-learning

I. Overview

With the increasing dependence on online services, phishing attacks have become a severe concern for individuals and organizations. Attackers create deceptive websites that look identical to legitimate ones, misleading users into sharing their credentials. Traditional security mechanisms, such as blacklists, struggle to keep up with the rapid evolution of phishing techniques. Therefore, a machine learning approach offers a more adaptive and robust solution.

A. *Objectives*

- *Develop a machine learning model to detect phishing websites based on their features.*
- *Analyze key website attributes such as URL length, presence of special characters, HTTPS usage, and domain age.*
- *Improve phishing detection accuracy compared to traditional rule-based methods.*

B. ***Project Plan***

1. **Literature Review:** Research existing phishing detection techniques and machine learning models.
2. **Dataset Collection:** Gather a labeled dataset of phishing and legitimate websites.
3. **Feature Extraction:** Identify distinguishing characteristics of phishing websites.
4. **Model Training:** Train machine learning models using supervised learning.
5. **Evaluation & Optimization:** Test and refine the model to improve accuracy.
6. **Deployment:** Develop a simple web-based interface for real-time phishing detection.

C. ***Project Deliverables***

- A trained machine learning model capable of detecting phishing websites.
- A dataset containing phishing and legitimate website features.
- A user-friendly interface for real-time URL analysis.
- A detailed project report and documentation.

II. Methodology.

A. ***Proposed Methodology***

- *Data Collection: Extract URLs from open-source datasets like PhishTank and Kaggle.*
- *Feature Engineering: Extract URL-based and domain-based features (e.g., HTTPS usage, domain age, IP addresses in URLs).*
- *Model Selection: Implement machine learning algorithms like Random Forest, Decision Trees, and Logistic Regression.*
- *Performance Evaluation: Use metrics such as accuracy, precision, recall, and F1-score to measure model effectiveness.*

B. ***Value Propositions***

- Provides an **automated** phishing detection system with high accuracy.
- Enhances **cybersecurity awareness** by educating users about risky websites.
- Helps **organizations** prevent credential theft and financial losses.
- Can be **integrated** into browsers or email security solutions.

C. ***Application***

- **Personal Use:** Users can check URLs before visiting websites.

- **Corporate Use:** Companies can integrate the model into their security infrastructure.

- **Browser Extensions:** Can be implemented as a real-time security browser extension.

D. *Tools Used*

- **Programming Language:** Python

- **Libraries:** Scikit-learn, Pandas, NumPy, BeautifulSoup, Selenium

- **Development Environment:** Jupyter Notebook / Google Colab

E. *Dataset Used*

- **PhishTank Dataset** (for phishing websites)

- **Alexa Top 1 Million Domains** (for legitimate websites)

III. More Information

A. *Domain*

- Cybersecurity

- Machine Learning

B. *Estimated Cost of Development*

- **Cloud Computing Costs (if deployed online):** $10-$50/month

- **Data Collection & Storage:** Free (using open-source datasets)

- **Development Costs:** Free (if self-developed using open-source tools)

IV. Progress Till Now (Timeline)

Task	**Status**	**Completion Date**
Literature Review	Completed	Week 1
Dataset Collection	Completed	Week 2
Feature Engineering	In Progress	Week 3

References

[1] Recent Research Paper (post-2022): Abutair, H. et al. (2023). "Machine Learning-Based Phishing Website Detection: A Review." *Cybersecurity Journal*. [DOI Link]

[2] PhishTank Database

[3] Kaggle Phishing Dataset