

Artificial Intelligence and Machine Learning for Intelligent Phishing Website Detection - A Comprehensive Analysis and Future Research Directions

Tejas Ravindra Bait¹, Aarti Sanjay Gawai²

¹Students, Department of Information Technology and Computer Science, D.G. Ruparel College, Mumbai, Maharashtra, India

²Assistant Professor, Department of Information Technology and Computer Science, D.G. Ruparel College, Mumbai, Maharashtra, India

¹tejasbait7@gmail.com

²aartigawai1611@gmail.com

Abstract - Phishing websites remain a persistent cybersecurity threat, as they imitate genuine online platforms to steal credentials, financial details, and other sensitive data. The fast growth of digital banking, e-commerce, cloud services, and social networking has widened the scope for such attacks. Techniques like URL obfuscation, typosquatting, homograph attacks, brand impersonation, malicious redirection, and AI-assisted webpage generation further make phishing sites hard to catch using conventional methods. This paper presents a structured review of Artificial Intelligence (AI) and Machine Learning (ML) approaches for phishing website detection, focusing on studies published between 2017 and 2025. It covers traditional machine learning, ensemble learning, deep learning, hybrid models, feature engineering, and Explainable AI (XAI). It also analyses common feature categories, datasets, evaluation practices, and deployment challenges reported in the literature. The reviewed work shows a clear shift from blacklist- and rule-based detection toward intelligent, multimodal, adaptive, and explainable systems. However, several limitations remain, including limited dataset diversity, lack of standard evaluation, weak real-time detection, poor interpretability, high computational cost, and low generalization to new attacks and AI-generated or adversarial phishing sites. Based on these findings, the paper highlights research gaps and suggests future directions for building accurate, adaptive, explainable, scalable, and real-time phishing detection systems.

Key Words:

Phishing Website Detection, Artificial Intelligence, Machine Learning, Explainable Artificial Intelligence, Cybersecurity, Feature Engineering.

1. INTRODUCTION

The rapid development of Information and Communication Technology (ICT) has transformed the way individuals, organisations, financial institutions, and governments communicate and provide digital services. Cloud computing, digital banking, e-commerce, social networking, online

education, and other web-based services have increased dependence on Internet-based platforms. This expansion has also created opportunities for cybercriminals to exploit users through deceptive online activities. Among these threats, phishing has emerged as a major form of cybercrime in which attackers imitate trusted organisations or legitimate websites to obtain sensitive information such as usernames, passwords, banking credentials, payment information, and personal data. Phishing websites are particularly challenging because attackers continuously modify their appearance and technical characteristics. Common techniques include domain spoofing, typosquatting, homograph attacks, URL obfuscation, brand impersonation, cloned login pages, malicious redirection, shortened URLs, and misuse of HTTPS. Recent developments in Artificial Intelligence (AI) and Generative AI have further increased the ability of attackers to produce realistic webpage content and sophisticated impersonation techniques. These developments make the detection of malicious websites increasingly difficult using conventional security mechanisms. Traditional phishing detection systems have primarily relied on blacklist databases, predefined rules, and heuristic analysis. Blacklist-based systems compare a website or URL against previously identified malicious resources, whereas rule-based and heuristic systems identify predefined suspicious characteristics. Although these approaches are relatively efficient, they have limited effectiveness against newly created or zero-day phishing websites that have not yet been identified. The short lifespan and continuous modification of phishing websites also make maintaining updated detection rules and blacklists challenging.

Artificial Intelligence and Machine Learning have therefore become important approaches for phishing website detection. Unlike static rule-based mechanisms, ML models can learn patterns from historical data and classify websites according to multiple characteristics. These characteristics may include lexical URL features, domain information, HTML and JavaScript structures, DNS information, SSL attributes, webpage content, and visual similarity. Algorithms such as Decision Tree, Random Forest, Support Vector Machine (SVM), Logistic Regression, Naïve Bayes, and K-Nearest Neighbours have been widely investigated, while ensemble methods such as Gradient Boosting, AdaBoost, and XGBoost combine multiple learning mechanisms to improve classification performance. Deep learning models, including

Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks, provide additional capabilities for learning complex representations from website and URL information.

Despite these developments, several challenges remain. Existing studies use different datasets, preprocessing techniques, feature extraction methods, algorithms, validation procedures, and evaluation metrics. These differences make direct comparison between approaches difficult. Furthermore, many studies depend on static benchmark datasets that may not adequately represent newly emerging phishing campaigns. Practical deployment challenges involving computational efficiency, inference latency, scalability, real-time detection, browser integration, cloud and mobile environments, explainability, and robustness against adversarial attacks also require further investigation.

