Explainable AI for IOT Devices and Robotic Communication Phishing Detection

A Machine Learning Approach Using LIME and SHAP

Zainab Fatima
Department of Software Engineering, NED University of Engineering and Technology, Karachi, Pakistan
zainab.ned@cloud.neduet.edu.pk

M. Hassan Tanveer
Department of Robotics and Mechatronics Engineering, Kennesaw State University, Marietta, Georgia, USA
mtanveer@kennesaw.edu (corresponding author)

Razvan C. Voicu
Department of Robotics and Mechatronics Engineering, Kennesaw State University, Marietta, Georgia, USA
rvoicu@kennesaw.edu (corresponding author)

Sumit Chakravarty
Department of Robotics and Mechatronics Engineering, Kennesaw State University, Marietta, Georgia, USA
schakra2@kennesaw.edu

Maria Ashfaq
Department of Software Engineering, NED University of Engineering and Technology, Karachi, Pakistan
ashfaq440707@cloud.neduet.edu.pk

Muazzam Khan
Department of Software Engineering, NED University of Engineering and Technology, Karachi, Pakistan
khan401085@cloud.neduet.edu.pk

Aqsa Zaib
Department of Software Engineering, NED University of Engineering and Technology, Karachi, Pakistan
zaib406737@cloud.neduet.edu.pk

Hazry Desa
Centre of Excellence for Unmanned Aerial Systems (COEUAS), Universiti Malaysia Perlis, Jalan Kangar-Alor Setar, Kangar, Malaysia
hazry@unimap.edy.my

Received: 19 April 2025 | Revised: 31 May 2025 | Accepted: 6 June 2025

 DOI: https://doi.org/10.48084/etasr.11595

ABSTRACT

Phishing is one of the most dangerous attacks in cybersecurity, which has increased since the introduction of IoT devices, involving attempts to trick users into handing over their passwords and sensitive data. Since most existing detection mechanisms are either nonintuitive or untrusted from the user's perspective, this project attempted to create a phishing detection system that relies on machine learning with explainable AI

(XAI). Considering the results of previous studies that stress the importance of accurate and understandable phishing detection models, a five-phase framework was adopted: data collection, data cleaning, data modeling, XAI, and design of an interactive mechanism. The PhishTank dataset was preprocessed to improve model performance by optimizing the feature set and eliminating noise. Random Forest (RF) was selected, which was the best in terms of accuracy, precision, recall, and F1 score compared to Logistic Regression (LR) and Decision Trees (DT) models. LIME and SHAP were used to offer interpretability and present feature importance at the instance and global levels, respectively. Through an engaging mechanism, users can input URLs, obtain predictions regarding possible phishing attempts, and even read explanations, promoting comprehension and trust. This research shows that including XAI can improve not only the efficacy of the phishing detection systems, but also the level of trust that users have in such systems and be the basis for even more robust and more explainable cybersecurity mechanisms.

Keywords-cybersecurity; phishing; IoT; URL classification; machine learning; random forest; explainable AI; interpretability; explainability; LIME; SHAP; feature importance

I. INTRODUCTION

Phishing is a type of attack that exploits specific human vulnerabilities to trick users into revealing sensitive information, such as credentials and financial information. It poses a threat to organizations and individuals by bypassing security measures. Phishing remains one of the most common attacks used by cybercriminals to infiltrate networks, often bypassing traditional security measures, and its prevalence has increased, especially since the COVID-19 pandemic, where people everywhere have shifted to working remotely [1]. In the past, phishing attacks targeted individuals and organizations using deceptive emails and harmful URLs. However, the growth of the Internet of Things (IoT) has brought new attack means, exploiting the vulnerabilities found in smart TVs, security cameras, and voice assistants, as non-technical users do not fully understand the security risks of using these smart devices. [2]

A. Phishing Detection Techniques

1) Existing Phishing Detection Methods

Phishing attacks exploit vulnerabilities that exist in both human behavior and technology systems. Considering the ever-changing nature of phishing strategies, several methods have been proposed for the detection and mitigation of such threats. Detection methods can be classified into email-based detection, URL-based detection, hybrid systems combining both [1], and IoT-based detection. For example, federated learning techniques have been used to detect phishing content by locally training models on IoT devices, thus preserving privacy and lowering latency [3].

2) Limitations of Traditional Phishing Detection Systems

One of the main demerits is black-box models, which involve complex algorithms with great accuracy but lacking transparency. Users are unable to understand why the model arrives at a certain decision, and thus confidence in system performance is deteriorating. When legitimate phishing alerts do not provide clear reasons for validation, end users and security analysts are likely to disengage, opening the road for security breaches. Traditional phishing detection systems are also plagued with problems in the form of false positives, which mark innocent emails as phishing, and false negatives, which do not identify phishing emails at all. While false negatives mean that phishing attempts evade detection, threatening user data, false positives can cause workflow disruption in the form of blocking crucial communications. When phishing attacks become more complex, recall must be balanced with precision, even for systems based on outdated machine learning models or static rule-based approaches [4].

