RESEARCH ARTICLE OPEN ACCESS

AI-Based Phishing URL Detection System: A Multi-Source Intelligence Approach with Machine Learning Classification

Selvamani C *, J. Savitha**
*(Department of Information Technology, Dr. N.G.P Arts and Science College, Coimbatore, Tamil Nadu, India
Email: cselvamani593@gmail.com)
** (Professor, Department of Information Technology, Dr. N.G.P Arts and Science College, Coimbatore, Tamil Nadu, India
Email: savithaj@drngpasc.ac.in)

--************************----------------------------------

ABSTRACT:

Phishing attacks remain one of the most pervasive cybersecurity threats, with the FBI's Internet Crime Complaint Centre reporting over 880,000 complaints and losses exceeding $12.5 billion in 2023 alone. This paper presents a comprehensive AI-based phishing URL detection system that combines machine learning classification with multi-source domain intelligence signals. The proposed system integrates four key analytical components: ML-based classification using ensemble methods, WHOIS domain analysis for registration intelligence, SSL certificate inspection for cryptographic validation, IP address geolocation and reputation analysis, and a unified risk scoring mechanism. By synthesizing these diverse data sources, the system achieves robust detection capabilities that address the limitations of traditional blacklist-based approaches and single-method detection systems. We evaluate the system's architecture, feature engineering methodology, and performance characteristics, demonstrating how multi-source intelligence fusion enables accurate, real-time phishing detection with low false positive rates.

Keywords — **Phishing Detection, Machine Learning, Ensemble Methods, WHOIS Analysis, SSL Certificate Validation, Risk Scoring, Cybersecurity.**

--************************----------------------------------

I. INTRODUCTION

The internet has become an essential societal utility, providing opportunities for both legitimate and illegitimate users, with cyberattacks, particularly phishing Uniform Resource Locator (URL) attacks, emerging as a significant cybersecurity concern especially with the increasing adoption of Artificial Intelligence (AI) by malicious actors, as the exponential growth of AI-driven phishing URL attacks presents new challenges for cyberspace security since attackers now leverage sophisticated techniques to create convincing and personalized fraudulent URLs that bypass traditional detection mechanisms. Phishing attacks exploit human vulnerabilities by tricking individuals into exposing sensitive information through deceptive URLs that mimic legitimate websites, primarily targeting personnel, e-commerce platforms, government agencies, healthcare units, and financial organizations, with IBM reporting that phishing remains the most common initial attack factor responsible for 41% of cybersecurity breaches globally, exemplified by a widely reported 2023 case where attackers used ChatGPT-like tools to generate sophisticated phishing emails that bypassed enterprise-grade filters causing a significant breach at a European financial institution. Traditional detection methods such as rule-based systems and blocklists are losing effectiveness as phishing threats become more sophisticated since these approaches are inherently reactive—they can only detect threats that have been previously identified and catalogued—and they fail to detect zero-day phishing attacks which are newly launched threats that have not yet been identified or blacklisted by existing security models. To address these limitations, researchers and practitioners have increasingly turned to machine learning approaches that can identify phishing attempts based on inherent characteristics rather than known signatures, and among various ML algorithms, Random Forest has emerged as particularly well-suited for phishing detection due to its ability to handle high-dimensional feature spaces, resistance to overfitting, built-in feature selection, and interpretability through feature importance metrics. This paper presents a phishing URL detection system that leverages Random Forest classification within a multi-source intelligence framework, combining Random Forest classification as an ensemble of decision trees analysing URL patterns and lexical features with automatic feature importance ranking, WHOIS domain analysis for registration intelligence including domain age, registrant information, and expiration dates, SSL certificate inspection for cryptographic validation examining certificate validity, issuer reputation, and chain integrity, IP address detection for geolocation and reputation analysis of hosting infrastructure, and a unified risk scoring mechanism that synthesizes signals from all components. By integrating these diverse intelligence sources with Random Forest's robust classification capabilities, the system achieves more reliable detection than any single method could provide alone while maintaining the interpretability necessary for user trust and security operations

II. RELATED WORK

Random Forest has been extensively validated as an effective algorithm for phishing URL detection, with the algorithm's ensemble nature of constructing multiple decision trees on bootstrapped training samples and aggregating their predictions providing inherent protection against overfitting while maintaining high accuracy on complex classification tasks. Tang and Mahmoud (2021) provided a comprehensive survey of ML-based solutions for phishing website detection, examining various algorithms including decision trees, random forests, support vector machines, and neural networks, with their analysis highlighting that Random Forest consistently ranks among top-performing algorithms particularly when feature interpretability is important. Recent comparative studies have demonstrated Random Forest's superiority in phishing detection contexts, with Phis Guard, an AI-powered detection system, reporting that while Boost achieved marginally higher accuracy of 96.2% versus 95.8%, Random Forest provided better interpretability through its native feature importance metrics and required less hyperparameter tuning, while Harnick et al. (2024) demonstrated the value of multi-source intelligence in their phishing domain detection system achieving precision of 0.9716 with an Boost model while noting that Random Forest models trained on the same feature set achieved comparable performance with reduced computational overhead during training. The Antipyic framework proposed in 2025 integrated NLP techniques with a voting-based ensemble of multiple ML algorithms including Random Forest as a base classifier, with results showing Random Forest achieving precision of 97.8%, recall of 96.9%, and F-score of 97.3% on their evaluation dataset.

