An Explainable Machine Learning Framework for URL-Based Phishing Detection

Ms. Kashish Mehra
B.Tech, Electronics & Comp. Eng.
Jaypee Institute of Info. Tech.
Noida, India
mehra.kashish455@gmail.com

Ms. Katya Chadha
B.Tech, Electronics & Comm. Eng.
Jaypee Institute of Info. Tech.
Noida, India
katyachadha5@gmail.com

Mr. Vinamra Agrawal
B.Tech, Electronics & Comp. Eng.
Jaypee Institute of Info. Tech.
Noida, India
agrawalvinamra12@gmail.com

***Abstract*—Phishing attacks are among the most common and effective cybercrime techniques, frequently exploiting malicious URLs to deceive users into revealing sensitive information. Although machine learning (ML) techniques have demonstrated high accuracy in phishing detection, many existing approaches operate as black-box models which, despite their effectiveness, offer limited insight into the decision-making process. This lack of transparency reduces user trust and limits the practical adoption of ML-based security systems. Recent research has introduced explainable artificial intelligence (XAI) to improve transparency in phishing-related tasks; however, much of this work focuses on user behavior analysis or email-based detection scenarios. In this work, we propose an explainable machine learning framework for URL-based phishing detection that prioritizes transparency alongside detection performance. The proposed framework integrates XAI techniques, namely Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP), to provide both local and global insights into feature importance.**

***Index Terms*—Phishing detection, explainable artificial intelligence, machine learning, URL analysis, cybersecurity**

I. INTRODUCTION

As the world is advancing through digitalization, phishing attacks remain one of the most common and effective forms of cybercrime, deceiving users into accessing fraudulent websites and disclosing sensitive information. According to existing research, many of these attacks require the victim to interact with such websites. Over the past years, numerous studies have demonstrated that machine learning (ML)–based approaches such as Random Forest, XGBoost, and Logistic Regression can achieve high accuracy in identifying phishing attacks by learning patterns from URL structures and features. Despite their effectiveness, many of these models operate as black-box systems. This creates a need for transparency and better understanding of the specific features being used for classification. Our study revolves around the application of Explainable Artificial Intelligence (XAI) in phishing detection models. According to the research "EXPLICATE: Enhancing Phishing Detection through Explainable AI and LLM-Powered Interpretability", Explainable AI is used to interpret machine learning models by explaining their predictions in a human-understandable manner, allowing users to comprehend and trust these decisions. Techniques such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) provide visibility into the features and patterns influencing phishing classification decisions. Despite the growing adoption of XAI, there is a lack of systematic comparative studies that evaluate multiple machine learning models alongside explainability techniques. Limited research explores the applicability of explainable frameworks across phishing detection scenarios involving URLs and emails, which creates a gap in the development of transparent and trustworthy security systems. This study aims to conduct a comparative analysis of commonly used machine learning models for phishing detection, with a particular focus on evaluating the effectiveness of explainable AI (XAI) techniques such as SHAP and LIME in improving model transparency and interpretability. By using both approaches in contrast with each other, the study seeks to identify an optimized and more transparent phishing detection solution.

II. METHODOLOGY

The proposed two-layer framework is inspired by prior research that combines machine learning-based detection with explainable artificial intelligence techniques to enhance transparency and user trust in phishing detection systems [1]. This lay-

ered structure is specifically designed to provide high-accuracy classification while offering human-understandable insights into the model's decision-making process.

II-A. Layer 1: Machine Learning-Based Classification

The first layer is responsible for identifying whether a given URL is legitimate or phishing. To achieve this, supervised machine learning classifiers are employed to distinguish between classes based on extracted features. This framework utilizes two distinct ML approaches:

- **Random Forest:** Employed as the primary classifier due to its robustness and ability to capture complex, non-linear patterns in URL-based features.
- **Logistic Regression:** Utilized as a baseline model owing to its simplicity and inherent interpretability.

Both models are selected for their compatibility with explainable AI techniques, ensuring a seamless transition between detection and interpretation.

