Advances in Artificial Intelligence and Machine Learning; Review 5 (3) 4115-4153 Received 08-05-2025; Accepted 1-08-2025; Published 07-08-2025

A Systematic Review of Artificial Intelligence Techniques for Phishing Detection

Anghelo Aguirre anghelo.aguirre@usil.pe
San Ignacio de Loyola University - Faculty of Engineering
Lima - Peru

Luis Salazar luis.salazarma@epg.usil.pe
San Ignacio de Loyola University - Faculty of Engineering
Lima - Peru

Corresponding Author: Anghelo Aguirre

Abstract

This systematic review investigates the application and effectiveness of Artificial Intelligence techniques such as Machine Learning (ML), Deep Learning (DL), and Generative Artificial Intelligence (Gen AI) in the detection and/or prevention of Phishing attacks. The analysed studies were categorised into four groups: ML-based models, DL-based models, hybrid models (ML, DL, others), and Large Language Models (LLMs). The results reveal that none of the models consistently outperforms the others (regardless of context); however, performance depends on factors such as the quality of the dataset used, algorithm architecture, hyperparameter tuning and deployment environment. However, a clear trend emerged: DL and Gen AI-based models tend to achieve higher accuracy and stability, particularly in complex scenarios and when working with large or unstructured datasets. Techniques such as convolutional neural networks (CNN), short and long-term memory (LSTM), and temporal convolutional networks (TCN) demonstrated exceptional performance, achieving accuracies above 99%. Similarly, tuned LLMs such as GPT-2-medium and Llama-3-8b-instruct showed strong classification capabilities in phishing detection tasks. In contrast, traditional ML algorithms such as Random Forest and XGBoost performed well in structured and balanced data environments, but exhibited limitations in adaptivity and semantic representation. The findings emphasise the importance of a holistic approach that considers not only the algorithm itself, but also the quality of the input data, computational resources and the practical context of the implementation. This review aims to provide relevant information on modern AI techniques to strengthen cybersecurity measures against evolving phishing threats.

Keywords: Phishing Detection, Phishing Prevention, Machine Learning, Deep Learning, Artificial Intelligence, Generative AI, Cybersecurity.

Citation: Anghelo Aguirre and Luis Salazar. A Systematic Review of Artificial Intelligence Techniques for Phishing Detection. **Advances in Artificial Intelligence and Machine Learning**. 2025;5(3):231.

1. INTRODUCTION

Phishing attacks have evolved into a global threat, with more than 298,000 incidents reported in 2023 and millions of dollars in losses associated with this crime [1]. Moreover, they accounted for 30% of remediated incidents in 2023, although their volume decreased by 44% compared to 2022. In contrast, the use of stolen valid credentials grew by 71%, making it the main access route for attackers. While attacks using generative AI are not yet a direct threat, phishing is expected to be one of the first cases where cybercriminals use this technology to generate malicious messages more quickly and convincingly [2]. Increasingly, traditional systems or tools for detecting these attacks are becoming obsolete, as cybercriminals employ more elaborate techniques or tools in the face of these proposals, such as the use of large-scale language models (LLMs) in order to generate fraudulent content.

In the face of this reality, the use of Artificial Iintelligence (AI) with its subsets: Machine Learning (ML) and Deep Learning (DL), have become supporting tools that fortify automated phishing detection and prevention. However, the rapid evolution of both attack methods and technological solutions has generated a growing diversity of approaches, hybrid models, and new techniques such as LLMs, metaheuristic optimisations and deep architectures that require systematic analysis.

Although there are individual studies that apply AI models to the phishing problem, there is currently no systematic review focusing on the most recent work (2023-2025) that allows a rigorous comparison of which ML and/or DL models are currently in use, which are the most effective according to key metrics such as accuracy or F1-score, and what factors impact the performance of such models, such as data type, feature selection or balancing techniques. *In this regard, this review offers researchers the following:*

- A systematic and updated review (covering scientific publications from 2023 to 2025) of prevention and detection of Phishing by using AI techniques (about the latest models of ML, DL, LLMs and Gen AIs).
- Comparative analysis and categorisation on AI models for this topic.
- The identification of critical factors that affect the phishing detection performance of the models and techniques described in this study.
- An analysis of the trending techniques and models highlighting their accuracy and stability.
- Detailed information of the performance of each model and technique covered in this review.

2. THEORETICAL FRAMEWORK

2.1 Phishing

A very common type of cyberattack where an email is sent pretending to be a real website, telling the victim to update their details. To do so, the victim is asked for their information and to log in to the imitated platform on a server of the criminal [3].

2.2 Machine Learning (ML)

A subfield of AI defined as the use of computational methods to improve performance and/or make predictions based on prior knowledge from collected data, which can be human-labelled training sets or information obtained through interaction with the environment. The quality of this data is essential for the effectiveness of the predictions made by the learning model [4].

2.3 Deep Learning (DL)

A subset of ML used for descriptive, prescriptive, and predictive purposes. In addition, it can be trained with supervised, unsupervised or reinforced approaches. DL differs from ML in the way it stores and processes relationships between features. While in other models the established relationships between features can be defined, in DL neural networks (the model on which it is based) automatically learn these relationships through different layers of processing [5].

2.4 Generative Artificial Intelligence

A subset of Deep Learning focused on the creation of new content from a set of data provided as input, taking advanced DL and neural network techniques. These Gen AI models can be trained to generate different types of outputs (text, images, music, video, etc.) [6].

2.5 Large Language Model

Model and subset of Gen AI trained with large amounts of data to understand and generate human language. LLMs are trained using architectures such as Transformers and have a significant capacity to perform tasks involving text (translation, classification, text generation, among others). Their performance lies in the amount of parameters and data they are trained with (allowing them to "learn" linguistic patterns on a large scale) [7].

2.6 Evaluation Metrics for AI Models.

The metrics used for this review are the following:

2.6.1 Confusion matrix:

A fundamental tool for evaluating the performance of a classifier with respect to test data. It is a two-dimensional matrix indexed in one dimension by the true class of an object (real class) and in another by the class it receives from the classifier (predicted class) [8] (see TABLE 1). It contains within itself the following values:

- True Positive (TP): The prediction result and the reality correspond to a positive case.
- False Positive (FP): The predicted outcome is positive, and what happened in reality is negative.
- True Negative (TN): The predicted outcome and the reality correspond to a negative case.
- False Negative (FN): The predicted outcome is negative, and what happened in reality is positive.

Table 1: Confusion Matrix

Class Designation		**Class with Actual Values**	
		True (T)	False (F)
Class with Predicated Values	Positive (P)	TP	FP
	Negative (N)	TN	FN

From this matrix (see TABLE 1) the terms Positive and Negative can be defined as follows:

$$Positive\ \ (P) = TP + FN$$
$$Negative\ \ (N) = FP + TN$$

2.6.2 Accuracy:

Calculated as the sum of two exact accuracies (TP + TN) and divided by the total number of data sets (P+ N). The best possible accuracy is 1.0 and the worst possible accuracy is 0.0 [9].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{P + N}$$

2.6.3 Precision:

Calculated as the number of correct positive predictions (TP), divided by the total number of positive predictions (TP + FP). The best accuracy is 1.0 and the worst is 0.0 [9].

$$Precision = \frac{TP}{TP + FP}$$

2.6.4 Recall:

Calculated as the ratio of exact positive predictions (TP) to the sum of the total number of positives (P). The best recall is 1.0 and the worst recall is 0.0 [9].

$$Recall = \frac{TP}{TP + FN}$$

2.6.5 F1 Score:

Calculated as the harmonic mean of accuracy and recall of a classifier [9].

$$F1Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

3. METHODOLOGY

3.1 General Approach

This systematic review was conducted on the basis of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) model in order to follow a structure that allows us to analyse how the use of AI, ML, DL and generative models (Gen AI) have been applied to the detection and prevention of phishing - in advances between the years 2023 and 2025. It is with this approach that 42 scientific articles were obtained (categorised and analysed) through a rigorous screening and whose data were also systematised with an analysis matrix.

To guide this review, the following research questions were posed:

- RQ1: What Machine Learning techniques have been used for phishing detection and/or prevention?
- RQ2: What Deep Learning techniques have been used for phishing detection and/or prevention?
- RQ3: Which hybrid techniques (ML + DL + others) have been used for phishing detection and/or prevention?
- RQ4: What LLMs and Gen AIs techniques have been used for phishing detection and/or prevention?

3.2 Information Sources and Databases

The articles included in this systematic review were identified by searching the Scopus database between April and June 2025. The time range considered was from January 2023 to June 2025, as this period represents a period of great dynamism in the development and application of artificial intelligence techniques, especially in relation to large-scale language models (LLMs) and generative AI (Gen AI).

However, it was observed that many studies published in 2023 were still in preliminary stages, mostly focused on conceptual explorations, models in the design phase, or proposals without clear empirical validation. Consequently, these articles did not provide sufficient evidence on their concrete applicability to phishing detection or prevention tasks.

For this reason, the review prioritised studies that, in addition to falling within the aforementioned period, presented solid empirical results, experimental validations and verifiable practical applications. We selected only research published in indexed scientific journals, especially those specialising in computer science, artificial intelligence and cybersecurity, thus guaranteeing both the scientific quality and the thematic relevance of the works analysed.

3.3 Inclusion and Exclusion Criteria

The selection was based on a screening process divided into three levels: title, abstract and full text of the article to be reviewed. At each of these levels, different specific inclusion and exclusion criteria were applied to ensure the quality and relevance of the studies included in this systematic review.

3.3.1 Level 1 - Title:

We included articles with titles indicating either empirical research or experimental evaluation related to techniques or the use of models based on AI, ML, DL and natural language processing (NLP) applied in the detection and/or prevention of phishing (see TABLE 2).

Table 2: Reference Filtering by Title

Inclusion Criteria (IC)	**Exclusion Criteria (EC)**
IC1. Empirical research or experimental evaluation.	EC1. Topic not related to phishing or without practical application.
IC2. IA techniques or models (ML, DL, NLP, etc.).	EC2. Reviews, editorials, abstracts, or non-original articles.
IC3. Specific application in phishing detection or prevention.	EC3. Published before 2023.
IC4. Suggested evaluation or metrics (accuracy, precision, performance).	EC4. Not in English.

Titles were excluded here if they were unrelated to phishing, determined to be review articles, editorials (books), abstracts, not original, published before 2023, and/or not written in English.

3.3.2 Level 2 - Abstract:

We included abstracts describing empirical studies or (real) experimental evaluations, where the use of AI with techniques applied to phishing was mentioned and quantitative results were considered (see TABLE 3).

Table 3: Filtering References by Abstract

Inclusion Criteria (IC)	**Exclusion Criteria (EC)**
IC1. Empirical study or actual experimental evaluation.	EC1. Not empirical or simulation only.
IC2. Use of AI techniques or models for phishing.	EC2. Does not use AI or does not apply to phishing.
IC3. Reports quantitative results.	EC3. Does not report quantitative results.
IC4. Published in 2023 or later.	EC4. Published before 2023.
IC5. Indexed and ranked journal article.	EC5. Not an indexed article.
IC6. In English.	EC6. Not in English.

