iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.79156

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Phishing Detection Techniques: A Systematic Review

Poorvi[1], Yogita Thareja[2]
[1]*Research Scholar, Vivekananda Institute of Professional Studies Technical Campus, Delhi*
[2]*Assistant Professor, Vivekananda Institute of Professional Studies Technical Campus, Delhi*

Abstract: Phishing attacks continue to function as the most widespread and destructive cyber threats which exist throughout today's digital environment. Cybercriminals use phishing attacks to trick users into disclosing their confidential information which includes their login details and financial information and their personal identification data. The advancement of phishing techniques through AI-generated content and deepfake technology and social engineering via multiple channels has rendered traditional detection methods which rely on blacklists and heuristics increasingly ineffective. This paper presents a systematic review of phishing detection techniques which authorities developed between 2015 and 2025. The study found that transformer-based models BERT and RoBERTa achieved a detection accuracy of 99 percent which is significantly better than classical machine learning baselines. The study identified several critical open challenges which include adversarial evasion and zero-day attacks and dataset limitations and multilingual coverage. The study identified specific directions for future research.
Keywords: Phishing Detection, Cybersecurity, Machine Learning, Deep Learning, URL Analysis, Email Security, Neural Networks.

I. INTRODUCTION

The internet has experienced rapid expansion which has resulted in a corresponding surge of cybercriminal activities. Among the many forms of cyber threats, phishing stands out as one of the oldest yet most persistently effective attack vectors. Phishing is a social engineering attack which enables an adversary to impersonate a trusted entity that includes banks government agencies and popular websites to deceive victims into revealing their private financial and personal data.

The scale of phishing is staggering. The Anti-Phishing Working Group (APWG) recorded over 1.3 million unique phishing attacks in Q1 2024 alone [1], and the Verizon Data Breach Investigations Report identifies phishing as the leading initial breach vector, which occurs in more than 36 percent of worldwide incidents [2]. The global cost of cybercrime is projected to exceed $10.5 trillion annually by 2025. Organizations across finance, healthcare, and government become daily targets which result in data breaches and financial losses and reputational damage.

Phishing becomes more dangerous because it can change its methods to defeat security systems. The initial phishing attacks used emails with bad writing and showed fake websites that were easy to identify as scams. Modern phishing attacks use authentic-looking spoofed websites together with AI-generated content and deepfake audio and video technology and spear-phishing which targets single individuals and smishing which uses SMS to steal information and vishing which uses voice calls to deceive users and multi-channel deception techniques. The development of new technologies has made traditional defense methods ineffective because they need to be used together with machine learning and deep learning detection systems which research organizations are currently developing according to [4][6].

The paper presents a systematic review which analyses various phishing detection methods to determine their detection performance while using common datasets and showing major obstacles which need to be addressed in upcoming research. The research objectives are: (1) survey and classify major phishing detection techniques from 2015–2025; (2) evaluate the advantages and limitations of each approach; (3) compare ML and DL model performance; (4) identify key datasets and evaluation metrics; and (5) highlight open research challenges. The rest of the paper is divided into multiple sections which include Section 2 about phishing taxonomy and real-world effects section 3 about detection methods section 4 about datasets and evaluation methods section 5 about comparative research and section 6 about existing problems and section 7 which provides the conclusion.

II. BACKGROUND AND PHISHING TAXONOMY

A. Definition and Types of Phishing Attacks

Phishing exploits cognitive biases — urgency, authority, and fear — to compel victims to act before critically evaluating the legitimacy of a communication [3]. Technically, attacks combine domain spoofing, content cloning, and social engineering. The major attack variants in the current taxonomy include:

