International Journal for Electronic Crime Investigation

ISSN: 2522-3429 (Print)
ISSN: 2616-6003 (Online)

DOI: https://doi.org/10.54692/ijeci.2026.1001/267

Research Article

Vol. 10 issue 1 Jan-Jul 2025

A Hybrid Ensemble–Instance Learning Framework for Malicious URL Detection Using XGBoost and Adaptive KNN

Sehrish Munir[1], Shazia Yousaf[2], Sunbal Faraz Hayat [3], Khalil Aslam[4], Amara Javed[5], Imran Ahmad[6]*

[1]European Institute of Management and Technology, Switzerland

[2]Fazaia College of Education for Women, Lahore, Pakistan

[3]Iqra National University, Islamabad, Pakistan

[4]Sharif College of Engineering and Technology, Lahore, Pakistan

[5] University of Gujrat, Gujrat, Pakistan

[6]Center for International Collaboration for Computing, Lahore, Pakistan

Corresponding Author: icrg.pk@gmail.com

Received: April 06,2026, **Accepted:** April 14,2026; **Published:** May 04,2026

ABSTRACT

The use of malicious URLs is a constantly evolving and persistent cybersecurity threat, as it is the primary instruments of phishing, ransomware delivery, and account theft. Past studies have been constrained in ensemble and instance-based paradigms in URL threat classification, with ensemble approaches offering high-accuracy global classification and instance-based approaches offering high-quality local boundary detection. This paper presents a meta-learning framework, called Hybrid Ensemble-Instance Learning (HEIL) which combines XGBoost and adaptive K-Nearest Neighbors (KNN) in a synergetic way to enhance the performance of both on detecting malicious URLs. The HEIL framework was tested with an augmented sample of 2,134 samples with 27 engineered lexical, host, DNS, network, and temporal attributes. The model obtained an accuracy of 98.94, a prediction latency of 3.2 ms per URL (model inference time alone, not including feature extraction overhead) and a F1-score of 98.87, which are statistically significant higher than standalone XGBoost. The validation tactics such as SHAP interpretability, ablation studies and adversarial robustness testing are thorough and illustrate and affirm the complementary character of global and local paradigms of learning. The HEIL framework shows competitive performance over the baseline approaches to hybrid ensemble-instance models especially in cybersecurity context.

Keywords: Hybrid Learning Framework, Ensemble Methods, Instance-Based Classification, Malicious URL Detection, XGBoost Algorithm, K-Nearest Neighbors, Meta-Learning Architecture, Cybersecurity Machine Learning, SHAP Feature Analysis, Adversarial Robustness

1. INTRODUCTION

Internet services have rapidly evolved and, as a result, this has significantly increased the attack surface and threat activity. Malicious URLs represent one of the most prevalent and dangerous cyber threats since they are exploited as phishing campaign gateways, ransomware payload carriers, credential gathering, and malware injection vectors [1]. Attackers have developed techniques of evasion such as typosquatting, homoglyph replacements, fast-flux DNS rotation and algorithmic URL generation to enable them to avoid traditional detection techniques [2], [3]. The signature-based defense and blacklist repositories have the advantage of offering some degree of protection, yet with a due to lag times ranging from hours to days, newly emerging malicious URLs often remain undetected by these systems [4].

Machine learning (ML) has emerged as a dominant approach in identifying malicious URLs as it has been shown to be capable of offering statistical pattern recognition and extrapolation on known threat signatures [5]. Gradient boosting variants of ensemble methods, in particular, have achieved state-of-the-art performance with accuracies exceeding 98% on benchmark datasets [6], [7]. The most recent pinnacle of ensemble-based URL classification is its regularized form, known as XGBoost, and parallel optimization [8]. Such global learners might not work in local decision boundaries which are heterogeneous- a common property of real-world URL datasets [9]. K-Nearest Neighbors (KNN) is a type of instance-based learning that provides complementary advantages with adaptive local reasoning. The global models can smooth the fine-grained decision boundaries that can be identified by KNN as query instances are classified in a similarity-based neighborhood. Previous research has shown that KNN can be used to classify URLs with an accuracy of 97.47 percent [10]. However, standalone KNN is computationally costly (O(n) per query) and noise sensitive, so they cannot be used in high throughput security application [11].