II-B. Layer 2: Explainable AI for Model Interpretability

While the machine learning layer provides the final classification, it often operates as a "black box," offering little insight into its reasoning. The second layer addresses this limitation by integrating Explainable Artificial Intelligence (XAI) techniques to provide clarity on model predictions. Two complementary XAI methods are utilized:

- **Local Interpretable Model-Agnostic Explanations (LIME):** Used to generate instance-level explanations by highlighting the specific URL features that most heavily influenced a single prediction.
- **SHapley Additive exPlanations (SHAP):** Provides both local and global explanations by quantifying the contribution of each feature across the entire dataset, ensuring a comprehensive understanding of model behavior.

III. EXPERIMENTS & RESULTS

At the current stage of this research, existing machine-learning approaches from previous studies are being implemented to establish a baseline for detection.These experiments are ongoing and primarily helps us to understand and evaluate feasibility and accuracy. Future work will focus on developing custom machine learning models trained on datasets and integrating them with Explainable artificial techniques, creating a layered approach for more efficient and user-favored model. We used the Kaggle dataset publicly available by [7]. The dataset contains 159244 phishing URLs 820 legitimate URLs. The dataset exhibits a significant class imbalance with 99.49 percent phishing URLs only 0.51 percent legitimate URLs. Therefore, accuracy alone can be misleading a naive model predicting "phishing" would get around 99.5 percent accuracy, necessitating the use of robust evaluation metrics beyond accuracy to ensure reliable model performance.

IV. CONCLUSION

This paper introduces a two-layer framework for catching phishing URLs that balances high-tech detection with actual human clarity. Most current phishing systems have a major flaw: they work like a "black box." Even when they're accurate, it's hard for users or security pros to trust a decision they don't understand. Our goal was to fix that lack of transparency to make these tools more practical for the real world. We designed the framework to handle two jobs separately. The first layer uses machine learning to flag suspicious URLs, while the second layer uses explainable AI tools—specifically LIME and SHAP—to show exactly why a site was flagged. This setup gives security teams the context they need to make confident, informed calls rather than just guessing. Right now, we've focused on the core design and methodology, but the next step is to put it through its paces. Our future work will involve testing the system against public datasets and measuring how it holds up under real-world pressure. Ultimately, this study shows that if we want better cybersecurity, we have to make our tools more transparent and user-friendly.

References

[1] B. V. Pavani, D. Mahitha, and B. U. Maheswari, "Enhancing online safety: Phishing URL detection using machine learning and explainable AI," in *Proc. 15th Int. Conf. on Computing Communication and*

Networking Technologies (ICCCNT), IEEE, 2024, doi: 10.1109/ICCCNT61001.2024.10723976.

[2] B. Lim, R. Huerta, A. Sotelo, A. Quintela, and P. Kumar, "EXPLICATE: Enhancing phishing detection through explainable AI and LLM-powered interpretability," IEEE, 2025.

[3] V. V. G. Varsha and P. A. Thomas, "Explainable AI for phishing detection: Techniques, challenges, and experimental validation," in *Proc. IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, IEEE, 2025, doi: 10.1109/RAICS66191.2025.11332593.

[4] Y. Aldoufani, A. Eleyan, and T. Bejaoui, "An intelligent phishing detection system using deep learning and explainable AI," in *Proc. IEEE Int. Symp. on Networks, Computers and Communications (ISNCC)*, IEEE, 2025, doi: 10.1109/ISNCC66965.2025.11250409.

[5] P. R. G. Hernandes Jr., *et al.*, "Phishing detection using URL-based XAI techniques," in *Proc. IEEE Symp. Series on Computational Intelligence (SSCI)*, IEEE, 2021, doi: 10.1109/SSCI50451.2021.9659981.

[6] A. S. Mandlik, M. S. Ganeshpure, C. N. Kadadas, L. Korra, J. U. Kidav, and M. A. Lavadkar, "AI-driven real-time threat detection for UPI transactions," in *Proc. 2025 Int. Conf. on Applications of Machine Intelligence and Data Analytics (ICAMIDA)*, IEEE, 2025, doi: 10.1109/ICAMIDA64673.2025.11209290.

[7] V. Adi, "Phishing URLs dataset with extracted features," Kaggle, 2020. [Online]. Available: https://www.kaggle.com/datasets/victusadi/phishing-urls-dataset-with-extracted-features/data