

FINAL RESEARCH PROPOSAL

An Explainable Machine Learning Framework for Phishing Website Detection Using URL and Content-Based Features

Course Title: Research Methodology in Computer Science

Course Code: AI800

Lecturer: Dr. Mohammed Anbar

Submission Date: 15 August 2026

Abstract

Phishing can be quite a persistent problem as the attacker can change the domain, page layout and social engineering very quickly. While machine-learning models can identify suspicious patterns that cannot be contained in static blacklists, they might result in high accuracy in detecting (but not understanding why) suspicious patterns. The proposal offers a quantitative experimental framework by integrating URL-based and lightweight features of web-content with explainable artificial intelligence (XAI). The four classifiers (Logistic Regression, Decision Tree, Random Forest, and XGBoost) will be tested on the same pipeline. To handle duplicates, missing values, etc., and take into account the particular class imbalance and leakage, preprocessing will be performed, and feature selection will find the most informative indicators. Measuring the performance will be based on accuracy, precision, recall, F1-score, ROC-AUC and confusion matrices. Global and case explanations will then be given using the best model with SHAP and LIME. The importance of feature quality, consistent evaluation and interpretability for reliable phishing detection has been emphasized in recent research by Kapan & Sora Gunal (2023), Kocyigit et al. (2024), and Nauman et al. (2025). The expected contribution is a repeatable, transparent framework that allows for achieving performance in a way that is understandable from a cybersecurity perspective.

Keywords: phishing detection, machine learning, URL analysis, cybersecurity, explainable artificial intelligence, SHAP, LIME.

Table of Contents

Abstract

Chapter 1: Introduction

- 1.1 Introduction
- 1.2 Background of the Study
- 1.3 Problem Statement
- 1.4 Research Gap
- 1.5 Research Questions
- 1.6 Research Objectives
- 1.7 Research Scope and Delimitations
- 1.8 Research Significance

Chapter 2: Literature Review

- 2.1 Phishing as an Evolving Cybersecurity Threat
- 2.2 Traditional Detection Approaches
- 2.3 Machine Learning for Phishing Detection
- 2.4 URL-Based Features
- 2.5 Content-Based Features
- 2.6 Feature Selection and Data Quality
- 2.7 Deep and Hybrid Learning
- 2.8 Explainable Artificial Intelligence in Cybersecurity
- 2.9 Critical Synthesis of Previous Studies
- 2.10 Research Gap Derived from the Literature
- 2.11 Literature-to-Objective Alignment

The literature directly supports the proposed objectives: Tamal et al. (2024) and Opara et al. (2024) guide feature identification; Alnemari and Alshammari (2023) and Almujaheed et al. (2024) support classifier comparison; and Kocyigit et al. (2024), Fan et al. (2024), and Nauman et al. (2025) motivate feature reduction and XAI.

- Feature identification → Tamal et al. (2024) and Opara et al. (2024) → extract URL and lightweight content indicators.
- Classifier comparison → Alnemari and Alshammari (2023) and Almujaheed et al. (2024) → use one evaluation protocol.
- Feature reduction and XAI → Kocyigit et al. (2024), Fan et al. (2024), and Nauman et al. (2025) → select features and apply SHAP/LIME.

Chapter 3: Research Methodology

- 3.1 Research Design
- 3.2 Proposed Research Framework
- 3.3 Data Source and Sampling
- 3.4 Data Preprocessing
- 3.5 Feature Engineering and Selection
- 3.6 Model Development
- 3.7 Experimental Procedure
- 3.8 Evaluation Metrics
- 3.9 Explainability Analysis
- 3.10 Tools and Technologies

3.11 Validity, Reliability, and Reproducibility
3.12 Ethical Considerations
3.13 Expected Outcomes
3.14 Research Milestones and Gantt Chart
3.15 Chapter Summary
References

Chapter 1: Introduction

1.1 Introduction

Phishing is one type of cyber deception technique that involves mimicking an authorized organization or service to trick users into giving personal data, financial information or other sensitive information. The rise of online banking, e-commerce, cloud services, and remote work has led to a rise in the attack surface and value of stolen accounts. Today's campaigns are able to register domain names for a short period of time, copy legitimate logon pages, and use confusing URL formats, making blacklist techniques less effective. In this context, recent reviews report phishing as an adaptive problem that needs data-driven approaches to identify new patterns that it hasn't encountered before (Thakur et al., 2023; Li et al., 2024; Popescul & Radu, 2025).