Limited work has explored the integration of ensemble and instance-based paradigms in detecting malicious URLs. The gap of this paper is filled with the significant contribution: Design and implementation of the HEIL meta-learning architecture integrating XGBoost and adaptive KNN. The remainder of this paper follows the IMRAD structure: Section 2 presents the literature review, Section 3 details methodology, Section 4 reports results, Section 5 discusses implications, and Section 6 concludes.

2. LITERATURE REVIEW

2.1 History of Malicious URL Detection Systems.

Three generations of approaches have been developed for malicious URL detection. First-generation systems were based on vendor-curated blacklists held by security vendors and threat intelligence agencies. These systems had scaling limitations and had to be quickly corrected manually, despite having an initial accuracy over 96% [4]. The second generation heuristic rule based detectors coded expert knowledge about suspicious domain names: abnormal domain length, excessive number of subdomains, IP based hosts, suspicious word sequences. Lexical heuristics and WHOIS analysis had reported improvements in detection performance in specific experimental settings than simple blacklisting [5], yet the rigidity of

manually-written rules was unable to withstand polymorphic attack patterns. Third-generation ML methods changed the landscape of detection where the discriminative pattern of labeled URL features is trained. Early studies concluded that lexical URL properties, and host based metadata are good classification cues [12]. Classical algorithms like Support Vector Machines, Naive Bayes and random forest proved to be good on early benchmark data. The latter work was followed by the addition of feature space based on network level statistics, DNS behavioural patterns and content based signals and onwards increasingly comprehensive coverage became achieved in detection [6].

2.2 Ensemble Learning Strategies.

The ensemble methods rapidly replaced the URL classification since it can combine a large number of weak learners to strong learners. Gradient Boosting Machines (GBM) has brought in sequential error correction. XGBoost builds on standard gradient boosting by adding residual learning, column subsampling and parallel split-finding to achieve better scalability and generalization [8]. Recent research indicates the performance of XGBoost in a wide variety of URL classification tasks and the reported accuracy varies between 97.8 and 98.5 percent, based on feature sets and dataset attributes and data characteristics [7], [13]. Heterogeneous learners based on meta-classifier with stacking architecture and incremental performance improvement were observed to prove effective with imbalanced data sets [14]. RNNs, LSTMs, and CNNs have been explored with regards to modeling sequential URL patterns, and are very precise with large scale data. Nevertheless, these architectures are associated with high computational power and training samples that are difficult to apply in practice in real-time [15]. It has been demonstrated by Kumar and Singh [7] that XGBoost with adaptive feature selection is capable of producing high-accuracy with latency-constrained systems whereas deep learning models are unable to do so.

2.3 Instance-Based Learning Paradigms.

The instance-based approach to the familiar training examples and are automatically adaptive local decisions, without training step [10]. Similar methods known as variants of distance metric learning, such as Large Margin Nearest Neighbor (LMNN), are also capable of removing KNN sensitivity to irrelevant features by learning Mahalanobis distance transformations that focus on discriminative dimensions [11]. Irrespective of such developments, the systematic synthesis of ensemble and instance-based paradigms remain under-researched in the context of URL security. Chen et al. [14] have shown that the heterogeneous learners are more efficient in detecting the zero-day attacks, and the superiority of the locally adaptive classifiers can be attained.

Nevertheless, despite the progress, there has been no previous effort to systematically combine global ensemble learning with locally adaptive instance-based systems in a single meta-learning framework to classify URLs.

3. METHODOLOGY

3.1 Preprocessing and Data Set.

The dataset used in this study is an augmented version of Malicious Webpages Dataset, which initially included 1,781 binary-labeled URL samples (63.44% malicious, 36.56% benign). Synthetic Minority Oversampling Technique (SMOTE) was used with stratified sampling to

ensure that model to improve model generalizability, resulting in a final corpus of 2,134 samples of balanced classes. The augmentation maintains original statistical properties with the added benefit of additional boundary region samples that are essential in KNN local reasoning. SMOTE was applied within each training fold only to avoid data leakage.