1) Email Phishing: Bulk emails impersonating reputable organizations — the most prevalent form.
2) Spear Phishing: Highly targeted, personalized attacks directed at specific individuals or organizations using prior reconnaissance, making them significantly more convincing than generic phishing.
3) Whaling: Spear phishing targeting senior executives or board members, with potentially severe strategic and financial consequences.
4) Smishing / Vishing: Attacks delivered via SMS text messages or voice/VoIP calls respectively, exploiting the informality and trust associated with these channels.
5) Pharming: DNS corruption or hosts file manipulation redirecting users from legitimate URLs to malicious sites transparently.
6) Clone Phishing: Replication of a previously legitimate email with authentic links replaced by malicious ones, re-sent from a spoofed address.
7) Browser-in-the-Browser (BiTB): Simulation of a browser pop-up window within the current page to mimic SSO portals, stealing credentials from victims who believe they are authenticating legitimately.

B. Attack Lifecycle and Real-World Impact

A phishing campaign follows a structured lifecycle: reconnaissance (OSINT-based target profiling), domain registration and infrastructure setup (lookalike domains, cloned pages, SSL certificates), lure deployment (emotionally manipulative messages via email or SMS), credential harvesting, and monetization via dark web sales or direct exploitation [3]. Notable case studies underscore the financial severity: a fraud campaign targeting Google and Facebook (2013–2015) caused losses exceeding $100 million. IBM's Cost of a Data Breach Report places the average phishing-related breach cost at $4.9 million. The COVID-19 pandemic and the widespread adoption of generative AI tools have dramatically accelerated attack volume and sophistication [13].

III. PHISHING DETECTION TECHNIQUES

A. List-Based Approaches

Blacklists — maintained by services such as Google Safe Browsing, PhishTank, and OpenPhish — store verified phishing URLs and enable near-instantaneous hash-based lookup, making them ideal for browser-level and email gateway integration [5]. Whitelists of trusted domains complement this by permitting known-good traffic. The fundamental limitation of list-based methods is their reactive nature: a phishing site must first be reported, verified, and indexed before it can be blocked — a process taking hours or days during which the site may victimize thousands of users. Zero-day attacks are inherently undetectable, and attackers routinely exploit rapid domain rotation and URL shorteners to evade blacklist coverage.

B. Heuristic-Based Approaches

Heuristic methods use manually engineered rules derived from known phishing characteristics to evaluate whether a URL or web page is malicious. Feature domains include URL analysis (string length, special character counts, subdomain depth, presence of IP addresses, HTTPS status), page content analysis (external resource loading, hidden iframes, suspicious form actions), and network metadata (WHOIS domain age, DNS resolution patterns, hosting provider reputation) [15]. Unlike blacklists, heuristic methods can detect novel phishing pages exhibiting known suspicious patterns without requiring prior reporting. However, they incur high false positive rates — newly registered legitimate domains may trigger rules incorrectly — and require continuous manual updating as attackers adapt specifically to evade documented heuristics.

C. Visual Similarity Approaches

Visual similarity detection is grounded in the observation that phishing websites are designed to visually replicate their target entities. Approaches include screenshot-based pixel comparison, brand logo detection using object recognition models, layout analysis via DOM structure comparison, and perceptual hashing algorithms robust to minor visual modifications [6]. These methods are effective against impersonation attacks targeting major financial institutions and social networks but are computationally expensive, require maintained and updated reference databases of legitimate site appearances, and are ineffective against text-based phishing vectors (e.g., phishing email bodies, vishing) that involve no visual web content.

D. Machine Learning Approaches

ML classifiers learn discriminative patterns from labelled datasets without requiring manual rule specification. The most studied algorithms in the phishing detection literature include Naive Bayes (efficient for text-heavy email classification), Decision Trees,

Random Forest (the most frequently used classifier, appearing in the majority of reviewed studies due to its robustness and accuracy [7]), Support Vector Machines (effective in high-dimensional feature spaces), and Gradient Boosting variants such as XGBoost and LightGBM [4]. Feature engineering draws from URL lexical properties, WHOIS and DNS metadata, HTML structural features, and network traffic behavioural patterns [10]. ML models achieve 97–99% accuracy on benchmark datasets and offer interpretability via feature importance scores. Their principal limitation is the reliance on manually engineered features, which requires ongoing maintenance as attack patterns evolve, and performance degradation on temporally shifted datasets.