This is a good application of machine learning since a classifier can be trained to learn all the relationships between lexical, structural and content features. But deployment of cybersecurity must also be held accountable – there must be understandable evidence that accompanies an alert. This proposal introduces XAI alongside predictive modeling to enable the final system to identify not just suspicious websites, but also the most significant features in its decision-making process (Achuthan et al., 2024; Fan et al., 2024).

1.2 Background of the Study

Early phishing defenses were based on blacklists, whitelists, heuristics that were manually defined, and visual similarity checks. In the world of malicious URLs, these tools are still useful but have their limitations when the bad guys use new URLs or don't use a known URL signature. Machine learning solves this issue by learning from labeled examples, and is able to be retrained as the attack behavior changes (Kapan & Sora Gunal, 2023). Lexical URL indicators, domain information, HTML structure, redirects, forms, and some combinations of these indicators have been taken up for study. The advantage of URL-only models is that they can work before a page is loaded completely, while content features provide context for more advanced imitations (Alnemari & Alshammari, 2023; Opara et al., 2024).

Deep and ensemble models have enhanced the ability to predict, but, it is also introducing a transparency challenge. A security analyst could reasonably ask the question, is a prediction based on domain structure, a suspicious redirect pattern, an

abnormal form or an artefact in the training data? These relationships can be uncovered using XAI techniques like SHAP and LIME, and can be used to help provide human oversight of automated decision-making (Nauman et al., 2025).

1.3 Problem Statement

Phishing detection research has reported high rates of success, but there are three caveats to note. Headlines accuracy values are not directly comparable to the others, as first, the studies often employ different sets of data, different feature sets, and different train-test procedures. Second, maximizing accuracy can mask false negatives, which can be particularly expensive if a malicious site is identified as legitimate. Third, high performing models can be black boxes that offer limited reasoning behind individual predictions (Almujahid et al., 2024; Kocyigit et al., 2024). The research problem is thus to develop a phishing-detection system that gives high and fairly evaluated performance and at the same time interpretable evidence of the features that accounted for the individual decisions.

1.4 Research Gap

URL classification has made significant advancements as demonstrated in the literature, however, fewer works combine URL classification, systematic explainability and ensemble learning and deep learning in a single compact framework. Some works focus on URL features, some focus on content features, and others focus on complicated architectures that might not be scalable to a smaller, more interpretable subset that would achieve similar performance. Recent XAI research has highlighted that explanation should not just be a final visualization, but also aid in feature selection and model validation (Fan et al., 2024; Alsuyayh et al., 2025; Nauman et al., 2025). This proposal does just that by incorporating URL and lightweight contents features, four supervised models under a single protocol, feature reduction and interpreting the final model globally and locally.

1.5 Research Questions

RQ1: What is the accuracy of a supervised machine-learning model in identifying phishing websites by using a combination of URL features and content-based features?

Which of Logistic Regression, Decision Tree, Random Forest, and XGBoost gives the best scores for Precision, Recall, F1-Score and ROC-AUC?

RQ3: What features are most likely to distinguish phishing websites from legitimate websites?

How can the SHAP and LIME methods enhance the understandability of phishing-detection judgements?

Is it possible to streamline a feature set while still maintaining competitive performance, and to make it more easily understood and more interpretable?

1.6 Research Objectives

General goal: Design and test an interpretable machine learning model for detecting phishing websites based on URL and web content characteristics.

Identifying relevant lexical, structural and content-based phishing indicators.

To preprocess the dataset using leakage-aware cleaning, transformation and class-distribution analysis.

To train and evaluate four supervised machine learning classifiers following same experiment design.

To assess the classifiers with other performance measures than accuracy.

To detect and evaluate a small number of features of value.

- To provide interpretation for the chosen model on both the global and local level through the use of SHAP and LIME.

1.7 Research Scope and Delimitations

The study will involve the binary classification of phishing which will be developed based on a public labeled dataset. It will highlight on URL, Lightweight content features and perform comparisons between Logistic Regression, Decision Tree, Random Forest and XGBoost. Voice/SMS phishing and Malware analysis and production deployment are not included. Results will be restricted to the chosen dataset and time range and don't represent universal zero-day detection.

1.8 Research Significance

The study is the introduction of a reproducible pipeline connecting feature quality, fair model comparison, and XAI. In practice, interpretable alerts can provide analysts with explanations as to why a page was flagged, aiding in accurate automation and human review in cybersecurity operations (Achuthan et al., 2024; Salem et al., 2024).

Chapter 2: Literature Review

2.1 Phishing as an Evolving Cybersecurity Threat

Phishing is the newest variant of the cybersecurity threat. Phishing is the latest form of cyber security threat. Phishing is a hybrid of technical and social engineering. A malicious page can be disguised as a trusted brand, contain fraudulent domain names, or cause the user to be redirected via multiple URLs to a credential page. These patterns are dynamic and phishing detection should be considered a classification problem rather than a static list-matching one. The trend towards machine-learning, deep-learning, and hybrid approaches is evident from the review papers by Li et al. (2024) and Popescul and Radu (2025), which both move away from static rules.

2.2 Traditional Detection Approaches

Blacklists can be effective for known threats, but will not work until a new domain is reported. While heuristic rules can detect suspicious traffic (e.g., IP addresses in URLs, too many subdomains, too many redirects), there may be false positives with fixed thresholds for legitimate websites these days. These constraints encourage learning-based methods over using a single manually written rule (Kapan & Sora Gunal, 2023).

2.3 Machine Learning for Phishing Detection

The supervised phishing detection is often implemented by Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, and Boosting. Alnemari and Alshammari (2023) and Almujaheed et al. (2024) agree on the comparative evaluation of algorithms. Therefore, the present study will use the same partition of the data for training the selected models and the same validation strategy and validation metrics for testing the models to determine the best fit.

2.4 URL-Based Features

URL features are appealing because they can be stripped away prior to a web page rendering. Common ones are: URL length; the presence of dots and hyphens; subdomains depth; the presence of special characters; the digit ratio; path length; the presence of an IP address; suspicious lexical tokens. Tamal et al. (2024) demonstrated that well-chosen URL features can help with successful supervised classification, and Haq et al. (2024) validated the usefulness of URL representation

for deep-learning detection. This efficiency will be maintained in the present study and test if content information is a measurable value.

2.5 Content-Based Features

Content-based features include data that can't be seen in the URL. These may include login forms, password fields, external-resource ratios, usage of iframes, redirects, suspicious scripts and characteristics of the document-object-model. Opara et al. (2024) showed the importance of raw URL and HTML features, while Murhej and Nallasivan (2025) delved into more content and component features. The results prove the necessity of a combined feature strategy, provided that the extraction of the content is carried out safely, without the need to run malicious code.

2.6 Feature Selection and Data Quality

Feature selection impacts the accuracy, computational expense, overfitting, and interpretation. Even in a large set of predictors, redundant or poor predictors can introduce noise. Kocyigit et al. (2024) proved that using reduced feature space is still competitive in phishing URL detection. Data quality also plays a key role: Duplicates can seep in between training and testing, missing data can skew models, and class imbalance can lead to misleading accuracy results. Hence, preprocessing and feature selection should be done only on the training data in a controlled pipeline.

2.7 Deep and Hybrid Learning

With deep architectures, representations can be learned directly from strings or page content. Alshingiti et al. (2023) tested the CNN, LSTM, and hybrid LSTM-CNN models, and Ujah-Ogbuagu et al. (2024) introduced a hybrid deep-learning approach for detecting spoofed-URLs in real-time. While the benefits of deep models are well-documented, as highlighted in systematic reviews, they also have higher computational demands and interpretability challenges (Kyaw et al., 2024). Because this proposal prioritizes transparent comparison and practical reproducibility, deep models are reviewed as state-of-the-art context rather than included in the main experiment.

2.8 Explainable Artificial Intelligence in Cybersecurity

XAI strives to make the behavior of a model comprehensible to humans. SHAP explains the contribution of each feature toward a prediction, and provides explanations both globally and at the level of the case. LIME aims to approximate a complex model around one observation and shows the features that led to this local

decision. Explainability can be used to build trust, error analysis and feature validation beyond the production of visual output, as demonstrated by Fan et al. (2024) and Nauman et al. (2025).

2.9 Critical Synthesis of Previous Studies

Study	Feature Focus	Method	Key Contribution	Research Opportunity
Kapan & Sora Gunal (2023)	Multiple	ML comparison	Classifier/feature evaluation	Add stronger XAI
Kocyigit et al. (2024)	URL	Feature selection + ML	Reduced feature space	Link selection to explanations
Tamal et al. (2024)	URL	Multiple supervised models	Large URL-feature analysis	Test content contribution
Opara et al. (2024)	URL + HTML	Deep model	Multi-source representation	Improve interpretability
Nauman et al. (2025)	Cybersecurity data	ML + XAI	Transparency-focused framework	Simplify for reproducibility

The synthesis reveals that there is a lack of focus on how predictive performance and interpretability are studied together. A more effective design would involve models compared within one protocol, testing whether features of the content have measurable value, and include explanations to explore influential features and classification errors.

2.10 Research Gap Derived from the Literature

The missing link is the absence of compact studies that can be reproduced with mixed features (URL/content), controlled feature reduction, multi-metric comparison, and global/local explanations. In this study, whether an interpretable reduced feature set can achieve full-model performance (Alsuqayh et al., 2025; Nauman et al., 2025) is tested.

2.11 Literature-to-Objective Alignment

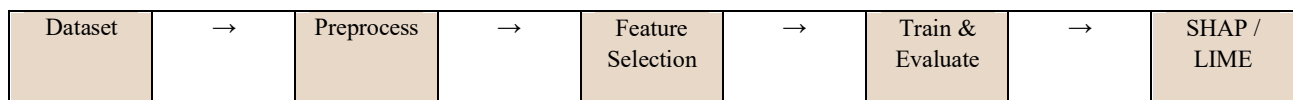
Prior studies directly support the objectives: feature identification (Tamal et al., 2024; Opara et al., 2024), classifier comparison (Almujahid et al., 2024), and feature reduction/XAI (Kocyigit et al., 2024; Nauman et al., 2025).

Chapter 3: Research Methodology

3.1 Research Design

The study will adopt a quantitative, experimental and comparative research design. Each observation will be a tagged URL or website. The preprocessing and evaluation methods will be standard across all models to ensure that any difference in performance is due to the algorithms and not due to different data handling. Inside the training partition, cross validation will be employed; an independent test partition will be left completely aside until final evaluation.

3.2 Proposed Research Framework



The framework spans from data quality control to feature engineering, model comparison, final testing and explanation. Features will be compared with a minimal feature set and, where possible, a URL only baseline during an ablation step. This design directly connects the research questions to measurable outputs of the experiment. Construct a data source and sample it. Build a data source and sample it.

3.3 Data Source and Sampling

An open labeled dataset of phishing and legitimate websites which has been recently released will be chosen, preferably published in peer-reviewed research. The data will be checked for provenance, class definitions, duplicate records, missing values, and availability of features. If the data set is too large to fit into computation resources, then the class proportions will be maintained using stratified sampling and the random seed will be stored. No personal credentials will be taken.

3.4 Data Preprocessing

- To minimize leakage, eliminate duplicates prior to splitting train-test sets.
- Estimate missing data and impute or remove justifiably.
- Apply transformations to categorical variables, using only those fitted on the training set.
- Scale numeric variables as necessary, especially for Logistic Regression.
- Check class balance and apply class weighting and/or training-only resampling if needed.
- Feature Engineering and Selection

3.5 Feature Engineering and Selection

Lexical and structural features like length, subdomain depth, digit ratio, special characters, path structure, and HTTPS usage along with suspicious tokens will be included in the URL variables. Selected indicators will be used as content variables, including forms, password fields, external links, redirects, iframes, and characteristics of the script itself. Highly redundant and very close to zero variance variables will be eliminated first. Mutual information or tree-based importance will then rank the features and later, SHAP results will serve as a separate interpretive check. Both full and reduced feature sets will be compared with regard to performance (Kocyigit et al., 2024).

3.6 Model Development

Model	Role	Strength	Main Limitation
Logistic Regression	Interpretable baseline	Fast and probabilistic	Limited non-linearity
Decision Tree	Rule-based baseline	Human-readable splits	Overfitting risk
Random Forest	Bagged ensemble	Robust non-linear modeling	Less transparent
XGBoost	Boosted ensemble	Strong predictive performance	Requires tuning

3.7 Experimental Procedure

The data will be split into 80:20 train-test stratified data split. To support the model comparison and tuning, 5-fold stratified cross validation inside the training data set will be used. Grid Search or Randomized Search will employ documented parameter range. All pre-processing and feature-selection processes will be done within the training pipeline, meaning that the test set will not affect the model construction. The configuration selected will then be tested once on the held-out test data.

3.8 Evaluation Metrics

The accuracy will provide an overall ratio of correctness; the precision will display the ratio of phishing cases that are correctly predicted; the recall will display the ratio of phishing cases that are detected and the F1-score will provide a balance of precision and recall. ROC-AUC will be used to analyze discrimination over thresholds and the confusion matrix will reveal false-positive and false-negative patterns. The final model will be chosen based on a balanced evaluation of these factors, and not just on accuracy, given that missed phishing sites can come at a price (Kocyigit et al., 2024), with special focus on recall.

3.9 Explainability Analysis

SHAP will be used to rank the importance of global features and explain individual predictions of the best model. A smaller, representative, sample of correct and incorrect classifications will use LIME for another explanation perspective locally. Explanations will be analyzed for its plausibility and consistency of the security. This will be particularly paid attention to when it comes to false positives and false negatives, as the explanations can provide insights into whether the model is capturing meaningful indicators of attacks, or artefacts specific to the acquired data (Fan et al., 2024; Nauman et al., 2025).

3.10 Tools and Technologies

The following tools will be used for implementation: Python 3.x, Jupyter Notebook, Pandas, NumPy, Scikit-learn, XGBoost, SHAP, LIME and Matplotlib. Versioning, random seeds, selected parameters and preprocessing decisions will be documented in order to be able to reproduce the experiment.

3.11 Validity, Reliability, and Reproducibility

To support internal validity, data leakage will be avoided, all classifiers will be evaluated in the same way and the test set will be independent. Cross validation will be used to evaluate the reliability. Fixed random seeds, saved settings and library versions will ensure reproducibility. The extent to which conclusions can be drawn beyond the study context will be constrained by the dataset and timeframe chosen (Li et al., 2024).

3.12 Ethical Considerations

No phishing campaigns against human subjects will be done with the project, and public research data will be used. It won't steal actual credentials, nor ask users to visit malicious websites. For content-based features, the study will make use of existing features, safe feature extraction methods and not run potentially harmful scripts. Data and sources will be documented, with emphasis on data licensing.

3.13 Expected Outcomes

- Reproducibility of a comparison of four supervised phishing-detection models.

Evidence of the effectiveness of combining URL/content features on the URL-only baseline.

A smaller feature set, with good performance where feasible.

A precision and recall based model selection using F1 score, ROC-AUC and confusion matrix evidences.

Global and local explanations involving influential phishing indicators and the explanation of model errors.

- A framework that has been documented that can be extended later to a browser or security-operations decision-support tool.

3.14 Research Milestones and Gantt Chart

Activity	W1	W2	W3	W4	W5	W6	W7	W8	Output
Topic & questions	■								Approved plan
Literature review	■	■	■						Synthesis matrix
Gap & objectives		■	■						Chapter 1 draft
Methodology			■	■					Experiment plan
Data preparation				■	■				Clean dataset
Model development					■	■			Initial results
Evaluation & XAI						■	■		Analysis
Writing & revision						■	■	■	Final proposal

3.15 Chapter Summary

The methodology translates every research question into a particular empirical task. Model comparison is about the effectiveness of detection and algorithm choice, while feature selection or feature ablation is about feature contribution and reduction, and SHAP/LIME is about interpretability. The design thus sought a direct link between the problem statement and literature gap and the objectives, methods, expected evidence, and final conclusions.

References

APA 7th edition. References are unnumbered, with a direct open-access/full-text link beneath each entry.

- Kapan, S., & Sora Gunal, E. (2023). Improved phishing attack detection with machine learning: A comprehensive evaluation of classifiers and features. *Applied Sciences*, 13(24), 13269. <https://doi.org/10.3390/app132413269>
<https://www.mdpi.com/2076-3417/13/24/13269/pdf>
- Alnemari, S., & Alshammari, M. (2023). Detecting phishing domains using machine learning. *Applied Sciences*, 13(8), 4649. <https://doi.org/10.3390/app13084649>
<https://www.mdpi.com/2076-3417/13/8/4649/pdf>
- Thakur, K., Ali, M. L., Obaidat, M. A., & Kamruzzaman, A. (2023). A systematic review on deep-learning-based phishing email detection. *Electronics*, 12(21), 4545. <https://doi.org/10.3390/electronics12214545>
<https://www.mdpi.com/2079-9292/12/21/4545/pdf>
- Atawneh, S., & Aljehani, H. (2023). Phishing email detection model using deep learning. *Electronics*, 12(20), 4261. <https://doi.org/10.3390/electronics12204261>
<https://www.mdpi.com/2079-9292/12/20/4261/pdf>
- Alshingiti, Z., Alaqel, R., Al-Muhtadi, J., Haq, Q. E. U., Saleem, K., & Faheem, M. H. (2023). A deep learning-based phishing detection system using CNN, LSTM, and LSTM-CNN. *Electronics*, 12(1), 232. <https://doi.org/10.3390/electronics12010232>
<https://www.mdpi.com/2079-9292/12/1/232/pdf>
- Prabakaran, M. K., Meenakshi Sundaram, P., & Chandrasekar, A. D. (2023). An enhanced deep learning-based phishing detection mechanism to effectively identify malicious URLs using variational autoencoders. *IET Information Security*, 17(3), 423–440. <https://doi.org/10.1049/ise2.12106>
<https://ietresearch.onlinelibrary.wiley.com/doi/pdfdirect/10.1049/ise2.12106>
- Kocyigit, E., Korkmaz, M., Sahingoz, O. K., & Diri, B. (2024). Enhanced feature selection using genetic algorithm for machine-learning-based phishing URL detection. *Applied Sciences*, 14(14), 6081. <https://doi.org/10.3390/app14146081>
<https://www.mdpi.com/2076-3417/14/14/6081/pdf>
- Haq, Q. E. U., Faheem, M. H., & Ahmad, I. (2024). Detecting phishing URLs based on a deep learning approach to prevent cyber-attacks. *Applied Sciences*, 14(22), 10086. <https://doi.org/10.3390/app142210086>
<https://www.mdpi.com/2076-3417/14/22/10086/pdf>
- Tamal, M. A., Islam, M. K., Bhuiyan, T., Sattar, A., & Prince, N. U. (2024). Unveiling suspicious phishing attacks: Enhancing detection with an optimal feature vectorization algorithm and supervised machine learning. *Frontiers in Computer Science*, 6, 1428013. <https://doi.org/10.3389/fcomp.2024.1428013>
<https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2024.1428013/pdf>

- Almujahid, N. F., Haq, M. A., & Alshehri, M. (2024). Comparative evaluation of machine learning algorithms for phishing site detection. *PeerJ Computer Science*, 10, e2131.
<https://doi.org/10.7717/peerj-cs.2131>
<https://peerj.com/articles/cs-2131.pdf>
- Ujah-Ogbuagu, B. C., Akande, O. N., & Ogbuju, E. (2024). A hybrid deep learning technique for spoofing website URL detection in real-time applications. *Journal of Electrical Systems and Information Technology*, 11, 7. <https://doi.org/10.1186/s43067-023-00128-8>
<https://link.springer.com/content/pdf/10.1186/s43067-023-00128-8.pdf>
- Kyaw, P. H., Aung, Y., & Tun, T. (2024). A systematic review of deep learning techniques for phishing detection. *Electronics*, 13(19), 3823. <https://doi.org/10.3390/electronics13193823>
<https://www.mdpi.com/2079-9292/13/19/3823/pdf>
- Meléndez, R., Ptaszynski, M., & Masui, F. (2024). Comparative investigation of traditional machine-learning models and transformer models for phishing email detection. *Electronics*, 13(24), 4877. <https://doi.org/10.3390/electronics13244877>
<https://www.mdpi.com/2079-9292/13/24/4877/pdf>
- Fan, Z., Li, W., Laskey, K. B., & Chang, K.-C. (2024). Investigation of phishing susceptibility with explainable artificial intelligence. *Future Internet*, 16(1), 31.
<https://doi.org/10.3390/fi16010031>
<https://www.mdpi.com/1999-5903/16/1/31/pdf>
- Achuthan, K., Ramanathan, S., Srinivas, S., & Raman, R. (2024). Advancing cybersecurity and privacy with artificial intelligence: Current trends and future research directions. *Frontiers in Big Data*, 7, 1497535. <https://doi.org/10.3389/fdata.2024.1497535>
<https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2024.1497535/pdf>
- Salem, A. H., Azzam, S. M., Emam, O. E., & Abohany, A. A. (2024). Advancing cybersecurity: A comprehensive review of AI-driven detection techniques. *Journal of Big Data*, 11, 105.
<https://doi.org/10.1186/s40537-024-00957-y>
<https://link.springer.com/content/pdf/10.1186/s40537-024-00957-y.pdf>
- Li, W., Manickam, S., Chong, Y.-W., Leng, W., & Nanda, P. (2024). A state-of-the-art review on phishing website detection techniques. *IEEE Access*, 12, 187976–188012.
<https://doi.org/10.1109/ACCESS.2024.3514972>
<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10788671>
- Jayaprakash, R., Natarajan, K., Daniel, J. A., Chinnappan, C. V., Giri, J., Qin, H., & Mallik, S. (2024). Heuristic machine learning approaches for identifying phishing threats across web and email platforms. *Frontiers in Artificial Intelligence*, 7, 1414122.
<https://doi.org/10.3389/frai.2024.1414122>
<https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1414122/pdf>
- Popescul, D., & Radu, L. D. (2025). AI in phishing detection: A bibliometric review. *Frontiers in Artificial Intelligence*, 8, 1496580. <https://doi.org/10.3389/frai.2025.1496580>
<https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1496580/pdf>

- Alhuzali, A., Alloqmani, A., Aljabri, M., & Alharbi, F. (2025). In-depth analysis of phishing email detection: Evaluating the performance of machine learning and deep learning models across multiple datasets. *Applied Sciences*, 15(6), 3396. <https://doi.org/10.3390/app15063396>
<https://www.mdpi.com/2076-3417/15/6/3396/pdf>
- Türk, F., & Kılıçaslan, M. (2025). Malicious URL detection with advanced machine learning and optimization-supported deep learning models. *Applied Sciences*, 15(18), 10090. <https://doi.org/10.3390/app151810090>
<https://www.mdpi.com/2076-3417/15/18/10090/pdf>
- Murhej, M., & Nallasivan, G. (2025). Component features based enhanced phishing website detection system using EfficientNet, FH-BERT, and SELU-CRNN methods. *Frontiers in Computer Science*, 7, 1582206. <https://doi.org/10.3389/fcomp.2025.1582206>
<https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2025.1582206/pdf>
- Nauman, M., Usman Akhtar, H. M., Gorbani, H., Hadi Ul Hassan, M., & Fayyaz, M. A. B. (2025). Transparent and trustworthy cybersecurity: An XAI-integrated big data framework for phishing attack detection. *Frontiers in Big Data*, 8, 1688091. <https://doi.org/10.3389/fdata.2025.1688091>
<https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2025.1688091/pdf>
- Wilk-Jakubowski, J. L. (2025). Machine learning and neural networks for phishing detection: A systematic review (2017–2024). *Electronics*, 14(18), 3744. <https://doi.org/10.3390/electronics14183744>
<https://www.mdpi.com/2079-9292/14/18/3744/pdf>
- Alsakarnah, R., Masoud, M. Z., & Ghababsheh, A. (2026). Hybridizing explainable AI (XAI) for intelligent feature extraction in phishing website detection. *Electronics*, 15(2), 350. <https://doi.org/10.3390/electronics15020350>
<https://www.mdpi.com/2079-9292/15/2/350/pdf>