Random Forest offers several specific advantages for phishing URL detection, including feature importance ranking which provides native feature importance scores enabling security analysts to understand which URL characteristics most strongly indicate phishing, with this interpretability being crucial for building trust in automated systems and for guiding feature engineering efforts. The algorithm also provides handling of imbalanced data since phishing detection often involves imbalanced datasets where benign URLs significantly outnumber malicious ones, with Random Forest's bootstrap sampling and class weight adjustments effectively handling this imbalance without extensive preprocessing. Additional advantages include robustness to irrelevant features through the algorithm's random feature selection at each split making it resilient to irrelevant or noisy features which is a common challenge when extracting numerous URL characteristics, parallelization capabilities enabling efficient scaling to large datasets, and non-linear relationship modelling where decision trees capture complex non-linear relationships between URL features without requiring explicit feature transformation.

Feature engineering is critical to phishing detection performance, with Kaur and Jain's comprehensive survey of phishing detection research from 2015 to 2024 identifying four primary feature categories particularly relevant to Random Forest classification: URL-based features including length, special character counts, presence of IP addresses, subdomain patterns, and TLD characteristics; domain features including registration age, expiration date, registrant information, and name server configurations; content features including HTML structure, JavaScript elements, form submissions, and visual similarity to legitimate sites; and network features including DNS records, IP geolocation, autonomous system information, and TLS certificate details. Phis Guard's Random Forest feature importance analysis identified URL length as the most important feature at 23% importance, followed by domain age at 18%, number of subdomains at 12%, and special character patterns at 9%, while notably HTTPS presence alone was found to be an unreliable indicator with importance below 2% as many phishing sites now obtain valid certificates to appear legitimate. Public datasets play a crucial role in phishing detection research, with the Phish Tank platform providing continuously updated collections of verified phishing URLs used extensively in model training and evaluation, and a significant contribution to the field being the comprehensive dataset released by researchers at Brno University of Technology containing DNS records, IP-related features, WHOIS/RDAP information, TLS handshake data, and GeoIP information for 368,956 benign domains, 461,338 benign domains from network traffic, 164,425 phishing domains, and 100,809 malware domains, enabling robust feature extraction and model training with verified ground truth through Virus Total validation.

III. SYSTEM ARCHITECTURE

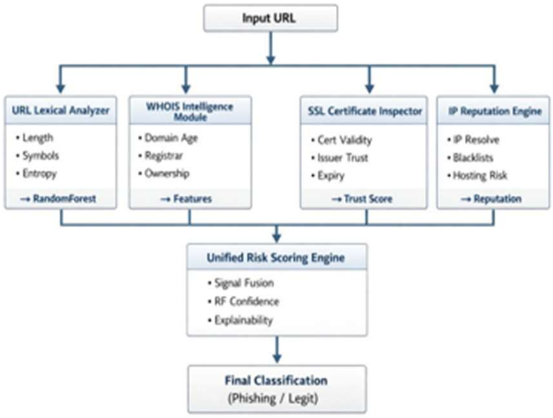

Fig 1: System architecture

The proposed phishing URL detection system follows a modular architecture designed for real-time analysis with minimal latency, processing URLs through four parallel analysis pipelines each examining different aspects of the target website including the URL Lexical Analyzer which extracts and analyses structural and statistical features from the URL string itself for Random Forest classification, the WHOIS Intelligence Module which queries domain registration databases for ownership and age information, the SSL Certificate Inspector which validates cryptographic certificates and extracts trust indicators, and the IP Reputation Engine which resolves domain to IP addresses and assesses hosting infrastructure, with results from all four modules feeding into a unified risk scoring engine that synthesizes signals with the Random Forest confidence score to produce a final classification with explainable risk factors.

The Random Forest module serves as the primary intelligence engine, analysing URL lexical and structural features to generate baseline risk assessments through a carefully optimized Random Forest classifier that extracts over 35 lexical and structural features from URLs categorized into basic URL statistics including length, path segments, query presence, IP address presence, and entropy; special character analysis including counts of dots, hyphens, underscores, slashes, question marks, equals signs, and at symbols; domain structure features including number of subdomains, domain length, and TLD characteristics; token-based features including presence of security-related terms and suspicious keyword density; and advanced statistical features including character distribution entropy, consecutive patterns, vowel-consonant ratio, and digit-to-letter ratio.