Modern phishing attacks, such as spear phishing and polymorphic phishing, are designed to avoid detection by traditional methods. In spear phishing, a phisher uses personal information to craft compelling emails that target an individual. Polymorphic phishing employs dynamic changes in its content to evade detection. Conventional systems use static filtering methods and outdated machine learning models, failing to keep up with such rapidly changing strategies. Attackers use contextual information and social engineering strategies to deceive users, which are quite challenging to detect. This exposes the need for phishing detection systems that can learn, adapt to new phishing tactics, minimize false positives, and convey interpretability so that users trust the system. This calls for the integration of Explainable AI (XAI) in phishing detection frameworks to address these shortcomings by providing insight into the rationale for detection decisions, improving user engagement, and enhancing overall system effectiveness [4]. In addition, conventional phishing detection systems tend to suffer from the inability to keep up with changing phishing strategies and handle the large volumes of data produced by IoT ecosystems.

IoT-based detection systems, on the other hand, offer more scalable and dynamic solutions using real-time monitoring and distributed learning mechanisms. In IoT networks, security frameworks can inspect large-scale traffic patterns, dynamically detect phishing attempts, and improve robustness through automated processes. For instance, a federated self-learning anomaly detection IoT system, called DÏoT, was shown to be highly effective in identifying compromised devices by learning and adapting to new threats independently, without collecting data centrally [5]. In contrast to traditional detection approaches, DÏoT strengthens security on IoT devices while minimizing the dependency on static rule-based systems.

B. IoT in Phishing Detection and Cybersecurity

1) Role of IoT in Phishing Detection and Cybersecurity

IoT cybersecurity systems can analyze threats in real-time and automatically take response actions. Network activity monitoring and phishing detection with alerts are possible

using a Cloud-based CPS that implements IoT sensors and data processing units [6].

- Combining IoT systems with ML algorithms can detect malicious patterns in phishing activities and distinguish between legitimate websites and their fraudulent counterparts [2].
- Behavioral analysis: IoT devices track user activities to recognize abnormal actions that may show signs of phishing incidents [7].
- Real-time monitoring and response: A security framework based on the IoT functionality can stop malicious websites immediately while also sending alerts and implementing authentication measures [8].

2) Automating Phishing Detection and Cybersecurity

IoT security operations are heavily dependent on technologies to achieve automation in their security functions. AI systems can examine network events while searching for phishing attacks, along with providing instant cyberattack responses.

- Automated cybersecurity agents: AI agents use their autonomous systems to inspect emails, as well as network traffic and URLs, in search of phishing attacks [7].
- Automated intrusion prevention systems: Security bots use AI to prevent phishing attacks by blocking malicious IP addresses and device isolation [9].

C. Explainable AI (XAI) for Cybersecurity

1) The Need for XAI in Cybersecurity

Currently, AI is nothing short of indispensable in cybersecurity, assisting in malware analysis, intrusion detection, and security operations, among many other important tasks. However, since many traditional models are unable to explain the process, there is an enormous gap in trust and understanding. It is tough for security analysts to respond appropriately to AI-driven alerts or threat detections due to the challenge of interpreting why the AI-based alert or detection took place, hence posing challenges. XAI answers these concerns by improving the transparency and accountability of the AI-based system, enabling better informed decision-making. With explainable insight into how threats are detected, XAI can reduce false positives, improve operational efficiency, and reduce the need to be constantly monitored by humans, thus optimizing cybersecurity operations [10].

2) Key XAI Techniques for Cybersecurity

Several key techniques have been developed to improve the explainability of AI models, with critical contributions in the domain of cybersecurity.

- LIME (Local Interpretable Model-Agnostic Explanations) helps explain the predictions that machine learning models make by approximating around a specific instance. LIME has been applied to interpret alerts resulting from intrusion detection systems [11].
- SHAP (SHapley Additive exPlanations) assigns importance scores to input features with insights on the amount each variable contributed to an AI model's output. SHAP applies well at both the global and local levels and, therefore, is a valuable tool to understand the predictions of the model about detected anomalies and threats [11].
- Layer-wise Relevance Propagation (LRP) provides scores of relevance on the input data components that significantly impact the model predictions. LRP was initially applied in areas such as computer vision; however, its potential in cybersecurity has been explored, specifically to identify patterns within malicious network traffic that influence decisions [10].

3) Purpose of XAI

The principal objective behind XAI is the creation of an explanation-rich framework for machine learning, ensuring that end-users understand, build trust, and effectively manage AI systems. This is most critical in a domain like cybersecurity, because outcomes are high-stakes and decisions by an AI system may have critical consequences. By explaining why an AI system reached a specific decision, XAI makes the system more transparent and trustworthy, and users feel much more confident about the actions taken [1].

4) Benefits of XAI

XAI has several key advantages that fit its suitability to be used in applications regarding cybersecurity. It can provide rational justifications for the decisions made by a model to be identified in terms of errors and thereby allow for increasing reliability and safety in AI systems. Moreover, it supports regulatory compliance toward greater security related to transparency and accountability, with strict implementation directly in fields that need tight data protection laws. XAI allows for a better understanding of AI reasoning related to fairness, bias, and adherence to ethical decisions made by AI systems. Lastly, XAI is critical in the validation of practitioners who work to ensure that AI systems meet the required standards and specifications. These benefits hold particularly true when XAI is introduced into phishing detection systems, as they not only improve threat detection accuracy but also promote the trust of users and cybersecurity professionals [1].