The growing volume of research in AI and ML-based phishing detection creates a need for a consolidated analysis of existing approaches. A review can help identify common methodologies, compare different feature categories and learning techniques, examine limitations, and determine areas requiring further investigation. This is particularly relevant as phishing techniques continue to evolve and attackers increasingly use sophisticated webpage generation and impersonation methods.

Accordingly, this paper reviews AI and ML-based phishing website detection research published between 2017 and 2025. The review covers traditional machine learning, ensemble learning, deep learning, hybrid approaches, feature engineering, Explainable Artificial Intelligence, datasets, evaluation methodologies, and deployment considerations. It further identifies research gaps and discusses future directions for developing more accurate, adaptive, explainable, scalable, and real-time phishing website detection systems.

The main contributions of this review are to analyse recent AI and ML-based phishing website detection approaches, examine commonly used feature categories and datasets, compare traditional, ensemble, deep learning, hybrid, and XAI-based approaches, identify limitations in existing research, and discuss future research opportunities. The remainder of this paper presents the literature review, comparative analysis, research gap, problem statement, research objectives, conceptual framework, methodology, expected outcomes, future research directions, and conclusion.

2. LITERATURE REVIEW

Phishing website detection has evolved from conventional blacklist and rule-based approaches toward machine learning, deep learning, ensemble, hybrid, and explainable AI-based techniques. The reviewed literature demonstrates that different approaches rely on different combinations of URL, host, content, visual, behavioural, and semantic features.

2.1 Traditional Machine Learning Approaches

Initial work on AI-driven phishing detection relied mainly on handcrafted features drawn from URLs and web page content, which were then fed to supervised classifiers that labelled each site as legitimate or malicious. A wide range of classical models has been tested for this task, including probabilistic classifiers such as Naïve Bayes, linear models such as Logistic

Regression, distance-based methods such as K-Nearest Neighbours, margin-based Support Vector Machines, and rule-based Decision Trees [1], [4], [6]. Ensemble methods, with Random Forest as a prominent example, have drawn particular interest because they aggregate the outputs of many individual trees and cope well with high-dimensional feature sets. Boosting techniques, among them AdaBoost, Gradient Boosting, and XGBoost, have also been explored. In every case, results depend heavily on how representative the training data is, how features are selected and preprocessed, and how the evaluation is designed. A practical advantage of these models is that they demand far less computing power than most deep learning architectures and can produce predictions quickly once suitable features are available. Their main weakness is reduced accuracy on phishing sites whose characteristics were absent or under-represented in the training set.

2.2 URL and Feature-Based Detection

URL analysis is one of the most commonly investigated approaches in phishing website detection. Lexical URL features may include URL length, number of special characters, subdomains, suspicious keywords, entropy, presence of an IP address, unusual symbols, and domain-related characteristics [3], [10].

Host-based features provide additional information about the domain and hosting environment. These may include domain age, DNS information, WHOIS information, SSL certificate characteristics, hosting information, and server reputation. Such information can provide useful evidence when determining whether a website is suspicious.

Content-based features examine webpage characteristics such as HTML structure, JavaScript behaviour, forms, iframes, links, redirects, and Document Object Model (DOM) information. These features can provide information beyond the URL itself and can help identify cloned or malicious webpages.

The literature suggests that no single feature category is sufficient for detecting every phishing scenario. Combining multiple categories can provide more comprehensive information, although feature extraction and processing may increase computational requirements.

2.3 Deep Learning Approaches

Deep learning has expanded phishing detection research by enabling models to learn complex feature representations. CNN-based approaches have been investigated for analysing webpage screenshots, visual layouts, and other image-based characteristics. Such approaches can be useful for identifying visual similarities between legitimate webpages and fraudulent copies [8].

RNN and LSTM architectures can process sequential information and have been investigated for URL and webpage-text analysis. Hybrid CNN-LSTM architectures combine visual and sequential information and may provide a broader representation of phishing characteristics.

Deep learning approaches can reduce dependence on manually designed features in some applications. However, they generally require appropriate datasets, greater computational resources, longer training processes, and careful model evaluation. Their black-box nature can also create challenges

when cybersecurity analysts need to understand why a particular website was classified as malicious.

2.4 Ensemble and Hybrid Approaches

Ensemble learning combines multiple classifiers or learning strategies to improve prediction robustness. Techniques including Random Forest, boosting algorithms, voting classifiers, and stacking approaches have been investigated for phishing detection [6], [11].

