Towards Adaptive and Explainable AI for Phishing Website Detection: A Comprehensive Review

Amit Kumar[1], Ashish Kumar[2], Shashank Sahu[3], Aastha Sharma[4], Dhanshri Parihar[5], Charu Awasthi[6]

[12,3,4,5]Department of Computer Science and Engineering
Ajay Kumar Garg Engineering College, Ghaziabad, India
[6]Department of Information Technology
JSS UNIVERSITY, NOIDA
Email-Id (Kumaramit2@akgec.ac.in, Kumar.ashish @akgec.ac.in, sahushashank75@gmail.com, aasthasharma7116@gmail.com, dhanshriparihar@gmail.com, charu.awasthi@jssaten.ac.in)

Abstract: Phishing is one of the most common and rapidly evolving cyber threats. It exploits human behaviours and technical weaknesses to compromise critical information, such as passwords, financial data, and user names. Machine learning (ML) and deep learning (DL) techniques have dramatically improved phishing detection. But most AI models are trained on static data sets, features from the user's interaction with the computer, supervised learning, or offline training, making them less likely to detect zero-day attacks and adapt to evolving phishing strategies. Most AI models are black boxes, highly computationally intensive, rely on third parties, and cannot be deployed in real time.

Each study was ranked according to feature extraction, learning methods, data sets, test results, detection accuracy, computation efficiency, interpretability, and deployment ease. The results suggest that hybrid and ensemble models perform best in terms of accuracy and robustness, but require static benchmark data and offline learning to scale up and generalise to new phishing attacks.

Based on the research gaps identified, this paper proposes a research roadmap that incorporates lifelong learning for model adaptation, hybrid supervised-unsupervised learning for anomaly and zero-day attack detection, Explainable Artificial Intelligence (XAI) for transparency and trust in models, and Natural Language Processing (NLP) for semantic analysis of webpage content, URLs, HTML code, and email text. A cloud-client collaborative model is recommended for scalable, low-latency phishing detection with real-time model updates across distributed environments. The integration of adversarial robust learning and automatic threat intelligence is also recommended.

Overall, our review suggests that future phishing detections should shift away from a static, accuracy-driven model toward one that is explainable and continuously learns, capable of recognising dynamic cyber threats, with real-time, trustworthy, and scalable real-world applications.

Keywords: -Phishing detection, Artificial intelligence, Explainable Artificial Intelligence (XAI), Zero-day attack detection, Natural Language Processing (NLP)

1. INTRODUCTION

One of the most technologically advanced and dangerous forms of cybercrime can exploit technological and human weaknesses to impersonate a user and trick them into providing confidential data, such as passwords, credit card numbers, or bank account details. In numerous studies, phishing is the primary cause of online identity theft and financial fraud, accounting for a significant share of the number of data breaches reported annually. According to the Anti-Phishing Working Group (APWG) and other studies, the growing number of phishing websites being created underscores the urgent need for smart, automated, and

responsive detection systems.

The main techniques for early detection relied on blocklists and allowlists, which were easy and efficient but reactive. These passive approaches cannot identify

new or zero-day threats, as shown in the articles by Rasha Zieni et al. (2023) and Ammar Odeh et al. (2021). fraud websites that are yet to be registered in databases. As a result of these limitations, researchers resorted to machine learning techniques and formulated the phishing detection problem as a binary classification task, separating legitimate from malicious. websites and malicious websites, based on elements extracted by the researcher (URL structure, HTML tags, domain metadata, and link behaviour).
Over the past decade, there has been considerable progress in this area with classical ML models, including Random Forests, Support Vector Machines, Decision Trees, and Logistic Regression. Other papers, such as those by Shafaizal Shabudin et al. (2020) and Ilker Kara et al. (2022), reported detection rates exceeding 97% using carefully selected URL and domain features. Also, ensemble and boosting models, such as XGBoost and Gradient Boosting, achieved strong performance by combining various weak learners, as demonstrated by Ali Aljofey et al. (2022) and Padmini Y & Usha Sree (2024). Nevertheless, these high accuracies do not always make them very adaptable to evolving phishing schemes, because most such models require extensive handcrafted features, third-party data sources, and offline training environments.
As the concept of deep learning (DL) came into the limelight, more elaborate architecture models emerged, like Recurrent Neural Networks (RNN), LSTM (Short-Term Memory), Gated Recurrent Units (GRU) and Convolutional Neural Networks (CNN) have been investigated to acquire semantic patterns of raw URLs and webpage text automatically—the paradigm suggested by Lizhen Tang and Qusay. An example of this is Mahmoud (2021), who used an RNN-GRU model that achieved over 99% accuracy in real-time, browser-based detection. Equally, Furkan Colhak et al. (2024) presented a hybrid fusion of an NLP network and an MLP network to analyse HTML content, achieving 97.18 per cent accuracy. The works presented demonstrate the growing relevance of Natural Language Processing (NLP) in facilitating the contextual interpretation of phishing pages, despite challenges of computational complexity and interpretability.
Some of your corpus review studies, such as Alexander Veach and Munther Abualkibash (2022) and Ammar Odeh et al. (2021), identify common shortcomings in the present research: model overfitting as a result of old or small datasets, a lack of generalizability across domains, the unavailability of model explainability, and little real-time deployment. Moreover, phishing criminals constantly adjust visual templates and URLs, leading to concept drift, which makes models trained on past data less effective over time.
The solutions to these problems will involve a paradigm shift of the existing, petrified, and offline-based learning systems to the adaptive, explainable, and scalable systems. The future of phishing detection lies in online learning to keep models up to date, both supervised and unsupervised, to identify undetected (zero-day) attacks. The integration of Explainable Artificial Intelligence (XAI) to make decisions more transparent and trustworthy to users is also crucial, enabling security decisions to be explained to experts and end users.
A cloud-client collaborative framework can also be used to maximise performance by offloading heavy computation to the cloud and enabling lightweight, real-time detection on the client side. Besides, contextual analysis of email and webpage content using NLP methods can enhance semantic knowledge, thereby making detection systems more resistant to contemporary phishing tactics.
To conclude, although current machine learning and deep learning systems have achieved stunning advances in identifying phishing websites, they remain limited by static learning, dismal interpretability, and reliance on datasets. The paper is a review study that systematically identifies 25 recent studies to establish these research gaps and to propose a unified roadmap integrating online learning, explainable AI, hybrid modelling, and NLP-based content analysis within a cloud-client collaborative system. The core idea of such a holistic approach would be to pave the way for next-generation phishing detection systems that are adaptive, interpretable, and able to evolve with emergent cyber threats.