E. Deep Learning Approaches

DL architectures learn hierarchical feature representations directly from raw input data, largely eliminating the need for manual feature engineering. The most prominent architectures in the phishing detection literature are: 1D-CNNs for character-level URL analysis, achieving 99.4% accuracy by learning local syntactic patterns indicative of malicious URLs [11]; LSTM and RNN variants for sequential modelling of URL strings and email text; CNN-LSTM hybrids that combine local and sequential pattern recognition; and transformer-based models (BERT, RoBERTa) for email classification. A systematic evaluation across 14 models and 10 datasets found BERT achieving 98.99% and RoBERTa 99.08% accuracy [9] — the highest reported results in the reviewed literature [12]. Despite their accuracy, DL models present practical challenges: they require large volumes of labelled training data, significant computational resources for training and inference, and offer limited interpretability, complicating enterprise deployment.

F. Emerging Techniques

Several novel approaches are reshaping the phishing detection landscape. Large Language Models (LLMs) enable zero-shot and few-shot classification without task-specific training data, potentially addressing the dataset dependency problem [13]. Generative Adversarial Networks (GANs) are used to synthesize adversarial phishing examples for adversarial robustness training, hardening classifiers against evasion attacks. Blockchain-based distributed blacklists offer tamper-resistant, decentralized phishing site repositories. Explainable AI (XAI) methods — particularly SHAP (SHapley Additive exPlanations) — attribute classification decisions to specific input features, improving transparency and auditability. A study published in Nature Scientific Reports [14] demonstrated that SHAP-based feature selection also yielded accuracy improvements by eliminating noisy features from URL-based classifiers.

IV. DATASETS AND EVALUATION METRICS

A. Commonly Used Datasets

High-quality labelled datasets are essential for developing and evaluating phishing detection systems. The most widely used sources in the reviewed literature include:

1) PhishTank: A community-driven repository of verified phishing URLs. A systematic review of 80 studies found PhishTank was used in 53 (66%) of them, making it the de facto standard phishing data source [7].
2) Alexa Top Sites: The most visited websites globally, used as the legitimate URL source in paired training datasets.
3) OpenPhish: A commercial automated phishing feed offering broader and more timely coverage than PhishTank, including campaigns not yet community-reported.
4) UCI ML Repository Phishing Dataset: A widely cited benchmark containing 11,055 instances with 30 engineered URL and webpage features, suitable for feature-based ML evaluation.
5) APWG eCrime Datasets: Timestamped historical phishing records made available to qualified researchers, enabling temporal analysis of attack evolution.

B. Evaluation Metrics

The standard evaluation metrics applied in phishing detection research are: Accuracy (proportion of correctly classified instances); Precision (proportion of phishing detections that are truly phishing); Recall (proportion of actual phishing instances correctly identified); F1-Score (harmonic mean of Precision and Recall, balancing false positives and false negatives); False Positive Rate (FPR — the proportion of legitimate URLs misclassified as phishing, critically important in security contexts where false positives block legitimate traffic and erode user trust); and AUC-ROC (Area Under the Receiver Operating Characteristic Curve, providing a threshold-independent summary of discriminative ability) [4][5].

C. Dataset Challenges

Despite available resources, systemic dataset challenges constrain the generalizability of reported results [5][7][8]. Class imbalance is pervasive: real-world traffic is dominated by legitimate URLs, yet many benchmarks are artificially balanced. Temporal staleness is a significant but underexplored issue — phishing URLs are ephemeral (sites often taken down within hours), meaning classifiers trained on static snapshots may not generalize to live deployments. Over-reliance on PhishTank introduces sampling bias, as its coverage is non-uniform across geographies, languages, and campaign types. Most critically, the near-complete absence of non-English phishing datasets renders trained models poorly suited to the global phishing landscape.

