An Explainable AI based Cyber Threat Detection Model to Minimize Phishing Attacks and Their Effect

Huraira Irfan[1], Kalsoom Safdar,[1,2*]
[1]Department of Computer Science & IT, University of Jhang.
[2]Faculty of Intelligent Computing, Universiti Malaysia Perlis, 02600 Arau, Perlis, Malaysia.
*Correspondence:kalsoombajwa11@gmail.com

Citation | Irfan. H, Safdar. K, "An Explainable AI based Cyber Threat Detection Model to Minimize Phishing Attacks and Their Effect", IJIST, Vol. 8 Issue 5 pp 2012- 2025, August 2026.
DOI | https://doi.org/10.33411/IJIST/1993
Received | July 29, 2026 **Revised** | Aug 15, 2026 **Accepted** | Aug 19, 2026 **Published** | Aug 20 2026.

Phishing and social engineering rank among the most significant facilitators of cyber-crime, which encompasses data breaches, cyber-attacks, ransomware schemes, and denial of service incidents. A robust threat detection model is essential for reducing the risk of phishing attacks, as these types of attacks are frequently discussed in dark web forums and are thus widely adopted. It has been noted that these attacks can have severe direct or indirect effects on our assets, with the resulting financial damage closely linked to their occurrence. In this research, two methodologies are introduced: the first is an explainable AI (XAI) model specifically designed to assess cyber-risks associated with correlated phishing threats. The second is a hybrid approach referred to as a classifier ensemble, which employs a combination of three top-performing machine learning models through ensemble learning and categorizes them based on several high-impact parameters related to model performance. Ultimately, both proposed models have been compared and evaluated across multiple factors. The proposed model in addition carried out a precision of 97.5%, consider of 98.7%, and F1-score of 98.1%, with an average development of about 2–3% in comparison with baseline models. SHAP evaluation recognized Page Rank, URL Length, Number of Hyperlinks, Domain Age, Google Index, and Phishing Hints because the maximum influential capabilities affecting phishing prediction, while LIME supplied instance-stage reasons to enhance transparency and analyst confidence. The experimental findings display that integrating explainable artificial intelligence with ensemble getting to know appreciably improves phishing detection accuracy whilst simultaneously improving version interpretability, making the proposed framework appropriate for deployment in realistic cybersecurity environments.

Keywords: Phishing and Cyber-attacks, Explainable AI, Threat Detection, Ensemble Learning

Introduction:

Nowadays, rapid expansion of digital communication, e-commerce platforms, and cloud-based services has significantly increased society's dependence on the Internet. However, this technological increase has moreover introduced crucial cybersecurity challenges, among which phishing attacks remain one of the most persistent and damaging threats [1][2]. Phishing is a social engineering approach in which attackers craft deceptive net web sites or URLs to trick clients into revealing sensitive information which consists of login credentials, monetary details, or non-public data. Unlike traditional malware attacks that make the maximum software program software vulnerabilities, phishing especially exploits human consideration, making it difficult to stumble on using conventional protection mechanisms [3]. Modern phishing attacks have superior beyond smooth fake net pages. Attackers now hire region spoofing, URL obfuscation, homograph attacks, and shortened links to skip blacklist-based totally absolutely and rule-based totally absolutely detection structures.

These strategies permit malicious URLs to carefully resemble valid internet sites, often fooling each customer and automatic filters. As a result, conventional detection techniques which encompass signature matching, static rule analysis, and blacklist databases are increasingly insufficient, considering they depend cautiously on recognized assault styles and can't effectively come across zero-day phishing URLs [4][5][6]. To cope with those limitations, system learning (ML) and deep learning (DL) fashions were notably discovered for phishing detection [7]. These techniques have a study lexical, host-primarily based totally absolutely, and behavioral functions of URLs to categorize them as benign or malicious. While ML fashions notably beautify detection accuracy, they are often characteristic as "black boxes," supplying predictions without clean reasoning.

In cybersecurity environments, this lack of transparency reduces character attention and makes it difficult for protection analysts to affirm or justify automated decisions, especially in high-chance packages which encompass monetary systems and important infrastructures. Explainable Artificial Intelligence (XAI) has emerged as a promising strategy to this trouble with the aid of using improving version transparency and interpretability. XAI strategies assist display because a version classifies a URL as phishing with the aid of using highlighting influential functions and selection styles [8]. This functionality now no longer most effective and will increase person self-belief however additionally assists cybersecurity specialists in expertise attacker conduct and enhancing protection strategies.

The subject of research is extremely critical and an important consideration in threat detection model. Phishing links should be addressed and detected in its initial stages to prevent any damage. To that effect, we require a powerful system that will be able to monitor and counter the phishing and cyber-attacks by the attackers.

