International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 3 (2026)

ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

EVALUATING AI-BASED TECHNIQUES FOR MALICIOUS DOMAIN DETECTION USING OPEN-SOURCE CYBERSECURITY DATASETS

1Bilal Ur Rehman, 2Anu Priya Jaishankar
1*Master of Artificial Intelligence, Airtics Education and Universidad Catolica San Antonio de Murcia*
2*Project Supervisor, Airtics Education and Universidad Catolica San Antonio de Murcia*

ABSTRACT

The rapid growth of internet-based services has increased reliance on domain name infrastructure and created new opportunities for phishing, spoofing, and malicious domain activity. This study evaluates artificial intelligence based techniques for malicious domain detection using an open-source phishing URL dataset containing 3,772 balanced samples and 89 engineered features. Logistic Regression, Random Forest, and XGBoost were trained and evaluated using an 80:20 stratified split, cross-validation, and metrics including accuracy, precision, recall, F1-score, and ROC-AUC. XGBoost achieved the strongest overall performance, with 96.29% accuracy, 98.14% phishing recall, and 0.994 ROC-AUC. SHAP analysis showed that reputation and structural URL features, including indexing status, page rank, domain age, and hyperlink patterns, were dominant contributors to model decisions. The findings indicate that ensemble learning combined with explainable AI can support accurate and transparent malicious domain detection.

Keywords: Malicious Domain Detection, Cybersecurity, Machine Learning, Phishing Detection, XGBoost, SHAP, Explainable AI

I. INTRODUCTION

Cybersecurity has become a major concern for institutions, governments, and individuals as digital systems continue to expand. Online banking, e-commerce, cloud services, and mobile applications rely heavily on domain name infrastructure, and attackers exploit this reliance through phishing pages, spoofed login portals, and malicious websites. These attacks often form the first stage of credential theft, financial fraud, ransomware infection, and broader data compromise.

Traditional defensive methods such as blacklists, signatures, and manually designed rules are reactive because they depend on threat indicators that have already been observed. Attackers can register new domains cheaply, alter URL structures, abuse temporary infrastructure, and rotate domains quickly. As a result, static detection systems often identify malicious domains only after harm has occurred.

Artificial intelligence offers a proactive alternative by learning patterns from labeled cybersecurity data and predicting whether unseen domains are likely to be malicious. Prior studies show that lexical, certificate, WHOIS, reputation, and network-based indicators can reveal subtle differences between benign and malicious domains. Hybrid feature approaches are particularly useful because malicious domains may appear legitimate in one feature category while still exposing suspicious behavior in another.

This paper, derived from the master's dissertation, investigates AI-based malicious domain detection using an open-source phishing URL dataset. The study compares **Logistic Regression**, **Random Forest**, and **XGBoost** to determine which approach provides the most reliable classification performance and whether explainable AI can improve trust in model outputs. The work is guided by the need for detection methods that

are accurate, reproducible, and interpretable for cybersecurity analysts.

II. METHODOLOGY

The research follows a quantitative experimental design because malicious domain detection is treated as a supervised binary classification problem. Each domain is labeled as either phishing or legitimate, and models are trained to classify unseen examples under the same experimental conditions.

The dataset used in the study is a publicly available phishing URL dataset named phishing_url.csv. It contains 3,772 instances with a balanced distribution of phishing and legitimate domains. The dataset includes 89 engineered feature columns representing reputation-based characteristics, structural URL attributes, and behavioral indicators. The target variable, status, identifies whether each domain is phishing or legitimate.

Preprocessing involved removing non-predictive columns such as the raw URL string and encoding the target variable into binary numerical form, where phishing domains were represented as 1 and legitimate domains as 0. Exploratory analysis confirmed that the dataset contained no missing values and that all predictive features were numerical. Standardization was applied for **Logistic Regression** because that model is sensitive to feature scale.