Here we excluded abstracts without empirical data applied to phishing, without quantitative results, that were published before 2023, that were not indexed in ranked journals - any quartile was considered here (Q1, Q2, Q3 and Q4), and/or that were not written in English.

3.3.3 Level 3 - Full text:

Articles that report empirical or experimental research (with real data), evaluate or propose AI models (-)specific to phishing detection and/or prevention), quantifiable metrics (described in the previous section) were included (see TABLE 4).

Table 4: Reference Filtering by Full Text

Inclusion Criteria (IC)	**Exclusion Criteria (EC)**
IC1. Empirical research with real data.	EC1. No empirical data or evaluation.
IC2. Evaluates or proposes AI models for phishing.	EC2. The technique is not AI nor is it applied to phishing.
IC3. Presents quantifiable metrics (accuracy, recall, F1, etc.).	EC3. No clear quantitative metrics.
IC4. Published in an indexed and ranked journal.	EC4. Not indexed or unclassified journal.
IC5. Published in 2023 or later.	EC5. Published before 2023.
IC6. Article in English.	EC6. Not in English.
IC7. Open access article.	EC7. An article that requires some subscription to be read.

Here, we excluded studies that do not present clear metrics, lack empirical data, have not been indexed, were published before 2023, and/or are not written in English.

3.4 Exclusive Use of Scopus and Justification for the Period 2023-2025

It was decided to search for these studies only in Scopus because it is a highly indexed database in computer science areas, to guarantee the quality and relevance of the articles included in this review.

In addition, it was decided to choose studies published within this time range (2023-2025) because of the advances they demonstrated in Large Language Models (LLMs) and Generative AI (Gen AI). As many of these initial studies were still in exploratory or preliminary stages, and knowing that 2023 marked the beginning of significant advances in these technologies, we prioritised those studies that offered solid empirical results with direct applicability to phishing detection and/or prevention.

3.5 Search Strategy

A search string incorporating the inclusion and exclusion criteria was used: "(TITLE ("phishing detection" OR "phishing prevention" OR "anti-phishing") AND TITLE-ABS-KEY ("machine learning" OR "deep learning" OR "artificial intelligence" OR "natural language processing" OR "neural network" OR "random forest" OR "support vector machine")) AND TITLE-ABS-KEY ("performance" OR "evaluation" OR "accuracy" OR "detection rate" OR "precision" OR "recall" OR "f1-score") AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (SUBJAREA , "COMP")) AND (LIMIT-TO (PUBYEAR , 2023) OR LIMIT-TO (PUBYEAR , 2024) OR LIMIT-TO (PUBYEAR , 2025)) AND (LIMIT-TO (LANGUAGE , "English"))".

3.6 Screening and Selection Process

Using the search string discussed in the previous section, 63 articles were initially identified. We then removed duplicates, reviewed article titles and abstracts, and assessed the full text of the shortlisted articles - studies that passed the second level of screening. After applying all filters, 42 articles were included in this systematic review and recorded in an Excel sheet for detailed analysis (see FIGURE 1).

4. RESULTS

In order to be able to answer the research questions proposed for this review, a thorough systematisation was carried out according to the information present in each of the 42 articles that were selected.

The extraction of data relevant to the case was categorised by the type of AI(s) used for phishing detection and the factors involved in the performance of the models presented or used. Four broad categories of AI techniques applied for phishing detection were identified, grouping diverse approaches with particular characteristics and results.

4.1 RQ1: Machine Learning Techniques Used in Phishing Detection and/or Prevention

TABLE 5 contains studies with techniques such as Random Forest, Decision Trees, SVM, and other classical classification methods.

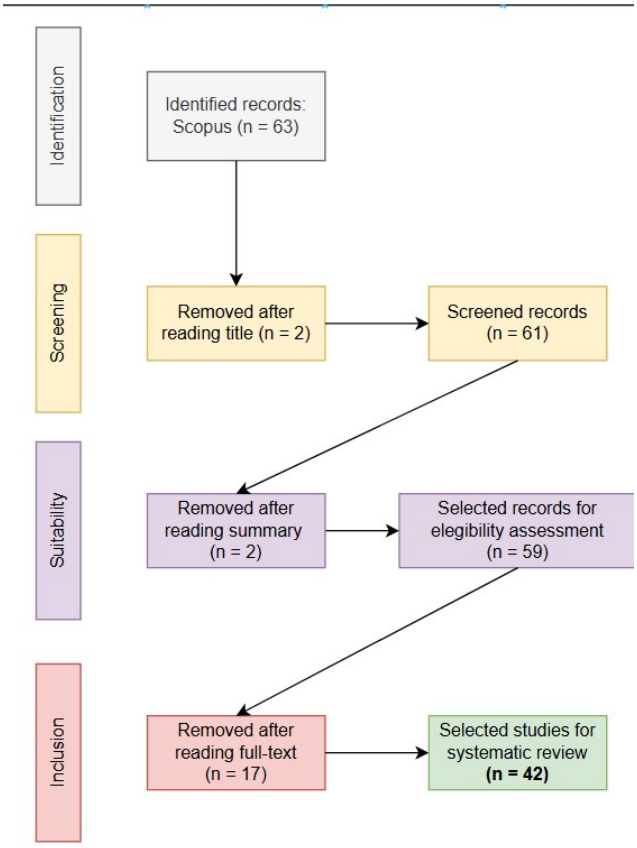

Figure 1: Process and number of filtered articles

Table 5: Studies based on Machine Learning

Ref.	Model/ Technique Used	Dataset Used	How was the model	Evaluation Metrics and Results					Additional Comments
				Model	Accuracy	Precision	Recall	F1-score	
[10]	- Random Forest (RF) - Decision Tree (DT) - Support Vector Machines (SVM) - Naive Bayes (NB) - K-Nearest Neighbors (KNN) - Gradient Boosting (GB)	- PhishTank Dataset, 58,645 URLs - Dataset Availability: Publicly Available	Different classical Machine Learning techniques (RF, DT, SVM, KNN, NB) were compared, evaluating their performance for detecting phishing based on URLs	DT	99.96%	100.00%	99.96%	99.98%	RF, DT and GB showed the best metrics. Ensemble learning models were more robust than individual models.
				RF	99.96%	99.96%	100.00%	99.98%	
				GB	99.96%	100.00%	99.96%	99.98%	

Continued on next page

[11]	- CatBoost - XGBoost - LightGBM	- Kaggle Malicious URLs Dataset, 522,214 URLs. - Dataset Availability: Publicly Available	They used CatBoost, XGBoost and LightGBM to compare their performance in detecting phishing URLs using gradient boosting techniques.	CatBoost XGBoost Light GBM	96.90% 92.10% 95.20%	98.00% 93.00% 96.00%	98.00% 97.00% 98.00%	98.00% 95.00% 97.00%	CatBoost showed the best performance in terms of accuracy and F1-score, with 96.9% accuracy, beating XGBoost and LightGBM in all key aspects.
[12]	- Decision Tree (DT) - Random Forest (RF) - Support Vector Machine (SVM)	- Mendeley Phishing Webpage Dataset, 11,430 URLs - Dataset Availability: Publicly Available	The study uses AWS SageMaker to preprocess the URL dataset and train DT, RF and SVM models to detect phishing on websites.	RF SVM DT	97.00% 94.00% 93.00%	97.00% 95.00% 93.00%	96.00% 96.00% 93.00%	97.00% 97.00% 93.00%	RF showed the best performance with 97% accuracy, using the Cyber Kill Chain concept to improve detection.
[13]	- Gaussian Naive Bayes (GaussianNB) - Random Forest (RF) - LightGBM (LGBM) - Gradient Boosting (GBoost)	- Webpage Phishing Detection Dataset, 24,200 URLs (PhishTank, Alexa, Mendeley Data, University of New Brunswick) - Dataset Availability: Publicly Available	The paper proposes a robust feature selection approach by analysing recent phishing behaviour. Models such as GaussianNB, RF, LightGBM, and Gradient Boosting were trained.	GBoost LGBM RF Gaussian NB	99.84% 99.84% 99.79% 98.67%	100.00% 100.00% 100.00% 97.00%	100.00% 99.00% 100.00% 90.00%	100.00% 100.00% 100.00% 98.00%	LightGBM and Gradient Boosting showed the best performance, with 99.84% accuracy, while GaussianNB was the most resistant to attacks without retraining with 98.67%.
[14]	- Multinomial Naive Bayes (MultNB) - Stochastic Gradient Descent (SGD) - Support Vector Classifier (SVC) - Logistic Regression (LR)	- Mendeley Phishing Website Dataset, 11,055 URLs - Dataset Availability: Publicly Available	The paper used several ML and Deep Learning classifiers, and made comparisons between execution modes (offline, batch, incremental) to detect phishing	MultNB SGD SVC LR	98.55% 98.80% 98.39% 98.26%	76.00% 98.79% 98.19% 99.00%	98.00% 91.00% 87.00% 88.00%	86.00% 95.00% 93.00% 93.00%	We compared the three execution modes: offline, batch, and incremental, finding that the incremental mode is ideal for real-time applications.

Continued on next page

[15] - Decision Tree (DT) - Random Forest (RF) - Extreme Gradient Boosting XGBoost) - Hybrid Model (DT + RF + XGB)	- Kaggle URL Phishing Dataset, 11,054 URLs - Dataset Availability: Publicly Available	Uses a hybrid machine learning model (DT, RF, XGBoost) and Canopy-based feature selection techniques. Evaluated for accuracy, recall, F1-score, and specificity.	DT	94.00%	94.00%	94.00%	94.00%	The hybrid assembly approach (Soft and Hard Voting) outperforms individual models, improving accuracy and reducing training time.
			RF	96.00%	92.00%	93.00%	92.00%	
			XGBoost	96.00%	96.00%	96.00%	96.00%	
			DT + RF + XGB	96.00%	97.00%	96.00%	96.00%	

This section presents the main Machine Learning models identified in the reviewed studies, which have been applied specifically for the detection and/or prevention of phishing attacks. These approaches are characterised by their computational efficiency, ease of implementation and ability to process large volumes of structured data, such as URLs or web page features. Through the comparative analysis of metrics such as accuracy, precision, recall and F1-score, the most effective models within this AI subtype were identified, as well as their training and application conditions. The aim is to highlight not only the most widely used algorithms, but also those that achieved outstanding performance in experimental and real-world contexts. A summary of graphs is presented on FIGURE 2.