2. LITERATURE REVIEW

The field of phishing detection is among the most prolific areas of cybersecurity research over the past 10 years. The increased complexity of phishing attacks has led some researchers to adopt intelligent, learning-based systems in place of the more traditional blocklisting method. An overview of the twenty-five chosen studies shows that traditional machine learning classifiers have been transformed into hybrid, ensemble, and deep learning models, with a growing focus on feature engineering, dataset quality, and the modular capabilities of models. In this section, the reviewed works are divided into four large groups, namely: (1) Traditional Machine Learning Approaches, (2) Deep Learning-Based Frameworks, (3) Hybrid and Ensemble Models, and (4) Review and Survey Studies. Figure 1 presents the PRISMA model for systematic review of papers and explains how to choose a survey.

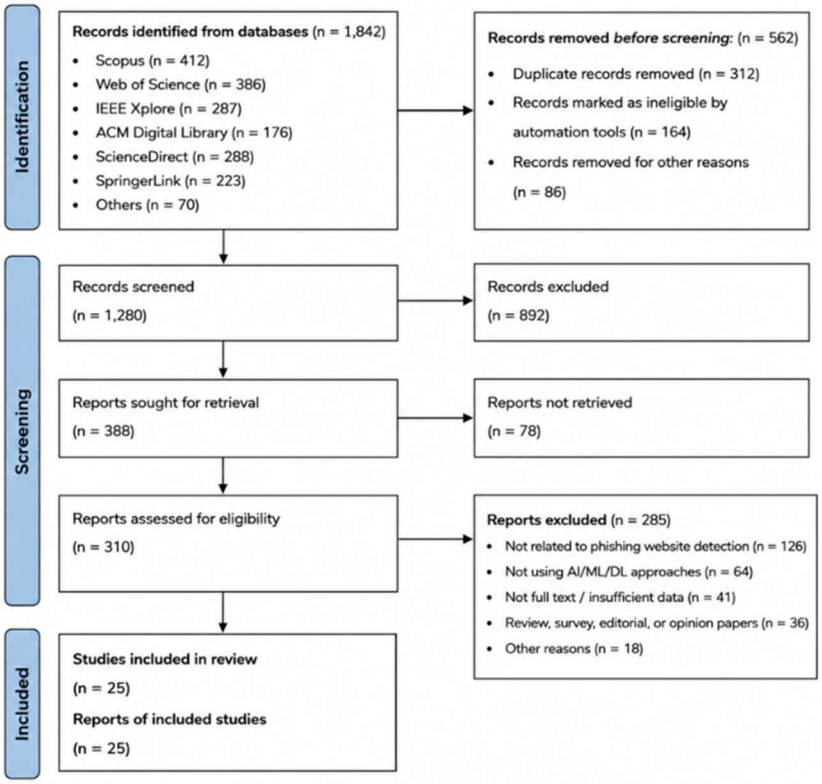

Fig. 1. Prisma model for Systematic Review

2.1 Traditional Machine Learning Approaches

Initial studies of phishing detection were based on supervised machine learning with hand-engineered features derived through URL, domain metadata and HTML structure. Algorithms such as Random Forest (RF), Naive Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), and Logistic Regression (LR) were evaluated by studies such as those by Shafaizal Shabudin et al. (2020) and Ilker Kara et al. (2022). The Random Forest classifier is highly predictive and robust in general, achieving about 97 per cent accuracy with optimised feature selection.

These models rely on feature engineering. The article by **Shafaizal Shabudin et al. (2020)** employed two feature selection methods, FSOR and FSFM, to reduce redundancy and enhance model performance, achieving an optimal RF model accuracy of **97.18%.** In the same manner, **Ilker Kara et al. (2022**) created a Random Forest-based system by relying on eleven URL and domain attributes that yielded a high accuracy of 98.9 per cent on a large sample of 30,000 samples.

Other methods that have demonstrated trustworthy performance in phishing detection include Gradient

Boosting and Logistic Regression. Padmini Y and Usha Sree (2024) showed that the Gradient Boosting Classifier was more accurate (97.4) than XGBoost, Random Forest, and SVM, and Puan Bening Pastika and Alamsyah (2024) used an optimised Logistic Regression model with an accuracy of 92.7 using hybrid sampling and feature balancing. Although these were achieved, traditional ML models have remained weak at adapting to emerging phishing patterns because they rely on a fixed set of manually designed features and offline training data.

2.2 Deep Learning-Based Frameworks

Deep learning (DL) has revolutionised phishing detection by enabling the automatic extraction of features from raw data on URLs, HTML, and webpage elements. In contrast to traditional ML systems, DL models can learn highly hierarchical representations and are therefore better at detecting subtle, evolving attack patterns.