The dataset was divided into training and testing subsets using an 80:20 stratified split to preserve class balance in both subsets. **Logistic Regression** was selected as an interpretable linear baseline, **Random Forest** represented a non-linear bagging ensemble, and **XGBoost** represented gradient boosting for structured tabular data. Fixed random seeds were used to support reproducibility.

Evaluation was conducted using accuracy, precision, recall, F1-score, and ROC-AUC. Recall for the phishing class was emphasized because false negatives represent malicious domains that remain undetected. A 5-fold stratified cross-validation strategy was also applied to assess generalization beyond a single train-test split. **SHAP** was used to interpret **XGBoost** predictions at global and local levels.

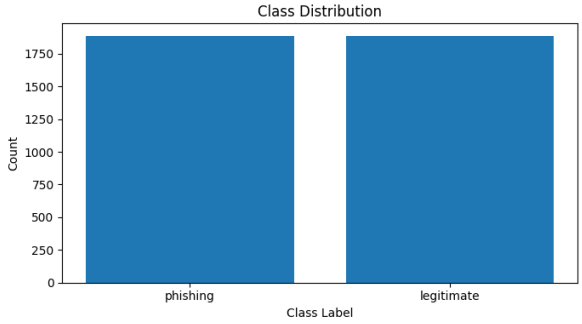

Figure 1: Class distribution of phishing and legitimate domains

Table 1: Comparative performance summary of evaluated models

Model	Accuracy	ROC-AUC	Recall	Interpretation
Logistic Regression	94.44%	0.9845	Noted as balanced	Strong interpretable baseline but limited by linear assumptions.
Random Forest	96.03%	0.9926	97.88%	Improved phishing detection through non-linear ensemble learning.
XGBoost	96.29%	0.9941	98.14%	Best overall model with strong recall and explainability support.

III. RESULTS AND DISCUSSION

All three models achieved strong classification performance, confirming that the engineered domain features contain substantial predictive information. **Logistic Regression** achieved 94.44% accuracy and 0.9845 ROC-AUC, establishing a strong baseline despite its linear decision boundary.

Random Forest improved detection performance by modeling non-linear relationships among features. It achieved 96.03% accuracy and 0.9926 ROC-AUC, with phishing recall of 97.88%. This improvement is important in cybersecurity settings because reducing false negatives lowers the probability that a malicious domain will be missed.

XGBoost produced the best overall results, achieving 96.29% accuracy, 98.14% phishing recall, and 0.9941 ROC-AUC. The reduction in false negatives compared with **Logistic Regression** and **Random Forest** demonstrates the advantage of gradient boosting for structured malicious-domain features.

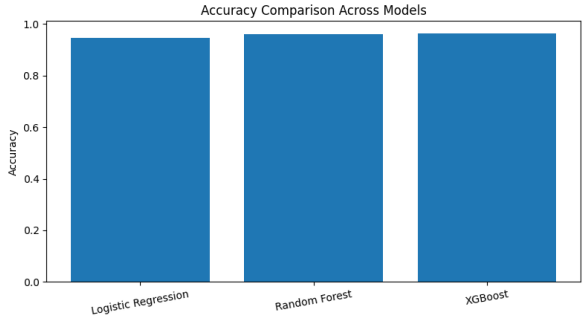

Figure 2: Accuracy comparison across evaluated models

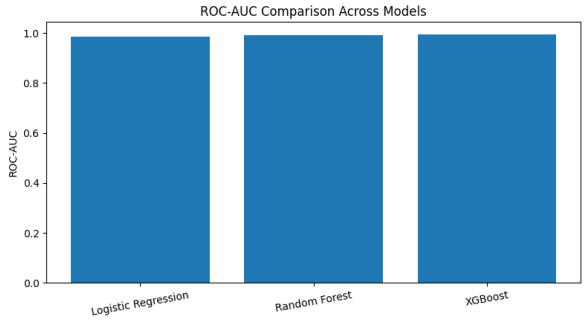

Figure 3: ROC-AUC comparison across evaluated models

Cross-validation confirmed the robustness of the ensemble methods. **Logistic Regression** achieved a mean ROC-AUC of 0.9831 +/- 0.0034, **Random Forest** achieved 0.9919 +/- 0.0021, and **XGBoost** achieved 0.9918 +/- 0.0031. These results show that the strongest models generalized consistently across folds rather than relying on a favorable train-test split.

