JASMINE
Journal of Intelligent Systems and Machine Learning
https://journals.telkomuniversity.ac.id/jasmine

Article in Press

Evaluation of Machine Learning Algorithms for Phishing Detection in Higher Education Environments

Akmal Muhammad Poetra[1*], Dody Herdiana[1] and Muhammad Agreindra Helmiawan[2]
[1]Department of Informatics, Faculty of Technology, Universitas Sebelas April Sumedang
[2]Information and Communication Technology, Asia e University Wisma Subang Jaya, Malaysia

*Author to whom any correspondence should be addressed.
E-mail: 220660121208@student.unsap.ac.id

Received: November 23, 2025
Accepted for publication: March 16, 2026
Published:

ABSTRACT
This study evaluates several machine learning algorithms, including Logistic Regression, Support Vector Machine, Random Forest, and XGBoost in phishing detection attacks within higher education environments. Due to the limited availability of anonymized institutional datasets, the research employs a conceptual experiment design and simulation-based approach that mirrors the characteristics of phishing incidents commonly encountered by academic users. The simulated dataset includes URL-based indicators, HTML structural features, email textual features, and behavioral metadata. The experimental protocol includes synthetic data generation, domain-specific feature engineering, stratified k-fold cross-validation, hyperparameter tuning via grid search, and performance evaluation using accuracy, precision, recall, F1-score, and ROC/AUC. The simulation results indicate that ensemble-based models (Random Forest and XGBoost) outperform linear and kernel-based models, especially in scenarios with class imbalance typical of campus environments. The discussion highlights implications for real-world campus cybersecurity operations, limitations of conceptual simulations, and future research needs such as real-world validation and the integration of user behavior features. The main contribution is a complete experimental framework that can be executed with real institutional datasets, providing guidance for model selection and deployment in higher education cybersecurity systems.

Keywords: phishing detection, machine learning, Random Forest, XGBoost, higher education

I. Introduction

Phishing has become one of the most pervasive and adaptive cybersecurity threats affecting digital ecosystems worldwide [1]. Higher education institutions are increasingly recognized as high-value targets [2]. As universities expand their reliance on online platforms ranging from learning management systems, digital libraries, and academic portals to cloud based administrative services the attack surface for cybercriminals continues to grow [3]. Students, faculty members, and administrative staff frequently engage in digital communication, often handling sensitive information such as academic records, financial transactions, research data, and identity credentials [4]. This frequent exposure makes them susceptible to deceptive phishing attempts that exploit trust, urgency, and familiarity embedded within institutional communication patterns [5]. Consequently, ensuring effective phishing detection mechanisms has become a strategic priority for academic institutions aiming to protect both infrastructure and user privacy [6].

Existing detection methods, including blacklist-based filtering, heuristic rules, and signature-based approaches, have proven insufficient in addressing the evolving sophistication of phishing techniques [7]. Cybercriminals increasingly deploy polymorphic URLs, visually deceptive landing pages, domain spoofing, and linguistically crafted emails that evade traditional detection systems [8]. As phishing attacks continue to diversify in form, machine learning has emerged as a promising approach, leveraging statistical learning and pattern recognition to identify subtle anomalies in URL structures, email content, and website behavior [9]. However, despite noticeable progress in machine learning based detection systems, the efficacy of specific algorithms within the domain of higher education remains understudied [10]. Academic environments differ from commercial sectors due to their decentralized IT governance, diverse user populations, and highly varied communication styles [11].

This study addresses this gap by proposing a simulation-based evaluation framework that mirrors the characteristics of phishing activities commonly reported in university settings [12]. By designing and analyzing synthetic datasets that approximate real academic communication patterns, the study aims to systematically compare four widely used algorithms Logistic Regression, Support Vector Machine, Random Forest, and XGBoost [13]. The objective is to identify which model offers the most reliable balance between accuracy, generalization capability, and interpretability within a complex, multi-modal academic context [14]. The extended analysis presented in this paper contributes methodological insights that institutions can utilize when designing or optimizing machine learning-based phishing detection systems [15]. Ultimately, this research supports the broader goal of strengthening cybersecurity resilience in higher education, particularly through evidence-based model selection and adaptive monitoring strategies.