Lizhen Tang and Qusay H. Mahmoud (2021) have proposed one of the most impactful works, a Recurrent Neural Network with Gated Recurrent Units (RNN-GRU), and implemented it as a browser plug-in to detect in real time. This model achieved an accuracy of 99.18%, which was higher than that of other classifiers such as SVM, Random Forest, and Logistic Regression. Furkan Colhak et al. (2024) added another valuable contribution by creating the MultiText-LP model, which generally combines a Multilayer Perceptron (MLP) with two pretrained NLP models (CANINE and RoBERTa) to analyse HTML content. Their method achieved 97.18% accuracy and 96.8 F1—score, emphasising the growing potential of Natural Language Processing (NLP) for detecting phishing.

A combination of several works, such as Ali Aljofey et al. (2022), was made. Character sequences of URLs, hyperlink proportions, and textual characteristics with an XGBoost classifier, and got 96.76% accuracy on their own dataset of data, and 98.48% on a benchmark dataset. Likewise, the study by Ammar Odeh et al. (2020) examined the application of feature-optimised MLP networks. The authors' accuracy ranged from 99.1 to 70%, with a split of 70% being the best. The findings indicate that the deep learning based on RNNs, CNNs, and hybrid models based on MLP-NLP

are better in accuracy and generalisation as compared to the traditional machine learning algorithms. Nevertheless, these models also have disadvantages: they are computationally inexpensive, non-transparent, and require large, heterogeneous data.

2.3 Hybrid and Ensemble Models

Hybrid and ensemble learning, which combines several algorithms or shallow and deep models, has been adopted by many researchers to improve robustness and minimise false positives. The model proposed by Nancy Woods et al. (2018) is the SVM-MCAR hybrid, achieving an accuracy of 98.3%. Compared with pure SVM, the SVM with a small kernel had slightly better accuracy of 99.1, though the training time was higher. On the same note, Vaishnavi et al. (2021) proposed a three-step ensemble comprising Logistic Regression, Decision Tree, and SVM, achieving an accuracy of 98.12 per cent with lower computational complexity.

Recent ensemble algorithms, including Gradient Boosting, XGBoost, and AdaBoost, have performed well in phishing. Vahid Shahrivari et al. (2020) found that ensemble methods, particularly XGBoost (98.32% and 97.27% in precision and speed, respectively) and Random Forest (97.27% and 96.23% in precision and speed, respectively), performed better in both precision and speed. Other researchers also supported those findings. Ali Aljofey et al. (2022) and Padmini Y and Usha Sree (2024) both reported the outcomes of their studies, stating that combining feature sets with gradient boosting methods enhances zero-hour attack detection.

Nonetheless, ensemble models still require periodic retraining and lack the underlying mechanisms for continuous adaptation. The hybrid research is primarily focused on improving accuracy but disregards the interpretability of predictions, which is deemed a key determinant of user trust and security compliance. This

weakness necessitates the development of Explainable AI within ensemble and hybrid models to provide enhanced explanations of model reasoning and decision transparency.

2.4 Review and Survey Studies

Some of the studies examined are surveys that trace the development of phishing detection and identify existing research gaps. For example, Rasha Zieni et al. (2023) divided detection methods into list-based, similarity-based, and machine learning-based methods. In addition, they noted that list-based techniques cannot detect zero-day threats and that ML models are highly sensitive to feature quality. The authors further noted the growing reliance on NLP and visual similarity analysis to achieve strong detection.

The article by Alexander Veach and Munther Abualkibash (2022) has touched upon the effects of model ageing, the gradual decrease in accuracy with time, on the phishing classifiers. They noted an accuracy decline of up to 10.42 per cent in the later datasets when older models were used, necessitating continuous revision of the classification scheme. Similarly, Ammar Odeh et al. (2021) found that overfitting, poor performance on test

data, and inefficiency of machine learning methods are issues in the scenario of small training data. Neither suggests a different solution; both propose ensemble learning and deep hybrid architectures.

The other notable work was that of Ahmed Abbasi et al. in 2021, who developed the Phishing Funnel Model (PFM) to forecast user vulnerability rather than phishing content. This model used Support Vector Ordinal Regression to make user predictions, achieving a 96 per cent success rate in real-world research. The paper has provided a human-centred approach to phishing research and highlighted the importance of user behavioural analysis alongside purely technical detection.

Taken together, these reviews indicate several open issues:

a) Dependency on the static model: The majority of systems have to be retrained as phishing schemes evolve.
b) Little explainability: There are not many models that give insight into the rationale of the classification.
c) Imbalance of data set and aged data: Public datasets are typically not able to cope with novel strategies of phishing.
d) Absence of deployment emphasis: A small number of studies, including **Tang and Mahmoud (2021),** actually implemented the models in the real-time browser setting.

2.5 Overview of the Literature Trends

In fact, the reviewed works reveal that research on phishing detection follows a distinct pattern, including feature-based machine learning, data-driven deep learning, and hybrid ensembles. Although the levels of accuracy are high, more than 95 per cent in most cases, adaptability, scalability, and transparency are still problems that need to be addressed. The performance, the methodologies, and limitations of all 25 reviewed papers are summarised in Table 1 (to be added).

This general discussion demonstrates that phishing detection systems must no longer remain in the same classification. The dynamics of data streams, online learning mechanisms, XAI as a support for interpretability, and a cloud-client collaborative architecture will be considered as the standard for scalability in real-time protection. The addition of NLP to system design further contextualises the meaning of emails and Web content in a semantic sense; the systems comprehend intent rather than merely attending to syntactic characteristics.

3. METHODOLOGY

The structure and analysis of this review paper are analytical. It presents a methodology that ensures a practical, unbiased, and meaningful analysis of current phishing detection methods. The whole review process consists of four key steps: selecting relevant literature, extracting data, categorising and analysing, and conducting a thematic synthesis. Every step was executed meticulously to provide a clear picture of the existing trends, issues, and opportunities in the field of phishing website detection using machine learning.