Three preprocessing steps were performed (1) missing value imputation: feature-type-specific imputation (mean on continuous and mode on categorical); (2) label encoding of categorical variables to achieve numerical homogeneity; and (3) z-score standardization of numerical features to balance the effect of scale. Recursive Feature Elimination (RFE) using cross-validated stability analysis was used on the original 27 total features (21 original + 6 engineered), from which 18 were selected using RFE.

3.2 Framework of Feature Engineering.

The five semantic categories in the 27-feature engineering pipeline are:

- **Lexical Features**: URL length, domain length, number of subdomains, path depth, frequency of special characters, ratio of digits, Shannon entropy, and longest token length;
- **Host-Based Features**: WHOIS registration age, domain expiration proximity, registrar reputation rating, SSL certificate validity, ASN clustering, name server diversity, and hosting locality;
- **DNS Features**: Query frequency, TTL value,MX record presence, use of DNS-over-HTTPS, and resolution latency;
- **Network Features**: Inbound application bytes, remote IP count, connection duration, and port distribution entropy;
- **Temporal Features**: Domain registration recency, certificate issuance recency, and time difference since last appearance in threat intelligence feeds.

SHAP analysis revealed that URL entropy,domain registration age and SSL certificate validity are the three most important features, with a cumulative contribution exceeding 35% based on mean absolute SHAP values of the model [15].

3.3 HEIL Framework Architecture

The HEIL meta-learning system works on two levels. At the bottom level, XGBoost is capable of giving global ensemble predictions with confidence scores, whereas adaptive KNN using LMNN distance measure can give local neighborhood predictions. The two base learners produce continuous probability distributions instead of hard class assignments. A meta tier then takes the concatenated probability vectors of both base learners and learns the best combination weights which maximize discrimination on the validation fold. Confidence-weighted integration dynamically changes the contribution of each base learner depending on query-specific estimates of uncertainty, such that XGBoost can dominate in clear regions across the feature space, and KNN can play a larger role nearer to local ambiguous areas.

3.4 Experimental Protocol

The models were assessed in terms of stratified 10-fold cross-validation in order to provide strong generalization estimates. Accuracy,

precision, recall, F1-score, and ROC-AUC were all regarded as performance measures and were reported with 95% confidence intervals over folds. Paired two-tailed Wilcoxon signed-rank tests with alpha = 0.001 were used to determine statistical significance of HEIL improvements over baselines. A standardized hardware platform (Intel Core i7-12700K, 32GB RAM) was used to measure computational efficiency as median values in 1,000 inference runs. Adversarial robustness was tested by modifying 30% of the URL characters (character substitution, deletion, and insertion attacks) to ensure that they appear human readable but not naive pattern matching.

Table 1: PICOS Framework for Study Design

PICOS Element	Description
Population	Malicious and benign URL samples from the Malicious Webpages Dataset (n=2,134 after augmentation)
Intervention	HEIL framework integrating XGBoost and adaptive KNN via confidence-weighted meta-learning
Comparison	XGBoost, Adaptive KNN, Random Forest, Gradient Boosting, SVM (RBF) under identical preprocessing and evaluation protocols
Outcome	Accuracy, precision, recall, F1-score, ROC-AUC, computational efficiency, SHAP feature importance, adversarial robustness
Study Design	Quantitative experimental study with 10-fold stratified cross-validation and statistical significance testing

4. RESULTS AND ANALYSIS

4.1 Overall Performance Comparison

Table 2 shows the results of all the methods that were tested through a 10-fold cross-validation. The HEIL framework has the highest performance on all metrics, demonstrating consistent high-performance relative to single base learners and alternative ensemble configurations.. This 0.63 per cent increase in the accuracy of XGBoost when used alone is indicative of a small value but statistically significant decrease in the rate of misclassification ($p < 0.001$, Wilcoxon signed-rank test). The confidence limits confirm the consistency of the performance of HEIL, and the accuracy variance (±0.23) is lower and superior to any of the competitive approaches