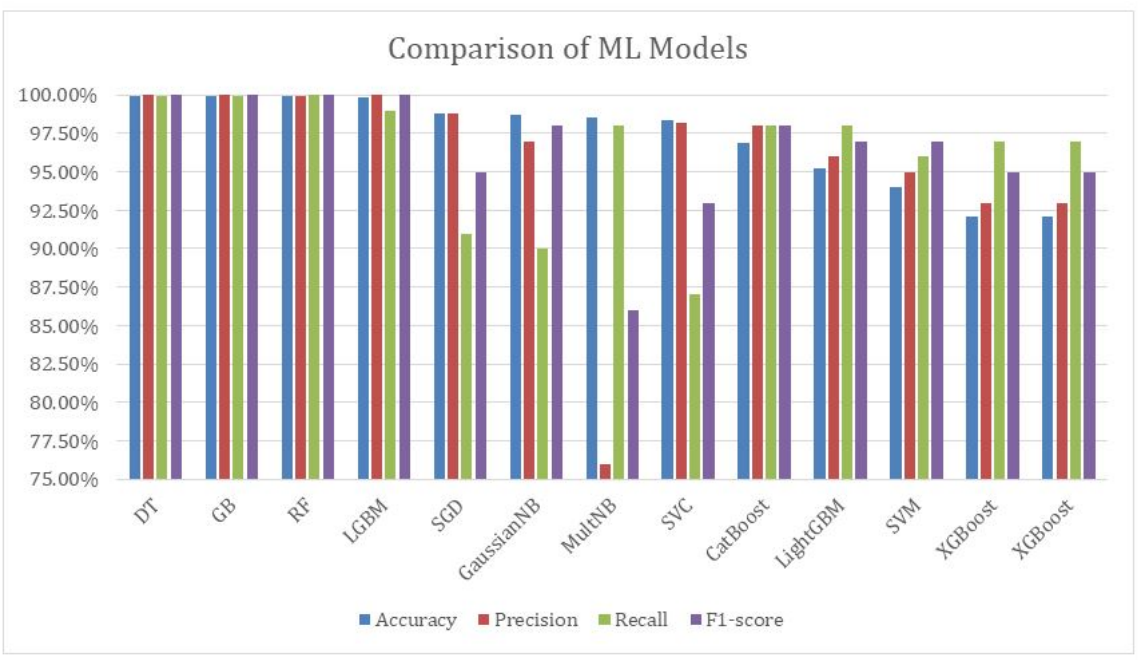

Figure 2: Comparison of ML Models (Sorted by Accuracy in descending order)

4.2 RQ2: Deep Learning Techniques Used in Phishing Detection and/or Prevention

The following section details the Deep Learning (DL) models used in phishing detection and prevention, as summarised in TABLE 6. This research focuses on advanced architectures such as convolutional neural networks (CNNs), long-term memory networks (LSTMs), hybrid models and attention mechanisms. Unlike traditional approaches, DL models have the ability to learn complex representations from unstructured data, such as text from emails, HTML content or temporal sequences of URLs. Thanks to this capability, they have demonstrated remarkable performance in identifying subtle and dynamic patterns that characterise modern phishing attacks. The most relevant models, their configurations, datasets used and the metrics that demonstrate their effectiveness are presented below. A summary of graphs is presented on FIGURE 3.

Table 6: Studies Based on Deep Learning

Ref.	Model/ Technique Used	Dataset Used	How was the model	Evaluation Metrics and Results					Additional Comments
				Model	Accuracy	Precision	Recall	F1-score	
[16]	Temporal Convolutional Network (TCN)	- Kaggle Phishing Data, 23,366 URLs - Dataset Availability: Publicly Available	TCN was used to extract features from URLs and HTML, combining embedding and hand-crafted features.	TCN	99.81%	99.91%	99.55%	99.73%	The TCN model demonstrated high accuracy compared to other DL approaches. Better performance when combining handcrafted features with embedding.
[17]	One Dimension Convolutional Neural Network (1D CNN)	- PhishTank Dataset, 20,000 URLs - Dataset Availability: Publicly Available	A CNN-based model was used for the accurate classification of URLs, differentiating between legitimate and phishing websites	1D CNN	98.77%	98.01%	98.01%	98.10%	1D CNN showed excellent performance in classifying phishing URLs with a well-structured 7-layer model.

Continued on next page

[18]	- Long Short-term Memory (LSTM) - Convolutional Neural Network (CNN) - LSTM-CNN	- PhishTank URL Phishing Dataset, 20,000 URLs - Dataset Availability: Publicly Available	Used a CNN-based approach to extract features from URLs, combined with LSTM and a hybrid LSTM-CNN model for classification	LSTM CNN LSTM-CNN	96.80% 99.20% 97.60%	95.90% 99.00% 96.90%	97.50% 99.20% 98.20%	96.80% 99.20% 97.60%	CNN showed the best performance in terms of accuracy, being more efficient and faster than LSTM and LSTM-CNN. The hybrid LSTM-CNN model also showed robust performance, but with higher computational consumption.
[19]	Bidirectional Encoder Representations from TransformersG and Convolutional Neural Networks (BERT - CNN)	- Kaggle Phishing Dataset, 18,650 URLs - Dataset Availability: Publicly Available (Kaggle)	BERT was used to extract linguistic features from the emails, and CNN for the classification of phishing and legitimate emails.	BERT-CNN	97.50%	97.00%	99.00%	98.00%	The BERT-CNN model showed 97.5% accuracy, demonstrating its effectiveness in detecting phishing emails in enterprise systems.
[20]	Variational Autoencoder and Deep Neural Network (VAE – DNN)	- Kaggle Dataset and ISCX-URL-2016 (99,658 URLs)-Dataset Availability: Publicly Available (Kaggle)	Used a hybrid VAE-DNN model for feature extraction from URLs, reducing dimensionality and then using a DNN for classification.	VAE-DNN	97.45%	97.89%	97.20%	97.54%	The VAE-DNN model showed an accuracy of 97.45%, higher than other DL models by integrating VAE for feature extraction and DNN for classification.

Continued on next page

[21]	Attention-based One-dimensional Convolutional Neural Network and Bidirectional Long Short-Term Memory (Attention-based 1D CNN – BiLSTM)	- KorCCVi v2 2,927 samples - Dataset Availability: Publicly Available	The model uses 1D CNN to extract features from speech transcripts, and then BiLSTM to learn temporal dependencies. In addition, an attention mechanism is implemented to improve the focus on the most relevant features.	1D CNN-BiLSTM	99.32%	99.32%	99.32%	99.31%	The combination of 1D-CNN and BiLSTM enhanced with attention allowed for high accuracy in voice phishing detection.
[22]	Deep Convolutional Neural Network (DCNN)	- UCI Machine Learning Repository Phishing Dataset, 1,353 URLs - Dataset Availability: Publicly Available	DCNN was used to convert feature vectors into images and classify phishing, legitimate and suspicious using a simple convolutional neural network.	DCNN	86.50%	-	-	-	DCNN showed adequate performance, with 86.5% accuracy in classifying phishing URLs, but did not achieve higher accuracy due to the simplicity of the model.
[23]	-Artificial Neural Networks (ANN) - Convolutional Neural Network (CNN) - Recurrent Neural Network (RNN) - Bi-directional RNN (BRNN) - Attention Networks (ATT)	- Phishing Attack Dataset, 5.1 million URLs - Dataset Availability: Publicly Available	Five different Deep Learning architectures: ANN, CNN, RNN, BRNN, and ATT, were used to detect phishing based on URLs	ANN	91.16%	-	-	-	CNN showed the best performance in terms of accuracy. The high quality of the dataset and the independent approach (without using third-party services) made the model efficient in phishing detection.
				CNN	98.74%	99%	99%	99%	
				RNN	95.10%	-	-	-	
				BRNN	94.70%	-	-	-	
				ATT	90.00%	-	-	-	

Continued on next page

[24]	Bidirectional Long Short-Term Memory (BLSTM)	- Enron Corpus Email Dataset, 619,446 emails - Dataset Availability: Publicly Available	A BLSTM model with FastText Word Embeddings was used for email phishing detection.	BLSTM	99.12%	98.43%	99.49%	98.96%	The BLSTM model showed high accuracy in email phishing detection. In addition, the use of FastText for word embeddings allowed for better detection of misspelled or out-of-vocabulary words.
[25]	ResNeXt method and embedded Gated Recurrent Unit through the Jaya optimization method (RNT-J)	- Kaggle Phishing Dataset, 25,539 emails - Dataset Availability: Publicly Available	Used ResNeXt with GRU (Gated Recurrent Unit) and SMOTE for preprocessing, and Jaya optimization for hyperparameter tuning.	RNT – J	98.00%	97.80%	99.00%	98.80%	The RNT-J model outperformed other methods, achieving an improvement of 11%-19% compared to traditional algorithms. The ability to adapt to new phishing attacks is one of its strengths.
[26]	Recurrent Neural Network with Long Short-Term Memory (RNN – LSTM)	- PhishTank Dataset, 10,000 URLs - Dataset Availability: Publicly Available	Used RNN and LSTM in a Deep Learning architecture for phishing detection in cloud environments. A sequential prediction approach was applied to URLs and emails.	RNN-LSTM	98.88%	98.20%	98.40%	98.40%	RNN-LSTM is highly effective for detecting phishing in cloud environments due to its ability to handle sequential and contextual data.

Continued on next page

[27]	Dense forward-backwards Long Short-Term Memory (d – FBLSTM)	- PhishTank MUPD Dataset, 4,574,786 URLs - Dataset Availability: Publicly Available	The model uses a hybrid approach of Dense Network combined with LSTM and CNN for phishing URL detection.	d-FBLSTM	98.45%	98.78%	99.02%	97.85%	The combination of LSTM and CNN in the Dense Network model provides robust phishing detection results. The d-FBLSTM model shows good performance even with synthetic URLs.
[28]	Recurrent Neural Network model post Whale Optimization Algorithm (RNN – WOA)	- Kaggle Phishing Website Detector Dataset, 11,000 URLs - Dataset Availability: Publicly Available	RNN was used for URL phishing detection, with optimisation of the RNN hyper-parameters using the Whale Optimization Algorithm (WOA).	RNN-WOA	92.00%	92.00%	94.00%	93.00%	The optimised model achieved an accuracy of 92%, with WOA showing efficient performance in optimising the RNN.
[29]	Optimised Convolutional Neural Network (CNN) through Adaptive Differential Evolution (JADE) and Chi-square feature selection	- Kaggle Phishing Website Dataset, 11,000 URLs - Dataset Availability: Publicly Available	JADE optimised the hyperparame-ters of the CNN, and Chi-square feature selection was used to select relevant features for phishing detection in e-commerce platforms.	CNN	94.00%	94.00%	94.00%	94.00%	CNN performed well, with a precision of 93%. The use of JADE optimization made the model stable, with lower loss values compared to other models.