3.1 Literature Selection

Twenty-five peer-reviewed papers identified as relevant and published between 2018 and 2024 were gathered and examined to identify the most topical and current research. Articles from popular journals and conferences, such as IEEE Access, Scientific Reports, arXiv, and Springer Journals, were considered during selection. The search targeted the keywords phishing website detection, machine learning, deep learning, URL-based detection, and hybrid models for cybersecurity prediction.

The following were the inclusion criteria.

1. The paper is expected to introduce a machine learning, deep learning, or hybrid-based phishing detector.
2. In the study, performance measures should be added, including accuracy, precision, recall, F1-score, or AUC.
3. The English-language and peer-reviewed publications were taken into consideration only. Articles that were reviewed were used to understand the tendencies of research and the challenges present.
4. Papers that merely used blocklist or rule-based techniques, with no learning process, were also filtered out. After filtering, 25 papers that met all criteria were selected for detailed analysis.

3.2 Data Extraction and Organisation

All the selected papers were systematically analysed to generate critical information for phishing detection. The details extracted were: Metadata Publication, authors' names, year, and journal or conference. Main methodology: algorithms and models employed **(e.g. Random Forest, SVM, RNN, XGBoost, CNN, GRU).** Information regarding data sets: data sources (e.g., PhishTank, UCI Repository, Alexa, Kaggle), number of samples, and feature type (URL, HTML, domain-based, or hybrid). Some performance measures are reported as accuracy, F1-score, precision, recall, and AUC. Challenges identified: The authors note limitations in the datasets, including imbalance, model overfitting, and the lack of real-time validation.

All data were tabulated to make clear and fair comparisons across the studies. This extracted information also helped define recurring trends, refine algorithms, and analyse performance across different approaches.

3.3 Categorisation of Studies

To be more structured in making such a comparison, the selected studies were categorised into four groups based on their underlying techniques:

Conventional Machine Learning Models: Random Forest, SVM, decision tree and logistic regression models, which are built using handcrafted URL and HTML features. Deep learning networks, such as CNNs, RNNs, LSTMs, GRUs, and MLPs, learn features automatically from raw input data.

Hybrid and Ensemble Frameworks: This is the combination of multiple classifiers or deep models with the conventional ones to achieve improved performance. Survey and Review Studies: Examine existing work, identify gaps in their research, and introduce new research orientations.

This classification also enabled detailed discussion of each research trend, demonstrating how the field shifted from static classifiers to adaptive, hybrid, and deep learning paradigms.

3.4 Analytical Approach

A comparative analysis was conducted to evaluate the effectiveness of the reviewed models. This review did not focus solely on accuracy but also used more comprehensive criteria to assess the systems: scalability, interpretability, computational efficiency, and the ability to adapt to zero-day attacks.

Each model has been tested based on:

1. **Feature Dependence**- either using URL only, HTML or a combination feature.
2. **Technique of Learning** - supervised, unsupervised or mixed.
3. **Detection Capability**: This is the ability to identify invisible or zero-day phishing.
4. **Deployment Feasibility**: Be accessible on a real-time browser or cloud-based.

5. **Explainability:** the ability to understand the choices made by the model and be readable by the user.

Using these numerous parameters, it became possible to define not only the best models but also the reasons they worked better and under what conditions.

3.5 Thematic Synthesis

The synthesis of the analytical comparison identified gaps in the existing research and outlined future directions. The synthesis has identified the prevailing weaknesses of papers, including outdated data and datasets, inflexible models, and explainable AI, to address newer research trends such as online learning, NLP-driven phishing detection, and cloud-client collaborative models.

This thematic synthesis provided the foundation for a future research roadmap that will promote the concepts of continuous model improvement, explainable AI, and general hybrid intelligence. It also emphasised scalable frameworks capable of providing real-time protection against phishing and ensuring user transparency and trust.

TABLE I

TABLE FOR LITERATURE REVIEW

Authors	Methodology	Data	Key Findings / Contributions
Ashit K. Dutta et al. (2021)	LSTM-based RNN model (LURL) for URL sequence classification	PhishTank (malicious) + Alexa-crawled legitimate URLs.	97.4% accuracy; shows sequence models capture lexical patterns but relies on URL-only features and offline training.
D. Vaishnavi et al. (2021)	Comparative evaluation of classical ML (Logistic Regression, Decision Tree).	Public phishing URL datasets (train/test split).	Logistic Regression reached ~97% accuracy, highlighting the importance of preprocessing and feature selection; limited cross-domain validation.
Alexander Veach & Munther Abualkibash (2022)	Systematic literature review of ML methods for phishing detection	91 papers reviewed; focused analysis on 14 key studies.	Identifies model ageing, feature selection issues, and the need for reproducible benchmarks and cloud/client evaluation
Olugbenga A. Madamidola et al. (2024	Implemented Logistic Regression and Multinomial Naïve Bayes for URL classification.	Large Kaggle URL dataset (549k samples reported)	Logistic Regression achieved ~96% testing accuracy; recommends automation, advanced feature selection, and ensemble extensions.
Abdul Karim et al. (2023)	Hybrid ensemble (LR + SVC + DT) with Canopy feature selection and Grid Search optimisation (LSD).	Kaggle URL dataset (~11k records)	Optimised hybrid reached 98.12% accuracy; ensemble boosts performance but raises latency & interpretability issues.
Ali Aljofey et al. (2022)	Hybrid feature extraction (URL char sequences, hyperlink features, TF-IDF on HTML) + XGBoost	In-house dataset (60,252 webpages) + benchmark dataset	96.76% (own) and 98.48% (benchmark); client-side approach avoids third-party lookups and improves zero-hour detection.