D. Explainable AI (XAI) for Phishing Detection Systems

Although the literature on XAI in phishing detection is sparse, feature selection methods with explainable features have been utilized in IoT attack detection [12]. In [13], SHAP was applied to the NSL-KDD dataset, explaining how different attributes contributed to the predictions made by the models, and the findings for the attack types were similar. In [14], it was observed that SHAP increased the accuracy of the model by 4.9% when applied in an industrial IoT framework. In [15], both SHAP and LIME were employed for validation. In [16], interpretable feature engineering was explored, using Pearson correlation and Recursive Feature Elimination (RFE), but without explicit XAI models. XAI techniques improve phishing detection systems by offering transparency along with actionable insights. The most used techniques are:

- Feature importance analysis: This would highlight the important features that are causing the decision to classify emails as threats, such as unusual sender addresses or some suspicious URLs. This, in turn, allows the user to understand why such a decision is made [12].
- LIME is meant to explain individual phishing predictions by generating explanations that support local feature contributions, or breaking down the decision components to determine the reasons a particular email was detected as a threat [12].
- SHAP calculates the effect of each feature on prediction so that users can see how different elements, such as keywords or URL structures, contribute to a phishing classification.
- Interpretable models: Models such as decision trees offer clear explanations of classification decisions, so that users can easily follow the recommendations of the system [12].
- Visualization tools: Heatmaps and attention maps visually represent risky parts of an email or website, so the user can easily notice which elements are risky [12].

E. Research Gaps and Motivation

Phishing is still a major threat despite all the progress in cybersecurity, mainly due to the continuous evolution of attack vectors and human factors that leave us vulnerable. Traditional AI models, especially deep learning-based systems, are very good at phishing detection but are black boxes, meaning that their decision-making process is complex and opaque, and end users cannot understand why certain emails are flagged as phishing. This lack of transparency is a problem, especially in high-stakes applications where trust and understanding of AI decisions are key [17]. In addition, blockchain-based approaches [18-19] to secure the IoT involve great cost and complexity, without offering explainability. Existing solutions do not provide adequate explanation frameworks for end users, which is a major gap in current research. Many phishing detection systems focus only on improving detection accuracy and ignoring the importance of user trust and behavior toward phishing warnings. Research on how users perceive and interact with these warnings is limited, and as a result, systems are less engaged and less effective. Without an explanation behind phishing alerts, users may not follow recommendations or misunderstand risks [17].

This work was motivated by the potential of XAI to improve phishing detection systems by providing transparent and human-understandable insights into AI-driven decisions. The integration of XAI aims to increase user trust and engagement in security software through human-readable explanations. XAI provides mechanisms to explain how AI models detect phishing, so users can understand the reasons behind specific alerts and, in turn, make better security decisions. By focusing on user experience, this approach tries to bridge the gap between technical performance and user trust, helping to better prevent phishing [12].

II. METHODOLOGY

The proposed framework entails five consecutive phases, in the form of a framework comprising data acquisition, data preprocessing, model selection, explainability integration, and an interactive mechanism for the end-user with respect to phishing prediction. This workflow streamlines the process of developing a sophisticated and interpretable phishing prediction system.

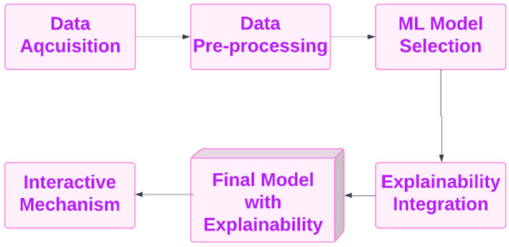

Fig. 1. Steps of the method.

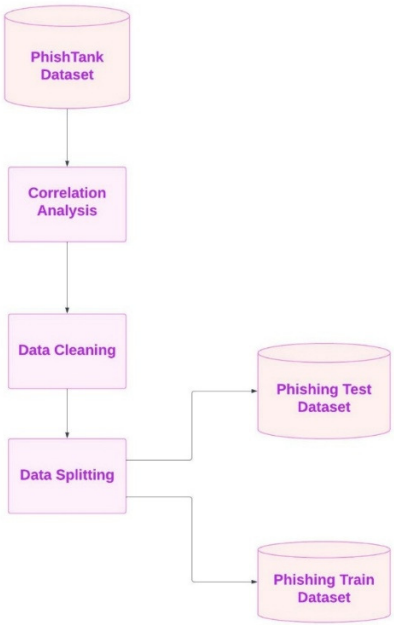

Fig. 2. Data preprocessing steps.