Continued on next page

[30] Convolutional Neural Network with a disorderly quantized attention (CNN – DQA) in a phishing detection based on hybrid features (PDHF)	- PhishTank, OpenPhish, Alexa, CNSTC:a. DATA1, 10,000 samplesb. DATA2, 1,172,577 URLs- Dataset Availability: Publicly Available	The PDHF model combines optimal artificial features with deep automatic features to detect phishing websites. A Random Forest classifier is used for classification.	PDHF	99.65%	99.42%	99.40%	99.41%	Compared favourably with other existing models in terms of accuracy and efficiency.
[31] Ensemble learning approach with: - Long Short-Term Memory (LSTM) - Bi-directional LSTM (Bi-LSTM) - Convolutional neural network LSTM (CNN-LSTM)	- GitHub Ethereum-lists dataset, 2,407 addresses - Dataset Availability: Publicly Available	The model uses LSTM, Bi-LSTM and CNN-LSTM to detect phishing attacks in real-time, with an ensemble voting approach. They are trained on a dataset of "x" transactions.	LSTM	98.75%	98.75%	100.00%	99.37%	The ensemble approach of LSTM, Bi-LSTM and CNN-LSTM performed better than other approaches in terms of accuracy and F1-Score.
			CNN-LSTM	98.75%	99.30%	99.86%	99.37%	
			Bi-LSTM	99.17%	99.30%	99.86%	99.58%	
			Ensemble	99.72%	99.86%	99.86%	99.86%	

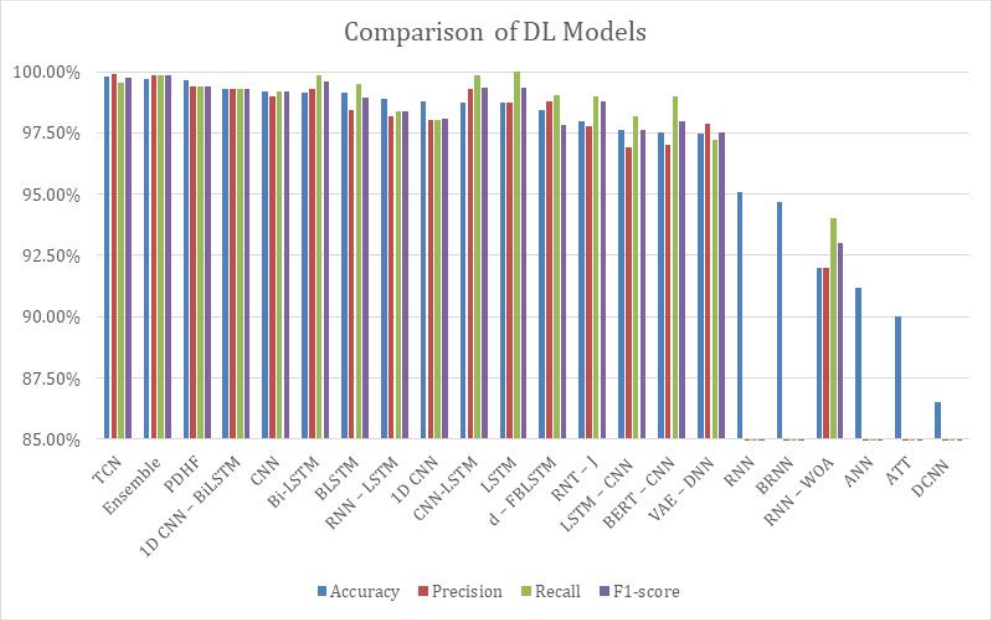

Figure 3: Comparison of DL Models (Sorted by Accuracy descending order)

4.3 RQ3: Hybrid techniques (ML + DL + others) used for phishing detection and/or prevention

This section covers studies that implement hybrid approaches, i.e., combinations of Machine Learning techniques, Deep Learning, and other complementary methods, such as evolutionary algorithms, swarm optimisation or reinforcement learning (see TABLE 7). These models seek to leverage the individual strengths of each approach to improve overall performance, especially in complex scenarios where data is diverse, unbalanced, or difficult to interpret. Through these combinations, more robust, adaptive, and accurate systems are achieved. In the studies analysed, there is a growing trend towards such solutions, which have shown competitive metrics and a high potential for real applications in phishing detection and prevention. The corresponding table summarises the models used, the datasets applied, and the results obtained in terms of effectiveness. A summary of graphs is presented on FIGURE 4.

Table 7: Studies Based on Hybrid Models (ML, DL or other)

Ref.	Model/ Technique Used	Dataset Used	How was the model	Evaluation Metrics and Results					Additional Comments
				Model	Accuracy	Precision	Recall	F1-score	
[32]	Back-Translation (BT) approach for Data Augmentation (DA)	- FSS, NIKL: KorCCVi v2 Dataset, 2,927 samples - Dataset Availability: Publicly Available	The model uses Back-Translation (BT) to augment the dataset and address the unbalanced class problem compared to SMOTE with various ML/DL algorithms for classification.	BT Method SMOTE Method	98.91% 97.23%	98.63% 96.10%	99.29% 98.45%	99.45% 98.65%	BT proved to be more effective than SMOTE in terms of contextual preservation. However, there was variability in performance depending on the language used in translation.
[33]	- Feedforward Neural Network (FNN) - Deep Neural Network (DNN) - Wide and Deep (WD) - TabNet	- Mendeley 2020 Dataset, 10,000 samples - Dataset Availability: Publicly Available	Used a Feedforward Neural Network (FNN) and DNN approach and integrated Wide and Deep Model and TabNet to improve detection performance.	DNN FNN TabNet WD	93.92% 94.27% 93.82% 91.39%	97.00% 97.00% 95.00% 97.00%	97.00% 97.00% 95.00% 97.00%	97.00% 97.00% 95.00% 97.00%	The FNN proved to be the most effective model with superior accuracy. The implementation of TabNet and Wide and Deep also showed good results, although with higher computational consumption.

Continued on next page

[34]	Multilayer Q-Learning with CaspNet on Logistic Bayesian Long Short-Term Memory (LB – LSTM - Mul_Q_capsnet) with Particle Swarm Optimization (PSO)	- Owner's custom dataset, 73,575 URLs - Dataset Availability: Publicly Available	Used a multi-layered Q-Learning approach combined with LB-LSTM to analyse malicious behaviour in social network URLs and optimisation with PSO	LB-LSTM - Mul_Q_capsnet	94.33%	88.67%	98.67%	94.34%	The Q-Learning and LB-LSTM approach was very effective in detecting phishing in social networks. Using PSO to optimise features showed a significant improvement.
[35]	Features selection using Reinforcement Learning method (RL) on:- Logistic Regression (LR)- Extreme Gradient Boosting (XGBoost)- Random Forest (RF)- Decision Tree (DT)	- Mendeley Phishing Dataset, 88,647 instances - Dataset Availability: Publicly Available	Reinforcement Learning (RL) was used to dynamically select the most relevant subset of features. Random Forest (RF) was used as a classifier to compare the performance of feature selection.	LR XGBoost RF DT	91.39% 92.45% 99.07% 95.49%	85.00% 94.00% 98.00% 98.00%	91.00% 95.00% 98.00% 97.00%	88.00% 95.00% 98.00% 98.00%	The use of RL for dynamic feature selection was shown to be highly effective (on the previous models), improving performance compared to other traditional feature selection methods.
[36]	Honey Badger Algorithm with Artificial Neural Network (GHBA-ANN)	- UCI Machine Learning Repository dataset, 1,100 instances - Dataset Availability: Publicly Available	The Honey Badger Algorithm (HBA) was used to optimise ANN training for phishing website detection by combining the optimisation algorithm with a neural network for classification.	HBA-ANN Single ANN	86.01% 83.70%	92.46% 91.37%	83.59% 80.63%	87.79% 85.61%	HBA-ANN showed better performance compared to Single ANN, achieving high accuracy and a remarkable F1-score. The HBA-ANN model proved to be suitable for neural network optimisation and website phishing detection.

Continued on next page

[37]	Chaotic Dragonfly Algorithm (CDA) on: - K-Nearest Neighbours (KNN) - Decision Tree (DT) - Support Vector Machine (SVM)	- UCI Repository Phishing Website Dataset, 11,055 websites - Dataset Availability: Publicly Available	CDA was used for automatic feature selection, followed by KNN, DT and SVM for phishing website classification.	KNN DT SVM	Without DA: 94.24% With DA: 95.32% Without DA: 94.24% With DA: 95.29% Without DA: 90.49% With DA: 94.07%	- - -	- - -	- - -	CDA was very effective in feature selection, improving the accuracy of these classifiers.
[38]	Pyramid Depth – wise Separable – MobileNetV3 and Deformable Convolutional Deformable Residual Neural Network (PDSMV3-DCRNN)	- IEEE: ISCX-URL phishing dataset, 114,400 URLs - Kaggle: URL-Based phishing dataset, 11,054 URLs - Mendeley 2020 dataset, 88,647 instances - Kaggle: Phishstorm, 60,000 URLs - Datasets Availability: Publicly Available	The model uses a combination of techniques such as CWGAN for data balancing, BGGOA for feature selection, and PyDS-MV3 and DCRNNN models for classification.	PDSMV3-DCRNN	99.21%	98.98%	99.05%	99.03%	The model exhibited fast training times (0.11 s) and achieved high accuracy on several datasets compared to existing methods.
[39]	Pheromone-based Graph Embedding Algorithm (PGEA)	- XBlock dataset, 1,259 nodes - Dataset Availability: Publicly Available	The model uses a pheromone-based sampling technique, optimised with a tabu list to improve the capture of temporal and repetitive transactional patterns. The embeddings are trained with word2vec and classified using SVM.	Deepwalk Node2vec Trans2vec Graph2vec PGEA	81.30% 81.30% 87.20% 83.80% 87.70%	70.30% 75.50% 79.40% 79.30% 81.50%	36.10% 38.10% 65.70% 48.36% 65.20%	47.30% 50.50% 71.80% 60.00% 75.40%	It was compared with other methods such as DeepWalk, Node2vec, Trans2vec, and Graph2vec, and showed a ranged 5-7% improvement in performance metrics.

Continued on next page

[40]	Combination of Logistic Regression (LR), Support Vector Machine (SVM) and Decision Tree (DT): LR+SVC+DT (LSD model)	- Kaggle Phishing Dataset, 11,054 URLs - Dataset Availability: Publicly Available	The models were trained on a dataset of phishing and legitimate URLs, with an ensemble learning approach using the hybrid LSD model to improve phishing detection.	DT RF LSD model	95.41% 96.77% 98.12%	95.80% 96.73% 97.31%	96.00% 97.51% 96.33%	95.91% 97.12% 95.89%	The hybrid LSD model with cross-validation and parameter optimisation outperformed the others in terms of accuracy and F1-Score.
[41]	Clustering and Classification Machine Learning methods (CMLM) applying K-Means methods on:- Deep Learning (DL)- Decision Tree (DT)	- OpenPhish Phishing Intelligence Dataset, 11,050 URLs - UCI Machine Learning Repository Dataset 10,000 URLs- Dataset Availability: Publicly Available	Hybrid model combines feature selection with clustering (K-means) and classification (DT and DL) methods.	DL DT	99.53% 97.20%	99.56% 100.00%	99.78% 96.70%	99.67% 98.32%	The hybrid approach showed exceptional results, outperforming other hybrid phishing detection methods in terms of accuracy and F1-score.
[42]	Particle Swarm Optimization (PSO) and Grey Wolf Optimizer (GWO) algorithms on XGBoost and Random Forest (RF)	- Mendeley Phishing Website Dataset, 88,647 instances- Dataset Availability: Publicly Available	The model uses PSO and GWO to reduce the dataset features, then applies XGBoost and RF for classification of phishing URLs.	RF XGBoost	92.00% 89.00%	83.60% 80.20%	85.20% 85.60%	84.40% 82.80%	The use of PSO and GWO allowed for a reduction in the number of features and improved model performance in terms of accuracy and processing time.