Hybrid approaches combine different feature categories or learning architectures. For example, a system may combine URL information with domain, HTML, textual, or visual characteristics. Similarly, hybrid deep learning models may combine CNN and LSTM architectures to analyse different forms of website information.

The primary advantage of these approaches is that multiple sources of evidence can be considered simultaneously. However, increased model complexity may result in higher computational requirements and more complicated deployment processes.

2.5 Explainable Artificial Intelligence

Explainable Artificial Intelligence has become an important research direction because high-performing AI models can be difficult for cybersecurity professionals to interpret. Techniques such as SHAP and LIME can provide information about the contribution of individual features to a model's prediction.

In phishing detection, explanations may identify the influence of URL characteristics, domain information, SSL attributes, webpage structure, or other features. Such explanations can support analyst decision-making, help investigate false predictions, and improve confidence in AI-based security systems.

However, XAI techniques can introduce additional computational overhead. The consistency and reliability of explanations also require careful evaluation. Therefore, explainability should not be considered only as an additional feature; it should be evaluated in relation to model performance, computational requirements, and practical usability.

2.6 Feature Engineering in Phishing Detection

Feature engineering remains an important component of many phishing detection systems. Common feature categories include lexical, host-based, content-based, visual, behavioural, and semantic features.

Lexical features are generally efficient to obtain and can support rapid URL-level classification. Host-based features provide information about domains and their infrastructure. Content-based features analyse HTML, JavaScript, forms, links, redirects, and webpage structures. Visual features examine screenshots, logos, layouts, typography, and similarity between webpages. Semantic features can analyse textual characteristics and webpage meaning.

The literature indicates that combining multiple feature categories can improve the representation of sophisticated phishing websites. However, such integration can increase feature extraction complexity and computational requirements.

A structured taxonomy summarizing these primary feature categories is illustrated in Fig. 1.

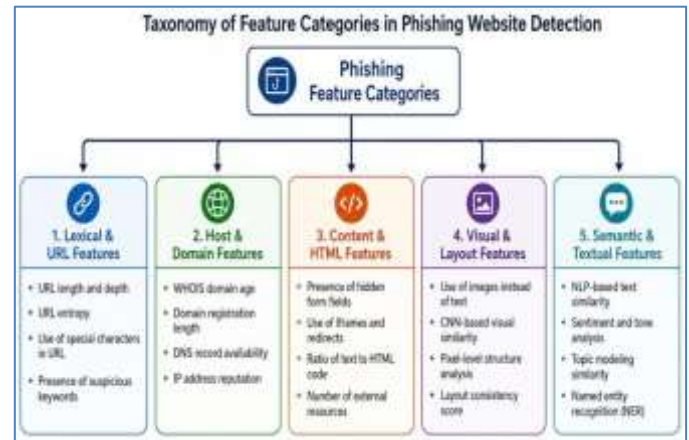


Fig -1: Taxonomy of Feature Categories Used in Intelligent Phishing Website Detection.

2.7 Critical Analysis of Existing Literature

The reviewed studies demonstrate a clear progression from traditional machine learning toward ensemble, deep learning, hybrid, and explainable AI approaches. Traditional machine learning remains useful because of its relatively low computational requirements and fast classification capabilities. Ensemble methods provide improved robustness by combining multiple classifiers, while deep learning provides greater capability for automatically learning complex representations. However, direct comparison between studies remains difficult because different researchers use different datasets, feature extraction procedures, preprocessing techniques, validation strategies, and performance measures.

Accuracy is frequently reported, whereas operational metrics such as inference latency, computational cost, scalability, false positive rate, zero-day robustness, and explainability are less consistently evaluated.

Another major limitation is the dependence on static benchmark datasets. Phishing websites continuously change their appearance, domain characteristics, content, and technical behaviour. Consequently, a model trained on historical data may not perform consistently against future phishing campaigns. These limitations indicate the need for diverse datasets, standardised evaluation procedures, adaptive learning, and real-world deployment studies.

3. COMPARATIVE ANALYSIS OF EXISTING APPROACHES

The reviewed literature covers a wide range of machine learning, deep learning, ensemble, and hybrid approaches. Table 1 summarises the main characteristics reported in the available literature.