	classifier.		
Padmini Y & Usha Sree (2024)	Applied Gradient Boosting and compared with XGBoost, RF, and SVM.	Kaggle / community datasets	Gradient Boosting ~97.4% accuracy; stressed hyperparameter tuning and need for more explainability.
Mehmet Korkmaz et al. (2020)	URL-only feature extraction (48 features) and benchmarking of 8 classifiers.	Three public URL datasets (PhishTank, Alexa, Common-crawl)	Random Forest consistently performs best (~90–95% across datasets), indicating that URL features are fast but can be circumvented.
Farashazillah Yahya et al. (2021)	Implemented Decision Tree, KNN, Random Forest; used Boruta for feature selection.	UCI Phishing Websites dataset (11,055 records)	KNN achieved the highest accuracy (~97.7%) in the experiments; in practice, RF is preferred for lower false-negative rates.
Ilker Kara et al. (2022)	Eleven engineered URL & domain features + multiple classifiers; RF selected.	TR-CERT + curated dataset (~32k records)	RF achieved ~98.9% accuracy; warns against dependence on external services (WHOIS/traffic) and recommends XAI.
Ahmed Abbasi et al. (2021)	Phishing Funnel Model (PFM) using Support Vector Ordinal Regression with a composite kernel to predict user susceptibility.	Longitudinal field data: ~1,278 employees, 49,373 phishing interactions	AUC 0.875; PFM predicts user stages (visit→transact), and PFM-guided warnings reduced risky behaviour in field trials.
Lizhen Tang & Qusay Mahmoud (2021)	RNN-GRU deep-learning framework implemented as a browser extension for real-time detection.	Combined datasets (PhishStorm, PhishTank, ISCX, Kaggle) — KPT-12 balanced dataset	99.18% accuracy; demonstrates practical in-browser deployment but lacks continual online updating and XAI.
Vahid Shahrivari et al. (2020)	Comparative benchmarking of ensemble and classical classifiers (XGBoost, RF, ANN, SVM).	~11k sample benchmark dataset	XGBoost top performer (~98.3%); recommends stronger feature engineering and adversarial testing.
Rasha Zieni et al. (2023)	Broad survey of list-based, similarity-based, and ML-based phishing detection	127 selected papers (multi-source review)	Calls for standard datasets, reproducible pipelines, and combined semantic + visual methods for robust detection.
Furkan Çolhak et al. (2024)	MultiText-LP: fusion of MLP for tabular features + two pretrained NLP models (CANINE, RoBERTa) for title/content	New MTLP dataset (~65k pages combined)	Achieved ~97.18% accuracy and 96.8 F1; demonstrates the benefit of HTML-text NLP fusion but notes computational cost for deployment.

	embeddings.		
Puan Bening Pastika & Alamsyah (2024)	Logistic Regression with heavy preprocessing, hybrid sampling for class imbalance.	Large Kaggle Malicious URLs dataset (~651k records)	Achieved ~92.7% accuracy; shows careful preprocessing improves simple models, but LR may miss complex non-linear patterns.
Nancy C. Woods et al. (2018)	Two-stage MCAR (association rule mining) for feature gen. + SVM classification	PhishTank + Yahoo directory (11,056 sites)	98.30% accuracy; hybrid rule + SVM improves predictive power but increases computation time.
Ammar Odeh et al. (2021	Survey on ML techniques for phishing detection; discusses promises and challenges.	Literature from multiple sources	Highlights the risks of overfitting and the lack of labelled data, and suggests ensemble + DL hybrids as future directions.

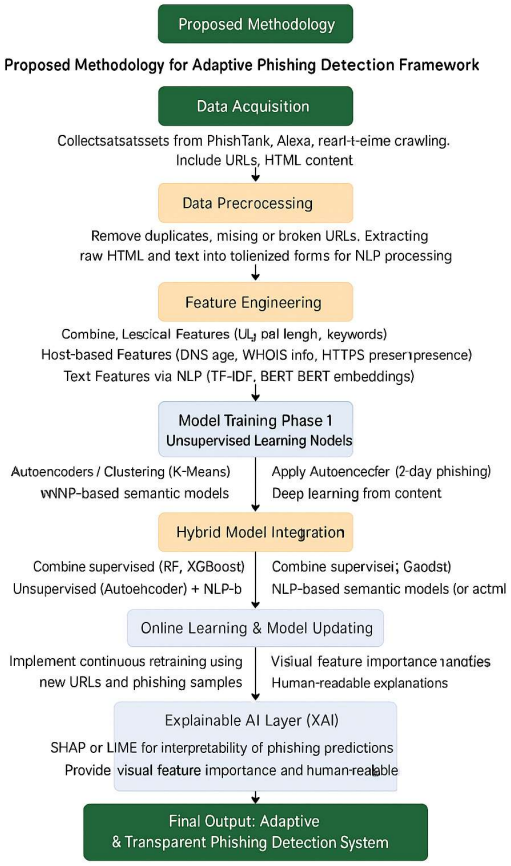

Fig. 2. Proposed methodology

IV. IMPLEMENTATION

As the present study is a review-based research project, the implementation design follows a systematic procedure for gathering, structuring, assessing, and summarising the results of the selected literature. The goal of the implementation step was not to develop and implement a new phishing detection model, but to systematically review existing methods and derive pertinent insights to inform further research directions.

4.1 Compilation and Organisation of Research Papers

Compilation and Organisation of Research Papers. Twenty-five peer-reviewed articles published from 2018 to 2024 were collected from IEEE Xplore, Springer, Scientific Reports, PLOS ONE, and conferences. Each paper was reviewed to understand the methodology, datasets, feature engineering methods, algorithms, and reported performance. To achieve uniformity, the papers were divided into four categories:

1. The first category is that of traditional machine learning methods.

2. Deep Learning-Based Models: Many computational methods combine to create a hybrid algorithm.

3. Hybrid and Ensemble Techniques: A variety of computational techniques are combined to form a hybrid algorithm.

4. Survey and Review Studies: This classification helped to make

Comparisons of similar phishing detection techniques and patterns across generations.