Continued on next page

[43] Convolutional Neural Network with Long Short-Term Memory (CNN-LSTM)	- PhishTank and Kaggle Malicious URL datasets, 651,191 URLs - Dataset Availability: Publicly Available	The browser extension sends URLs to the server where they are classified using a deep learning model (CNN) and machine learning (LR, DT, RF, SVM).	LR DT RF SVM CNN CNN-LSTM	95.61% 96.30% 96.45% 96.18% 98.07% 94.23%	97.02% 96.93% 97.71% 97.84% 98.07% 94.23%	89.95% 92.13% 91.79% 90.87% 98.07% 94.23%	93.35% 94.47% 94.66% 94.23% 98.07% 94.23%	Proposed extension installed in browsers such as Google Chrome and Microsoft Edge to detect malicious URLs in real time, using CNN for classification.
[44] Securing ENcrypted mulTIparty computatIoN for Enhanced data privacY and phishing detection (SENTINEY)	- Owner's custom phishing email dataset, 10,000,000 e-mails - Dataset Availability: Publicly Available	The approach uses SMPC (Secure Multi-Party Computation) together with ML and string-matching techniques to detect phishing in encrypted emails.	SENTINEY	99.50%	94.20%	96.50%	97.10%	The SENTINEY system combines SMPC with unsupervised learning and string matching to ensure privacy and efficiency in detecting phishing attacks.
[45] Three layers hybrid framework stacking: - Decision Trees (DT) - Random Forest (RF) - Logistic Regression (LogReg)	- Mendeley Phishing Website Dataset, 80,000 websites - Dataset Availability: Publicly Available	The research uses hybrid models with multiple website features (URL, HTML, DOM Tree) and evaluates the use of *stacking* functions to improve robustness and detection effectiveness.	Hybrid Frame-work	97.44%	96.32%	96.81%	96.56%	The hybrid approach showed faster and more effective performance, highlighting its robustness against evasion attacks.

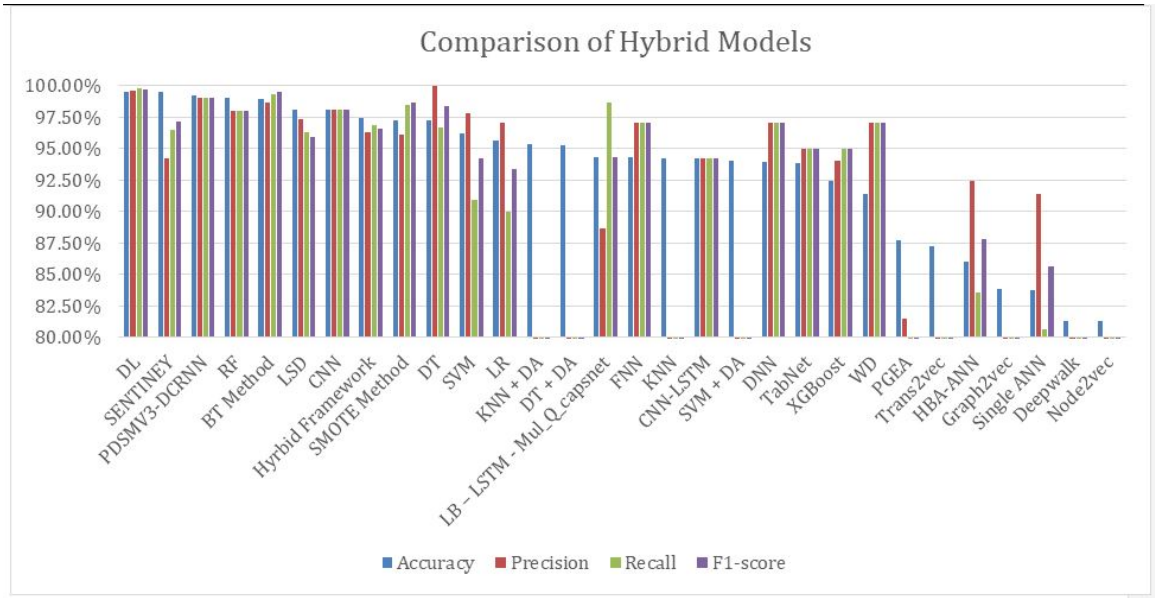

Figure 4: Comparison of Hybrid Models (Sorted by Accuracy in descending order)

4.4 RQ4: LLMs and Gen AIs Llms and Gen Ais Techniques Used for Phishing Detection and/or Prevention

The last category analysed corresponds to large-scale language models (LLMs) and generative artificial intelligence (Gen AI) techniques, whose application in phishing detection has gained prominence in recent years. As shown in TABLE 8, these models (including variants such as GPT-3.5, Claude 2, BERT, and MobileBERT) have been used to process natural language, identify malicious content in emails, generate synthetic data to improve the training of other models, and even detect semantic patterns in URLs. Unlike traditional approaches, LLMs have an advanced ability to understand linguistic context, which makes them promising tools for detecting increasingly sophisticated attacks that are difficult to classify using conventional rules. This section describes the main studies that have employed these techniques, the application contexts and the results achieved. A summary of graphs is presented on FIGURE 5.

Table 8: Studies using LLM/Gen AI

Ref.	Model/ Technique Used	Dataset Used	How was the model	Evaluation Metrics and Results					Additional Comments
				Model	Accuracy	Precision	Recall	F1-score	
[46]	Large Language Models (LLMs) as: - Llama-3-8b-instruct - Gemini 1.5 Pro - Phi-3-medium-4k-instruct - Llama-3-70b-instruct - GPT-3.5	- Owner's custom phishing Email Dataset, 40 emails - Dataset Availability: Proprietary	Various LLMs were employed to detect phishing in AI-generated emails, with a focus on models such as Llama-3	llama-3-8b-instruct	97.50%	95.24%	100.00%	97.56%	Llama-3-8b-instruct proved to be the most effective model, with 97.50% accuracy and good performance in detecting AI-generated phishing emails.
				Gemini 1.5 Pro	92.50%	86.96%	100.00%	93.02%	
				Phi-3-medium-4k-instruct	92.50%	90.48%	95.00%	92.68%	
				llama-3-70b-instruct	90.00%	90.00%	90.00%	90.00%	
				GPT-3.5	90.00%	94.44%	85.00%	89.47%	
[47]	Methodologies: - Term Frequency-Inverse Document Frequency (TF-IDF) - Word2Vec (W2V) - Bidirectional Encoder Representations from Transformers (BERT) To assess performance on: - Logistic Regression (LogReg) - Decision Tree (DT) - Random Forest (RF) - Multilayer Perceptron (MLP)	- Kaggle phishing Mail Dataset, 18,650 emails - Dataset Availability: Publicly Available	BERT was used for phishing email classification with advanced NLP capabilities; Word2Vec and TF-IDF were used for feature extraction.	TF-IDF+ LR	97.00%	97.00%	98.00%	98.00%	BERT proved to be the most accurate model, outperforming Word2Vec and TF-IDF in email phishing detection tasks.
				TF-IDF+ DT	95.00%	97.00%	93.00%	98.00%	
				TF-IDF+ RF	97.00%	98.00%	97.00%	98.00%	
				TF-IDF+ MLP	98.00%	98.00%	98.00%	98.00%	
				W2V+ LR	96.00%	97.00%	97.00%	97.00%	
				W2V+ DT	92.00%	93.00%	94.00%	93.00%	
				W2V+ RF	97.00%	97.00%	98.00%	97.00%	
				W2V+ MLP	98.00%	98.00%	98.00%	98.00%	
				BERT	99.00%	98.00%	99.00%	99.00%	

Continued on next page

[48]	MobileBERT with Covariance Matrix Adaptation Evolution Strategy (CMA-ES)	- Kaggle phishing email dataset, 3,508? emails - Dataset Availability: Publicly Available	MobileBERT was used to extract deep semantic features, while CMA-ES optimised the hyperparame-ters of the model to improve accuracy in email classification	Mobile-BERT + CMA-ES	95.00%	97.00%	95.00%	96.00%	MobileBERT and optimisation with CMA-ES proved to be effective in the context of Semantic Web, achieving good performance with the Kaggle dataset.
[49]	Framework for Large Language Model (LLM) – Guided Phishing Text Generation and Detection on: - Transformer (Tr) - Multimodal Transformer (MmTr) - Multimodal Transformer with cross–attention (MmTrCA)	- DeepAI Phishing Email Dataset, 865 emails - Dataset Availability: Publicly Available	Used LLM to generate phishing text with prompt chaining and a heuristic algorithm to improve the quality of the generated text, and used a transformer model with cross-attention for multimodal analysis.	Tr MmTr MmTrCA	92.35% 94.73% 96.91%	95.14% 97.98% 96.47%	92.31% 89.29% 96.47%	93.70% 93.43% 96.47%	The generative approach using LLM enhanced with prompt chaining and heuristics showed high effectiveness in phishing detection in dynamic scenarios and in conjunction with multimodal analysis of text and URLs.
[50]	Bidirectional Encoder Representations from Transformers (BERT)	- Kaggle Phishing Dataset, 18,650 emails - Dataset Availability: Publicly Available	BERT was used for contextual and semantic feature extraction, and Semantic Web technologies were integrated to improve phishing detection in virtual reality environments.	BERT	95.00%	96.00%	97.00%	96.50%	The integration of BERT with semantic technologies has shown excellent performance in detecting phishing in dynamic environments such as virtual realities.