A. Data Acquisition and Preprocessing

The first stage involved collecting data from PhishTank [20], which is a major source of phishing URL datasets. PhishTank is a reputable community-driven database that contains phishing URLs that have been reported. This dataset has many phishing and non-phishing URLs and helped in constructing the phishing detection model. A thorough data preprocessing phase, comprising feature extraction, data cleansing, and data segregation, was performed to maintain the quality and usefulness of the dataset for phishing detection. This stage was key in improving the performance and reliability of the model, as it allowed for enhancing the most important features and ensuring consistency of the training and test data.

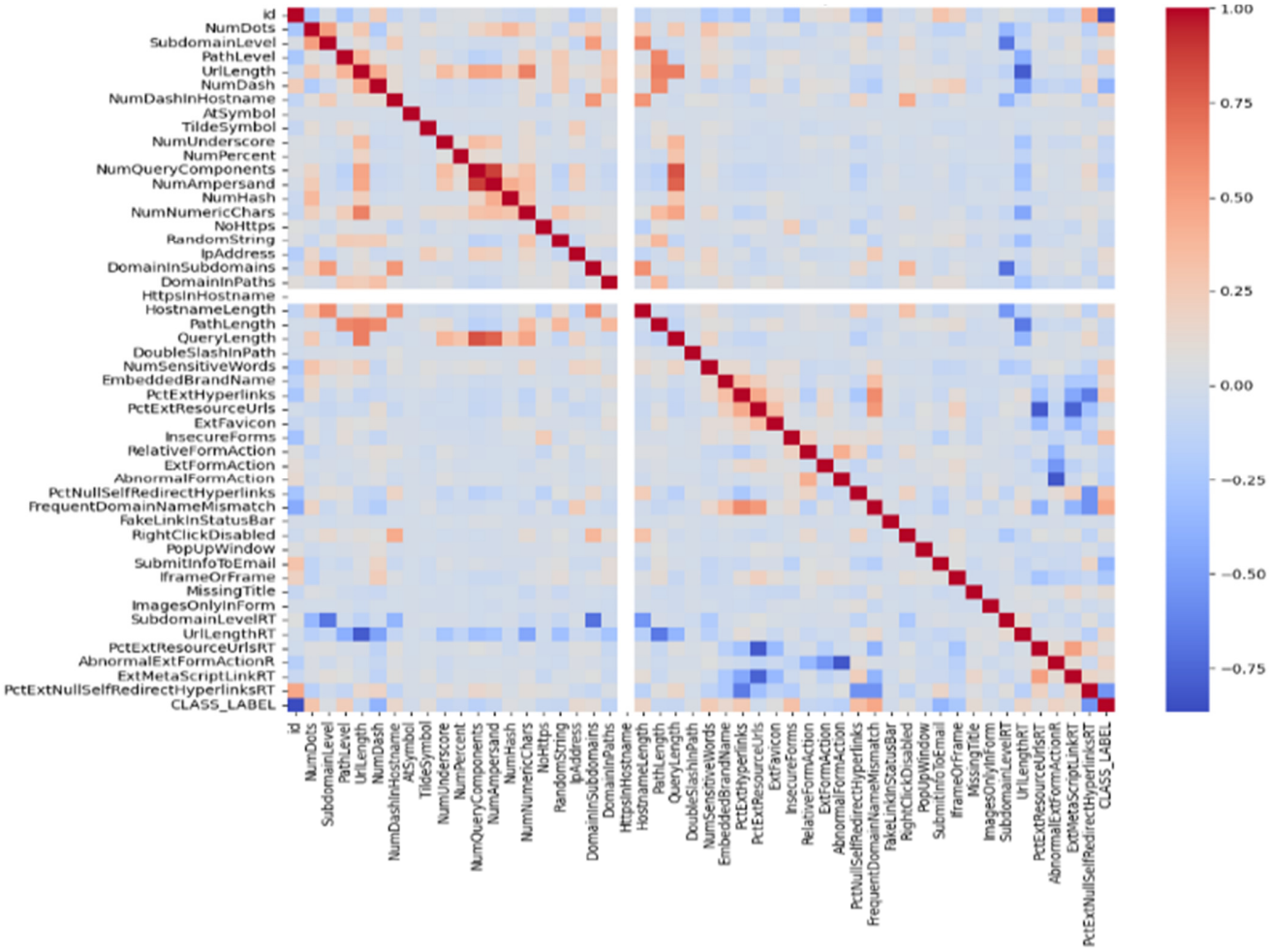

Fig. 3. Correlation heatmap.

A correlation heat map was used to study the dependencies between each feature and the target (non-phishing or phishing). Through the strength of the correlations, features that were weak or even negatively correlated with the target label were discarded. The id feature was removed, and the AtSymbol and HttpsInHostname features were omitted because their correlations with the target label were weak. This helped improve the model's performance and reduce the likelihood of overfitting to only the most predictive features.

After feature selection, data cleaning was implemented to have a consistent dataset. This involved removing missing values, duplicates, and irrelevant entries that would potentially disrupt model training. The remaining dataset was appropriate for machine learning activities, expediting model creation and improving prediction accuracy levels. Lastly, data splitting was performed to assist in training and evaluation. After cleaning, the dataset was split into training and test datasets in an 80:20 ratio so that ample data could be used to create the model. This ratio enabled effective learning of the model, but also allowed testing of the model with data that it had not seen, offering reliable results to test the generalization of the model.