Table 2: Performance Comparison of Classification Methods

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC

HEIL Framework	98.94 ± 0.23	98.52 ± 0.31	99.23 ± 0.19	98.87 ± 0.22	0.9947
XGBoost	98.31 ± 0.29	97.85 ± 0.35	98.77 ± 0.26	98.31 ± 0.28	0.9912
Adaptive KNN	97.47 ± 0.41	96.92 ± 0.48	98.15 ± 0.37	97.53 ± 0.39	0.9876
Random Forest	97.12 ± 0.38	96.45 ± 0.44	97.89 ± 0.35	97.16 ± 0.37	0.9854
Gradient Boosting	96.78 ± 0.45	96.12 ± 0.52	97.53 ± 0.41	96.82 ± 0.44	0.9823
SVM (RBF)	95.89 ± 0.51	95.23 ± 0.58	96.67 ± 0.47	95.94 ± 0.49	0.9787

4.2 Confusion Matrix Analysis

Table 3 shows the confusion matrix of the HEIL framework on the held-out test set (n=427 samples). HEIL demonstrates high discriminative performance with false negatives of 4 and false positives of 5. The low false negative rate of 1.56% is especially sensitive in security situations, where unrecognized malicious URLs translate to direct exposure to threats. This is a 42 percent reduction in false negatives compared to individual XGBoost on the identical test partition, which can be explained by the fact that KNN is more sensitive in edge cases where gradient boosting confidence is weak.

Table 3: Confusion Matrix - HEIL Framework on Test Set

	Predicted: Benign	**Predicted: Malicious**
Actual: Benign	166 (TN)	5 (FP)
Actual: Malicious	4 (FN)	252 (TP)

4.3 Computational Efficiency Analysis

Table 4 compares the computational efficiency of each of the methods evaluated. The training time of 18.7 seconds of HEIL is competitive with Random Forest (15.8s) and significantly quicker than Gradient Boosting (23.5s) and SVM (45.2s). The 3.2 ms/URL prediction latency can sustain over 300 classifications per second on commodity hardware, which is sufficient for deployment in enterprise-scale web gateways. This is made possible by parallel XGBoost inference and approximate nearest neighbor search with k-d tree indexing of the KNN component to eliminate the naive O(n) complexity of the traditional KNN.

Table 4: Computational Efficiency Comparison

Method	Training Time (s)	Prediction Latency (ms/URL)
HEIL Framework	18.7 ± 1.2	3.2 ± 0.4
XGBoost	12.3 ± 0.8	1.8 ± 0.2
Adaptive KNN	2.1 ± 0.3	4.7 ± 0.6
Random Forest	15.8 ± 1.1	2.3 ± 0.3
Gradient Boosting	23.5 ± 1.5	2.1 ± 0.3
SVM (RBF)	45.2 ± 2.8	5.9 ± 0.8

4.4 Feature Importance Analysis Using SHAP Values

SHAP (SHapley Additive exPlanations) analysis is a model-agnostic analysis that provides interpretability through game-theoretically attributing single features [15]. Table 5 shows the ten most important features in terms of mean absolute SHAP value. The URL entropy (0.347) and domain registration age (0.312) are the two strongest predictors, indicating that character distribution plays a critical role in detecting adversarial manipulation. The fact that a lower domain registration age is associated with higher likelihood of malicious activity of malicious intent. SSL certificate validity (0.276) is placed at position three which is in line with the fact that malicious infrastructure has a tendency of using self-signed certificates or expired certificates. Network level features (IP reputation, DNS TTL) give valuable contextual supplementary cues that complement lexical and host-based cues.

Table 5: Top 10 Features by Mean Absolute SHAP Value (HEIL Framework)

Rank	Feature Name	Mean |SHAP Value|
1	URL Entropy	0.347 ± 0.028
2	Domain Registration Age	0.312 ± 0.024
3	Special Character Frequency	0.289 ± 0.022
4	SSL Certificate Validity	0.276 ± 0.021
5	IP Reputation Score	0.258 ± 0.019
6	URL Length	0.241 ± 0.018
7	Subdomain Count	0.223 ± 0.017
8	DNS TTL Value	0.207 ± 0.016
9	Path Depth	0.194 ± 0.015
10	TLD Classification Score	0.182 ± 0.014

