



PREDICTIVE PHISHING WEBSITE DETECTION THROUGH MACHINE LEARNING

RASITHABANU ABDHUL RAHIM

RESEARCH CONTEXT

- Phishing attacks are a major cybersecurity threat, targeting individuals and organizations by stealing sensitive information through deceptive websites.
- Traditional detection methods (e.g., rule-based, blacklists) are limited in handling evolving phishing tactics.
- Machine Learning (ML) offers a dynamic solution by learning patterns from data to detect phishing websites effectively.

RESEARCH PROBLEMS

- Phishing websites are becoming increasingly sophisticated, making it difficult for traditional methods to detect them accurately.
- Existing approaches often suffer from high false positives, low adaptability, and inefficiency in detecting new phishing techniques.
- There is a need for a more robust, accurate, and adaptive phishing detection system.

AIMS AND OBJECTIVES

Aim:

- To develop a predictive phishing detection system using Machine Learning models to classify websites as phishing or legitimate.

Objectives:

- Identify key features of phishing websites from URLs.
- Evaluate the performance of various ML algorithms (e.g., Random Forest, SVM, Logistic Regression).
- Develop a hybrid approach for improved accuracy in phishing detection.

LITERATURE REVIEW

•Key Studies:

- Yang et al. (2019): Used deep learning for phishing detection [1] .
- Zhu et al. (2020): Proposed DTOF-ANN model with 99.1% accuracy [2] .
- Adeyemo et al. (2020): Used ensemble learning for phishing detection with 99.4% accuracy [3] .
- Zamir et al. (2020): Achieved 97.4% accuracy using Stacking1 approach [4] .

•Findings:

- ML models, especially ensemble methods, show high accuracy in phishing detection.
- Feature selection and hybrid approaches improve detection performance.

LITERATURE GAP

Gaps Identified:

- **Data Imbalance:** The issue of **imbalanced datasets** (more legitimate sites than phishing sites) is not adequately addressed in many studies [1].
- **Hybrid Approaches:** Insufficient research on hybrid (rule-based + ML) models compared to pure ML methods, particularly regarding complexity and resource use [3].

METHODOLOGY

- **Dataset Collection:** Gathered phishing website dataset with 30 key attributes.
- **Feature Selection:** Extracted attributes such as URL length, SSL state, domain age, Google Index, etc.
- **Machine Learning Algorithms:** Trained models: Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), Decision Tree, Random Forest
- **Performance Metrics:** Evaluated using Accuracy, Precision, Recall, F2-score, and Confusion Matrix.
- **Result Analysis:** Advanced models achieve higher accuracy, balancing performance and efficiency.

PROJECT PLANNING

Week 1-2

- Literature review and dataset collection

Week 3-4

- Feature extraction and data preprocessing.

Week 5-6

- Model training and evaluation.

Week 7-8

- Hybrid model development and testing.

Week 9-10

- Final analysis and report writing.

PROJECT IMPACT AND LIMITATIONS

Impact:

- Improves cybersecurity by detecting phishing sites.
- Uses ML for accurate classification.
- Prevents identity theft & fraud.
- Automated & scalable detection.

Limitations:

- Struggles with new phishing tactics.
- Needs regular updates.
- Some models have lower accuracy.
- High computation for complex models.

REFERENCES

1. Yang et al. (2019): Phishing detection using deep learning. Yang, P., Zhao, G., & Zeng, P. (2019). *IEEE Access*, 7, 15196-15209 [1] **(Classification of Phishing Websites using Machine Learning Models)**.
2. Zhu et al. (2020): DTOF-ANN model for phishing detection. Zhu, E. J., Chen, Z., Liu, F., & Fang, X. (2020). *Applied Soft Computing*, 95, 106505 [2] **(Classification of Phishing Websites using Machine Learning Models)**.
3. Adeyemo et al. (2020): Ensemble-based logistic model trees for phishing detection. Adeyemo, V. E. et al. (2020). *ACCS 2020*, 627-641 [3] **(Classification of Phishing Websites using Machine Learning Models)**.
4. Zamir et al. (2020): Stacking1 approach for phishing URL detection. Zamir, A. et al. (2020). *The Electronic Library*, 38(1), 65-80 [4] **(Classification of Phishing Websites using Machine Learning Models)**.
5. Abdulhamit Subasi, E. K. (2020). Comparison of Adaboost with MultiBoosting for Phishing Website Detection. *Procedia Computer Science*, 168, 272 – 278 **(Online Phishing Detection Using Machine Learning)**.

REFERENCES

6. Ping Yi, Y. G. (2018). Web Phishing Detection Using a Deep Learning Framework. *Wireless Communications and Mobile Computing* (**Online Phishing Detection Using Machine Learning**).
7. Q. Abu Al-Haija and M. Al-Fayoumi, "An intelligent identification and classification system for malicious uniform resource locators (URLs)," *Neural Comput. Appl.*, pp. 1–17, 2023 (**Phishing Site Detection Using Machine Learning**).
8. Q. Abu Al-Haija, M. Alohaiy, and A. Odeh, "A lightweight double-stage scheme to identify malicious DNS over HTTPS traffic using a hybrid learning approach," *Sensors*, vol. 23, no. 7, pp. 3489, 2023 (**Phishing Site Detection Using Machine Learning**).
9. R. S. Rao and A. R. Pais, "Two level filtering mechanism to detect phishing sites using lightweight visual similarity approach," *J. Ambient Intell. Humanized Comput.*, vol. 11, no. 9, pp. 3853–3872, Sep 2020 (**Phishing Site Detection Using Machine Learning**).

REFERENCES

10. Mouton, M. Malan, L. Leenen, and H.S. Venter, "Social Engineer Attack Framework," *IEEE Conf. on Information Security for South Africa*, pp. 1-9, 2014 **(Phishing Attack Detection using Python and Machine Learning)**.
11. S. Gupta, A. Singhal, and A. Kapoor, "A literature survey on social engineering attacks: Phishing attack," *International Conf. on Computing, Communication and Automation (ICCCA)*, pp. 537-540, 2016 **(Phishing Attack Detection using Python and Machine Learning)**.
12. K. Suda Muthami, "TFAD: TCP flooding attack detection in software-defined networking using proxy-based and machine learning-based mechanisms," *Cluster Computing*, 26.2, pp. 1461-1477, 2023 **(Phishing Websites Detection using Machine Learning)**.
13. P. Naga Srinivasa, R. Panigrahi, A. Singh, and A. K. Bhoi, "Probabilistic buckshot-driven cluster head identification and accumulative data encryption in WSN," *Journal of Circuits, Systems and Computers*, 31(17), 2250303, 2022 **(Phishing Websites Detection using Machine Learning)**.



THANK YOU

RASITHABANU ABDHUL RAHIM