Table -1: Comparative Analysis of Reviewed Phishing Website Detection Studies

Ref	Approach	Main Features / Data	Key Observation
[1]	Machine Learning	URL / website attributes	Effective for distinguishing phishing and legitimate sites.
[3]	URL & Domain Analysis	URL & domain metrics	Domain traits provide strong cues for detection.
[6]	Comparative ML	Multi-feature sets	Performance varies by feature selection and dataset.
[8]	CNN + BiLSTM	Visual & sequential data	Combines spatial layout and sequence analysis.
[10]	Deep Learning	Lexical URL features	DL can process raw URL strings directly.
[11]	Stacking Ensemble	Multi-source features	Combines weak models to improve classification robustness.
[12]	ML Framework	Phishing dataset features	Dataset diversity heavily impacts detection accuracy.
[15]	ML Review	Feature categories	Reviews current progress and key ML drawbacks.
[18]	Systematic Review	ML & Neural Networks	Maps open challenges and emerging research needs.
[19]	Large Benchmark	Real-world web data	Offers an updated dataset for benchmarking models.

The comparison indicates that traditional ML methods are suitable for feature-based classification and relatively efficient applications, while ensemble approaches can improve robustness by combining multiple models. Deep learning and hybrid approaches provide greater flexibility for processing complex visual, sequential, and multimodal information but generally require higher computational resources.

The reviewed studies also demonstrate that URL, domain, host, content, and visual features provide different types of information for phishing detection. Combining multiple feature categories can improve the representation of sophisticated phishing websites; however, it may also increase feature extraction complexity and computational requirements.

Dataset characteristics are another important factor affecting reported performance. Differences in dataset size, source, feature availability, class distribution, and collection period make direct comparison between studies difficult. A model performing well on a static benchmark dataset may not necessarily maintain the same performance against newly emerging phishing websites.

Overall, the comparison shows that no single approach is universally optimal. Future phishing detection systems should consider not only classification performance but also

generalisation, computational efficiency, real-time detection, scalability, and explainability. Standardised datasets and evaluation metrics are therefore required to enable more meaningful comparison of existing and future phishing detection approaches.

4. RESEARCH GAP

4.1 Existing Challenges

Phishing website detection remains challenging because attackers continuously modify domain names, URLs, webpage content, visual appearance, and other characteristics. Techniques such as brand impersonation, typosquatting, homograph attacks, URL obfuscation, malicious redirects, cloned interfaces, and AI-assisted webpage generation can reduce the effectiveness of conventional detection mechanisms.

AI and ML models improve detection by analysing multiple website characteristics. Nevertheless, how well these models perform is largely shaped by the training data itself particularly its diversity, currency, and overall quality. Static and imbalanced datasets may not adequately represent newly emerging phishing websites, multilingual content, AI-generated webpages, or sophisticated brand impersonation. Practical deployment is another challenge. Complex feature extraction and advanced models may increase computational requirements and detection latency. Therefore, an effective real-world solution must balance accuracy with computational efficiency, scalability, adaptability, interpretability, and response time.

4.2 Limitations of Current Research

Existing studies frequently emphasise classification performance while giving comparatively less attention to practical deployment requirements. Different datasets, preprocessing methods, feature sets, algorithms, and evaluation procedures make direct comparison difficult.

Deep learning and ensemble models can provide strong predictive capabilities but may operate as black-box systems. XAI techniques can improve interpretability, but their computational cost and reliability require additional investigation.

Another limitation is the reliance on offline experiments using static datasets. Comparatively fewer studies examine continuous model updating, concept drift, real-time detection, browser integration, cloud and mobile deployment, and long-term operational performance.

4.3 Identified Research Gap

The literature indicates several areas requiring further investigation. First, standardised comparative evaluation remains limited because improvements may result from differences in datasets, features, preprocessing, or algorithms rather than the model itself.

Second, many studies focus on individual feature categories. Modern phishing websites may combine multiple deception techniques, suggesting a need for multidimensional feature

integration involving URL, domain, HTML, textual, visual, and other relevant characteristics.

Third, practical deployment requires greater attention to computational efficiency, inference latency, scalability, adaptability, and explainability. High experimental performance alone does not guarantee effective real-world operation.

Finally, emerging AI-assisted phishing techniques are not consistently represented in existing datasets and detection frameworks. Adaptive approaches capable of learning from evolving phishing behaviour therefore represent an important future research opportunity.

5. PROBLEM STATEMENT

Existing phishing website detection systems have progressed from blacklist and rule-based mechanisms to machine learning, deep learning, ensemble, hybrid, and explainable AI approaches. However, many existing approaches remain dependent on static datasets, manually selected features, inconsistent evaluation methodologies, and controlled experimental environments. These limitations can reduce their ability to generalise to newly emerging phishing websites and sophisticated attacks. The absence of standardised evaluation, adaptive learning, multimodal feature integration, practical deployment analysis, and reliable explainability creates a continuing challenge for real-world phishing website detection. Therefore, further research is required to identify and evaluate approaches capable of providing accurate, adaptive, explainable, computationally efficient, and scalable phishing website detection.