B. Model Selection:

Several machine learning models were evaluated based on accuracy, precision, recall, and F1 score. This aimed to find a model that generates results that are not only accurate but also consistent and interpretable. The models considered were Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF). As a start for quickly examining the dataset's generalities, a linear LR was selected as an interpretable baseline. DT was selected because of its ability to visualize the decision process. DT is inherently interpretable with a structured way of making decisions. RF, which is an ensemble model that contains several DTs, was added because of its great accuracy and the ability to control for overfitting that comes from combining the results of many trees. Since the RF performance was the best, offering highly accurate and stable predictions, it was chosen for deeper testing and further implementation attempts to integrate XAI.

C. Explainability with LIME and SHAP

LIME and SHAP were used to improve the interpretability of the RF, allowing local and global interpretation of the predictions. These methods were selected to provide insights into how the model arrived at certain decisions and specifically

how certain features influence the phishing classifier outputs. LIME understands individual predictions by simulating the model close to the neighborhood of an instance proposition. This technique comes in handy in situations dealing with various classifications, such as a website phishing or non-phishing types. LIME explains its predictions by creating a prediction model and showing the features that were most likely active during the prediction.

SHAP, on the other hand, offers an insight that is different from the model interpretation method and tries to objectively assess the importance of each feature concerning the overall prediction. As suggested by the SHAP summary plot, interpretation is made possible by identifying the most important features responsible for the output of the model, which further enables understanding why the model made the decision. Feature importance analysis is a great tool that helps in interpreting the model and showing the combination of features that form the predictions, thus explaining the general decision-making structure.

D. Interactive Mechanism for User Prediction and Explainability

To improve accessibility and interactivity, an interactive component was added to the phishing detection system that enables the user to input a URL and receive a prediction about it, along with an appropriate assessment of the prediction made. This improves system usability, as users have immediate access to the predictions of the model, while numerous explanatory outputs contribute to the sense of explicability. The model first predicts whether a given URL is phishing or non-phishing. In addition to the prediction, the model also offers an explanation of how the prediction was made. This explanation indicates the features in the data that the model used to reach a conclusion about the prediction. This feature improves the transparency and trust of users. By making predictions interpretable, the mechanism promotes informed decision-making and greater confidence.

III. RESULTS AND COMPARATIVE ANALYSIS

A. Model Selection

Three machine learning models, namely LR, DT, and RF, were tested and evaluated to determine the target model for phishing detection. Table I shows a comparison between the models using accuracy, precision, recall, and F1 score. RF was superior compared to the others in all metrics. This is mainly because RF is a composite of a multitude of DTs, making it more powerful and thus more accurate in predictions. Thus, for further investigation and deployment into the phishing detection system, RF was used as the base model.

TABLE I. PERFORMANCE RESULTS FOR EACH MODEL

	Accuracy	**Precision**	**Recall**	**F1-Score**
LR	0.9390	0.9391	0.9390	0.93899
DT	0.9740	0.9740	0.9740	0.97400
RF	0.9835	0.9835	0.9835	0.98350

B. Outputs of LIME and SHAP

Once RF was selected, LIME and SHAP were implemented and compared in providing transparency and addressing concerns about interpretability. These explainability techniques help to gain knowledge and provide insight about the local (individual) and global (structure) prediction of the model, leading to increased explainability and trust from the end user perspective.

1) LIME Analysis

LIME was able to explain individual predictions through the building of local surrogate models that try to imitate the behaviour of the RF classifier for a particular sample. As shown in Figures 4 and 5, LIME shows the importance of every feature for a given instance, as well as establishes which features were most useful in making a prediction (such as predicting that a URL is a phishing site). The features that were among the most important contributors included FrequentDomainNameMismatch and NumSensitiveWords, which exhibited a strong correlation with phishing predictions and therefore aided the users in understanding the reasons behind a certain classification.

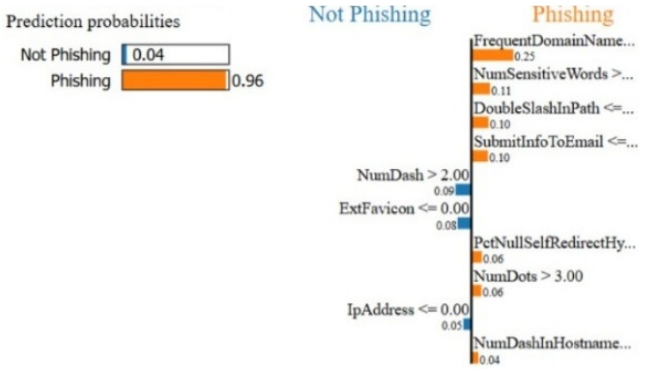

Fig. 4. LIME output showing the probability and feature contributions in a bar chart for a specific sample classified as phishing.