Phishing is a type of web-based fraud, whereby the individuals who commit such frauds attempt to obtain personal details concerning them or their organisations, through the social engineering process. Attackers use psychological vulnerabilities to take advantage of phishing emails, malicious URLs, or deceitful text messages to make victims reveal sensitive information, including credit card account log-in credentials, banking details, social media log-in credentials, or medical portal log-in credentials [9]. The consequences of phishing are much more than those that affect individual users; they tend to lead to numerous other cyber-incidents, such as ransomware attacks, financial frauds, and major data breaches[10].The COVID-19 pandemic is a traditional instance of the sort of trend, as cybercriminals used the confusion withinside the international to broaden phishing campaigns that have been masqueraded as authorities comfort packages (like U.S. mortgage packages or U.K. economic assistance). Such campaigns incurred greater than USD 2.07 billion in losses and secondary threats which include tax scams and commercial enterprise e

mail comprise [11]. Though phishing assaults are technically carried out, facts systems (IS) students declare that they're ordinarily precipitated and tormented by a financial factor. As an instance, it has additionally been said that proactive phishing prevention in South Korean businesses stored USD 7.6 million 12 months with the aid of using averting productiveness losses [12]. The European Union has said on cyber-coverage that the dangers related to phishing have to be counter measured proactively and in particular with state-of-the-art analytics to counter country of gadget learning (ML) and explainable synthetic intelligence (XAI) [13][14]. This financial angle shows that during this example phishing isn't always a technical hitch, however a systemic risk, which calls for strategic economic planning, which includes cyber-coverage and optimized safety investment.

Explainable Artificial Intelligence (XAI) has emerged as a promising option to this trouble via means of improving version transparency and interpretability [14]. XAI strategies assist with the screen because a version classifies a URL as phishing via means of highlighting influential capabilities and choice patterns. This functionality now no longer best will increase consumer self-belief however additionally assist cybersecurity specialists in information attacker conduct and enhancing defense strategies.

Recent advances in phishing detection have more and more targeted on integrating explainable artificial intelligence (XAI) with machine learning to enhance each predictive overall performance and version transparency [15]. Contemporary research has established that XAI strategies together with SHAP and LIME can display the contribution of character features, allowing cybersecurity analysts to higher recognize version selections and growing accept as true with in automatic detection systems. Recent investigations have additionally explored transformer-primarily based totally architectures, hybrid ensemble models, and feature-choice frameworks for phishing detection, reporting enhancements in interpretability without appreciably compromising category overall performance [16][17]. However, lots of those strategies typically emphasize both prediction accuracy and clarification quality, whilst constrained interest has been given to reaching a powerful stability among each goal in realistic cybersecurity environments.

Despite significant progress in explainable phishing detection, several research challenges remain unresolved. Existing studies frequently evaluate their models using a single dataset, limiting the assessment of model generalizability across different phishing scenarios [18]. Furthermore, many explainable phishing detection frameworks rely exclusively on a single classifier, making them vulnerable to variations in feature distributions and evolving phishing strategies [19]. Although SHAP and LIME have become popular explanation techniques, they are often presented only as post-hoc visualization tools rather than being incorporated into an integrated decision-support framework capable of improving both detection effectiveness and analyst confidence [20]. In addition, relatively few studies compare interpretable models with hybrid ensemble approaches using consistent evaluation criteria across multiple datasets [21][22]. Motivated by these limitations, this study proposes two complementary phishing detection frameworks. The first framework employs an Explainable Artificial Intelligence (XAI)-enhanced XGBoost classifier integrated with SHAP and LIME to provide both global and local interpretability while maintaining high predictive accuracy. The second framework introduces a hybrid ensemble model based on multiple machine learning classifiers combined through weighted voting to improve robustness and classification performance. By evaluating both frameworks using multiple phishing datasets and comprehensive performance metrics, this research aims to provide a practical, transparent, and reliable phishing detection solution suitable for deployment in real-world cybersecurity environments.

The contribution of this paper is listed below:

To enhance an interpretable cybersecurity threat detection model based on XAI

To create an effective model system that uses a collection of machine learning algorit hms to identify phishing attempts.

Moreover, the paper is prepared as follows: Firstly, covers the associated work, discussing in advance techniques for Threat detection and Phishing URLs and system mastering-primarily based totally approaches. Secondly, method is discussed, wherein we proposed approaches; one is the usage of XAI and the other one is hybrid model primarily based totally on aggregate of system mastering algorithms. Thirdly, it showcases the results, comparing the fashions primarily based totally on accuracy, precision, recall, and F1 scores, with a focal point on how the ensemble version carried out throughout one-of-a-kind datasets. Finally, it wraps up the paper through summarizing the primary findings and suggesting destiny studies paths, in particular in integrating XAI into system mastering fashions lessen phishing attacks.

Novelty of the Study:

Despite the huge development carried out in phishing detection the usage of device mastering and explainable synthetic intelligence (XAI), current research usually awareness on both maximizing type of accuracy or enhancing version interpretability, whilst restrained interest has been given to integrating each goal inside a unified and sensible cybersecurity framework [23]. Many current XAI-primarily based totally phishing detection techniques depend on an unmarried type set of rules and offer post-hoc motives without comparing their effectiveness along hybrid ensemble mastering techniques [24], as shown in Table 1. Furthermore, comparative analyses throughout more than one phishing dataset stay noticeably restrained, decreasing self-assurance withinside the generalizability of said results. Overall, the proposed framework contributes to the advancement of explainable cybersecurity by demonstrating that high phishing detection performance and model interpretability can be achieved simultaneously, providing a practical foundation for deploying trustworthy AI-based phishing detection systems.

Table 1. Comparison of Existing Studies

Feature	Existing XAI-based Studies	Proposed Study
Explainable AI	✓	✓
SHAP Analysis	✓	✓
LIME Analysis	Some studies	✓
Hybrid Ensemble Learning	Limited	✓
Weighted Voting Ensemble	Rare	✓
Practical Cybersecurity Framework	Partial	✓

Materials and Methods:

One of the frequent cybersecurity risks, which exploits trust in people to access private data, such as financial data, logins, and personal data, is phishing. These attacks often make use of deceptive URLs in order to deceive the victims from inputting their details in fake websites [25][26]. The Phishing URL Detection system will help to minimize this risk by developing a machine learning model that is capable of distinguishing between the phishing and legitimate URLs [27]. Those who are using the system can enter the URLs in the graphical user interface and get direct confirmation of their existence.