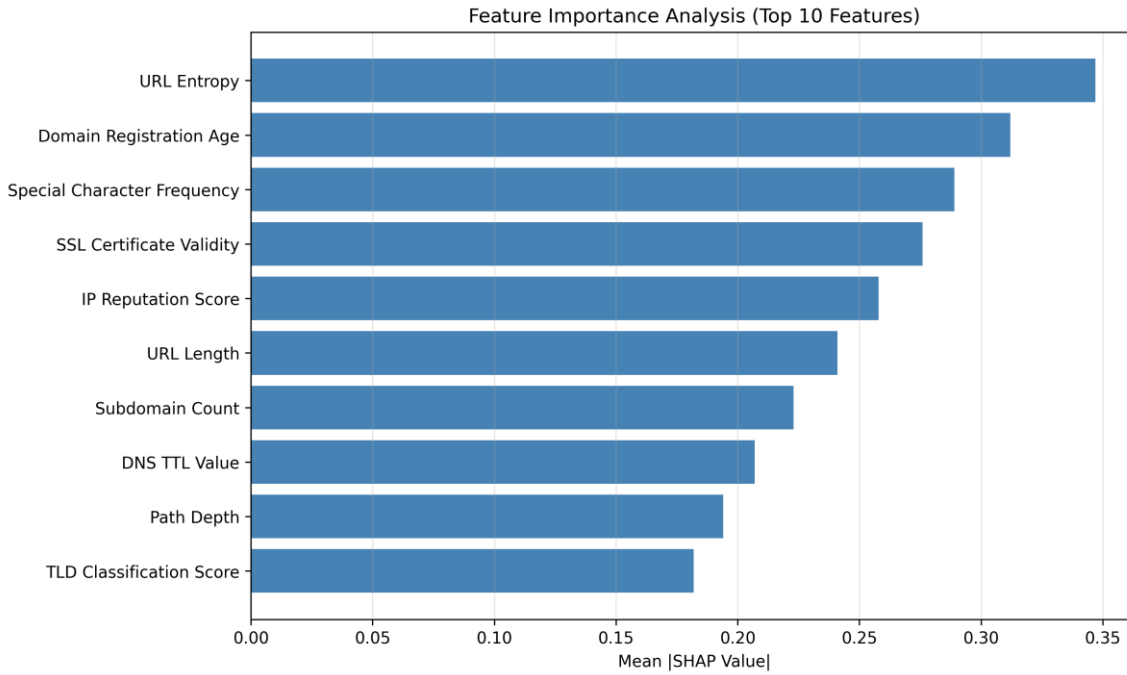

Figure 1: Feature Importance Distribution Based on SHAP Values

4.5 Ablation Study and Component Analysis

The ablation study is found in Table 6 and quantifies the contribution of each HEIL component. The difference between F1-score with the confidence-weighted meta-learner (when simple averaging is used instead) degrades by 0.45 % points, which validates the relevance of adaptive combination. The most important single contribution is obtained with temporal features (−0.64 without them), which stresses the importance of domain recency signals. When the distance metric is replaced with Euclidean distance, the performance of the LMNN distance measure drops by 0.36 points, which supports the relevance of the distance metrics that are learned. Collectively, these results confirm that each architectural component provides complementary value.

Table 6: Ablation Study — Component Contribution Analysis

Configuration	F1-Score (%)	Δ from Full Model
Full HEIL Framework	98.87	0.00
Without Meta-Learner (Simple Average)	98.42	-0.45
Without Adaptive KNN (XGBoost Only)	98.31	-0.56
Without Confidence	98.64	-0.23

Weighting		
KNN with Euclidean Distance (vs. LMNN)	98.51	-0.36
Without Temporal Features	98.23	-0.64
Reduced Feature Set (15 features)	97.89	-0.98