V. COMPARATIVE ANALYSIS

Table 1 presents a structured comparison of the major detection categories across five practical dimensions: detection accuracy, processing speed, zero-day attack coverage, computational cost, and interpretability. These dimensions reflect the key trade-offs that determine suitability for different deployment contexts.

Table 1: Comparative Analysis of Phishing Detection Techniques

Technique	Accuracy	Speed	Zero-Day	Comp. Cost	Interpretability
Blacklist	Medium	Very High	None	Very Low	High
Heuristic	Moderate	High	Low	Low	High
Visual Similarity	High	Low	Medium	High	Medium
Random Forest (ML)	97–99%	High	Medium	Medium	Medium
SVM (ML)	96–98%	Medium	Medium	Medium	Medium
CNN (DL)	~99%+	Medium	High	High	Low
BERT/RoBERTa (DL)	99%+	Low–Med	High	Very High	Very Low
LLM-Based	Very High	Low	Very High	Very High	Very Low

The analysis reveals a clear performance-cost trade-off. Blacklist and heuristic approaches offer minimal overhead and maximum interpretability, suitable for high-throughput browser-level filtering, but are wholly inadequate as standalone defences given their inability to handle zero-day attacks. Classical ML approaches — particularly Random Forest — represent the optimal practical middle ground: high accuracy (97–99%), interpretability via feature importance, and moderate resource requirements, making them well-suited for organizational-scale deployments. DL models, especially BERT and RoBERTa, achieve peak accuracy and strong generalization to novel attacks but demand substantial computational infrastructure and offer limited auditability. Effective real-world systems typically employ layered architectures: blacklists provide speed for known threats, while ML or DL models handle novel and zero-day cases. Emerging LLM-based and adversarially trained systems are promising for the most challenging scenarios — zero-day and multilingual attacks — but remain early-stage in terms of production readiness.

VI. CHALLENGES AND FUTURE RESEARCH DIRECTIONS

A. Adversarial Evasion

As ML-based detectors have proliferated, attackers have begun crafting inputs specifically designed to evade them — inserting innocuous characters, altering whitespace, or embedding obfuscation layers to reduce classifier confidence scores without altering the phishing intent of the page [13]. This adversarial arms race represents one of the most pressing open challenges in the field. Proposed defences include adversarial training (augmenting datasets with GAN-generated adversarial examples), ensemble detection architectures (combining multiple independent classifiers whose joint outputs are harder to simultaneously fool), and adversarial perturbation detection as a secondary classification task. Future research should establish standardized adversarial benchmarks to enable rigorous comparative evaluation of robustness.

B. Zero-Day Attacks and URL Obfuscation

Newly registered, single-use phishing domains and URL shorteners (bit.ly, tinyurl) defeat both blacklists and URL feature-based classifiers by concealing the true destination until resolution [5]. Effective responses include real-time content analysis post-rendering, behavioural features derived from page DOM and JavaScript execution, and temporal domain features (registration recency, domain age) as strong predictors of phishing intent. Lightweight classifiers capable of analysing previously unseen domains in real-time with minimal latency overhead represent a key design target for future research.

C. AI-Powered and Multilingual Phishing

The widespread adoption of generative AI tools has eliminated the grammatical errors and generic phrasing that previously served as reliable phishing indicators which now enable convincing personalized phishing at industrial scale [13]. The detection methods need to develop new capabilities which can use stylometric analysis and authorship attribution to identify AI-generated content. The dataset ecosystem which focuses exclusively on English creates challenges for detecting phishing attacks that target users who speak all major languages. The research priorities should focus on using language-agnostic features which include URL structure and visual layout and network metadata together with XLM-RoBERTa multilingual pre-trained models. International research collaboration will be essential for assembling representative multilingual phishing corpora [8].