Furthermore, modern phishing attacks increasingly employ sophisticated evasion techniques, including dynamic URL redirection, automated webpage generation, and cloaking, which can bypass conventional detection methods. Although deep learning and ensemble approaches can improve detection performance, their computational requirements and model complexity may create challenges for real-time deployment. Additionally, the limited availability of continuously updated benchmark datasets containing emerging phishing campaigns can affect long-term model reliability and contribute to **concept drift** as attack patterns evolve.

Another important challenge is the limited integration of heterogeneous information sources within a unified detection process. URL and domain characteristics, webpage structure, HTML content, textual information, and visual features can provide complementary evidence for phishing classification. However, existing studies often focus on specific feature categories or individual techniques. Consequently, achieving a suitable balance between detection accuracy, computational efficiency, adaptability, scalability, and interpretability remains challenging. This research gap motivates the development of the **conceptual AI/ML-based phishing website detection framework** proposed in this study.

6. RESEARCH OBJECTIVES

6.1 Main Objective

To critically review and analyse Artificial Intelligence and Machine Learning approaches for phishing website detection and identify research gaps and future directions for developing an accurate, adaptive, explainable, and practical detection system.

6.2 Specific Objectives

1. To analyze ML, ensemble, deep learning, hybrid, and XAI approaches for phishing website detection.
2. To examine the major feature categories used for phishing website classification, including URL, domain, host, HTML, content, visual, and semantic features.
3. To compare existing detection approaches based on datasets, features, algorithms, evaluation methods, strengths, and limitations.
4. To identify limitations related to dataset quality, generalisation, computational efficiency, explainability, real-time detection, and practical deployment.
5. To examine emerging challenges associated with AI-assisted, adversarial, and continuously evolving phishing websites.
6. To identify future research directions for developing adaptive, multimodal, explainable, scalable, and real-time phishing website detection system

7. PROPOSED CONCEPTUAL FRAMEWORK

This study does not implement or experimentally evaluate a new phishing detection system. Instead, based on a comprehensive review of existing literature, a conceptual AI/ML-based phishing website detection framework is proposed to illustrate how diverse information sources and advanced learning techniques can be effectively integrated. Modern phishing attacks increasingly employ sophisticated evasion techniques, making single-source detection mechanisms insufficient for reliable classification. To address this limitation, the proposed conceptual framework synthesizes multi-modal data sources, ranging from structural URL and domain properties to visual and webpage content attributes, within a unified detection workflow. This multi-layered architecture aims to bridge the gap between traditional feature engineering and modern deep learning approaches by providing a structured blueprint for future experimental implementations. The overall architecture of the proposed conceptual framework is organized into a cohesive two-stage operational pipeline, as illustrated in **Fig. 2(a)** and **Fig. 2(b)**.

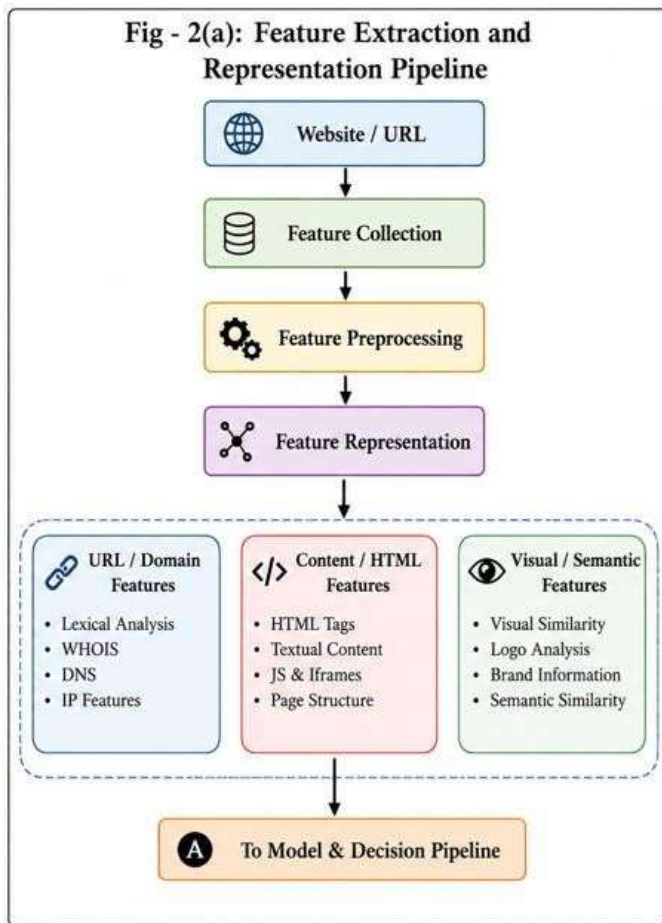


Fig -2(a): Feature Extraction and Representation Pipeline.