The comparative results indicate a clear progression from the linear baseline to more advanced ensemble approaches. **Logistic Regression** remains valuable where transparency and low computational cost are required, but **Random Forest** and **XGBoost** provide stronger security performance by capturing non-linear feature interactions.

The **SHAP** analysis improved interpretability by showing which features contributed most strongly to the **XGBoost** predictions. Reputation-related features such as google_index, page_rank, and domain_age had high influence, indicating that indexed, older, and higher-reputation domains were more likely to be legitimate. Structural indicators such as hyperlink counts, repeated 'www' patterns, dots, hyphens, and phishing hints also contributed to distinguishing malicious domains.

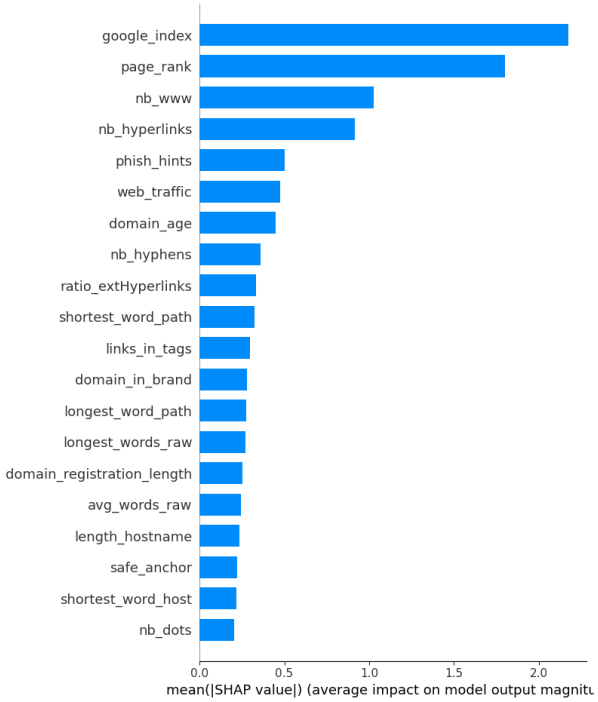

Figure 4: Global feature importance based on mean SHAP values for the XGBoost model

From a practical perspective, the combination of high-performing ensemble models and explainable AI is valuable because security analysts need to understand why an alert has been generated. **SHAP** explanations provide evidence for model decisions and help analysts validate suspicious domains before taking operational action.

IV. CONCLUSION

This study evaluated AI-based techniques for malicious domain detection using an open-source phishing URL dataset. The findings show that machine learning models can classify malicious and legitimate domains with high accuracy when trained on well-engineered reputation, structural, and behavioral features.

XGBoost was the most effective model, achieving the highest accuracy, phishing recall, and ROC-AUC. **Random Forest** also performed strongly and improved phishing detection compared with **Logistic Regression**. The results confirm that ensemble-based approaches are better suited than linear models

for capturing complex feature relationships in malicious-domain classification tasks.

The integration of **SHAP** strengthened the study by making the best-performing model interpretable. Feature attribution showed that model decisions were driven by meaningful cybersecurity indicators rather than arbitrary correlations. This is important for real-world deployment, where analysts need transparent explanations before trusting automated detection outputs.

The study is limited by its reliance on a single pre-engineered dataset and offline classification rather than live detection. Future work should evaluate cross-dataset generalization, incorporate live DNS or certificate transparency data, and explore deep learning approaches on raw URL strings or sequential domain data. Human-in-the-loop evaluation should also be considered to assess how analysts use explanations in operational environments.

REFERENCES

[1] Alshehri, M. D. and Pasquier, M. (2020) A systematic literature review of graph neural networks for cybersecurity. IEEE Access, 8, pp. 170-188.