The proposed method applies Explainable Artificial Intelligence (XAI) in a hybrid version to become aware of and mitigate the cyber threats that phishing assaults entail as shown in Figure 1. The framework includes additives which might be mitigation via way of Means of Cyber Insurance (MCI) and phishing chance container (PRC).

The second module (Probability of Phishing Attacks) determines the probability of phishing attacks despite the IT security measures and compliance with rules [28].

Generalised Linear Models (GLMs) are used to examine survey data in the past, such as documents of the Poniman Institute and the CSI-FBI. The model also incorporates some things that can explain such as the duration of time between security deployments and regulatory actions. This module demonstrates the effectiveness of external influences and security investments made by organisations to reduce the phishing attack risk.

The third module (Detection of Phishing Attacks) involves the machine learning classifiers to determine the probability that a phishing URL will occur [29]. The phishing sample provided by Sahingoz et al. has an excellent ratio of authentic and malicious URLs. The characteristics are identified and transformed into a vector, such as statistics on the site, abnormalities in the URL structure, and address bar features. The work of several classifiers is estimated, among which are naïve Bayes classifiers, Support Vector Machines (SVM), and Decision Trees (DT). Once the maximum accuracy has been reached the bagging ensemble of decision trees is selected and further analyzed. The strategies of explainability such as feature importance score are employed to attract attention to the relevant factors which influence the findings of phishing detection.

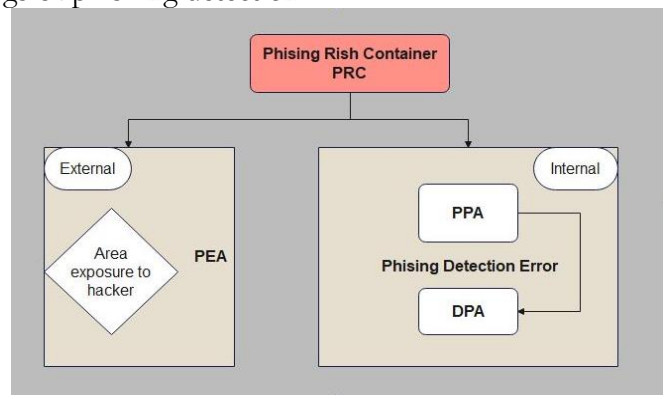


Figure 1. Phishing Risk Proposed Framework

Following the evaluation of the facts preprocessing a part of the study, the studies embraces numerous system studying methodologies pertinent to anomaly-primarily based totally Intrusion Detection Systems (IDS) [30]. These fashions consist of techniques of statistical evaluation, choice trees, classification, and clustering collectively with supervised and unsupervised system studying and probable a aggregate of those studying fashions. The accuracy, precision, recall, and F1 rating metrics are used at the fashions to decide how every of the fashions plays in successfully figuring out the anomalies even as on the equal time minimizing the fake positives as proven in Figure 2 and Figure 3.

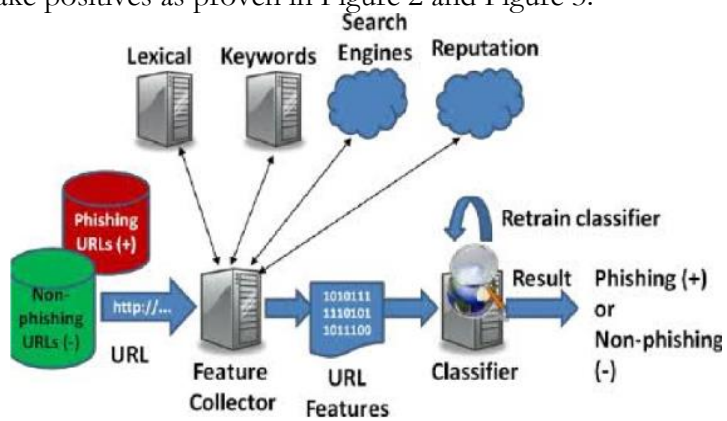


Figure 2. Flow of detecting phishing URLS

Overall Workflow of the Proposed Framework:

The proposed phishing detection framework follows a scientific multi-level workflow, as illustrated in Figure 1 and Figure 2:

Stage 1: Dataset Acquisition:

The first level includes gathering phishing URL datasets from publicly to be had repositories. Two benchmark datasets were hired on this exam to assess the robustness and generalization functionality of the proposed fashions.

Stage 2: Data Preprocessing:

Before version training, the datasets go through numerous preprocessing operations to enhance records exceptional and overall performance. Missing values are recognized and eliminated wherein necessary, reproduction samples are eliminated, and beside the point attributes are excluded.

Prior to model training, the datasets underwent the following preprocessing steps:

Duplicate URL records were identified and removed.

Missing or incomplete feature values were handled appropriately.

Class labels were encoded into binary values (Legitimate = 0, Phishing = 1).

Numerical features were normalized to reduce scale differences.

Irrelevant or redundant attributes were excluded.

The processed datasets were partitioned into training and testing subsets for model development.