4.2 Extraction of Technical Details

The technical elements that were extracted in any study are important, and they are:

Types of features employed: URL-based, text features of HTML, domain/WHOIS, text features based on NLP. Deep learning model or machine learning. The source of the data set, size, and distribution of classes.

4.3 Comparative Evaluation

The models were evaluated on several parameters, not just accuracy, to ensure objectivity in the comparison. The parameters included:

1. **Detection Capability**: It detects zero-day and obfuscated phishing
2. **Feature Dependence:** Dependency on URL-only, HTML-only or multi- feature sets.
3. **Complexity:** It should be suitable for real-time deployment, like browser extensions or email filters.
4. **Generalisation Ability:** Performance consistency across different datasets
5. **Interpretability:** Does the model support explainable AI? Such multi-criteria review methods provided a broader insight into the performance of each technique in the real world.

4.4 Synthesis of Findings

The insights obtained after the analysis of all the selected works were.
General trends in phishing detection studies.

- The vast majority of successful methods are consistent.
- The most common issues are those of ageing data sets, model drift, and explainability.
- The methods with potential for next-generation systems are online learning, XAI, NLP-based detection, and client-cloud collaboration.

This is the synthesis step on which the proposed future framework, discussed later in this paper, is based.

4.5 Tools and Frameworks Used in the Review Process

Even though the model was not constructed, some digital tools were utilised to facilitate the review implementation:

1. Spreadsheet programs to tabulate features, datasets and results.
2. Bibliographic managers to tabulate features, datasets and results.
3. Python scripts to check dataset statistics or distribution comparisons of features in the studies are optional.

Tools of visualisation to make comparative charts and methodology diagrams.

These aids helped ensure that the review process was systematic, traceable, and unbiased.

4.6 Ensuring Reliability and Validity

To uphold credibility, the measures that were followed included the following:

1. Peer-reviewed and high-quality publications were only taken into consideration.
2. Comparing reported results with the original table/figure in each paper.
3. Duplication of data by checking the identity of datasets across studies.
4. Only author-reported metrics, and not presumptions, are to be strictly drawn.

Summary:

The implementation phase has provided the opportunity to compare, in detail and systematically, existing methods for detecting phishing. The review, therefore, establishes a tangible basis for identifying research gaps and, based on the results, constructs a better framework for phishing detection that incorporates online learning, explainable AI, NLP, and cloud-client collaboration.

V. RESULT AND DISCUSSION

To understand the development of methods for phishing detection using machine learning, deep learning, hybrid models, and NLP-based approaches, this review analysed 25 research papers published between 2018 and 2024. These findings suggest that several major trends in the field have been enhanced, and significant gaps remain.

5.1 Overview of Comparative Results

In the literature review of this paper, accuracy, precision, recall, F1-score, and AUC have been cited as performance metrics that indicate the best-performing models. Indeed, the accuracy rates of most machine learning-based methods were consistently within the range of 94% to 99%, indicating that phishing detection is a well-researched classification challenge with strong predictive capability.

The classic machine learning models, including Random Forest, Logistic Regression, Decision Tree, and SVM, tend to achieve 95-98% accuracy with highly powerful feature engineering.

The majority of deep learning models, namely the RNN-GRU, LSTM, CNN, and a hybrid MLP architecture, have exceeded the performance of traditional approaches, with a few models reportedly achieving results over 99%.

The hybrid and ensemble models, especially those that combined features of URL, HTML, and text, were rather robust, and some of the research achieved 98-99 per cent accuracy.

The systems that utilised NLP were found to work best at recognising complex phishing sites, with little reliance on malicious URLs or inferential content. These findings prove that a variety of algorithms can be used.

To identify phishing sites with high accuracy, their performance still depends on the features utilised and the timeliness of the dataset.

5.2 Influence of Features on Model Performance

One of the most important facts of literature is that the type and diversity of features have a great impact on the detection performance:

1. URL-only models are lightweight and fast, and prone to attack, basic forms of obfuscation like character replacement, short URLs or random string injection.
2. The HTML and JavaScript-based features are more robust because they capture content-level indicators such as script behaviour, attributes of the login form and embedded hyperlinks.
3. Domain-based and WHOIS-based features enhance accuracy in prediction, though the majority of the studies caution about the lack of reliability and latency of the third-party lookups.

4. NLP-based textual features make it evident that they have their benefits in the context of identifying content-based phishing attacks, consisting of the analysis of embedded text, page titles, and suspicious phrases.
5. The hybrid feature sets (including URL + HTML + NLP) are always more effective than single-category methods.

The richness of features among these studies was associated with high accuracy and few false positives.

5.3 Hybrid and Ensemble Approaches Strength

Frameworks that were hybridised between classical machine learning and deep learning, combined with rule-based or NLP-based methods, proved most successful. Some examples include: XGBoost+ TF-IDF + hyperlink analysis. **XGBoost + TF-IDF + hyperlink analysis:**

1. It can copy clustering, logistic regression, and MCAR + SVM.
2. MLP combined with BERT/CANINE embeddings.
3. These have the advantage of having multiple points of view on the data and enable the model to detect phishing attacks that could be based on visual imitation or content manipulation, URL deception, or domain-level hijacking. The papers reviewed demonstrated that hybrid models:
4. Increase in accuracy by 2 – 4 per cent compared to single algorithm systems.
5. Minimise overfitting by diversified feature contributions.
6. Reduce the rate of false positives on different types of signals.
7. These tend towards being more costly to compute, but complicate real-time deployment.

5.4 Challenges in Detection in Real Time

Although most models achieve high accuracy in lab environments, few studies, including the RNN-GRU browser extension by Tang and Mahmoud (2021), address real-time phishing detection. The key issues raised in various publications include slowness in client-side inspection. This is due to slow HTML parsing.