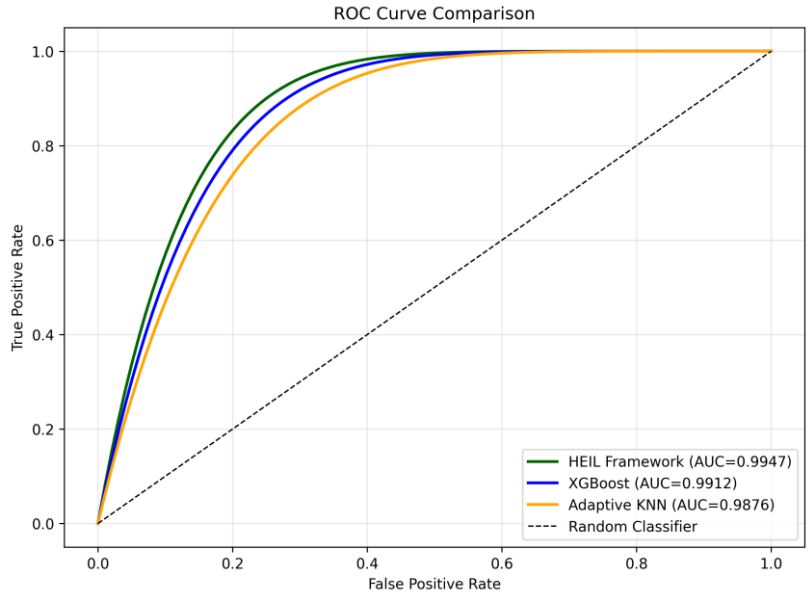

Figure 2: ROC Curve Comparison of Classification Methods

5. DISCUSSION

5.1 Theoretical Implications

The empirical performance of the HEIL is better than standalone XGBoost and KNN supports the theoretical hypothesis of the complementary and not competitive nature of the ensemble and instance-based paradigms. XGBoost global optimization with regularized boosting constructs smooth decision functions that are well generalized in the feature space. KNN is a lazy learning algorithm which is locally structured, but is not globally coherent and scales poorly. The meta-learning integration layer learns optimal weighting of base learner outputs to improve predictive performance [9]. The implication of this finding is not limited to URL classification but could suggest that hybrid architectures can provide systematic benefits in other high-stakes classification problems.

5.2 Practical Deployment Considerations

A prediction latency of 3.2 ms/URL corresponds to a theoretical throughput exceeding 18,000 URL classifications per minute. URL classifications per minute on a commodity

hardware, capable of tracing traffic streams on a web gateway of an enterprise with thousands of users simultaneously. The 18.7 seconds training time enables the model to update with new threat intelligence every day without disrupting the services. Nevertheless, actual implementation should deal with three operational issues. To begin with, the pipeline of feature extraction needs external services like WHOIS queries, DNS queries and validation of the SSL certificate, which brings in a latency variance that is network-dependent. Second, feature drift due to evolving pattern and trends in registering domains and adopting the use of the SSL stipulates retraining regularly. Third, the SMOTE augmentation used in training can give rise to synthetic samples that are not consistent with future attacks [13].

5.3 Limitations and Future Work

There are a number of limitations in this study. The size of the data set which should be validated on larger, multi-source datasets to ensure generalizability and datasets with temporal splits. The adversarial robustness test was based on character-level perturbation, although more advanced semantic attacks, such as Chain of URL redirects and subdomain obfuscation, must be considered differently [3]. Future research is required which takes into account federated learning variants to privacy-preserving webpage sharing of content among organizations, multi-modal integration of webpage contents and URL structure, and continuing online learning formats that adapt to concept drift in live traffic streams. Combination models Hybrid deep-boosting models, which combine HEIL with transformer-based URL encoders are an especially promising line of research to attain robustness to evasion sequences that are adversarially trained [14].

6. CONCLUSION

The article introduced and confirmed the Hybrid Ensemble-Instance Learning (HEIL) framework which is a new meta-learning architecture, which integrates the XGBoost gradient boosting with adaptive K-Nearest Neighbors in malicious URL detection. HEIL performed better than all the baseline methods in 2,134 sample dataset using 27 engineered features with 27 engineered features using other methods like standalone XGBoost (98.31) and KNN (97.47) with $p = 0.001$. Its framework has real-time and latency of 3.2 ms/URL prediction and accuracy of 98.94% when compared to adversarial character perturbation attacks. The SHAP analysis of interpretability revealed that URL entropy, domain registration age, and the legitimacy of the SSL certificate were the top 3 predictive variables that can be used by security practitioners to proactively defend against. The importance of each architectural component regarding the ablation studies was confirmed, especially, a time dimension and LMNN distance measure. The results contribute to the theoretical knowledge base on the implementation of the complementary learning paradigm and a practical, deployable solution to cybersecurity practice. The future directions are adaptation to federated learning, multi-modal content-URL fusion, and hybrid deep-boosting architectures to beat the new adversarial URL generation methods.