Feature	Value
FrequentDomainNameMismatch	1.00
NumSensitiveWords	1.00
DoubleSlashInPath	0.00
SubmitInfoToEmail	0.00
NumDash	4.00
ExtFavicon	0.00
PctNullSelfRedirectHyperlinks	0.07
NumDots	4.00
IpAddress	0.00
NumDashInHostname	2.00

Fig. 5. LIME output showing the feature contributions in a table for a specific sample classified as phishing.

2) SHAP Analysis

SHAP calculates the contribution of each feature to the prediction of all possible classes and thus provides an overall perspective. SHAP summary plots, as shown in Figures 6 and 7, list the importance of these features in descending order,

with the most important ones being PctNullSelfRedirectHyperlinksRT, PctExtHyperlinks, and FrequentDomain. The SHAP bar plot (Figure 7) reorders the features according to importance, and the violin plot takes a step further showing the effect that single features have on the prediction values. Positive or high SHAP values specify that the feature increases the tendency of phishing, while negative and low feature values indicate a non-phishing tendency.

The distribution of SHAP values clarifies how high values for these features draw stronger predictions for phishing detection, evidenced by the accumulation of red dots on the positive side (Figure 6). SHAP gives reasoning for all features in the dataset, in contrast to LIME, which seeks local explanations of individual instances. This general or global interpretability is complementary to LIME and is useful in explaining the reasoning behind the model predictions. In addition, global integrability can explain how specific features, such as IframeOrFrame, can change the prediction in both directions depending on the context.

PctNullSelfRedirectHyperlinksRT demonstrates a critical role in phishing classification. The SHAP analysis also reveals features with very little power in the model's decisions, such as UrlLength, NumAmpersand, or PathLength, which have very low SHAP values, indicating that they have minimal effect on the ability to tell whether a URL is phishing. This knowledge can assist in future selection processes by allowing the influence of various attributes that do not help in determining the accuracy of the model to be eliminated.

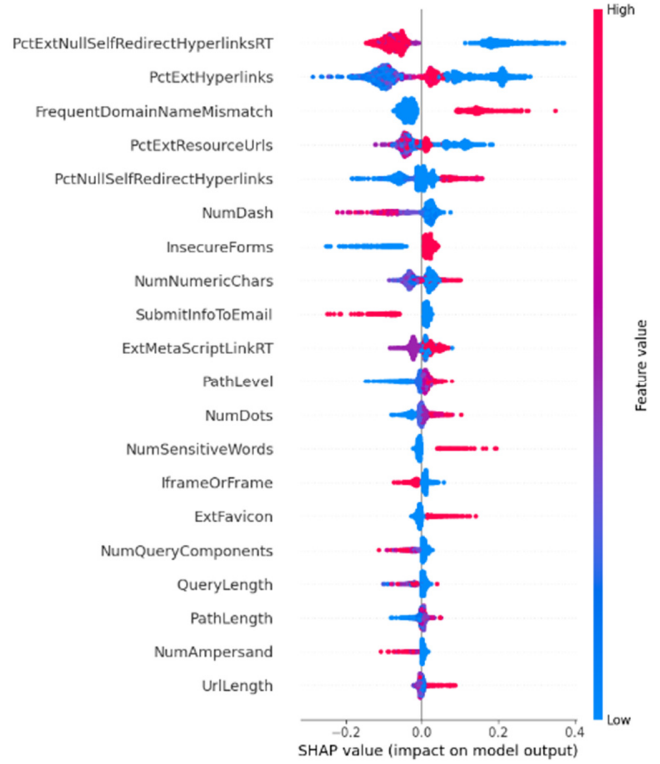

Fig. 6. SHAP summary plot showing global feature importance across all samples in the dataset.

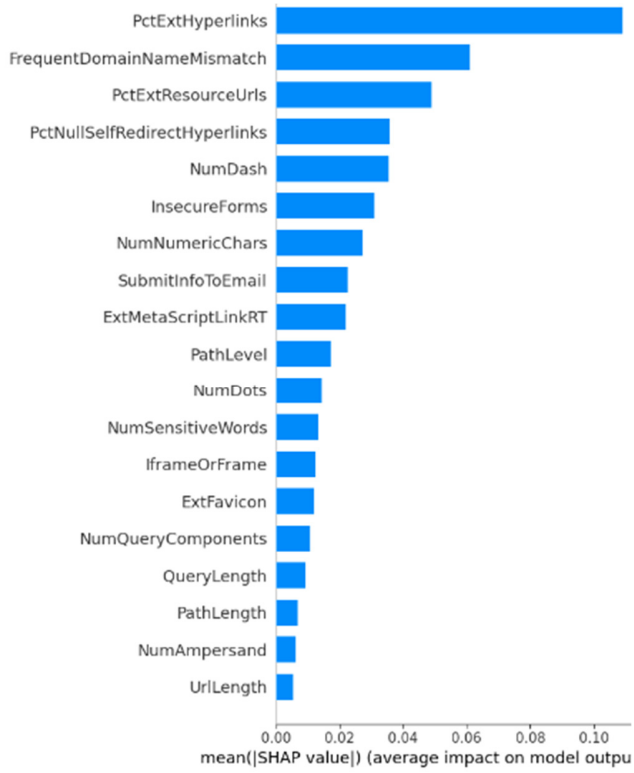

Fig. 7. SHAP bar summary plot showing all features in order of their average impact on model predictions.