Continued on next page

[51] - Claude 2 – Prompt engineering Fine-tuning LLMs: - bloom–560m - distilgpt2 - openai–gpt - baby–llama–58m - gpt2 - gpt2–medium	- Mendeley Phishing URL Dataset, 11,430 URLs - Dataset Availability: Publicly Available	LLMs prompt engineering and fine-tuning techniques are compared for the detection of phishing URLs.	bloom–560m	92.40%	92.06%	92.80%	92.43%	The proposal with Fine-tuned GPT-2 performed better than the proposal made with Claude 2-prompt engineering.
			Claude 2	92.90%	94.78%	90.80%	92.74%	
			distilgpt2	95.90%	94.22%	97.80%	95.98%	
			openai–gpt	96.10%	96.01%	96.20%	96.10%	
			baby–llama–58m	96.60%	96.05%	97.20%	96.62%	
			gpt2	96.60%	95.87%	97.40%	96.63%	
			gp2–medium	97.30%	97.78%	96.80%	97.29%	

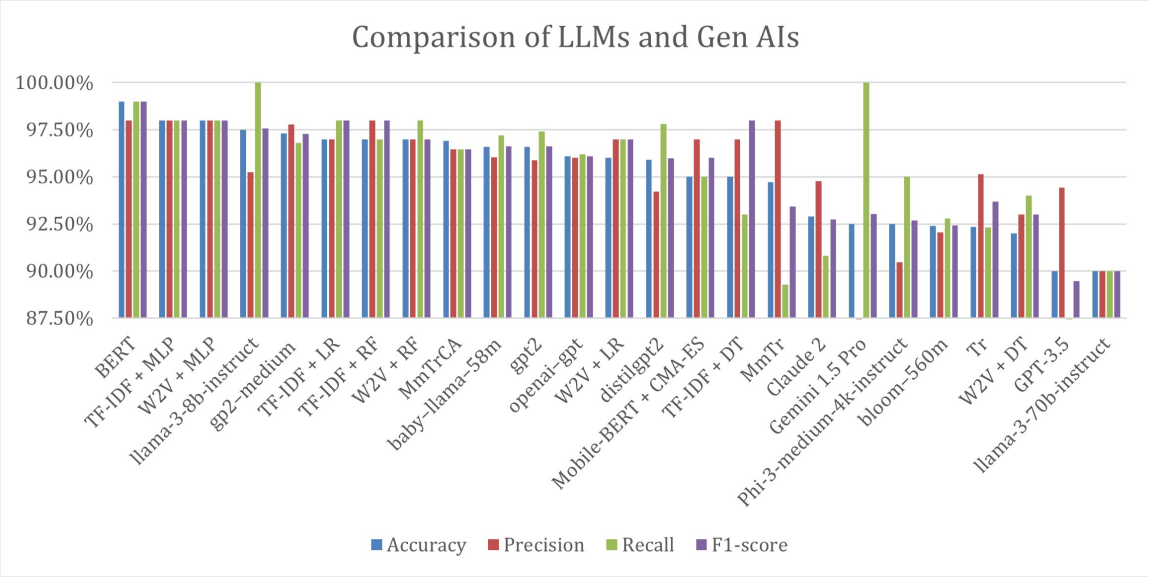

Figure 5: Comparison of LLMs and Gen AIs (Sorted by Accuracy descending order)

5. DISCUSSION

The present systematic review aims to analyse the current landscape on the use of Artificial Intelligence techniques, including Machine Learning, Deep Learning, and Generative Models, in the detection and prevention of phishing attacks. The results obtained, organised according to the research questions posed, are presented below.

5.1 Relevant Models

A wide variety of approaches were identified that employ both classical ML algorithms and more complex DL architectures and state-of-the-art language models. The most frequently used models include:

5.1.1 Classical ML algorithms:

Such as Random Forest, Support Vector Machines (SVM), Decision Trees, XGBoost, and Ligth-GBM. These models are still widely used due to their computational efficiency and ease of implementation, especially in resource-constrained contexts [10], [12], [13], [35].

5.1.2 Deep learning models:

Convolutional neural networks (CNN), recurrent networks (RNN), LSTM, BiLSTM, and hybrid models such as CNN-LSTM stand out. These architectures show a high learning capacity in URL classification tasks, emails, or malicious behaviour patterns [23], [31].

5.1.3 Hybrid models:

The integration of ML and DL has been used to improve overall accuracy by combining the best of both approaches. An example of this is the URLGuard model, which achieves an accuracy of 96% by integrating various classifiers [15].

5.1.4 Generative models (Gen AI/LLMs):

An emerging trend has been identified in the use of language models such as GPT-2, Llama or gpt-2-medium, both in their fine-tuned form and through prompt engineering. These models have been used to detect fraudulent text, generate synthetic examples and enrich unbalanced datasets, obtaining F1-scores above 97% in some cases [46], [51]

This finding suggests that the field (anti-phishing AI techniques) is evolving, and that researchers are combining different techniques and methodologies to adapt to the increasing sophistication of phishing attacks.

5.2 Comparison of Techniques and Results

Regarding the most effective AI algorithms and techniques for detecting and preventing phishing attacks, one of the most relevant findings of this systematic review is that there is no single model that can be considered superior in all contexts. The effectiveness of each approach varies significantly depending on the type of data analysed, the application environment and the methodology employed in each research. However, when looking at the most consistent results, a clear trend is identified: models based on Deep Learning and Generative AI tend to outperform traditional algorithms, especially in complex tasks and with large volumes of data.

Among the techniques that stand out for their high performance are convolutional neural networks (CNNs), Long Short-Term Memory (LSTM) networks and hybrid models that combine both architectures. The study by [38] achieved an accuracy of 99.21% using an optimised DCRNN architec-

ture, while [30] reported an accuracy of up to 99.65% with PDHF model, indicating a highly reliable detection capability in real-world contexts.

Also worth mentioning is the TCN model implemented by [16], which achieved an accuracy of 99.81%, demonstrating that even less common architectures within Deep Learning can deliver remarkable results when properly designed and trained with well-structured datasets.

On the other hand, more recent advances in Gen AIs or LLMs have also begun to stand out. The study by [51] shows that a gpt-2-medium model tuned specifically for the task of phishing detection can achieve an F1-score of 97.29%, outperforming even untrained versions that only use prompt engineering. This shows the potential of Gen AI not only for content generation tasks, but also as a robust classification and detection tool.

On the other hand, traditional Machine Learning algorithms, such as Random Forest, XGBoost or LightGBM, while still performing well, tend to have certain limitations in more complex or unstructured scenarios. However, in contexts where well-labelled and balanced data is available, these algorithms can be highly effective. A representative example is that of [10], who achieved 99.96% accuracy using Random Forest and Gradient Boosting, with significantly shorter training times compared to deep learning models.

We determined that the most effective models share certain characteristics: they are able to capture complex relationships between variables, they adapt well to different types of input (text, URL, transactions, audio), and they have been validated in both simulated and real environments. Moreover, many of the best-performing models not only stand out for their accuracy but also for their stability and adaptability, which is essential in such a dynamic environment as cybercrime.

Here, it is confirmed that the most promising approach is one that integrates multiple techniques, leverages recent advances in generative AI, and adapts to the specific context of the problem. The key is not only the algorithm per se, but its strategic implementation and the quality of the dataset on which it is trained.

The top three models/techniques with the best accuracy were compiled per table (taken from TABLE 5 - TABLE 8) and grouped by AI type. These are shown below in TABLE 9 - TABLE 12 and visually represented in FIGURE 6 - FIGURE 9.

Table 9: Top 3 ML Models

Ref	Model	Accuracy	Precision	Recall	F1-Score
[11]	CatBoost	96,90%	98,00%	98,00%	98,00%
[10]	DT	99,96%	100,00%	99,96%	99,98%
[13]	GaussianNB	98,67%	97,00%	90,00%	98,00%

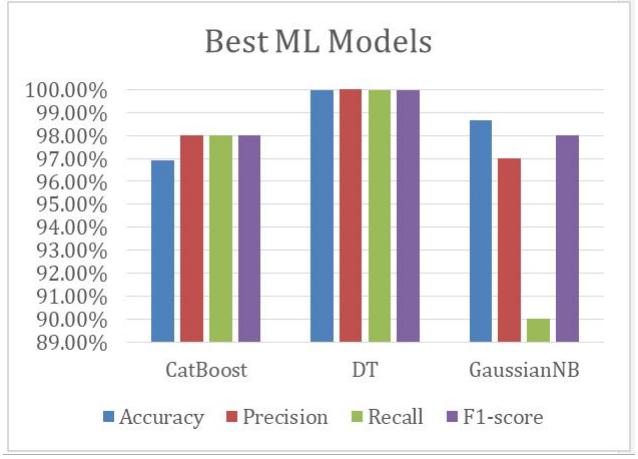

Figure 6: Graphic representation of TABLE 9

Table 10: Top 3 DL Models

Ref	Model	Accuracy	Precision	Recall	F1-Score
[16]	TCN	99,81%	99,91%	99,55%	99,73%
[31]	Ensemble	99,72%	99,86%	99,86%	99,86%
[30]	PDHF	99,65%	99,42%	99,40%	99,41%

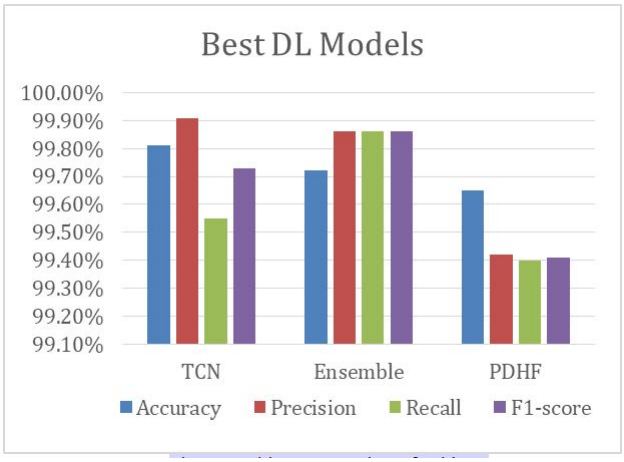

Figure 7: Graphic representation of TABLE 10

Table 11: Top 3 Hybrid Models

Ref	Model	Accuracy	Precision	Recall	F1-Score
[41]	DL	99,53%	99,56%	99,78%	99,67%
[44]	SENTINEY	99,50%	94,20%	96,50%	97,10%
[38]	PDSMV3-DCRNN	99,21%	98,98%	99,05%	99,03%

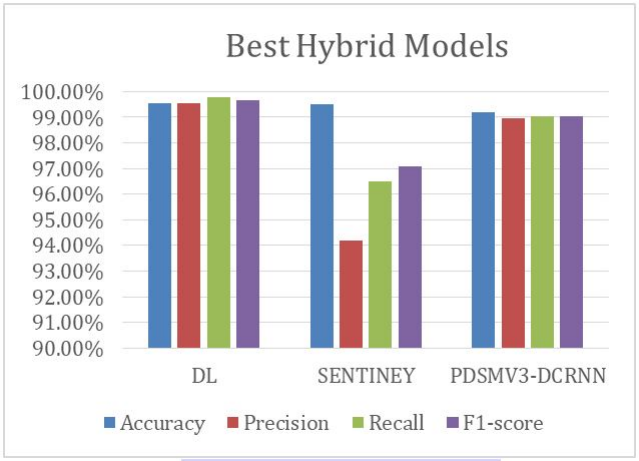

Figure 8: Graphic representation of TABLE 11

Table 12: Top 3 LLMs and Gen AIs

Ref	Model	Accuracy	Precision	Recall	F1-Score
[47]	BERT	99,00%	98,00%	99,00%	99,00%
	TF-IDF + MLP	98,00%	98,00%	98,00%	98,00%
	W2V + MLP	98,00%	98,00%	98,00%	98,00%

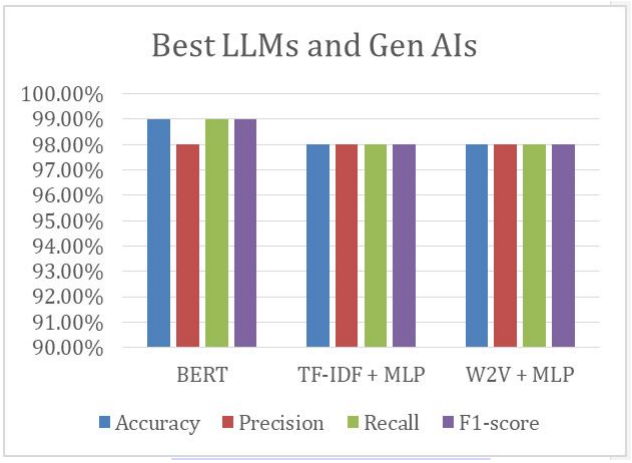

Figure 9: Graphic representation of TABLE 12

5.3 Factors Affecting Model Performance

The systematic review showed that the performance of artificial intelligence models used for phishing attack detection does not depend solely on the type of algorithm implemented. Research agrees that there are multiple structural, contextual, and technical factors that have a significant impact on their effectiveness. Recognising and understanding these elements is essential not only to optimise current results but also to guide the development of more robust, adaptable, and sustainable solutions over time.