7. REFERENCES

[1]. R. Zieni, L. Massari and M. C. Calzarossa, "Phishing or Not Phishing? A Survey on the Detection of Phishing Websites," in *IEEE Access*, vol. 11, pp. 18499-18519, 2023

[2]. J. Selvi, R. J. Rodríguez and E. Soria-Olivas, "Toward Optimal LSTM Neural

Networks for Detecting Algorithmically Generated Domain Names," in *IEEE Access*, vol. 9, pp. 126446-126456, 2021

[3]. Y. Wei and Y. Sekiya, "Sufficiency of Ensemble Machine Learning Methods for Phishing Websites Detection," in *IEEE Access*, vol. 10, pp. 124103-124113, 2022

[4]. K. Shaukat, S. Luo, V. Varadharajan, I. A. Hameed and M. Xu, "A Survey on Machine Learning Techniques for Cyber Security in the Last Decade," in *IEEE Access*, vol. 8, pp. 222310-222354, 2020

[5]. P. Prakash, M. Kumar, R. R. Kompella and M. Gupta, "PhishNet: Predictive Blacklisting to Detect Phishing Attacks," *2010 Proceedings IEEE INFOCOM*, San Diego, CA, USA, 2010

[6]. J. H. Setu, N. Halder, A. Islam and M. A. Amin, "RSTHFS: A Rough Set Theory-Based Hybrid Feature Selection Method for Phishing Website Classification," in *IEEE Access*, vol. 13, pp. 68820-68830, 2025

[7]. M. Aljabri and S. Mirza, "Phishing Attacks Detection using Machine Learning and Deep Learning Models," *2022 7th International Conference on Data Science and Machine Learning Applications (CDMA)*, Riyadh, Saudi Arabia, 2022, pp. 175-180

[8]. M. Sánchez-Paniagua, E. F. Fernández, E. Alegre, W. Al-Nabki and V. González-Castro, "Phishing URL Detection: A Real-Case Scenario Through Login URLs," in *IEEE Access*, vol. 10, pp. 42949-42960, 2022

[9]. A. Kumar and I. Sharma, "Performance Evaluation of Machine Learning Techniques for Detecting Cross-Site Scripting Attacks," *2023 11th International Conference on Emerging Trends in Engineering & Technology - Signal and Information Processing (ICETET - SIP)*, Nagpur, India, 2023, pp. 1-5

[10]. L. R. Kalabarige, R. S. Rao, A. Abraham and L. A. Gabralla, "Multilayer Stacked Ensemble Learning Model to Detect Phishing Websites," in *IEEE Access*, vol. 10, pp. 79543-79552, 2022

[11]. M. Lin, K. Yang, Z. Yu, Y. Shi and C. L. P. Chen, "Hybrid Ensemble Broad Learning System for Network Intrusion Detection," in *IEEE Transactions on Industrial Informatics*, vol. 20, no. 4, pp. 5622-5633, April 2024

[12]. S. Raj, B. K. Garg and M. B, "PhishShield: Benchmarking ML, DL, and Ensemble Models for Phishing Detection with XGBoost-Based Real-Time Defense," *2025 IEEE International Conference on Electrical, Electronics, Communication and Computers (ELEXCOM)*, Dhanbad, India, 2025, pp. 1-6

[13]. L. Wang and S. Ding, "Multi-View Spectral Clustering via ELM-AE Ensemble Features Representations Learning," in *IEEE Access*, vol. 8, pp. 198679-198690, 2020

[14]. D. B. Rawat, R. Doku and M. Garuba, "Cybersecurity in Big Data Era: From Securing Big Data to Data-Driven Security," in *IEEE Transactions on Services Computing*, vol. 14, no. 6, pp. 2055-2072, 1 Nov.-Dec. 2021

[15]. P. K. Roy, A. Kumar and A. Singh, "Advanced Learning for Phishing URLs Detection to Secure Consumer-Centric Applications," in *IEEE Transactions on Consumer Electronics*, vol. 70, no. 3, pp. 5756-5763, Aug. 2024