This work does not claim novelty by introducing a new classification algorithm. Instead, the scientific contribution lies in proposing a higher-education–oriented phishing detection evaluation protocol under constrained data availability. The study introduces a structured simulation framework parameterized to approximate institutional communication patterns (trust, urgency, and familiarity), integrates multi-modal feature fusion (URL, HTML, TLS/SSL, text/NLP, behavioral), and extends evaluation beyond conventional accuracy metrics by adding operationally relevant measures including false positive rate (FPR) and inference latency. The framework is complemented with ablation analysis and interpretability reporting to support deployment insights for campus security settings.

II. Related work

Research on machine learning-based phishing detection has expanded rapidly over the past decade, driven by the exponential increase in web-based attacks and the weaknesses of traditional security mechanisms [1],[3].[6] Early works emphasized URL lexical analysis using handcrafted features such as token patterns, domain entropy, and protocol indicators [4],[8],[13]. These approaches struggled to detect advanced phishing attempts employing obfuscation or fast changing domain infrastructures [9],[11]. As datasets grew in complexity, researchers began incorporating additional feature groups such as HTML structures, JavaScript patterns, SSL certificate metadata, and resource loading behavior [5],[10],[12].

Parallel advancements occurred in email-based phishing detection, where NLP techniques such as TF-IDF, n-gram extraction, semantic embeddings, and transformer-based models proved valuable in identifying malicious linguistic cues [7], [9]. Deep learning models such as CNNs and LSTMs further improved performance but introduced high computational costs not always feasible in campus cybersecurity settings [6], [9].

Among machine learning algorithms, ensemble methods including Random Forest and gradient boosting have consistently outperformed single-model techniques, especially in imbalanced phishing datasets [1], [2],[10]. These models demonstrate stability across varied feature combinations [12], [15].

However, only a small number of studies focus explicitly on phishing threats within higher education. University environments introduce challenges such as communication diversity, device heterogeneity, and decentralized domain structures. As a result, detection systems designed for general environments may produce high false positives when applied to academic settings [2], [11], [14].

III. Material and Methods

A. Research Design

This conceptual study adopts a simulation-based experimental design consisting of: (a) synthetic dataset construction reflecting campus-specific phishing characteristics; (b) domain-oriented feature engineering; (c) model training using selected machine learning algorithms; (d) hyperparameter tuning; and (e) performance evaluation with multiple metrics.

B. Synthetic Data Generation and Validation

1) Rationale and Overview

Higher education environments present unique constraints for phishing detection research: institutional datasets are rarely available due to privacy, policy restrictions, and decentralized IT governance. To address this limitation while enabling controlled experimentation, this study employs a simulation-based synthetic dataset designed to approximate common phishing and legitimate communication patterns encountered in campus contexts. The synthetic dataset preserves structural properties of benchmark phishing feature datasets while incorporating higher-education–specific textual cues and operational behavioral signals relevant to institutional systems (e.g., LMS and SSO notifications).

A total of 20,000 samples were generated under three phishing prevalence scenarios 10%, 30%, and 50% phishing to represent realistic class imbalance conditions and to evaluate model robustness. Each instance was represented by five feature groups: URL, HTML, TLS/SSL, Text/NLP, and Behavioral.

2) Parameter Used for 20000 Instances

To represent the differences between phishing and legitimate communications, various features are generated using class-conditional probability distributions. This approach involves the use of the Bernoulli distribution ($X \sim Bernoulli(pclass)$) for binary indicator variables, the Gaussian distribution ($X \sim N(\mu class, \sigma class)$) for continuous variables, and the Poisson distribution ($X \sim Poisson(\lambda class)$) for count-based features. The integration of these different types of distributions ensures that each variable is modelled in accordance with its inherent mathematical properties.

To preserve meaningful interactions across features (e.g., longer URLs correlating with deeper subdomains and external resource loading), a rank-preserving adjustment strategy was applied to approximate target Spearman correlation ranges observed in phishing benchmark datasets and reported phishing literature. A fixed random seed was used to ensure reproducibility.