One of the most relevant factors (and repeatedly pointed out in the studies analysed) is the quality and balancing of the dataset used. Many models have very high accuracy metrics, but when analysing the recall or false negative rate, it is found that they have difficulties in correctly detecting phishing attempts in unbalanced classes. To counter this limitation, some researchers applied oversampling techniques, such as SMOTE or synthetic data generation using generative models. The work of [32] is a prominent example: they used multilingual back-translation and oversampling to improve detection in voice messages, achieving an F1-score of 98.91%.

Another critical factor is the type of data used to train and evaluate the models. The best performances were observed in models applied to malicious URLs and emails, which are structured and widely available data. In contrast, emerging modalities such as vishing (voice phishing) or fraud in blockchain environments still present significant challenges, largely due to the limited availability of representative datasets and the higher level of complexity of these signals. For example, the work of [31], while achieving high accuracy (99.86%) in blockchain, did so with an expensive ensemble model and very specific data, making it difficult for widespread application.

Model architecture and hyperparameter tuning also play a determining role. Models such as LSTMs or transformers can achieve exceptional levels of accuracy, but they require careful training, proper variable selection, and fine-tuning of their parameters to avoid problems of overfitting or underestimation. In this regard, [46] showed that a fine-tuned llama-3-8b-instruct model outperformed a simple prompt engineering model by a wide margin, even working with the same dataset.

Another important factor is the model implementation environment. Some studies evaluated their proposals in real or semi-real conditions, integrating them into browser extensions, cloud systems or messaging platforms. These works showed not only good results in traditional metrics but also greater stability and scalability. [43], for example, implemented their CNN model in a web browser, achieving an accuracy of 98.07% in real time, which represents an added value in terms of practical applicability.

Finally, the level of computational resources available must be considered, an aspect that often goes unnoticed but can be decisive for the actual adoption of a technology. More complex models, such as those based on graphs or large transformers, tend to require specialised hardware and long training times. On this point, studies such as [39] warn that, although their model achieved an accuracy close to 88%, its applicability in limited environments is low due to their technical demands.

Taken together, these factors allow us to understand that the effectiveness of AI models against phishing is not limited to choosing “the best algorithm”, but is the result of an ecosystem where multiple dimensions are involved: the data, the infrastructure, the methodological approach and the

context of use. As these variables align, it is possible to move towards detection systems that are more robust, dynamic, and adaptable to the new forms of attack that are constantly emerging.

When comparing the best performing models in this systematic review, a clear trend emerges: Deep Learning (DL)-based models slightly outperformed traditional Machine Learning (ML) models in terms of accuracy. This difference is especially evident in the top models identified:

The TCN (Temporal Convolutional Network) model achieved an accuracy of 99.81%, while the best ML model, Random Forest, obtained 99.96%, although the latter was applied on a highly balanced and structured dataset.

However, the DL models also showed greater stability in more complex contexts, such as HTML, rich text, and sequential data processing, which could explain their competitive or superior performance in several cases.

This relative superiority of DL models could be explained by several factors observed in the studies:

5.3.1 Volume and complexity of the data used:

Several DL models were applied to large or multimodal datasets, such as:

- PhishTank and Kaggle, with tens or hundreds of thousands of URLs.
- Enterprise email datasets, with long text, semi-structured or HTML content.

Deep architectures, such as CNN, LSTM or TCN, have a greater capacity to learn hierarchical and non-linear representations from large volumes of data, while classical Machine Learning algorithms can become inefficient or less accurate when the relationships between features are complex or implicit.

A clear example is the model applied by [23] (CNN) that used more than 5.1 million URLs from the PhishTank dataset, and managed to stabilise its performance on multiple metrics with an accuracy of 98.74%. This size allowed it to avoid overfitting, unlike other studies that used datasets with less than 10,000 samples, where the accuracy was similar or even higher (such as TCN with 20,000 URLs and 99.81%), which may hide a lack of generalisability.

5.3.2 Machine learning capability of temporal and contextual relations:

CNN and TCN models showed better performance on URLs because they contain repetitive patterns and predictable structures (such as length, symbol usage, domain position), which can be efficiently captured by convolutions.

In contrast, models applied to email analysis (such as BLSTM with FastText or BERT) excelled at interpreting contextual semantics, typos, or covert intentions, which is typical of free text.

5.3.3 Integration of embeddings and advanced representation techniques:

While many ML models worked with traditional vectors (such as TF-IDF or one-hot encoding), DL models used dense embeddings (such as Word2Vec, FastText or proprietary embeddings), which better encode the semantics and structure of the language or URL. This was especially evident in studies where techniques such as:

- Character-level embedding + TCN.
- Word embeddings + BLSTM.
- CNN applied on vectors converted into images.

These techniques allow the model to better "understand" the content and generalise even in the face of syntactic or formatting variations in the input data.

5.3.4 Inherent limitations of classical ML models:

Although models such as Random Forest, XGBoost or LightGBM performed excellently on structured and balanced datasets, they have limitations when dealing with:

- Unstructured or high-dimensional data (such as HTML, free text, or speech).
- Need for manual feature extraction (feature engineering), which may not capture deep relationships between variables.
- Less adaptability in contexts where patterns evolve dynamically (e.g., AI-generated phishing emails).

5.3.5 Fine tuning and regularisation techniques:

DL studies frequently applied hyperparameter optimisation techniques, regularisation (Dropout, BatchNorm), and ensemble strategies that contributed to mitigating overfitting and improve performance in cross-validations. In contrast, many ML studies applied basic settings, without advanced training techniques or robust validation.

In several studies, it is not reported whether cross-validation or testing in domains other than the original dataset was applied. For example, the TCN model of [16] achieved 99.81% accuracy with a balanced dataset of only 20,000 URLs, which raises doubts about its generalisability. Without clear evidence of regularisation or robust splitting of the dataset, it is reasonable to assume that there is a risk of overfitting, especially if not validated out-of-domain or with recent phishing data.

Finally, the studies reviewed and their results highlights different important areas constituting current challenges and possible future studies on the identification and impediment of phishing by using AI such as:

5.4 Quality Limitations and Datasets Balance

Though there are high reached precisions by some models, there is still a challenge in false-negative detections (specially in unbalanced datasets). It is needed deeper research on strong techniques to handle unbalance and creation of better representative and varied datasets [16], [39], [42].

5.5 Limited Available Data of Emerging Modalities on phishing

There is a remarkable lack of representative datasets for new phishing modalities such as vishing or blockchain environment frauds, making it harder to develop and validate efficient models in these areas [31].

5.6 Robustness and Adaptability against Evolving Threats

The nature of cybersecurity dynamic, especially on the emergence of phishing AI-generated generated, demands better adaptability models and their semantic representation for detecting new continuous attacking ways. The variability of models' performance in function of external factors such as language translation for the data augmentation is a challenge too [21].

5.7 Optimization and Model Generalisation Complexity

Achieving complex models (DL, LLms) optimal performance and reliable generalization requires a careful truning of hyperparameters and advanced regularization techniques to avoid overfitting. The lack of reports on cross-validation or external domains tests in some of the studies, propose question marks on the true models' generalizability [16], [18], [31], [46].

5.8 Challenges on Practice and Real-Time Implementation

Though there have been explored browser extensions and cloud systems for real-time detection, the fluent integration in real operative environments of complex models, guaranteeing big-scale stability and efficiency, still being an area that requires more development and research [26], [43].

6. CONCLUSIONS

From the analysis of the 42 selected studies between 2023 and 2025, it is clear that artificial intelligence has become a key tool to tackle phishing. More than 50 models applied to this problem were identified, and the most striking feature is that 85% of them achieved an accuracy of more than 95%, and around 30% even exceeded 99%. This shows that while there is no perfect model, there are approaches that are making a difference. For example, [16] achieved an *accuracy* of 99.81%

with TCN and the BERT model of [47], applied to AI-generated phishing emails, achieved 99%, which has outperformed other studies.

We also saw that more traditional models such as Random Forest or CatBoost, are still useful, especially when working with well-organised data. However, when it comes to analysing emails, URLs with more complex structures or new forms of phishing, such as voice or blockchain, Deep Learning and generative models (such as gpt-2-medium or BERT) proved to be more efficient and adaptable.

An important point we found is that the performance of these models depends not only on the algorithm, but on several factors: how the data is constructed, whether it is balanced, how well variables are selected, or even whether modern techniques such as embeddings or hyperparameter settings are used. The best results often come from clever combinations of all of these.

Finally, while the results reported in recent literature are promising, it is important to note that extremely high metrics should be interpreted with caution. Many of them could be influenced by optimistic experimental practices and do not necessarily reflect the performance of models in dynamic and non-equilibrium environments. The research community should reinforce the use of more robust validations, cross-domain tests and datasets with conditions closer to operational reality.

7. RECOMMENDATIONS

According to the findings compiled in the presented systematic review, the use of AI techniques for phishing detection is constantly growing; however, there are still important technical gaps that justify further development work in order to increase the effectiveness, applicability, and understanding of current models. First, it is advisable to conduct research on emerging phishing modalities, such as vishing, smishing, and attacks on social networks and blockchain, which exhibit minimal coverage in the current literature. This requires the generation of multi-channel datasets, as well as the development of multi-modal deep learning models capable of operating in real time, especially on mobile platforms. For the latter, the use of CNN+BiLSTM is recommended, which should be trained with synthetic data generated by generative AI. Secondly, it is essential to address the development of interpretable versions of AI because many of the current models are black boxes. We recommend the use of tools such as LIME and SHAP to justify their predictions, as well as the design of user-evaluable interfaces to enable audits. Likewise, we suggest the exploration of generative models that are not only restricted to detection, but also contribute to the simulation of advanced phishing campaigns, enabling the generation of information for the production of more and better data, the development of robustness tests and the strengthening of training. Another priority recommendation is the creation of public, balanced and multilingual datasets, as the scarcity of data in languages other than English limits the use of the models. The generation of inter-institutional agreements for the construction of datasets in underrepresented languages, such as Spanish, is recommended in order to encourage use and development in local contexts. On the other hand, the development of lightweight and efficient models that facilitate the use of biometrics in organisations with low infrastructure is recommended, as they allow the evaluation of performance in real operational conditions. Finally, the addition of language generation models to validation evaluators is suggested to analyse the behaviour of classifiers against LLM-generated mails emulating fraudulent mails.