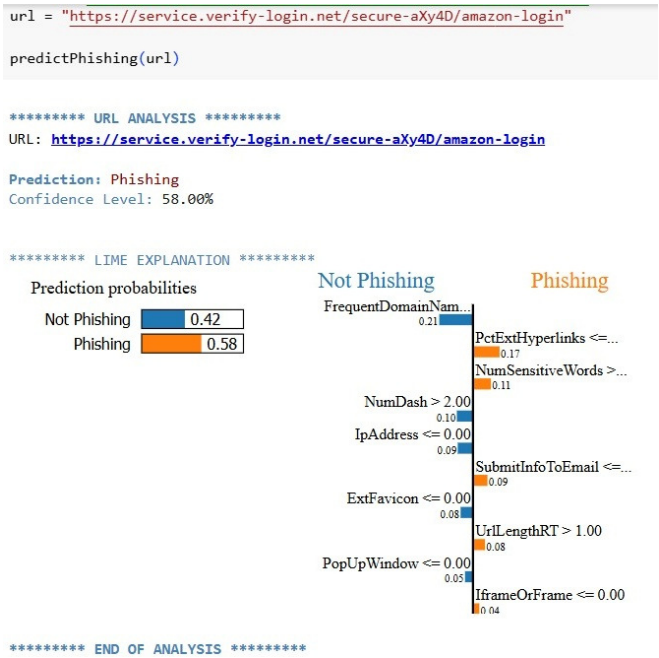

Fig. 8. Overall analysis.

Feature	Value
FrequentDomainNameMismatch	0.00
PctExtHyperlinks	0.00
NumSensitiveWords	2.00
NumDash	3.00
IpAddress	0.00
SubmitInfoToEmail	0.00
ExtFavicon	0.00
UrlLengthRT	58.00
PopUpWindow	0.00
IframeOrFrame	0.00

Fig. 9. Sample output of the interactive mechanism.

C. Interactive Mechanism

The phishing detection system was enhanced with an interactive component that allowed end users to enter a URL and interactively receive a prediction and an explanation. Whenever a user inputs a URL, the system first classifies it as phishing or non-phishing and then displays the confidence level of its prediction. Furthermore, since it is an individual URL, the LIME results help explain why it received such a label and highlight the most important attributes for the assessment. In this way, users can better understand the model and its policies and rules. This enhances trust and allows users to make more informed decisions, thus improving the overall experience and effectiveness of the phishing detection system.

IV. CONCLUSION

This study presented a complete and explainable phishing detection system using machine learning along with an XAI layer. Using a well-explored PhishTank dataset and rigorous data preprocessing, such as feature evaluation and data cleaning, the model was able to maintain high levels of reliability and generalizability. The robustness of the RF model was utilized due to its performance in terms of accuracy, precision, recall, and F1-score, compared to LR and DT. Explainability was one of the primary objectives of this project. Using both LIME and SHAP explained how the model works from both local and general perspectives, providing confidence in the system.

LIME demonstrated the effects of different features for different predictions, enabling users to clearly understand the reason behind the prediction of each classified URL. In contrast, SHAP provided insight on the global feature importance, identifying factors such as PctNullSelfRedirectHyperlinksRT and PctExtHyperlinks that were the most predictive of phishing. IoT devices have given rise to phishing attacks beyond traditional endpoints, which introduces new vulnerabilities. The proposed framework provides explainable phishing detection that can be extended for IoT security, bringing transparency in AI-driven threat mitigation in connected environments. An interactive component was integrated into the system to enable end-users to interact with the model more closely, as they can receive instant predictions and explanations for the URLs submitted. Such a user-oriented structure instills confidence through understanding the reasoning behind every classification. This, in turn, allows users to make reasonable decisions, which solves both the technical and practical demands for an efficient phishing detection system.

In subsequent studies, the system can be improved with other machine learning models and mechanisms that allow live feedback to refine the design. Other enhancement processes can integrate different types of phishing data and broaden the feature space. On top of that, the development of more sophisticated XAI methods might allow a better understanding of the model's performance, and therefore improve user experience as well as system performance. This work provides an encouraging starting point for the development of high-trust and high-accuracy phishing detection tools that are also interpretable. This research can also bridge the gap between AI-driven cybersecurity, IoT security, and user transparency, and highly accurate, trustworthy, and interpretable phishing detection tools.

REFERENCES

[1] M. Alanezi, "Phishing Detection Methods: A Review," *Technium: Romanian Journal of Applied Sciences and Technology*, vol. 3, no. 9, pp. 19–35, Oct. 2021, https://doi.org/10.47577/technium.v3i9.4973.

[2] N. Fatima *et al.*, "AI-Powered Phishing Detection and Mitigation for IoT-Based Smart Home Security," *Journal of Computing & Biomedical Informatics*, vol. 8, no. 1, Oct. 2024.