Stage 3: Feature Engineering:

Feature engineering transforms the raw URL statistics into significant predictive variables. The extracted functions encompass lexical characteristics (which includes URL period and unique characters), area-associated statistics (which includes area age and web page rank), link statistics, Google indexing statistics, and phishing indicator functions [31]. These engineered attribute's function inputs for the device gaining knowledge of classifiers and the explainability module.

Stage 4: Machine Learning Model Construction:

Seven supervised ML algorithms are used here to categories URLs into phishing and valid categories. The evaluated classifiers encompass Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Gradient Boosting, and XGBoost. In addition, a hybrid ensemble classifier is built with the aid of combining more than one high-acting classifiers the usage of a weighted balloting method to enhance prediction robustness.

Stage 5: Explainable Artificial Intelligence:

To enhance transparency, Explainable Artificial Intelligence strategies are integrated into the proposed framework. SHAP (Shapley Additive Explanations) is used to offer international characteristic significance with the aid of using figuring out the variables that maximum strongly affect phishing detection throughout the complete dataset. LIME (Local Interpretable Model-Agnostic Explanations) enhances SHAP with the aid of using producing neighborhood motives for man or woman URL predictions, permitting cybersecurity analysts to apprehend why a selected URL is assessed as phishing or valid.

Stage 6: Performance Evaluation:

Finally, the skilled fashions are evaluated the usage of Accuracy, Precision, Recall, F1-score, and confusion matrix evaluation. Comparative overall performance evaluation is carried out to pick out the only phishing detection version. The explainability consequences received through SHAP and LIME are analyzed collectively with the class overall performance to evaluate each predictive functionality and interpretability.

Accordingly, it outlines the three-step process utilized in this paper, which includes the metrics used to evaluate the algorithms' performance, the machine learning techniques used to build the models, and the datasets used. In this case, the first step is to describe the features and brief descriptions of the dataset, along with the number of occurrences and website instances that are categorized as either authentic or phishing. The second stage

describes the construction of models, and the third step covers the performance evaluation metrics: accuracy, F1-score, recall, and precision.

Here, we have made use of the two primary Excel-formatted datasets—one for XAI model and the other for the hybrid model, each of which has unique characteristics, numbers, and features.

Dataset 1:

Dataset 1 contains more than 10,000 website instances represented by 83 predictive features. Each instance is categorized into one of two classes:

Legitimate Website

Phishing Website

The extracted features primarily describe webpage content, URL characteristics, hyperlink behavior, domain information, and security-related attributes that are commonly associated with phishing attacks.

Dataset 2:

Dataset 2 contains more than 10,000 URL samples represented by 50 extracted features. Like Dataset 1, each sample belongs to either the legitimate or phishing class. The dataset consists of lexical URL characteristics and structural properties that enable machine learning algorithms to distinguish malicious URLs from benign ones.

Each of the datasets was run through seven classification machine learning algorithms to determine whether a web site is genuine or a phishing one. Patterns in data can be located by a computer, and the information can be used to make decisions, predictions or even actions but this would not need explicit programming. This is made possible by a machine learning algorithm which is a computation model. Such algorithms are the building blocks of the current AI technologies and are applicable in a variety of applications, such as the recommendation systems, fraud detection, phishing sites detection, natural language processing, image and speech recognition, and even a self-driving car.

The classifiers Random Forest (RF), XGBoost (XGB), Gradient Boosting, Decision Trees (DT), Multilayer Perceptron (MLP), and AdaBoost are all integrated in our ensemble model. Because of variations in structure, hyperparameters, and learning methods, each model offers distinct advantages. After these classifiers were trained independently, we aggregated their outputs using Weighted Voting, which greatly improved the identification of harmful behaviour in IoT scenarios.

XGBoost Hyperparameters:

The XGBoost classifier was selected because the number one explainable learning model due to its capacity to capture complicated nonlinear relationships at the same time as keeping excessive predictive overall performance. The model becomes configured the usage of the hyperparameters shown in Table 2 below.

Table 2. Configuration Settings of Hyperparameters

Hyperparameter	Value
Objective	binary: logistic
Booster	gbtree
Number of Trees (n_estimators)	200
Learning Rate	0.05
Maximum Depth	6
Regularization Lambda	1
Evaluation Metric	Log loss
Random State	42

Ensemble Learning Configuration: The proposed hybrid model combines more than one independently trained classifier with the usage of a weighted vote casting strategy. Each

classifier generates prediction probability, and the very last class is decided via way of aggregating those possibilities in step with predefined classifier weights. The weights had been assigned primarily based totally at the man or woman validation overall performance of every classifier to enhance robustness and decrease the impact of weaker learners. For global model interpretation, SHAP (Shapley Additive explanations) was employed using the Tree Explainer specifically designed for tree-based models. Local explanations were generated using the LIME Tabular Explainer.

Results:

Here we have utilized two datasets from the credible sources that have been used in the training and testing of both of approaches. Every step is organized and explained in detail using some comments too. In addition to the results, here we have presented the analysis of complete study, related works have been compared with each other based on different factors and then conclusion is drawn based on that analysis. Here we are presenting these as follows:

Here are the Factors:

- Distribution of data
- Dataset size and volume
- Underfitting and overfitting
- Training and testing size
- Parameters
- Model/ technique selection
- Hyper tuning of the parameters

Table 3. Accuracy Results of Models Before XAI

Model	Best Score (%)
Logistic Regression	93.21
Random Forest	94.19
Gradient Boosting	**95.89**
SVM	95.65
KNN	95.10

Results of Models:

AI Role:

The XAI-enhanced XGBoost model completed about 2–3% development in type performance as compared to baseline approaches as shown in Table 3.