As shown in **Fig -2(a)**, the workflow begins with a website URL or webpage as input. The initial processing stage involves feature collection, feature preprocessing, and feature representation across three core categories: URL/domain attributes (lexical analysis, WHOIS details, DNS records, IP features), content/HTML characteristics (tags, textual content, scripts, iframes, page structure), and visual/semantic properties (visual similarity, logo analysis, brand information, semantic similarity).

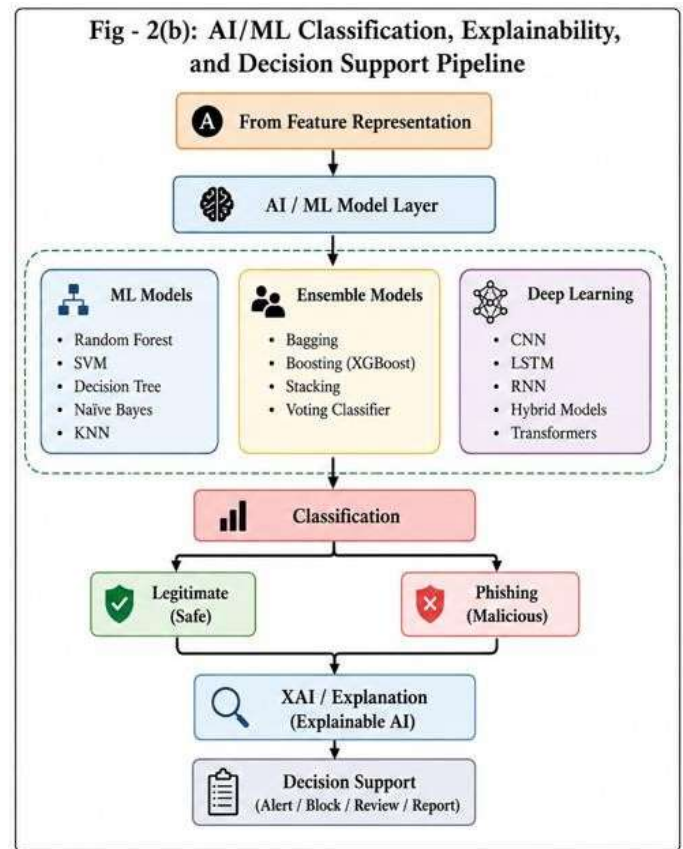


Fig -2(b): AI/ML Classification, Explainability, and Decision Support Pipeline.

As detailed in **Fig -2(b)**, the extracted feature representations are routed to the model layer, which encompasses traditional machine learning (e.g., Random Forest, SVM, Decision Trees, Naïve Bayes, KNN), ensemble methods (e.g., Bagging, XGBoost, Stacking, Voting Classifiers), and deep learning architectures (e.g., CNN, LSTM, RNN, Hybrid Models, Transformers) depending on data modalities. The classification stage outputs a prediction identifying whether the website is legitimate (safe) or phishing (malicious). An explainability layer (XAI) then interprets feature contributions to provide actionable decision support (alerts, blocking, manual review, or reporting) for security systems and analysts.

This framework is conceptual and is derived from the approaches identified in the reviewed literature. It does not represent an implemented or experimentally validated system.

8. RESEARCH METHODOLOGY

This study uses a structured literature-review methodology to examine AI and ML approaches for phishing website detection. The methodology focuses on identifying relevant scholarly research, organising the literature into thematic categories, comparing existing approaches, and identifying research gaps.

8.1 Literature Identification

Relevant research studies were identified based on their relationship to phishing website detection, Artificial Intelligence, Machine Learning, Deep Learning, ensemble

learning, feature engineering, and Explainable Artificial Intelligence.

Search concepts included terms such as:

- Phishing website detection
- AI-based phishing detection
- Machine learning phishing detection
- Deep learning phishing detection
- Phishing URL classification
- Explainable AI phishing detection
- Intelligent phishing detection

The literature considered in this review primarily covers research published between 2017 and 2025.

8.2 Study Selection

The selected studies were examined according to their relevance to phishing website detection and their contribution to understanding detection approaches, datasets, features, algorithms, evaluation methods, and limitations.

Studies primarily focused on unrelated cybersecurity problems or phishing activities outside the scope of website detection were not considered central to this review.

8.3 Thematic Classification

The selected literature was organised into major research categories based on the detection techniques and research focus discussed in the studies.

The main categories include:

- Traditional Machine Learning
- Ensemble Learning
- URL and Feature-Based Detection
- Deep Learning
- Hybrid Approaches
- Explainable Artificial Intelligence (XAI)
- Dataset and Benchmarking Research
- Real-Time and Adaptive Detection

This classification provides a structured basis for analysing and comparing different approaches to phishing website detection.

8.4 Data Extraction and Comparative Analysis

Relevant information from selected studies was extracted based on their detection approach, major features, machine learning techniques, datasets, and key observations. A comparative analysis of representative studies was then performed to identify the strengths, limitations, and differences among existing approaches. The comparison was based only on information available in the reviewed sources.

8.5 Research Gap Identification

The findings from the reviewed literature were synthesised to identify common research gaps and challenges. Particular attention was given to dataset diversity, model generalisation, real-time detection, explainability, computational efficiency,

adaptive learning, and detection of evolving and AI-assisted phishing attacks.

9. EXPECTED OUTCOMES

As this paper is a review study and does not involve experimental implementation, the expected outcomes are analytical rather than numerical.

The review is expected to provide:

- A consolidated understanding of AI and ML techniques used for phishing website detection.
- A thematic classification of URL, domain, host, content, visual, and semantic features used in existing approaches.
- A comparative analysis of traditional ML, ensemble, deep learning, hybrid, and XAI-based approaches.
- Identification of limitations associated with datasets, feature engineering, model generalisation, computational efficiency, explainability, and practical deployment.
- A conceptual framework showing how multiple website information sources and AI/ML techniques can be integrated for phishing website detection.
- Future research directions for adaptive, multimodal, explainable, scalable, and real-time phishing website detection.

10. FUTURE RESEARCH DIRECTIONS

10.1 Adaptive and Continuous Learning

Future phishing detection systems should investigate adaptive learning mechanisms capable of responding to changing phishing characteristics. Continuous model updating may help address changes in URL structures, webpage content, domain behaviour, and attack patterns.

10.2 Multimodal Feature Integration

Future research should investigate the integration of URL, domain, HTML, webpage content, visual, and semantic features. Combining multiple sources of evidence may improve robustness against phishing websites that deliberately modify one or more characteristics.

10.3 Explainable AI

XAI should be further investigated to provide understandable explanations for phishing classification decisions. Future research should evaluate explanation reliability, consistency, computational overhead, and usefulness to cybersecurity analysts.

10.4 Real-Time Detection

Research should focus on real-time detection systems capable of analysing websites with low inference latency. Browser, cloud, and mobile deployment environments should be considered while maintaining an appropriate balance between accuracy and computational efficiency.

10.5 AI-Generated and Adversarial Phishing

The increasing use of AI-assisted webpage generation creates a need for datasets and detection methods that represent realistic AI-generated phishing content. Future research should also evaluate the robustness of detection models against adversarial modifications designed to evade classification.

10.6 Standardised Benchmarking

Future studies should establish more consistent datasets, feature definitions, validation procedures, and evaluation metrics. Standardised benchmarking would enable more meaningful comparison between different phishing detection approaches.

10.7 Lightweight and Scalable Models

Advanced models should be evaluated for computational efficiency and scalability. Lightweight models may be particularly useful for practical environments where processing resources and response time are constrained.

11. CONCLUSIONS

Phishing websites remain a significant cybersecurity challenge because attackers continuously modify their techniques to imitate legitimate websites and evade conventional detection mechanisms. This review examined the development of AI and ML-based phishing website detection, covering traditional machine learning, ensemble learning, deep learning, hybrid approaches, feature engineering, and Explainable Artificial Intelligence.

The reviewed literature demonstrates a transition from static blacklist and rule-based detection toward intelligent approaches capable of analysing multiple website characteristics. Traditional machine learning remains valuable because of its relatively efficient classification capabilities, while ensemble and deep learning approaches provide additional capabilities for handling complex feature representations. Hybrid and multimodal approaches further provide opportunities to combine different sources of website evidence.

However, the review also identifies important limitations. Existing research frequently depends on static datasets, inconsistent evaluation methodologies, and controlled experimental environments. Dataset diversity, model generalisation, computational efficiency, real-time detection, explainability, adaptive learning, and robustness against evolving and AI-assisted phishing techniques remain important challenges.

The findings indicate that future phishing detection research should move beyond isolated improvements in classification

performance and focus on systems that are adaptive, multimodal, explainable, computationally efficient, and suitable for real-world deployment. Standardised benchmarking and continuously updated datasets will also be important for evaluating the long-term effectiveness of detection models.