D. Explainability and Enterprise Trust

Deployment of complex ML and DL models in enterprise security environments requires interpretable auditable classification decisions which help analysts respond to security incidents while meeting regulatory standards and establishing operational trust. SHAP and LIME provide post-hoc feature attribution which remains absent from most phishing detection systems as their primary design elements according to reference [14]. Future systems should treat explainability as a primary design objective — not an afterthought — and explore human-in-the-loop architectures that combine high model accuracy with analyst oversight for ambiguous or high-stakes decisions.

VII. CONCLUSION

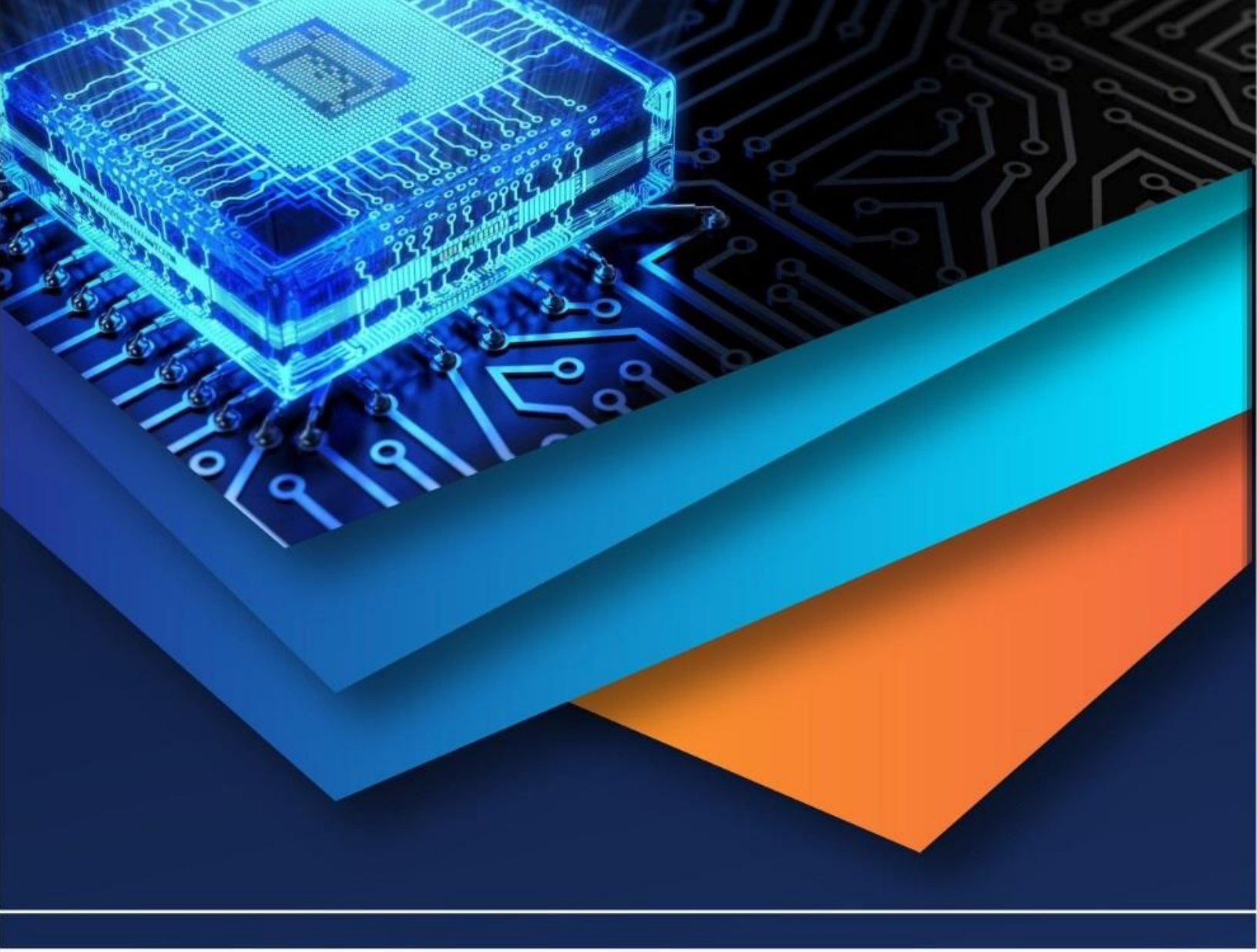
The paper conducted an extensive systematic review of all phishing detection methods which include rule-based systems and heuristic systems and visual systems and machine learning systems and deep learning systems. The key findings are threefold. First, organizations need to implement hybrid detection systems which combine blacklist speed with machine learning and deep learning capabilities because no single detection system can handle all practical situations. Second, transformer-based deep learning models which include BERT and RoBERTa have achieved state-of-the-art accuracy because they reach accuracy levels between 98 and 99 percent on benchmark phishing email datasets. Third, the most critical open challenges are adversarial evasion, the rapid adoption of generative AI by phishing actors, multilingual dataset scarcity, and the absence of explainability as a first-class system requirement.