References

[1] https://www.ic3.gov/AnnualReport/Reports/2023_IC3Report.pdf

[2] https://www.ibm.com/think/x-force/2024-x-force-threat-intelligence-index

[3] N. Rébé, Artificial Intelligence: Crime, War, and Justice. Ethics International Press. 2023.

[4] Mohri M, Rostamizadeh A, Talwalkar A. Foundations of Machine Learning. 2nd ed. MIT Press. 2018.

[5] Berghout E, Fijneman R, Hendriks L, de Boer M, Butijn BJ. Advanced Digital Auditing. Cham: Springer International Publishing. 2023.

[6] Chakraborty U, Roy S, Kumar S. Rise of Generative AI and ChatGPT: Understand how Generative AI and ChatGPT are transforming and reshaping the business world (English Edition). BPB Publications. 2023.

[7] Alammar J, Grootendorst M, Large HO. Language Models: Language Understanding and Generation. 1st ed. O'Reilly Media Incorporated. 2024.

[8] Ting KM. Confusion Matrix. In: Sammut C, Webb GI, editors. Encyclopedia of Machine Learning. US. Boston: Springer. 2011:209.

[9] Vujovic ŽĐ. Classification Model Evaluation Metrics. Int J Adv Comput Sci Appl. 2021;12:599-606.

[10] Albishri AA, Dessouky MM. A Comparative Analysis of Machine Learning Techniques for URL Phishing Detection. Eng Technol Appl Sci Res. 2024;14:18495-18501.

[11] Odeh A, Al-Haija QA, Aref A, Taleb AA. Comparative Study of CatBoost, XGBoost, and LightGBM for Enhanced URL Phishing Detection: A Performance Assessment. J Internet Serv Inf Sec. 2023;13:1-11.

[12] Rashid SH, Abdullah WD. Enhanced Website Phishing Detection Based on the Cyber Kill Chain and Cloud Computing. Indones J Electr Eng Comput Sci. 2023;32:517-529.

[13] Omar AR, Taie S, Shaheen M. From Phishing Behavior Analysis and Feature Selection to Enhance Prediction Rate in Phishing Detection. Int J Adv Comput Sci. Appl. 2023;14.

[14] Jennifer Dsouza DJ, Rodrigues AP, Fernandes R. Multi-Modal Comparative Analysis on Execution of Phishing Detection Using Artificial Intelligence. IEEE Access. 2024;12:163016-163041.

[15] Paithane PM. URL Guard: A Holistic Hybrid Machine Learning Approach for Phishing Detection. Int J Inf Eng Electron Bus. 2025;17:95-110.

[16] Aljofey A, Bello SA, Lu J, Xu C. Comprehensive Phishing Detection: A Multi-Channel Approach With Variants TCN Fusion Leveraging URL and HTML Features. J Netw Comput Appl. 2025;238:104170.

[17] Aldakheel EA, Zakariah M, Gashgari GA, Almarshad FA, Alzahrani AI. A Deep Learning-Based Innovative Technique for Phishing Detection in Modern Security With Uniform Resource Locators. Sensors. 2023;23:4403.

[18] Alshingiti Z, Alaqel R, Al-Muhtadi J, Haq QE, Saleem K, et al. A Deep Learning-Based Phishing Detection System Using CNN LSTM and LSTM-CNN. Electronics. 2023;12:232.

[19] Gupta BB, Gaurav A, Arya V, Attar RW, Bansal S, et al. Advanced Bert and CNN-Based Computational Model for Phishing Detection in Enterprise Systems. CMES Comput Model Eng Sci. 2024;141:2165-2183.

[20] Prabakaran MK, Meenakshi Sundaram PM, Chandrasekar AD. An Enhanced Deep Learning-Based Phishing Detection Mechanism to Effectively Identify Malicious URLs Using Variational Autoencoders. IET Inf Secur. 2023;17:423-440.

[21] Boussougou MMK, Park DJ. Attention-Based 1D CNN-BiLSTM Hybrid Model Enhanced With Fast Text Word Embedding for Korean Voice Phishing Detection. Mathematics. 2023;11:3217.

[22] Kulkarni AD. Convolution Neural Networks for Phishing Detection. Int J Adv Comput Sci Appl. 2023;14:15-19.

[23] Sahingoz OK, BUBEr E, Kugu E. DEPHIDES: Deep Learning Based Phishing Detection System. IEEE Access. 2024;12:8052-8070.

[24] Wolert R, Rawski M. Email Phishing Detection With BLSTM and Word Embeddings. Int J Electron Telecommun. 2023;69:485-491.

[25] Alsubaei FS, Almazroi AA, Ayub N. Enhancing Phishing Detection: A Novel Hybrid Deep Learning Framework for Cybercrime Forensics. IEEE Access. 2024;12:8373-8389.

[26] Senouci O, Benaouda N. Enhancing Phishing Detection in Cloud Environments Using RNN-LSTM in a Deep Learning Framework. J Telecommun Inf Technol. 2025:1-9.

[27] Aldo Tenis AA, Santhosh R. Modelling an Efficient URL Phishing Detection Approach Based on a Dense Network Model. Comput Syst Sci Eng. 2023;47:2625-2641.

[28] Gupta BB, Gaurav A, Attar RW, Arya V, Alhomoud A, et al. Optimized Phishing Detection With Recurrent Neural Network and Whale Optimizer Algorithm. Comput Mater Continua. 2024;80:4895-4916.

[29] Gaurav A, Chui KT, Arya V, Attar RW, Bansal S, et al. Optimized AI-Driven Semantic Web Approach for Enhancing Phishing Detection in E-Commerce Platforms. Int J Semant Web Inf Syst. 2024;20:13.

[30] Zhu E, Cheng K, Zhang Z, Wang H. PDHF: Effective Phishing Detection Model Combining Optimal Artificial and Automatic Deep Features. Comput Secur. 2024;136:103561.

[31] Ogundokun RO, Arowolo MO, Damaševičius R, Misra S. Phishing Detection in Blockchain Transaction Networks Using Ensemble Learning. Telecom. 2023;4:279-297.

[32] Boussougou MKM, Hamandawana P, Park DJ. Enhancing Voice Phishing Detection Using Multilingual Back-Translation and SMOTE: An Empirical Study. IEEE Access. 2025;13:37946-37965.

[33] Nayak GS, Muniyal B, Belavagi MC. Enhancing Phishing Detection: A Machine Learning Approach With Feature Selection and Deep Learning Models. IEEE Access. 2025;13:33308-33320.

[34] Khan AI, Unhelkar B. An Enhanced Anti-Phishing Technique for Social Media Users: A Multilayer Q-Learning Approach. Int J Adv Comput Sci Appl. 2024;15:18-28.

[35] Patil SS, Shekokar NM, Iyer SC. Design of Intelligent Feature Selection Technique for Phishing Detection. IIUM Eng J. Jan. 2025;26:254-277.

[36] Mohamad MA, Ahmad MA, Mustaffa Z. Hybrid Honey Badger Algorithm With Artificial Neural Network (HBA-ANN) for Website Phishing Detection. Iraqi J Comput Sci Math. 2024;5:671-682.

[37] Alshammari G, Alshammari M, Almurayziq TS, Alshammari A, Alsaffar M. Hybrid Phishing Detection Based on Automated Feature Selection Using the Chaotic Dragonfly Algorithm. Electronics. 2023;12:2823.

[38] Prasad YB, Dondeti V. PDSMV3-DCRNN: A Novel Ensemble Deep Learning Framework for Enhancing Phishing Detection and URL Extraction. Comput Secur. 2025;148:104123.

[39] Xiao S, Zhang L, Tian Z, Su S, Qiu J, et al. Pheromone-Based Graph Embedding Algorithm for Ethereum Phishing Detection. Comput Netw. Apr. 2025;260:111123.

[40] Karim A, Shahroz M, Mustofa K, Belhaouari SB, Joga SR. Phishing Detection System Through Hybrid Machine Learning Based on URL. IEEE Access. 2023;11:36805-36822.

[41] Jazyah YH, Al Shalabi L. Phishing Detection Using Clustering and Machine Learning. IAES Int J Artif Intell. 2024;13:4526-4536.

[42] Hamdan A, Tahboush M, Adawy M, Alwada'n T, Ghwanmeh S, et al. Phishing Detection Using Grey Wolf and Particle Swarm Optimizer. Int J Electr Comput Eng syst. 2024;14:5961-5969

[43] Dam ML, Hung HD, Minh CH, Sy Vu Q, Tran TN. Real-Time Phishing Detection Using Deep Learning Methods by Extensions. Int J Electr Comput Eng. 2024;14:3021-3035.

[44] Hendaoui F, Hendaoui S. SENTINEY: Securing Encrypted Multi-Party Computation for Enhanced Data Privacy and Phishing Detection. Expert Syst Appl. 2024;256:124896.

[45] van Geest RJ, Cascavilla G, Hulstijn J, Zannone N. The Applicability of a Hybrid Framework for Automated Phishing Detection. Comput Secur. 2024;139:103736.

[46] Zhang J, Wu P, London J, Tenney D. Benchmarking and Evaluating Large Language Models in Phishing Detection for Small and Midsize Enterprises: A Comprehensive Analysis. IEEE Access. 2025;13:28335-28352.

[47] Al Tawil AA, Almazaydeh L, Qawasmeh D, Qawasmeh B, Alshinwan M, et al. Comparative Analysis of Machine Learning Algorithms for Email Phishing Detection Using TF-IDF Word2Vec and BERT. Comput Mater Continua. 2024;81:3395-3412.

[48] Zhou L, Gaurav A, Arya V, Attar RW, Bansal S, et al. Enhancing Phishing Detection in Semantic Web Systems Using Optimized Deep Learning Models. Int J Semant Web Inf Syst. 2024;20:1-13.

[49] Shammi L, Shyni CE. Generative AI-Based Phishing Text Generation Using Hybrid Prompt Design With Heuristic Algorithm for Multimodal Phishing Detection. Int J Intell Eng Syst. 2025;18:485-503.

[50] Zhou L, Gaurav A, Alhalabi WA, Arya V, Alharbi E. Integrating AI and Semantic Web Technologies for Robust Phishing Detection in Virtual Realities. Int J Semant Web Inf Syst. 2025;21:1-19.

[51] Trad F, Chehab A. Prompt Engineering or Fine-Tuning? A Case Study on Phishing Detection With Large Language Models. Mach Learn Knowl Extr. 2024;6:367-384.