Overall, AI and ML provide substantial potential for improving phishing website detection. However, achieving dependable real-world protection requires continued research into adaptability, transparency, robustness, scalability, and practical deployment. Addressing these challenges can contribute to the development of next-generation intelligent phishing website detection systems and strengthen cybersecurity in increasingly digital environments.

REFERENCES

- [1] J. Kumar, A. Santhanavijayan, B. Janet, B. Rajendran, and B. S. Bindhumadhava, "Phishing Website Classification and Detection Using Machine Learning," in Proc. Int. Conf. Computational Collective Intelligence, 2020, doi: 10.1109/ICCCI48352.2020.9104161.
- [2] V. B. Wable and P. Thakur, "CheckPhish: Leveraging a Machine Learning Approach for Detecting Phishing Websites," in Proc. IEEE Int. Conf. Computing, Communication and Networking Technologies, 2023, doi: 10.1109/ICCCNT56998.2023.10308156.
- [3] A. Kara, A. Ok, and A. Ozaday, "Characteristics of Understanding URLs and Domain Names Features: The Basis for Phishing Detection," IEEE Access, 2022, doi: 10.1109/ACCESS.2022.3223111.
- [4] "Detection of Phishing Websites Using an Efficient Feature-Based Machine Learning Framework," International Journal of Engineering and Advanced Technology, 2020, doi: 10.35940/ijeat.C5909.029320.
- [5] H. Kansagara et al., "A Hybrid Approach for Phishing Website Detection Using Machine Learning," 2021.
- [6] K. Omari, "Comparative Study of Machine Learning Algorithms for Phishing Website Detection," International Journal of Advanced Computer Science and Applications, 2023, doi: 10.14569/IJACSA.2023.0140945.
- [7] "Phishing Website Detection Using Deep Learning," in Proc. ACM International Conference, 2022, doi: 10.1145/3542954.3542967.
- [8] "Analysis of Phishing Website Detection Using CNN and Bidirectional LSTM," in Proc. International Conference on Electronics, Communication and Aerospace Technology, 2020, doi: 10.1109/ICECA49313.2020.9297395.
- [9] D.-J. Liu et al., "AntiPhishNN: Hybrid Neural Networks Using URL and Domain Information for Phishing Detection," 2023.
- [10] "A Deep Learning-Based Innovative Technique for Phishing Detection in URLs," Sensors, vol. 23, no. 9, 2023, doi: 10.3390/s23094403.

- [11] "Phishing Website Detection Based on Multi-Feature Stacking," in Proc. International Conference on Artificial Intelligence and Computer Engineering, 2021, doi: 10.1109/ICAICE54393.2021.00140.
- [12] "Phishing Websites Detection Using a Novel Multipurpose Dataset and Machine Learning Methods," Expert Systems with Applications, 2022, doi: 10.1016/j.eswa.2022.118010.
- [13] "Leverage Website Favicon to Detect Phishing Websites," Journal of Electrical and Computer Engineering, 2018, doi: 10.1155/2018/7251750.
- [14] "Machine Learning Based Phishing Web Sites Detection," in Computer and Information Sciences, 2016, doi: 10.1007/978-3-319-27247-4_11.
- [15] A. S. Sengar et al., "A Review on Phishing Websites Revealing Through Machine Learning," in Proc. IEEE International Conference, 2021, doi: 10.1109/SMART52563.2021.9676288.
- [16] "Advancements in Phishing Website Detection: A Comprehensive Analysis," in Proc. IEEE Signal Processing and Information Technology Conference, 2024, doi: 10.1109/SIU61531.2024.10601036.
- [17] "Dataset of Suspicious Phishing URL Detection," Frontiers in Computer Science, 2024, doi: 10.3389/fcomp.2024.1308634.
- [18] J. Wilk-Jakubowski, L. Pawlik, G. Wilk-Jakubowski, and A. Sikora, "Machine Learning and Neural Networks for Phishing Detection: A Systematic Review (2017–2024)," Electronics, vol. 14, no. 18, Art. no. 3744, 2025.
- [19] T. Dalton et al., "PhreshPhish: A Real-World, High-Quality, Large-Scale Phishing Website Dataset and Benchmark," 2025, arXiv:2507.10854.
- [20] F. Maliha, G. T. S. Jahin, U. Habiba, A. N. Chy, and M. J. A. Patwary, "Explainable Comprehensive Analysis of Cross-Dataset Phishing Detection Models," 2026.
- [21] N. Thalpage and E. Jayarathne, "Explainable AI Approaches for Detecting and Mitigating Phishing Attacks: A Review," 2025.