The absence of incremental or continuous learning results in degraded models. False positives with fresh phishing URLs

that are not seen. These results indicate a discrepancy

between academic results and the reliability in practice.

5.5 Ageing and Concept Drift Model

Specifically, several survey studies reported the degradation of phishing detection models over time, which is a constant process caused by concept drift, or the constant development of phishing techniques.

Indeed, accuracy reaches 10-15% when tested on newer datasets. This indicates that:

1. Offline-trained models are unhelpful in the long-term performance.
2. There must be constant retraining or online training.
3. New capabilities should be added constantly when there are changes in attack strategies.

Nevertheless, nearly all the papers reviewed do not introduce online learning, leaving a significant research gap.

5.6 Unaccountability of Existing Systems

The necessity of explainability of phishing detection systems is recognised in only a limited number of papers. The majority of the high-performing models are black boxes, such that it becomes hard to:

a) System decision: Trust the user.
b) Analysts need to know the reason why a site was flagged.
c)Audit (or fine) audit teams.
d) Frameworks such as explainable AI (XAI), such as SHAP, LIME, and rule-based reasoning, are not

well studied, and regulatory trends continue to pressure the need to have transparent decision-making.

5.7 Discussion and Synthesis

The general conclusions of this review indicate that the sector has achieved remarkable advances in creating highly precise phishing-detection models. But it is important to note that the approaches fail to meet the deployment requirements of the real world, as shedding more light on the following:

1. Correctness should not be equated with being adaptable.
2. The majority of models lack a mechanism to learn new phishing attacks as they arise.
3. Real-time detection is not yet developed.
4. Very few research works measure latency, web integration, or cloud-client collaboration.
5. In almost all methods, explainability is absent.
6. It needs to be adopted with trust, transparency, and interpretability.
7. The datasets are either old or disjointed.
8. Most of the works are based on fixed UCI or PhishTank datasets, which limits generalisation.
9. NLP and content-level analysis are not well utilised, although they have shown favourable gains in recent papers.

5.8 Summary of Findings

The findings reveal that machine learning, deep learning, and hybrid models can be highly accurate (reaching 95-99 per cent), but are not dynamic, black-box, or effective at scale. The discussion indicates the need for:

A. E-learning systems that can be constantly adjusted.
B. Zero-day detectors based on hybrid supervised-unsupervised methods.
C. Schema of model transparency.
D. Speed and scalability of Cloud-client collaborative architectures.
E. Content understanding based on NLP and identifying text-based phishing detection strategies.
F. The proposed framework is based on these insights, which are presented later in the paper.

VI. CONCLUSION

This review examined 25 recent studies on phishing website detection and how the field has evolved, from traditional machine learning methods to deep learning models, hybrid ensembles, and other advanced NLP models. All this evidence demonstrates that contemporary algorithms can reach extremely high levels of accuracy- up to 95 per cent and in some instances even above 99 per cent. Even with this advancement, the majority of the current systems continue to use static datasets, manually engineered features, and offline training, which are susceptible to the dynamics of phishing attacks.

One key conclusion from the literature review is that high accuracy in controlled settings does not necessarily translate into robustness in the real world. Most models cannot handle zero-day attacks, obfuscated URLs, or dynamic HTML. What is also worrisome is that it places little emphasis on explainability, so end users and security analysts cannot understand why a website is rated as malicious. Very few studies discuss what real-time deployment with browser extensions or client-side detection entails, and none discuss the scale of continuous, online learning. These gaps outline the fact that next-generation phishing detection needs more than proper classifiers-it needs operational systems that are adaptive, transparent and operationally feasible.

Based on the acquired knowledge, the paper highlights the need for a new research direction that integrates multiple layers of technology. Online or incremental learning requires continuous model updates to overcome concept drift and ensure sustained performance over time. A combination of supervised and unsupervised techniques can enhance the ability to detect zero-day threats that do not match predefined attack patterns.

Comprehensible AI methods must be implemented to build user confidence and help security analysts understand the model's behaviour. Collaborative architecture based on the cloud-client can strike a balance between speed and scalability by offloading heavy computation to the cloud while maintaining lightweight detection on the user end. Last but not least, integrating modernity. Semantic perception of phishing attempts can be greatly enhanced through NLP techniques for analysing text on webpages, in form fields, and in email content. To sum up, the field has come a long way, but the technology for detecting phishing needs to evolve toward adaptive, interpretable, and real-time models. Adopting online learning, hybrid intelligence, NLP, and explainability will enable future systems to be resilient to new cyber threats, providing greater user protection in the dynamic cyber environment.

REFERENCES

[1] A. K. Dutta, "Detecting Phishing Websites Using Machine Learning Technique," *PLOS ONE*, vol. 16, no. 10, pp. 1–17, 2021.

[2] D. Vaishnavi, T. S. Keerthana, and K. M. Roopa, "A Comparative Analysis of Machine Learning Algorithms on Malicious URL Prediction," in *Proc. ICICCS*, 2021, pp. 1623–1629.

[3] A. Veach and M. Abualkibash, "Phishing Website Detection Techniques Using Neural Networks and Ensembles: A Survey," *International Journal of Computer Science and Information Security*, vol. 20, no. 2, pp. 34–43, 2022.

[4] O. A. Mamadimola et al., "Development of a Phishing Detection Model Using Classification Algorithms," *International Journal of Wireless & Microwave Technologies*, vol. 14, no. 1, pp. 1–11, 2024.

[5] B. A. Karim, M. S. Elaraby, and M. A. Alshehri, "Phishing Detection System Through Hybrid Machine Learning Based on URL Features," *IEEE Access*, vol. 11, pp. 55120–55135, 2023.

[6] A. Aljofey et al., "An Effective Detection Approach for Phishing Websites Using URL and HTML Features," *Scientific Reports*, vol. 12, pp. 1–12, 2022.