Confusion Matrix and Error Analysis:

The confusion matrix demonstrates an excessive stage of type accuracy with minimum fake positives and fake negatives as provided in Table 4. The confusion matrix demonstrates that the proposed framework achieves an excessive stage of category accuracy with enormously few Fake Positive (FP) and Fake Negative (FN) predictions. Although the general overall performance is satisfactory, an in-depth evaluation of those misclassifications gives precious insights into the sensible conduct and boundaries of the proposed phishing detection model.

Table 4. Confusion Matrix and Error Analysis

Error Type	Cause	Practical Impact
False Positive	Long URLs, multiple subdomains, URL shortening, many hyperlinks	Legitimate websites may be incorrectly blocked
False Negative	Advanced phishing pages using HTTPS, domain spoofing, carefully designed URLs	Malicious websites may evade detection, increasing cybersecurity risk

SHAP Interpretability Results:

SHAP analysis identified critical features influencing phishing detection, including:

Page Rank

URL Length

Number of Hyperlinks

Domain Age

Google Index

Phishing Hints

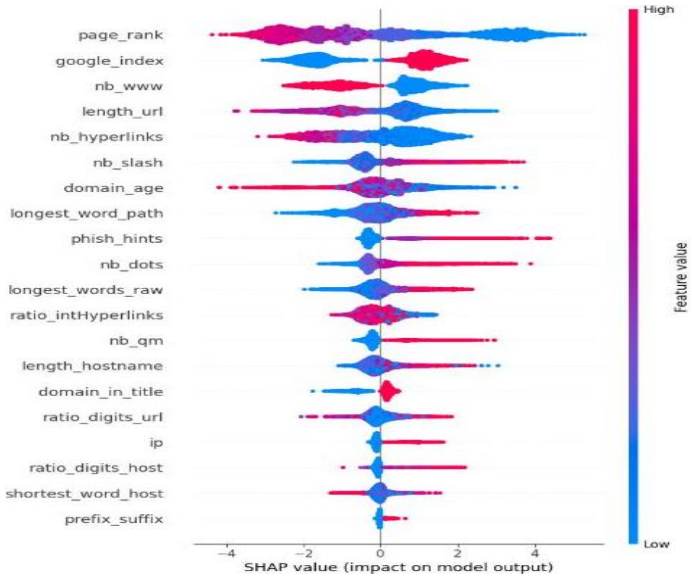


Figure 3. Chart of Global interpretability using SHAP

Above we've proven the Global Interpretability chart the usage of SHAP Algorithm: its miles used to pick out key capabilities usually withinside the dataset. The strategies are used right here for the usage of XAI integration with the XGBoost classifier. The LIME output lists the pinnacle capabilities (and their weights) that contributed maximum to the version`s choice for that precise URL.

Positive weights imply capabilities pushing the prediction in the direction of the phishing class, and bad weights in the direction of legitimate.

This neighborhood perception allows customers to recognize why the version flagged a specific URL, complementing the worldwide SHAP analysis.

In actuality, we would display those elements or incorporate them into the user interface so that end users could see which URL capabilities influenced the decision rather than "ust "phi"hing" ve"sus "l"git." We elevate an efficient ensemble classifier known for its exceptional performance on phishing detection by incorporating XGBoost as:

SHAP Algorithm:

Next, we use SHAP to determine global interpretability (i.e., overall essential features).

LIME Algorithm:

Lime is used for local explanations in the model only. Thus, these improvements improve the phishing URL classifier's effectiveness and transparency, which leads to an increase in accuracy of two to three percent in our model as shown in Figure 4 and Table 5, a comparison of all used ML models.

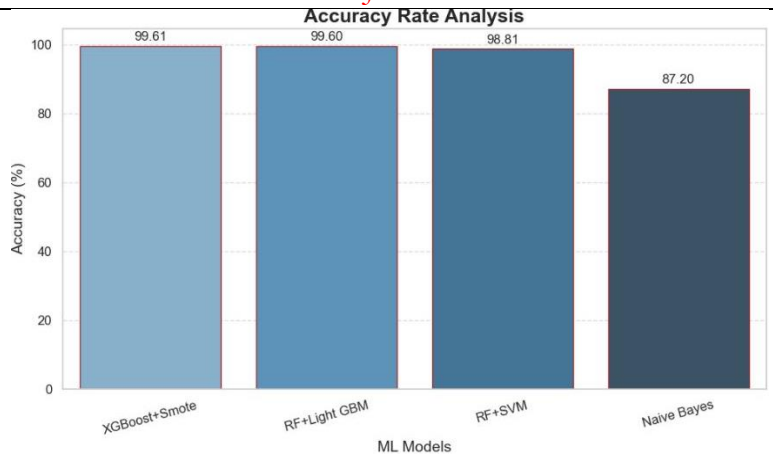


Figure 4. Bar Plot for Accuracy Rates after XAI Integration

Table 5. Accuracy Results after XAI Integration

Model	Accuracy (%)
XGBoost + SMOTE	99.61
Random Forest + LightGBM	99.60
Random Forest + SVM	98.81
Naïve Bayes	87.20

Quantitative Assessment of Explainability:

To complement the qualitative visualization supplied with the aid of using SHAP and LIME, the explainability characteristics of the proposed framework have been evaluated the use of extensively normal interpretability standards mentioned with inside the XAI literature. Rather than counting on a single rationalization technique, the proposed framework integrates SHAP for worldwide version interpretation and LIME for nearby instance stage reasons. This complementary approach improves the transparency of phishing detection choices with the aid of supplying each standard characteristic significance and precise reasons for character predictions. The explainability characteristics of the proposed framework are summarized in Table 6.