[3] "Phishing Threats in IoT-Based Systems: Detection and Mitigation Techniques," *Insights2Techinfo*, Nov. 20, 2024. https://insights2techinfo.com/phishing-threats-in-iot-based-systems-detection-and-mitigation-techniques/.

[4] I. Vayansky and S. Kumar, "Phishing – challenges and solutions," *Computer Fraud & Security*, vol. 2018, no. 1, pp. 15–20, Jan. 2018, https://doi.org/10.1016/S1361-3723(18)30007-1.

[5] T. D. Nguyen, S. Marchal, M. Miettinen, H. Fereidooni, N. Asokan, and A. R. Sadeghi, "DÏoT: A Federated Self-learning Anomaly Detection System for IoT," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, Dallas, TX, USA, Jul. 2019, pp. 756–767, https://doi.org/10.1109/icdcs.2019.00080.

[6] S. Krishnaveni, T. M. Chen, M. Sathiyanarayanan, and B. Amutha, "CyberDefender: an integrated intelligent defense framework for digital-twin-based industrial cyber-physical systems," *Cluster Computing*, vol. 27, no. 6, pp. 7273–7306, Sep. 2024, https://doi.org/10.1007/s10586-024-04320-x.

[7] S. R. Alotaibi *et al.*, "Explainable artificial intelligence in web phishing classification on secure IoT with cloud-based cyber-physical systems," *Alexandria Engineering Journal*, vol. 110, pp. 490–505, Jan. 2025, https://doi.org/10.1016/j.aej.2024.09.115.

[8] S. Sivamohan, S. S. Sridhar, and S. Krishnaveni, "TEA-EKHO-IDS: An intrusion detection system for industrial CPS with trustworthy explainable AI and enhanced krill herd optimization," *Peer-to-Peer Networking and Applications*, vol. 16, no. 4, pp. 1993–2021, Aug. 2023, https://doi.org/10.1007/s12083-023-01507-8.

[9] S. Naaz, "Detection of Phishing in Internet of Things Using Machine Learning Approach," *International Journal of Digital Crime and Forensics (IJDCF)*, vol. 13, no. 2, pp. 1–15, 2021, https://doi.org/10.4018/IJDCF.2021030101.

[10] N. Capuano, G. Fenza, V. Loia, and C. Stanzione, "Explainable Artificial Intelligence in CyberSecurity: A Survey," *IEEE Access*, vol. 10, pp. 93575–93600, 2022, https://doi.org/10.1109/ACCESS.2022.3204171.

[11] G. Srivastava *et al.*, "XAI for Cybersecurity: State of the Art, Challenges, Open Issues and Future Directions." arXiv, Jun. 03, 2022, https://doi.org/10.48550/arXiv.2206.03585.

[12] S. S. Shafin, "An explainable feature selection framework for web phishing detection with machine learning," *Data Science and Management*, vol. 8, no. 2, pp. 127–136, Jun. 2025, https://doi.org/10.1016/j.dsm.2024.08.004.

[13] M. Wang, K. Zheng, Y. Yang, and X. Wang, "An Explainable Machine Learning Framework for Intrusion Detection Systems," *IEEE Access*, vol. 8, pp. 73127–73141, 2020, https://doi.org/10.1109/ACCESS.2020.2988359.

[14] S. Bahadoripour, H. Karimipour, A. N. Jahromi, and A. Islam, "An explainable multi-modal model for advanced cyber-attack detection in industrial control systems," *Internet of Things*, vol. 25, Apr. 2024, Art. no. 101092, https://doi.org/10.1016/j.iot.2024.101092.

[15] S. Akintade, S. Kim, and K. Roy, "Explaining Machine Learning-Based Feature Selection of IDS for IoT and CPS Devices," in *Artificial Intelligence Applications and Innovations*, 2023, pp. 69–80, https://doi.org/10.1007/978-3-031-34107-6_6.

[16] B. Wu, S. Yu, L. Peng, and L. Wang, "Interpretable wind speed forecasting with meteorological feature exploring and two-stage decomposition," *Energy*, vol. 294, May 2024, Art. no. 130782, https://doi.org/10.1016/j.energy.2024.130782.

[17] G. J, "The Role of Explainable AI in Understanding Phishing Susceptibility," *Journal of Recent Trends in Computer Science and Engineering*, vol. 12, no. 1, pp. 1–6, Mar. 2024.

[18] B. E. Sabir, M. Youssfi, O. Bouattane, and H. Allali, "Towards a New Model to Secure IoT-based Smart Home Mobile Agents using Blockchain Technology," *Engineering, Technology & Applied Science Research*, vol. 10, no. 2, pp. 5441–5447, Apr. 2020, https://doi.org/10.48084/etasr.3394.

[19] N. A. Alsharif, S. Mishra, and M. Alshehri, "IDS in IoT using Machine Learning and Blockchain," *Engineering, Technology & Applied Science Research*, vol. 13, no. 4, pp. 11197–11203, Aug. 2023, https://doi.org/10.48084/etasr.5992.

[20] S. Tiwari, "Phishing Dataset for Machine Learning." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/shashwatwork/phishing-dataset-for-machine-learning.