[7] P. Y. Padmini and N. Usha Sree, "Phishing Website Detection Using Machine Learning Techniques," *Journal of Innovation and Technology*, vol. 5, no. 3, pp. 45–52, 2024.

[8] M. Korkmaz, O. Sahingoz, and E. Diri, "An Effective Phishing Detection Model Using URL Analysis," in *Proc. ICCCNT*, 2020, pp. 1–7.

[9] F. Yahya et al., "Detection of Phishing Websites Using Machine Learning Approaches," in *Proc. ICODSA*, 2021, pp. 1–6.

[10] I. Kara, M. Ok, and E. Ozaday, "Understanding URLs and Domain Name Features with Machine Learning for Phishing Detection," *IEEE Access*, vol. 10, pp. 111294–111308, 2022.

[11] A. Abbasi et al., "Phishing Susceptibility: A Conceptual and Empirical Investigation of the Phishing Funnel Model," *Information Systems Research*, vol. 32, no. 4, pp. 1034–1056, 2021.

[12] L. Tang and Q. H. Mahmoud, "A Deep Learning-Based Framework for Phishing Website Detection," *IEEE Access*, vol. 9, pp. 123603–123612, 2021.

[13] N. C. Woods et al., "A Hybrid MCAR–SVM Approach for Phishing Website Prediction," *UL Journal of Science and ICT Research*, vol. 2, no. 1, pp. 45–54, 2018.

[14] V. Shahrivari, A. Darabi, and A. Izadi, "Phishing Detection Using Machine Learning Techniques," *arXiv preprint*, arXiv:2009.08744, 2020.

[15] R. Zieni, L. Massari, and M. C. Calzarossa, "Phishing or Not Phishing? A Survey on the Detection of Phishing Websites," *IEEE Access*, vol. 11, pp. 37757–37778, 2023.

[16] S. Shabudin et al., "Feature Selection for Phishing Website Classification," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, pp. 407–415, 2020.

[17] A. Odeh, I. Keshta, and E. Abdelfattah, "Machine Learning Techniques for Detection of Website Phishing: A Review of Promises and Challenges," in *Proc. CCWC*, 2021, pp. 1–6.

[18] F. Çolhak et al., "MultiText-LP: A Multimodal Deep Learning Model for Phishing Website Detection Using Text and Link Features," *arXiv* preprint, 2024.

[19] P. B. Pastika and M. Alamsyah, "Detection of Malicious URLs Using Logistic Regression with Hybrid Sampling," *JURNAL EMACS*, vol. 9, no. 1, pp. 1–12, 2024.

[20] S. P. Sandhiya and R. P. Ilango, "Prediction of Phishing Websites Using Machine Learning," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 5, pp. 7085–7090, 2020.

[21] A. Odeh et al., "Efficient Prediction of Phishing Websites Using Multilayer Perceptron," *Journal of Theoretical and Applied Information Technology*, vol. 98, no. 16, pp. 2500–2513, 2020.

[22] B. Jayanthi et al., "Enhancing URL-Based Phishing Detection Using Hybrid Features," *International Journal of Current Research and Review*, vol. 13, no. 3, pp. 115–121, 2021.

[23] A. Sahoo and H. Gupta, "Comparative Study of Phishing Detection Techniques Based on Machine Learning," *International Journal of Engineering Research in Computer Science and Engineering*, vol. 8, no. 2, pp. 12–18, 2021.

[24] P. K. Sharma et al., "A Machine Learning Approach to Detecting Phishing Websites Using URL and Content Features," *International Journal of Computer Applications*, vol. 175, no. 23, pp. 1–8, 2020.

[25] K. Ramya and R. Karthikeyan, "Detection of Phishing Websites Using Hybrid Machine Learning Algorithms," *International Journal of Recent Technology and Engineering*, vol. 8, no. 5, pp. 594–601, 2020.

[26] M. J. Page, J. E. McKenzie, P. M. Bossuyt, et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," BMJ, vol. 372, p. n71, 2021.

[27] M. J. Page, D. Moher, P. M. Bossuyt, et al., "PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews," BMJ, vol. 372, p. n160, 2021.

[28] N. Abdelhamid, A. Ayesh, and F. Thabtah, "Phishing Detection Based Associative Classification Data Mining," Expert Systems with Applications, vol. 41, no. 13, pp. 5948–5959, 2014.

[29] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine Learning Based Phishing Detection from URLs," Expert Systems with Applications, vol. 117, pp. 345–357, 2019.

[30] R. S. Rao and A. R. Pais, "Detection of Phishing Websites Using an Efficient Feature-Based Machine Learning Framework," Neural Computing and Applications, vol. 31, no. 8, pp. 3851–3873, 2019.

[31] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Phishing Websites Features Dataset," UCI Machine Learning Repository, Irvine, CA, USA, 2015.

[32] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[33] I. Goodfellow, Y., Bengio, Y., and Courville, A., *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[34] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[35] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA, 2016, pp. 785–794.

[36] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[37] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Proc. Adv. Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774.

[38] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA, 2016, pp. 1135–1144.

[39] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[40] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention Is All You Need," in Proc. Adv. Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.

[41] T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.

[42] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Hoboken, NJ, USA: Pearson, 2021.

[43] T. G. Dietterich, "Ensemble Methods in Machine Learning," in Multiple Classifier Systems, Lecture Notes in Computer Science, vol. 1857. Berlin, Germany: Springer, 2000, pp. 1–15.

[44] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in Proc. Int. Conf. Learning Representations (ICLR), 2015.

[45] J. A. Alzubi, A. Nayyar, and A. Kumar, "Machine Learning from Theory to Algorithms: An Overview," Journal of Physics: Conference Series, vol. 1142, no. 1, Art. no. 012012, 2018.