Table 6. Characteristics of the Proposed Framework

Explainability Criterion	SHAP	LIME	Proposed Framework
Global Interpretability	✓	✗	✓
Local Interpretability	✗	✓	✓
Feature Importance Ranking	✓	Partial	✓
Instance-Level Explanation	Partial	✓	✓
Model-Agnostic	✗	✓	Partial
Tree-Based Optimization	✓	✗	✓
Analyst Decision Support	Moderate	Moderate	High

Statistical Significance Analysis:

To decide whether the determined overall performance enhancements are statistically huge, statistical speculation trying outturned in to carried out the use of the type consequences acquired at some stage in the cross-validation experiments. The null speculation assumes that there's no huge overall performance distinction among the proposed framework and the baseline classifiers.

An importance degree of $\alpha = 0.05$ turned into follow edit some point of the experiments.

Therefore, $p < 0.05$ shows a statistically huge improvement.

$p \geq 0.05$ shows that the determined distinction can be because of random variation.

Comparative Analysis with Recent Deep Learning-Based Phishing Detection Models:

Recent advances in phishing detection have followed deep studying architectures, especially Transformer-primarily based totally fashions and massive language model (LLM)-assisted frameworks, to seize complicated semantic and contextual traits of phishing URLs and emails [32]. These techniques have proven to be robust predictive overall performance with the aid of studying wealthy characteristic representations without massive guide characteristic engineering. However, many deep studying fashions require giant computational resources, massive-scale labelled datasets, and specialized hardware for each schooling and inference. In contrast, the proposed framework combines explainable device studying with ensemble studying to acquire excessive detection overall performance even as keeping computational performance and obvious decision-making.

Discussion:

The findings exhibit that integrating Explainable Artificial Intelligence with system mastering not ably complements cybersecurity danger detection. While traditional ML fashions offer excessive predictive performance, they regularly fail to explain decision-making processes. XAI addresses this quandary via the means of imparting obvious and interpretable outputs.

The proposed explainable phishing detection framework is designed to assist deployment in sensible cybersecurity environments in which each detection accuracy and version transparency are essential. The framework may be incorporated into a couple of protection layers, along with stable electronic mail gateways, internet browsers, business enterprise proxy servers, cloud-primarily based totally protection platforms, Security Information and Event Management (SIEM) systems, and Security Operations Centers (SOCs). Incoming URLs may be evaluated earlier than consumer access, permitting companies to perceive doubtlessly malicious web sites in actual time whilst concurrently offering interpretable motives that help protection analysts for the duration of incident investigation.

The hybrid ensemble technique similarly progressed detection accuracy via means of leveraging the strengths of a couple of classifiers. The aggregate of predictive accuracy and interpretability makes the proposed framework especially appropriate for real-international cybersecurity environments in which accept as true with and transparency are critical.

The cross-validation results demonstrate that the proposed models consistently maintain high predictive performance across different data partitions, indicating strong generalization capability. Furthermore, statistical hypothesis testing confirms that the performance improvements over the baseline classifiers are statistically significant at the 95% confidence level, suggesting that the observed gains are not attributable to random variation but rather to the proposed explainable and ensemble learning strategies.

Although Transformer-primarily based totally phishing detection fashions have validated super predictive capability, they commonly contain better computational fees and require drastically large datasets for powerful training. In contrast, the proposed framework achieves aggressive detection overall performance via means of integrating XGBoost with SHAP, LIME, and a hybrid ensemble method even as retaining version transparency and decreasing computational requirements. Furthermore, in contrast to many deep getting to know strategies that are characteristic as black-container fashions, the proposed framework gives each worldwide and neighborhood explanations, permitting cybersecurity analysts to apprehend the motivate the back of person phishing predictions. These traits make the proposed framework appropriate for deployment in realistic cybersecurity environments in which interpretability, efficiency, and accept as true with are essential. Although the

proposed technique executed promising results, demanding situations stay concerning scalability, computational overhead, and version to rising phishing techniques. Future paintings must inspect deep mastering architectures and automatic function engineering techniques to enhance detection performance.

Conclusions:

In conclusion, we estimate the chances of the occurrence of phishing attacks which may negatively impact a business by cybercriminals. We applied a binary classifier that was like the anti-phishing technology (or filters) applied by companies to prevent phishing URLs. Subsequently, we recommended IT risks mitigation measures that were aligned with the needs of the company, which included third-party cyber insurance or investments on cybersecurity, or both. Lastly, our scenario analyses demonstrated that a third-party insurer would be delighted to cover the phishing risk in case the probability of an attack was reduced to a reasonable level by applying the appropriate security technologies.

The second method, which has been introduced in this paper, is a hybrid model, which utilizes more than one machine learning model, and relying on the combination of these models, we have utilized various datasets toclassification model. The machine learning algorithms can detect the common traits of most phishing attacks. In this paper, various machine learning approaches have been employed to analyze two phishing datasets. It is aimed at determining what type of websites they are and whether they are ordinary. We shall then employ the normalisation process on the dataset to ensure all the features take the same scale. The findings indicated that XGB algorithm is more efficient in predictions as compared to others. It was also accurate with 0.978, précised with 0.975, recalled with 0.987 and F1-scored 0.981 in the first dataset. The results of all the algorithms in the second dataset through all the four evaluation measures took the same results implying that they were equally good at prediction.

In our future work, we intend to

1) apply the deep learning algorithms to the dataset that we have just discussed and another machine learning algorithm. The recommendations include
2) feature selection methods,
3) generalize the ensemble model, and
4) investigate the impact of other methods of data pre-processing.

Acknowledgments: We extend our heartfelt gratitude to Dr. Kalsoom who played an important role in improving the research article. Also, the manuscript has not been published or submitted to other journals previously.

Conflicts of Interest: The authors declared no conflict of interest.

References:

[1] K. Kalasampath, K. N. Spoorthi, S. Sajeev, S. S. Kuppa, K. Ajay, and A. Maruthamuthu, "A Literature Review on Applications of Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 13, pp. 41111–41140, 2025, doi: 10.1109/ACCESS.2025.3546681.

[2] Z. Cheng, Y. Wu, Y. Li, L. Cai, and B. Ihnaini, "A Comprehensive Review of Explainable Artificial Intelligence (XAI) in Computer Vision," *Sensors 2025, Vol. 25, Page 4166*, vol. 25, no. 13, p. 4166, Jul. 2025, doi: 10.3390/S25134166.

[3] O. Kaya, A. F. M. S. Shah, M. A. Karabulut, S. N. Karahan, M. S. Osmanca, and N. Acir, "Explainable Artificial Intelligence (XAI): Concepts, Applications, Challenges, and Future Perspectives," *IEEE Access*, vol. 14, pp. 27394–27417, 2026, doi: 10.1109/ACCESS.2026.3663161.

[4] F. Herrera, "Reflections and attentiveness on eXplainable Artificial Intelligence (XAI). The journey ahead from criticisms to human–AI collaboration," *Inf. Fusion*, vol. 121, p. 103133, Sep. 2025, doi: 10.1016/J.INFFUS.2025.103133.

[5] A. Søgaard, "Externalist XAI?," *Theor.*, vol. 91, no. 2, p. e12581, Apr. 2025, doi: 10.1111/THEO.12581

[6] T. Kehkashan *et al.*, "Explainable phishing website detection for secure and sustainable cyber infrastructure," *Sci. Reports 2025 151*, vol. 15, no. 1, pp. 41751-, Nov. 2025, doi: 10.1038/s41598-025-27984-w.

[7] R. Ahmed and S. EP, "An intelligent phishing email detection system using ensemble methods and explainable AI," *Knowledge-Based Syst.*, vol. 337, p. 115388, Mar. 2026, doi: 10.1016/J.KNOSYS.2026.115388.

[8] L. Sawe, J. Gikandi, J. Kamau, and D. Njuguna, "Sentence Level Analysis Model for Phishing Detection Using KNN," *J. Cyber Secur.*, vol. 6, no. 1, pp. 25–39, Jan. 2024, doi: 10.32604/JCS.2023.045859.

[9] R. J. van Geest, G. Cascavilla, J. Hulstijn, and N. Zannone, "The applicability of a hybrid framework for automated phishing detection," *Comput. Secur.*, vol. 139, p. 103736, Apr. 2024, doi: 10.1016/J.COSE.2024.103736.

[10] N. Biswas, K. M. M. Uddin, S. T. Rikta, and S. K. Dey, "A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach," *Healthc. Anal.*, vol. 2, p. 100116, Nov. 2022, doi: 10.1016/J.HEALTH.2022.100116.

[11] I. Markoulidakis and G. Markoulidakis, "Probabilistic Confusion Matrix: A Novel Method for Machine Learning Algorithm Generalized Performance Analysis," *Technol. 2024, Vol. 12, Page 113*, vol. 12, no. 7, p. 113, Jul. 2024, doi: 10.3390/TECHNOLOGIES12070113.

[12] B. Wu, S. Yu, L. Peng, and L. Wang, "Interpretable wind speed forecasting with meteorological feature exploring and two-stage decomposition," *Energy*, vol. 294, p. 130782, May 2024, doi: 10.1016/J.ENERGY.2024.130782.

[13] S. S. Shafin, "An explainable feature selection framework for web phishing detection with machine learning," *Data Sci. Manag.*, vol. 8, no. 2, pp. 127–136, Jun. 2025, doi: 10.1016/J.DSM.2024.08.004.

[14] L. Mansour, N. Hilal, N. Abbas, and S. Kadry, "Generalized Explainable AI Framework for Phishing Detection on Heterogeneous Textual Data," *Comput. Decis. Mak. An Int. J.*, vol. 3, pp. 859–875, Jul. 2026, doi: 10.59543/COMDEM.V3I.18329.

[15] M. A. Uddin, M. Mahiuddin, and I. H. Sarker, "An explainable transformer-based model for phishing email detection: A large language model approach," *Comput. Networks*, vol. 277, p. 112061, Mar. 2026, doi: 10.1016/J.COMNET.2026.112061.

[16] Z. Fan, W. Li, K. B. Laskey, and K. C. Chang, "Investigation of Phishing Susceptibility with Explainable Artificial Intelligence," *Futur. Internet 2024, Vol. 16, Page 31*, vol. 16, no. 1, p. 31, Jan. 2024, doi: 10.3390/FI16010031.

[17] E. A. El Abdellaoui Alaoui, A. Filali, A. Sallah, M. Hajhouj, A. Hessane, and M. Merras, "Towards Transparent Cybersecurity: The Role of Explainable AI in Mitigating Spam Threats," *Procedia Comput. Sci.*, vol. 236, pp. 394–401, Jan. 2024, doi: 10.1016/J.PROCS.2024.05.046.

[18] S. J. Almheiri, A. A. Shah, S. Abbas, M. Ahmad, and M. A. Khan, "Smart sustainable cyber security: modelling an interpretable and transparent threat detection with explainable artificial intelligence," *Discov. Sustain. 2025 61*, vol. 6, no. 1, pp. 442-, May 2025, doi: 10.1007/S43621-025-01280-Z.

[19] K. S. Alketbi and A. Mehmood, "A Comprehensive Survey of Explainable Artificial Intelligence Techniques for Malicious Insider Threat Detection," *IEEE Access*, vol. 13, pp. 121772–121798, 2025, doi: 10.1109/ACCESS.2025.3587114.

[20] H. Alamro *et al.*, "Enhanced intrusion detection in cybersecurity through dimensionality reduction and explainable artificial intelligence," *Sci. Reports 2025 151*, vol. 15, no. 1, pp. 33848-, Sep. 2025, doi: 10.1038/s41598-025-06761-9.

[21] Noor Hazlina Abdul Mutalib, Aznul Qalid Md Sabri, Ainuddin Wahid Abdul Wahab, Erma Rahayu Mohd Faizal Abdullah, "Explainable deep learning approach for advanced persistent threats (APTs) detection in cybersecurity: a review," *Artif. Intell. Rev.*, vol. 57, no. 297, 2024, [Online]. Available: https://link.springer.com/article/10.1007/s10462-024-10890-4

[22] J. Sivakumar, N. Rafid Salman, F. Rafid Salman, H. R. Salimova, and E. Ghimire, "Ai-Driven Cyber Threat Detection: Enhancing Security Through Intelligent Engineering Systems," *J. Inf. Syst. Eng. Manag.*, vol. 10, no. 19s, pp. 790–798, Mar. 2025, doi: 10.52783/JISEM.V10I19S.3116.

[23] A. V. Turukmane, S. Patil, K. Varshinee, L. Yadav, and G. Khekare, "Leveraging Explainable AI for threat detection in Industrial IIoT-based intrusion prevention systems," *Explain. Artif. Intell. Ind. Internet Things Technol. Appl.*, pp. 211–221, Jan. 2025, doi: 10.1201/9781003537380-11.

[24] I. Elgarhy, M. M. Badr, M. Mahmoud, J. Ni, M. Alsabaan, and T. Alshawi, "Investigation of the Robustness of XAI-Based Federated Learning Against Adversarial Attacks for Smart Grid False Data Detection," *IEEE Internet Things J.*, vol. 12, no. 15, pp. 32179–32192, 2025, doi: 10.1109/JIOT.2025.3576225.

[25] S. Kumar, "Explainable AI for Intrusion Detection Systems: Enhancing Trust, Transparency, and Real-Time Threat Response," *Int. J. AI, BigData, Comput. Manag.*, vol. 7, no. 2, pp. 115–123, Apr. 2026, doi: 10.63282/3050-9416.IJAIBDCMS-V7I2P119.

[26] J. Shin, "Feasibility of local interpretable model-agnostic explanations (LIME) algorithm as an effective and interpretable feature selection method: comparative fNIRS study," *Biomed. Eng. Lett. 2023 134*, vol. 13, no. 4, pp. 689–703, Jun. 2023, doi: 10.1007/S13534-023-00291-X.

[27] "Phishing Statistics 2024: Trends and Prevention Strategies for 2025." Accessed: Aug. 10, 2026. [Online]. Available:https://www.sangfor.com/blog/cybersecurity/phishing-statistics-and-how-to-prevent-phishing

[28] B. V. Pavani, D. Mahitha, and B. U. Maheswari, "Enhancing Online Safety: Phishing URL Detection Using Machine Learning and Explainable AI," *2024 15th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2024*, 2024, doi: 10.1109/ICCCNT61001.2024.10723976.

[29] D. M. Belete and M. D. Huchaiah, "Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results," *Int. J. Comput. Appl.*, vol. 44, no. 9, pp. 875–886, Sep. 2022, doi: 10.1080/1206212X.2021.1974663.

[30] E. J. Williams and A. N. Joinson, "Developing a measure of information seeking about phishing," *J. Cybersecurity*, vol. 6, no. 1, pp. 1–16, Jan. 2020, doi: 10.1093/CYBSEC/TYAA001.

[31] S. T. Rikta, K. M. M. Uddin, N. Biswas, R. Mostafiz, F. Sharmin, and S. K. Dey, "XML-GBM lung: An explainable machine learning-based application for the diagnosis of lung cancer," *J. Pathol. Inform.*, vol. 14, p. 100307, Jan. 2023, doi: 10.1016/J.JPI.2023.100307.

[32] T. Sutter, A. S. Bozkir, B. Gehring, and P. Berlich, "Avoiding the Hook: Influential Factors of Phishing Awareness Training on Click-Rates and a Data-Driven Approach to Predict Email Difficulty Perception," *IEEE Access*, vol. 10, pp. 100540–100565, 2022, doi: 10.1109/ACCESS.2022.3207272.

Copyright © by authors and 50Sea. This work is licensed under Creative Commons Attribution 4.0 International License.