The research objectives from Section 1, which include five research goals, have been accomplished through multiple research activities. The research team completed three tasks which involved surveying major detection methods and assessing their respective strengths and weaknesses and the study performed a detailed comparison of machine learning and deep learning results and the research team identified essential datasets and evaluation metrics and they specified ongoing research areas that need exploration. The security research community needs to focus on creating five types of detection systems because generative AI technology enables criminals to conduct advanced phishing attacks which require protective measures for our digital world.

REFERENCES

[1] APWG. Phishing Activity Trends Report, Quarterly 2024. Available: apwg.org
[2] Verizon. Data Breach Investigations Report, 2023. Available: verizon.com/business/resources/reports/dbir/
[3] Z. Alkhalil et al., "Phishing Attacks: A Recent Comprehensive Study and a New Anatomy," Front. Comput. Sci., vol. 3, p. 563060, 2021.
[4] A. Sharma & N. Gupta, "A Comprehensive Survey on ML-Based Phishing Detection," ACM Comput. Surv., vol. 56, no. 4, pp. 1–28, 2024.
[5] Journal of Applied Security Research, 2025. "A Survey on Phishing Attack Taxonomy, Detection Techniques, Datasets, and Security Measures." tandfonline.com
[6] Springer AI Review, Dec 2024. "Staying Ahead of Phishers: A Review of Recent Advances in Phishing Detection." link.springer.com
[7] Journal of King Saud University – Computer and Information Sciences, 2023. "A Systematic Literature Review on Phishing Website Detection Techniques." sciencedirect.com
[8] Frontiers in AI, 2025. "AI in Phishing Detection: A Bibliometric Review." frontiersin.org
[9] Applied Sciences, 2025. "In-Depth Analysis of Phishing Email Detection: Evaluating ML and DL Models Across Multiple Datasets." mdpi.com
[10] M. A. Tamal et al., "Unveiling Suspicious Phishing Attacks," Front. Comput. Sci., vol. 6, p. 1428013, 2024.
[11] I. Haq et al., "Lightweight 1D-CNN Model for URL-Based Phishing Detection," IEEE Access, vol. 12, pp. 102394–102406, 2024.

[12] S. Kumar et al., "Transformer-Based Approaches for Email Phishing Detection," IEEE Access, vol. 12, pp. 109823–109836, 2024.
[13] arXiv, 2025. "Evolution of Phishing Detection with AI: A Comparative Review." arXiv:2507.07406
[14] Nature Scientific Reports, 2025. "Explainable Phishing Website Detection for Secure and Sustainable Cyber Infrastructure."
[15] A. Basit et al., "A ML Approach for Phishing Detection Using URL Features," Comput. Secur., vol. 132, pp. 102948–102957, 2023.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY