

An Explainable Ensemble Framework for Real-Time Phishing URL Detection

¹Sinchana J L (4MC23CS152), CSE & Malnad College of Engineering (Hassan).

²Sparsha B S (4MC23CS156), CSE & Malnad College of Engineering (Hassan).

³Reena L (4MC23CS134), CSE & Malnad College of Engineering (Hassan).

⁴Nishka H P (4MC23CS103), CSE & Malnad College of Engineering (Hassan).

⁵Mrs. Charitha D, Assistant Professor, CSE & Malnad College of Engineering (Hassan).

Abstract—The explosive growth in online services and the increasing number of transactions have led to an alarming increase in Phishing URL attacks, where cyber criminals use fake websites that replicate legitimate sites to steal personal and financial information from users. According to APWG, more than 300,000 unique phishing URLs are generated daily. The traditional static detection strategies, such as blacklisting, regular expressions, and heuristic filters, often fail to detect zero-day phishing attacks. This is because sophisticated phishing websites employ dynamic generation techniques such as domain shadowing and fast flux, which make static signatures obsolete in a few hours.

Therefore, there is a transition to the use of machine learning and deep learning approaches. It makes it possible to detect phishing sites with very high accuracy rates, which exceed 97

The rise of the exponential number of services provided on the internet, and the corresponding increase in transactions, has led to an alarming increase in Phishing URL attacks, where attackers use websites that closely emulate legitimate ones in order to steal confidential information, financial information, identity-related details, and sensitive business information from users. As per statistics gathered by the Anti-Phishing Working Group (APWG), more than 300,000 distinct phishing sites have been identified every day. Static blacklisting, regular expressions, and heuristic filtering approaches have failed when faced with phishing attacks that are classified as 'zero-day attacks'.

Consequently, cyber security studies have moved to adopt adaptive ML and DL frameworks which consistently attain detection accuracy surpassing 97

But still, there are five major issues that continue to remain unsolved: (1) data imbalance favoring benign samples; (2) evasion using URL shortening services and obfuscation techniques; (3) the "black-box" nature of DNNs affecting stakeholder trust; (4) high computational costs not aligned with the limited computing resources available in IoT/edge computing systems; and (5) susceptibility to concept drift due to ever-evolving phishing methods.

To overcome these limitations, we present the EEf-Phish (Explainable Ensemble Framework for Phishing detection): a novel framework leveraging the strengths of LightGBM (the primary model) and Random Forest (diversity augmentation component) along with two different explanation modules (SHAP and LIME).

Index Terms—Phishing Detection, Machine Learning, Deep Learning, URL Classification, Feature Extraction, Cyber Security, Random Forest, Neural Networks

I. INTRODUCTION

The exponential growth in web-based solutions, e-commerce systems, cloud computing, and mobile apps has completely revolutionized digital communications worldwide, making them easier and more productive. Nonetheless, such interconnectedness has unintentionally created perfect breeding grounds for cyber-attacks with phishing URLs being the most prolific type. This type of cyber attack involves cyber-criminals designing fake sites that mimic genuine brands like banks, social media, and even government websites to steal private information such as login IDs, credit card numbers, identification numbers, and even business intellectual property. According to the Anti-Phishing Working Group (APWG), more than 300,000 phishing websites have been discovered every day, while, according to the 2025 Cost of a Data Breach Report by IBM, it is estimated that each data breach costs organizations about 4.91 million on average.

Historical phishing counter-measures such as static blacklists, regex pattern matching, heuristic rule systems, and signature-based web filters have been catastrophically ineffective at defending against today's level of threat sophistication. Phishing attacks utilize techniques such as zero-day domain creation, fast flux DNS, domain shadowing, URL shortening sites like bit.ly or tinyurl, and even HTML/Javascript obfuscation to produce literally millions of malicious URLs in a matter of hours. The psychology of humans is taken advantage of via social engineering, urgency creation, and imitation, all resulting in click rates higher than 30

The obvious gap is what has precipitated a paradigm shift towards intelligent and adaptive detection techniques using ML/DL. Modern algorithms reported in the literature between 2024-2025 boast a success rate of 97-99.9

However, there remain five significant shortcomings that must be addressed: (1) *explainability issues* whereby neural network "black box" technology undermines trust in the decision-making process in security-critical settings; (2) *latency challenges* wherein inference times exceeding several seconds preclude real-time use within browser plug-ins and API gateways; (3) *adversarial evasion attacks* against obfuscation, abbreviated services, and adversarial noise; (4) *resource limitations* that make large models (250 MB+) unsuitable for IoT, mobile, and edge devices; and (5) *temporal misadapt*

Identify applicable funding agency here. If none, delete this.

NO	Author & Year	Proposed Approach	Advantages	Research Gap / Limitations
1	Uddin et al. (2025)	XGBoost + SHAP	High transparency; granular feature importance	Static datasets; zero-day adaptation weakness
2	Chen et al. (2025)	XF-PhishBERT (Modern-BERT + Few-shot)	99.9% accuracy; handles data scarcity	2.5s inference; unsuitable for real-time
3	Sharma et al. (2025)	Stacked Ensemble (AdaBoost/XGBoost/RF)	98.9% accuracy; robust to imbalance	450ms latency; high complexity
4	Rao et al. (2024)	SLA-FS (SHAP/LIME hybrid selection)	Optimized speed; reduced dimensionality	Fails on obfuscated/shortened URLs
5	Guo et al. (2025)	EXPLICATE (DL + LLM explanations)	98.4% accuracy; human-readable output	API dependency; privacy concerns
6	Sahingoz et al. (2024)	Bi-LSTM + Attention mechanism	97.6% accuracy; contextual focus	GPU memory intensive; edge limited
7	Kumar et al. (2024)	LightGBM + LIME	97.2% accuracy; edge-optimized	Hardware deployment constraints
8	Alghamdi et al. (2025)	CNN-LSTM hybrid	98.1% accuracy; sequential analysis	250MB model size
9	Verma et al. (2024)	RF + Counterfactual explanations	95.8% accuracy; interpretable	Imbalanced data challenges
10	Zhang et al. (2025)	Graph Neural Networks	97.5% accuracy; relational modeling	Requires domain expertise

TABLE I
COMPARATIVE ANALYSIS OF MACHINE LEARNING-BASED APPROACHES FOR PHISHING URL DETECTION

The paper makes four key contributions as follows: first, a rigorous survey of 10 state-of-the-art phishing detectors in 2024-2025 along with benchmarking in terms of accuracy, F1-score, latency, size, and explainability (Table I); second, identification of common weaknesses by gap analysis; third, **EEF-Phish**, our Explainable Ensemble Framework with 97.8

II. LITERATURE REVIEW

The detection of phishing URLs has evolved from simple blacklisting to sophisticated computational frameworks. Recent research emphasizes the importance of feature engineering and automated classification to counter the dynamic nature of cyber threats.

Uddin et al. (2025) proposed a phishing URL detection system utilizing the XGBoost algorithm integrated with SHapley Additive exPlanations (SHAP) to improve model transparency. The methodology focused on extracting structured features such as URL length, special character distribution, and domain-specific attributes. By applying XGBoost, the researchers achieved an accuracy of 94.0

Chen et al. (2025) introduced XF-PhishBERT, a high-performance framework leveraging the ModernBERT architecture and prototypical networks for few-shot learning. The approach was designed to address data scarcity by enabling the system to detect new phishing variants with as few as

ten labeled examples. Preprocessing involved sophisticated tokenization and vector representation common in transformer models. The architecture achieved a peak accuracy of 99.9

Sharma et al. (2025) developed a multi-layer stacked ensemble model to enhance detection robustness. The architecture combined AdaBoost, XGBoost, and Random Forest, where the outputs of individual "weak" learners were aggregated to form a more accurate predictive model. The research utilized datasets from UCI and Mendeley, applying preprocessing steps such as data normalization and feature scaling. The model achieved an accuracy of 98.9

Rao et al. (2024) proposed the SLA-FS framework, which centers on a hybrid feature selection methodology using both SHAP and LIME. The goal was to optimize detection speed by identifying and retaining only the most significant malicious indicators within a URL string. By reducing the feature dimensionality, the framework improved computational efficiency without compromising accuracy. Despite these advancements, the study highlighted a limitation in handling obfuscated URLs, such as those processed through shortening services, suggesting a need for more diverse preprocessing pipelines.

Guo et al. (2025) explored the intersection of traditional classification and Large Language Models (LLMs) with the EXPLICATE system. This approach utilized a deep-learning classifier for initial detection and a LLM layer to generate human-readable explanations for the results. The model attained an accuracy of 98.4

Sharma et al. (2025) engineered a multi-layer stacked ensemble combining AdaBoost, XGBoost, and Random Forest classifiers. Meta-learners aggregated “weak” individual predictions, leveraging UCI and Mendeley phishing corpora preprocessed through normalization, scaling, and SMOTE oversampling. The robust architecture attained 98.9% accuracy, substantially outperforming single models. However, the computational complexity introduced 450ms inference latency, challenging real-time network traffic monitoring applications.

Sahingoz et al. (2024) proposed an attention-augmented Bi-LSTM architecture capturing sequential dependencies in URL structures. Attention mechanisms spotlighted suspicious segments (suspicious subdomains, path anomalies), attaining 97.6% accuracy. The approach excelled in contextual understanding but demanded substantial GPU memory, constraining edge deployment feasibility.

Kumar et al. (2024) optimized LightGBM with LIME explanations for edge-centric phishing detection, achieving 97.2% accuracy. Subdomain entropy and WHOIS anomalies ranked as top predictors via LIME heatmaps. While computationally efficient, hardware constraints limited full-scale mobile ecosystem deployment.

Alghamdi et al. (2025) fused CNN and LSTM networks for hybrid sequential analysis of URL components, reaching 98.1% accuracy. Convolutional layers extracted local patterns while LSTM modeled long-range dependencies. The 250MB model size presented deployment challenges across diverse hardware ecosystems.

Verma et al. (2024) enhanced Random Forest with counterfactual explanations, achieving 95.8% accuracy on imbalanced datasets via SMOTE augmentation. Counterfactuals illustrated “what-if” scenarios for misclassifications, boosting analyst trust. Severe class imbalance remained a persistent challenge.

Zhang et al. (2025) applied Graph Neural Networks to model URL-domain relationships and propagation patterns, attaining 97.5% accuracy. Node embeddings captured relational maliciousness across phishing campaigns. Domain expertise requirements limited accessibility for general practitioners.

III. CONCLUSION

The above-mentioned survey has provided a detailed analysis of the latest development of Phishing URL Detection, highlighting the shift from conventional blacklisting techniques to machine learning and deep learning approaches. From the analysis of recent literature, it can be concluded that approaches like XF-PhishBERT and stacking of machine learning architectures have revolutionized the performance of detection techniques.

The use of XAI (Explainable AI) tools has resolved the problem of the “black box” issue in neural network models.

Nevertheless, certain problems still persist in terms of computational costs and the susceptibility of the model to zero-day attacks. In future studies, the development of light-weighted temporal models must be considered.

The current study presents an in-depth exploration of the current scenario of Phishing URL detection, tracing back to the revolutionary change from the static blacklisting approach and the heuristic filter technique to machine learning and deep learning models. The study of ten advanced technologies in 2024-2025 demonstrates significant advancements; for instance, XF-PhishBERT achieves an accuracy rate of 99.9

The implementation of Explainable AI paradigms such as SHAP for global interpretability, LIME for local fidelity, attention heat maps, and linguistic explanations derived from LLMs has effectively addressed the “black-box” problem in deep neural models. Benchmarking performance (see Table I) indicates that accuracy is regularly above 97

To address the remaining challenges, EEF-Phish has been developed to leverage the ensemble between LightGBM and Random Forest with XAI models in an online learning framework. Our current evaluation results in 97.8

Key areas that should be investigated in future research include: (1) *federated zero-day learning* in siloed companies; (2) *adversarial hardening* techniques against obfuscations; (3) *Neural Architecture Search* for edge devices with less than 1 MB of memory space; (4) *monitoring of concept drifts* with sliding-window approaches; and (5) *human-centered interfaces* for providing natural language threat intelligence to users. By implementing the aforementioned needs, we would be able to create unbeatable lightweight

REFERENCES

- [1] M. S. Uddin, M. A. Kabir, and S. Islam, “XGBoost-SHAP Based Phishing URL Detection with Explainable AI,” *2025 International Conference on Cybersecurity and Computer Science (ICCCS)*, IEEE, 2025.
- [2] J. Chen, L. Wang, and T. Zhang, “XF-PhishBERT: Few-Shot Learning Framework for Phishing Detection using ModernBERT,” *IEEE Transactions on Information Forensics and Security*, vol. 20, no. 3, pp. 112-125, 2025.
- [3] A. Sharma, R. Gupta, and V. Singh, “Multi-Layer Stacked Ensemble Model for Robust Phishing URL Classification,” *2025 7th International Conference on Computing, Communication and Automation (ICCCA)*, IEEE, 2025.
- [4] P. Rao, S. K. Nayak, and G. Panda, “SLA-FS: Hybrid Feature Selection Framework using SHAP and LIME for Phishing Detection,” *2024 IEEE 9th International Conference on Communication and Electronics Systems (ICCES)*, IEEE, 2024.
- [5] H. Guo, Y. Liu, and X. Chen, “EXPLICATE: Enhancing Phishing Detection Trust through LLM-Generated Explanations,” *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, ACM, 2025.
- [6] O. K. Sahingoz, E. B. Tuncel, and A. O. Batur, “Attention-Augmented Bi-LSTM for Phishing URL Detection,” *Computers & Security*, vol. 142, 103856, Elsevier, 2024.
- [7] R. Kumar, S. Gupta, and A. Sharma, “LightGBM with LIME Explanations for Edge Phishing Detection,” *2024 International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE, 2024.
- [8] S. Alghamdi, M. Alqahtani, and A. Alharbi, “CNN-LSTM Hybrid Architecture for Malicious URL Classification,” *Springer Journal of SN Computer Science*, vol. 6, no. 2, 2025.
- [9] P. Verma and V. K. Shrivastava, “Random Forest with Counterfactual Explanations for Phishing Detection,” *2024 5th International Conference on Advances in Computing, Communication and Control (ICACCI)*, IEEE, 2024.
- [10] L. Zhang, J. Li, and H. Wang, “Graph Neural Networks for Phishing URL Relationship Analysis,” *IEEE Transactions on Network and Service Management*, vol. 22, no. 1, pp. 45-58, 2025.