[2] Bahnsen, A. C., Bohorquez, E., Villegas, S., Vargas, J. and Gonzalez, F. A. (2018) Classifying malicious hosts based on TLS client hello messages. IEEE International Conference on Data Mining Workshops, pp. 172-181.

[3] Basit, A., Zafar, M., Liu, X., Javed, A., Jalil, Z. and Song, H. (2021) A comprehensive survey of AI-driven phishing detection techniques. IEEE Access, 9, pp. 112-131.

[4] Capuano, N., Erra, U. and Gaeta, M. (2020) Explainable machine learning in cybersecurity: A systematic review. IEEE Access, 8, pp. 145-164.

[5] Das, A., Lakhina, A. and Tabriz, P. (2019) Real-time phishing URL detection using lexical analysis. Proceedings of the 12th ACM Conference on Security and Privacy in Wireless and Mobile Networks, pp. 40-51.

[6] Fahl, S., Harbach, M., Muders, T., Baumann, N., Freisleben, B. and Smith, M. (2012) Why Eve and Mallory love Android: An analysis of Android SSL (in)security. Proceedings of the 2012 ACM Conference on Computer and Communications Security, pp. 50-61.

[7] Kim, J., Lee, H. and Kim, H. K. (2019) Detecting phishing websites via graph-based analysis of host-domain relationships. IEEE Access, 7, pp. 103138-103154.

[8] Le, T., Markopoulou, A. and Faloutsos, M. (2018) PhishDef: URL names say it all. IEEE INFOCOM 2018 Conference on Computer Communications, pp. 2388-2396.

[9] Mohammad, R. M., Thabtah, F. and McCluskey, L. (2015) Predicting phishing websites based on self-structuring neural network. Neural Computing and Applications, 25(2), pp. 443-458.

[10] Ring, M., Wunderlich, S., Grüdl, D., Landes, D. and Hotho, A. (2019) A survey of network-based intrusion detection data sets. Computers & Security, 86, pp. 147-167.

[11] Rodríguez, R. J., Christin, N. and Caballero, J. (2021) A survey of domain name system-based malicious domain detection. ACM Computing Surveys, 54(6), pp. 1-36.

[12] Sánchez-Paniagua, M., Sanz, B., Moore, T., García, S. and Bringas, P. (2019) A hybrid approach for detecting phishing sites using URL, SSL, and content-based features. International Journal of Information Security, 18(4), pp. 1-16.

[13] Sharafaldin, I., Lashkari, A. H. and Ghorbani, A. A. (2018) Toward generating a new intrusion detection dataset and intrusion traffic characterization. ICISSP 2018 - 4th International Conference on Information Systems Security and Privacy, pp. 108-116.

[14] Whittaker, C., Ryner, B. and Nazif, M. (2010) Large-scale automatic classification of phishing pages. Proceedings of the 17th Network and Distributed System Security Symposium (NDSS).

[15] Zhang, C., Xie, Y., Bai, H., Yu, B. and Li, Y. (2021) A survey on federated learning for data privacy preservation. IEEE Transactions on Neural Networks and Learning Systems, 33(4), pp. 1544-1562.

[16] Zhang, J., Xiang, Y., Zhou, W. and Wang, Y. (2018) Detection of malicious domains using graph-based associations. Future Generation Computer Systems, 79, pp. 263-271.

[17] Zhang, Z., Chen, Y. and Li, X. (2022) Explainable AI for intelligent cybersecurity systems. IEEE Transactions on Emerging Topics in Computational Intelligence, 6(2), pp. 1-13.

[18] Breiman, L. (2001) 'Random forests', Machine Learning, 45(1), pp. 5-32.

[19] Chen, T. and Guestrin, C. (2016) 'XGBoost: A scalable tree boosting system', Proceedings of the

22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794.

[20] Lundberg, S.M. and Lee, S.-I. (2017) 'A unified approach to interpreting model predictions', Advances in Neural Information Processing Systems (NeurIPS), 30, pp. 4765-4774.