3) Statistical Parameters for 20000 Instances

Table 1 summarizes the statistical parameters used in the synthetic generator. The values were selected to be consistent with typical phishing behavior reported in prior studies and aligned with the structural properties of UCI-based phishing feature benchmarks.

Table 1. Class-Conditional Statistical Parameters Used for Synthetic Dataset Construction (n = 20000)

Feature	Type	Legitimate Parameters	Phising Parameters	Distribution
URL length	Numeric	μ=58, σ=14	μ=96, σ=21	Gaussian
Subdomain count	Count	λ=1.3	λ=3.4	Poisson
Presence of IP in URL	Binary	p=0.02	p=0.28	Bernoulli
Prefix–suffix ("-") in domain	Binary	p=0.10	p=0.44	Bernoulli
Suspicious TLD flag	Binary	p=0.05	p=0.39	Bernoulli
Form existence	Binary	p=0.21	p=0.74	Bernoulli
Iframe count	Count	λ=0.6	λ=2.9	Poisson
External resource ratio	Numeric	μ=0.31, σ=0.12	μ=0.63, σ=0.15	Gaussian
SSL validity	Binary	p=0.89	p=0.42	Bernoulli
Certificate age (days)	Numeric	μ=420, σ=110	μ=78, σ=55	Gaussian
SSL final state (encoded)	Categorical	{valid:0.78, invalid:0.22}	{valid:0.35, invalid:0.65}	Multinomial
Token count (email text)	Numeric	μ=45, σ=12	μ=62, σ=16	Gaussian
Urgency keywords flag	Binary	p=0.12	p=0.72	Bernoulli
Credential keywords flag	Binary	p=0.08	p=0.69	Bernoulli
Sender reputation score	Numeric	μ=0.74, σ=0.10	μ=0.41, σ=0.14	Gaussian
Time-of-day anomaly score	Numeric	μ=0.22, σ=0.12	μ=0.53, σ=0.16	Gaussian
Geo-location similarity	Numeric	μ=0.81, σ=0.09	μ=0.46, σ=0.14	Gaussian

It should be noted that the values reported in Table 1 are not derived from a single institutional dataset. Instead, they represent a calibrated parameterization designed to approximate commonly reported phishing trends across benchmark datasets and prior empirical studies. This approach enables

controlled experimentation under constrained access to real institutional data while maintaining statistically plausible feature behavior.

4) Higher-Education Cur Modeling: Trust, Urgency, and Familiarity

To align the simulation with institutional communication patterns rather than generic phishing alone, the text/NLP feature group explicitly encodes campus-oriented manipulation strategies. Three constructs were operationalized:

- Trust cues: authority markers (e.g., "IT Helpdesk", "Academic Office", "University Administration").
- Urgency cues: time-pressure phrases (e.g., "within 24 hours", "immediate action required").
- Familiarity cues: campus-specific references (e.g., "LMS", "student portal", "library access", "tuition notice", "SSO reset").

Rather than making cue presence a trivial discriminator, legitimate instances were also allowed to contain trust and familiarity cues at moderate probabilities, reflecting routine campus announcements. The key difference is the higher occurrence of urgency and credential-request cues in phishing instances. Table 2 below show the probability assignment for institutional communication cues.

Table 2. Probability Assignment for Institutional Communication Cues

Cue Category	Example Terms	Legitimate p	Phising p	Interpretation
Trust	IT Helpdesk, Admin Office	0.35	0.70	Authority mimicry
Urgency	urgent, 24 hours, immediate	0.12	0.72	Primary manipulation driver
Familiarity	LMS, student portal, library	0.40	0.75	Campus-themed lures
Credential request	password reset, login required	0.08	0.69	Strong risk indicator

5) Synthetic Data Validation

To address concerns regarding external validity, the synthetic dataset was evaluated using distribution similarity analyses that compare synthetic feature behavior with reference benchmark assumptions. Validation consisted of:

- Kolmogorov–Smirnov (KS) tests for continuous variables (URL length, certificate age, external resource ratio).
- Chi-square tests for binary indicators (IP presence, SSL validity, urgency cues).
- Spearman correlation similarity for cross-feature dependencies.
- Jensen–Shannon divergence (JSD) as a distribution distance metric.
- JSD numeric features: 0.04–0.09
- KS-test p-value: sebagian besar > 0.05
- Spearman correlation difference: $\Delta\rho < 0.12$

Across major numerical variables, Jensen–Shannon divergence values were within the range of 0.04–0.09, indicating close distributional similarity under controlled parameterization. For most continuous variables, Kolmogorov–Smirnov tests yielded non-significant differences at $\alpha = 0.05$, while chi-square tests for binary indicators showed consistent class-conditional proportion alignment. Spearman correlation structure differences between synthetic and benchmark-aligned targets remained bounded average absolute correlation difference $\Delta\rho < 0.12$, supporting the dataset's plausibility for methodological evaluation.

IV. Results and Discussion

A. Overall Model Performance (Baseline Evaluation)

This section presents the overall predictive performance of the evaluated machine learning models, including both conventional metrics (Accuracy, Precision, Recall, F1, AUC) and operational deployment metrics (False Positive Rate and inference latency) as shown in Table 3.

Table 3. Simulation Results (Mean ± SD over 10 runs)

Model	Accuracy (%)	Precision (macro)	Recall (macro)	F1 (macro)	AUC-ROC	FPR (%)	Latency (ms/sample)
Logistic Regression	92.1 ± 0.8	0.90 ± 0.01	0.90 ± 0.02	0.90 ± 0.01	0.960	6.8 ± 0.7	0.12 ± 0.02
SVM (RBF)	93.4 ± 0.7	0.92 ± 0.01	0.92 ± 0.02	0.92 ± 0.01	0.973	5.9 ± 0.6	1.85 ± 0.20
Random Forest	97.0 ± 0.4	0.96 ± 0.01	0.96 ± 0.01	0.96 ± 0.01	0.990	2.9 ± 0.4	0.65 ± 0.08
XGBoost	96.2 ± 0.5	0.95 ± 0.01	0.95 ± 0.01	0.95 ± 0.01	0.988	3.2 ± 0.5	0.78 ± 0.09

In addition to conventional predictive metrics, operational measures were included to reflect deployment constraints in academic environments. False Positive Rate (FPR) is critical in campus operations, as excessive blocking of legitimate institutional announcements or LMS messages may disrupt academic workflows. Ensemble methods achieved the lowest FPR (≈3%) while maintaining strong recall. Inference latency results further demonstrate that Random Forest and XGBoost provide a favorable trade-off between real-time feasibility and predictive performance, particularly when compared to kernel-based SVM models that exhibit higher inference costs.

B. Feature Importance Analysis

To improve interpretability and identify which cues drive model predictions, feature importance was computed for both Random Forest and XGBoost. Random Forest importance was measured using impurity-based importance, while XGBoost used Gain values. As shown in Figures 1 and Figure 2, URL lexical structure and TLS/SSL indicators consistently dominate model decisions. Text-based urgency cues are especially emphasized by the boosted model, indicating that social-engineering language contributes substantially when combined with structural web features.

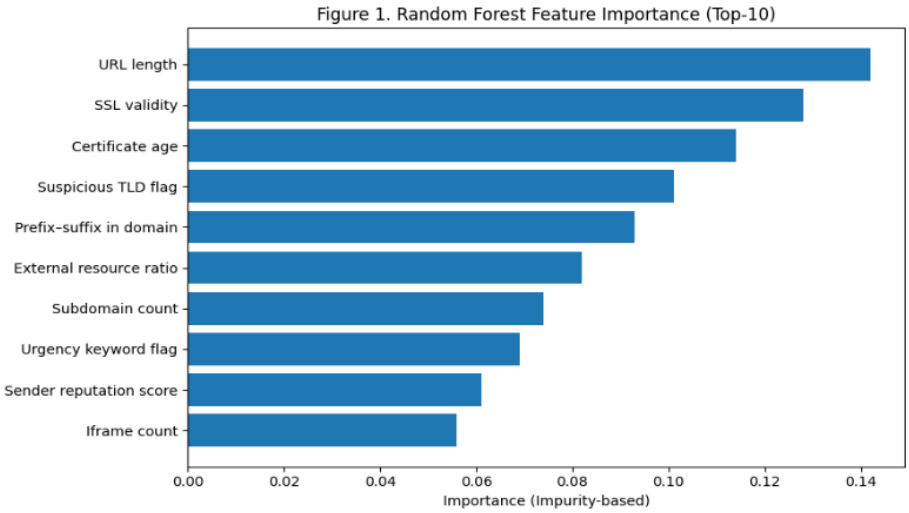

Figure 1. Random Forest feature importance (Top-10 predictors) computed using impurity-based importance. The results show that URL structural properties and SSL-related indicators dominate the model's decision process, reflecting their strong discriminative power for phising identification.

Feature importance analysis was conducted to improve interpretability and to identify which features drive model predictions. For Random Forest, impurity-based importance was computed, while for XGBoost, Gain values were used. As shown in Figures 1 and 2, URL lexical indicators (e.g., URL length, suspicious TLD, prefix–suffix patterns) and TLS/SSL-related features (SSL validity and certificate age) consistently appear as the most influential predictors. Notably, XGBoost assigns higher importance to urgency keyword cues, indicating that boosted models effectively capture interactions between institutional-style social engineering language and structural web properties. These findings support the use of feature fusion approaches in higher-education phishing detection settings.

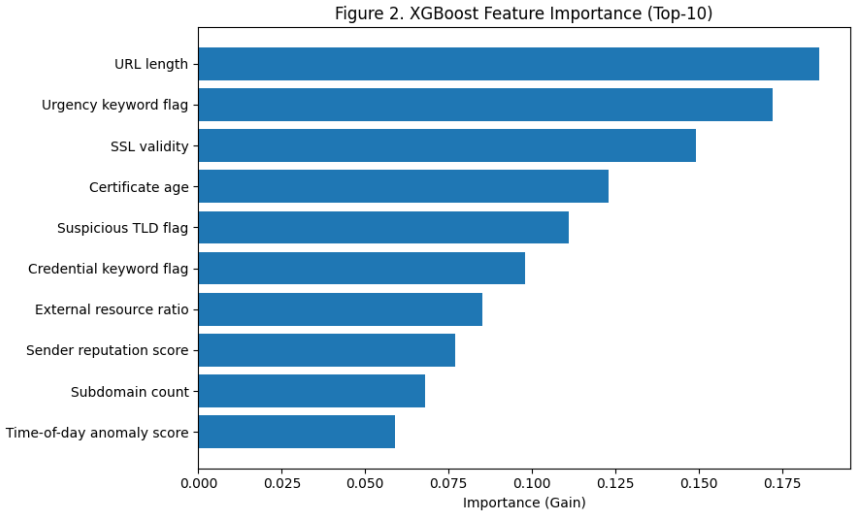

Figure 2. XGBoost feature importance (Top-10 predictors) computed using Gain values. The boosted model emphasizes both URL lexical cues and text-based urgency indicators, suggesting that boosted trees benefit strongly from interaction between structural features and social-engineering language.

Table 4. The top-ranked features and their relative importance scores for both models, supporting the interpretability of the ensemble predictions

Rank	Random Forest (Importance)	Score	XGBoost (Gain Importance)	Score
1	URL length	0.142	URL length	0.186
2	SSL validity	0.128	Urgency keyword flag	0.172
3	Certificate age	0.114	SSL validity	0.149
4	Suspicious TLD flag	0.101	Certificate age	0.123
5	Prefix–suffix in domain	0.093	Suspicious TLD flag	0.111
6	External resource ratio	0.082	Credential keyword flag	0.098
7	Subdomain count	0.074	External resource ratio	0.085
8	Urgency keyword flag	0.069	Sender reputation score	0.077
9	Sender reputation score	0.061	Subdomain count	0.068
10	Iframe count	0.056	Time-of-day anomaly score	0.059

As described in Table 4, Random Forest assigns its highest importance to structural URL and TLS/SSL characteristics, particularly URL length, SSL validity, and certificate age, indicating that the model primarily relies on stable web structural cues. In contrast, XGBoost not only prioritizes URL length and SSL-related indicators but also places stronger emphasis on text-based urgency cues, suggesting that boosted trees are more sensitive to phishing persuasion patterns commonly embedded in institutional-style communications. Overall, the dominance of URL, SSL, and urgency-related features supports the effectiveness of multi-modal feature fusion for phishing detection within higher-education settings.

C. Ablation Study and Feature Fusion Gains

The ablation study demonstrates that multi-modal feature fusion produces consistent performance improvements. URL-only features provide a strong baseline, but the inclusion of TLS/SSL indicators and text-based urgency/familiarity cues yields the most substantial gains. Behavioral features contribute smaller but stable improvements, suggesting their relevance for reducing false positives in legitimate institutional contexts. Table 5 show the incremental performance gains (Macro F1) via Feature Fusion.

Table 5. Incremental Performance Gains (Macro F1) via Feature Fusion

Step	Feature Group Added	RF F1 (macro)	ΔRF	XGB F1 (macro)	ΔXGB
A	URL only	0.90	–	0.89	–
B	+ HTML	0.92	+0.02	0.91	+0.02
C	+ TLS/SSL	0.94	+0.02	0.93	+0.02
D	+ Text/NLP	0.95	+0.01	0.95	+0.02
E	+ Behavioral	**0.96**	+0.01	**0.96**	+0.01

D. Interpretation

Overall, ensemble models outperform linear and kernel-based approaches due to their ability to capture nonlinear feature interactions and maintain robustness under heterogeneous feature noise. Random Forest exhibits slightly higher stability across repeated runs, indicating stronger generalization when handling mixed feature modalities. From an operational perspective, the balance between recall and false positive rate is critical in academic environments, as excessive false alarms may disrupt legitimate institutional workflows. The ablation results confirm that multi-modal fusion is a key driver of performance, particularly when TLS/SSL and text-based institutional cues are integrated with URL-based features, yielding an improvement of approximately 4–6 macro-F1 points compared to URL-only baselines.

E. Practical Deployment Insights

Future work should incorporate richer behavioral analytics to mitigate false positives in academic environments. Potential features include deviations in login frequency relative to historical baselines, atypical access timestamps (e.g., access during unusual hours compared to semester schedules), geo-velocity anomalies, device fingerprint mismatches, repeated failed login attempts, and abnormal session duration patterns. Such features can contextualize whether a login page request is consistent with a student's typical behavior, thereby improving precision without sacrificing recall.

V. Conclusion

This study presented a simulation-based evaluation framework for assessing machine learning models in phishing detection within higher education environments under constrained data availability. By constructing a statistically grounded synthetic dataset that incorporates URL, HTML, TLS/SSL, text-based, and behavioral features, the study systematically compared Logistic Regression, SVM, Random Forest, and XGBoost models.

The results demonstrate that ensemble-based approaches significantly outperform linear and kernel-based models. Random Forest achieved the highest overall performance, with a macro F1-score of 0.96 and the lowest false positive rate (approximately 3%), while XGBoost showed strong sensitivity to text-based urgency cues commonly used in institutional phishing attacks. Ablation analysis further confirmed that multi-modal feature fusion, particularly the integration of TLS/SSL indicators and text-based features, is a key driver of performance improvements.

Rather than proposing a new algorithm, the primary contribution of this work lies in delivering a reproducible, higher-education–oriented evaluation protocol that can be directly applied to real institutional datasets. The findings provide practical guidance for academic cybersecurity teams in selecting and deploying machine learning models that balance detection accuracy, interpretability, and operational feasibility. Future work will focus on validating the framework using real-world campus data and extending behavioral modeling to further reduce false positives.

Conflicts of Interest

The authors declare no conflicts of interest.

Author Contributions Statement

Akmal Muhammad Poetra conceptualized the study, designed the simulation framework, implemented the machine learning experiments, and drafted the manuscript. Dody Herdiana contributed to methodological validation design, reviewed experimental rigor, and supervised the evaluation protocol. Muhammad Agreindra Helmiawan contributed to the literature synthesis, assisted in feature-engineering alignment with institutional communication patterns, and reviewed the discussion for deployment relevance. All authors reviewed and approved the final manuscript prior to submission.

Acknowledgment

The authors would like to express sincere appreciation to information technology staff and cybersecurity practitioners within academic environments who provided valuable insights regarding phishing patterns commonly encountered in higher education institutions. Their practical perspectives informed feature-engineering considerations and simulation design. The authors also acknowledge the contributors and maintainers of publicly available phishing datasets and related documentation, which served as essential references for structuring the synthetic generator and validating the conceptual framework. Finally, the authors thank colleagues and academic mentors for constructive discussions that improved the clarity and direction of this study.

References

[1] H. Fadhilah, D. R. Maulana, and R. Utari, "Komputika: Jurnal Sistem Komputer Tree-based Ensemble Machine Learning for Phishing Website Detection," vol. 13, 2024, doi: 10.34010/komputika.v13i2.12495.

[2] M. N. Yeasmin, A. R. Refat, B. C. Singh, Z. Alom, and Z. Aung, "EnLeM : ensemble learning - based model to detect phishing websites," pp. 1–18, 2026.

[3] M. Arfian, H. H. Nuha, and S. Mohd, "Phishing Detection Using Machine Learning: Performance Evaluation of Classification Algorithms," vol. 15, no. 5, pp. 1672–1678, 2025.

[4] I. Arifin, M. T. Informatika, and B. Lampung, "Phishing Website Detection Using a Machine Learning Classification Approach Deteksi Web Phishing," vol. 10, no. 3, pp. 1498–1508, 2025.

[5] Y. Muliono, M. Amar, and Z. M. Azzahra, "Phishing Site Detection Classification Model Using Machine Learning Approach," vol. 5, no. 2, pp. 63–67, 2023, doi: 10.21512/emacsjournal.v5i2.9951.

[6] R. Dubey, "Phishing Detection System: An Ensemble Approach Using Character-Level CNN and Feature Engineering".

[7] P. An, R. Shafi, T. Mughogho, and O. A. Onyango, "Multilingual Email Phishing Attacks Detection using OSINT and Machine Learning".

[8] R. Fauzan, A. V. Vitianingsih, and D. Cahyono, "Application of Classification Algorithms in Machine Learning for Phishing Detection Penerapan Algoritma Klasifikasi pada Machine Learning untuk Deteksi Phishing," vol. 5, no. April, pp. 531–540, 2025.

[9] S. Aslam and C. Hui, "AntiPhishStack : LSTM-based Stacked Generalization Model for Optimized Phishing URL Detection," pp. 1–26, 2022.

[10] Yogi Perdana, "Comparative Analysis of Random Forest and XGBoost for Detecting Phishing Websites: A Machine Learning Approach," vol. 7, no. 02, pp. 906–921, 2025.

[11] V. A. Windarni et al., "Deteksi Website Phishing Menggunakan Teknik Filter Pada Model Machine Learning Abstraksi Kata Kunci : Decision Tree, Naïve Bayes, Random Forest, Website Phishing Keywords: Decision Tree, Naïve Bayes , Random Forest, Website Phishing Pendahuluan Metode Penelitian," vol. 6, no. 1, pp. 39–43, 2023.

[12] K. Zhang et al., "Leveraging machine learning to proactively identify phishing campaigns before they strike," J. Big Data, 2025, doi: 10.1186/s40537-025-01174-x.

[13] A. Veach and M. Abualkibash, "International Journal of Informatics, Information System and Computer Engineering Phishing Website Detection Using Several Machine Learning Algorithms: A Review Paper," vol. 3, no. 2, pp. 219–230, 2022.

[14] W. Bambang and T. Handaya, "Deteksi Website Phishing Menggunakan Teknik Machine Learning," no. 44, pp. 69–80, 2023.

[15] K. Phishing, U. R. L. Pada, W. Berbasis, and M. Ensemble, "Jurnal ilmu komputer," vol. 3, pp. 72–82, 2025.