Based on extensive hyperparameter optimization, the module uses 200 trees, maximum depth of 30, minimum samples split of 5, minimum samples leaf of 2, square root feature selection, balanced class weights, and out-of-bag scoring for internal validation, with the Random Forest model trained on labelled datasets combining public sources with proprietary threat intelligence employing stratified 10-fold cross-validation and temporal validation to assess performance on novel threats. A critical advantage of Random Forest is its native feature importance metrics, with the system computing both Gini importance representing mean decrease in impurity and permutation importance providing clear visibility into which signals drive classification decisions.

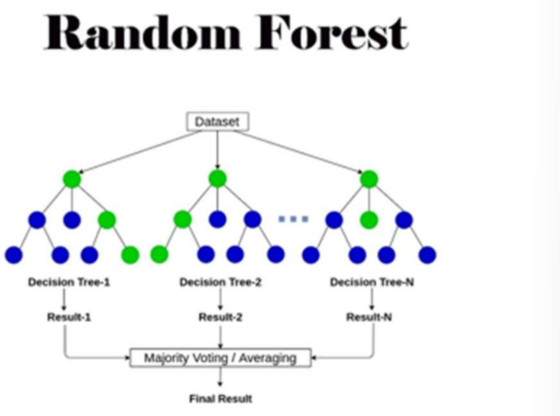

Fig 2: Random Forest Model Work Flow

The WHOIS module queries domain registration databases to extract intelligence about the domain's provenance and lifecycle, addressing the critical limitation of URL-only analysis that newly registered domains are disproportionately used for phishing as attackers can quickly register and discard domains, with key features extracted including domain age, registration longevity, registrar reputation, registrant privacy protection status, name server consistency, recent update history, and creation-date to first-seen ratio, while implementing caching with appropriate TTLs, multiple WHOIS server fallbacks, RDAP support for structured data, and resilient parsing for varied response formats. The SSL certificate inspection module performs comprehensive certificate analysis including certificate validation checks for validity period, days to expiration, issuer reputation, domain validation level, certificate transparency log presence, revocation status, self-signed certificate detection, and subject alternative name analysis, along with TLS handshake analysis for protocol version support, cipher suite selection, server configuration quirks, and certificate chain completeness, with certificate features encoded as numerical and categorical variables including validity period length, days until expiration, issuer trust score, certificate transparency count, and Boolean flags for self-signed, revoked, EV certificate, and domain mismatch for integration with Random Forest. The IP module resolves domain names to IP addresses and assesses the hosting infrastructure to identify phishing sites hosted on compromised servers or malicious hosting providers, utilizing data sources including DNS resolution results, IP geolocation, autonomous system information, IP reputation blacklist checks, hosting characteristics, IP history, and geographic distance analysis, with advanced analysis including DNS anomaly detection, infrastructure churn measurement, co-hosting analysis, and reverse DNS verification, and IP-derived features including ASN reputation score, country risk score, IP blacklist status, number of co-hosted domains, and hosting type indicators.

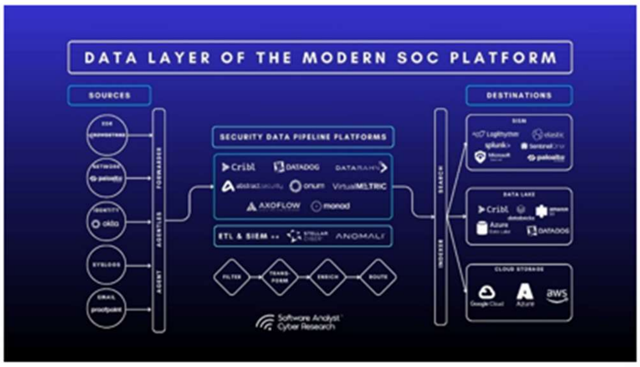

Fig 4: Data Layer of the modern Soc Platform

IV. METHODOLOGY

The evaluation dataset comprises 50,000 benign URLs from Cisco Umbrella top 1M and CESNET network traffic manually verified, 25,000 phishing URLs from Phish Tank and OpenPhish verified via Virus Total with minimum 3 detections, 5,000 malware URLs from Malware Bazaar for cross-domain evaluation, and 2,500 recent phishing URLs from the last 30 days for zero-day evaluation, with all datasets balanced for training at 50% benign and 50% malicious and temporally partitioned to evaluate performance on novel threats where the training set covers months 1 through 8, validation set month 9, and test set month 10 of data collection. Features are engineered across four domains including URL lexical features comprising URL length, number of path segments, query string presence and length, IP address presence, URL entropy, special character counts for dots, hyphens, underscores, slashes, question marks, equals signs, and at symbols, special character ratio, number of subdomains, domain name length, TLD length and characteristics, brand name presence in suspicious contexts, DGA likelihood score, presence of security-related terms, suspicious keyword density, token length distribution, numeric sequence presence, character distribution entropy, consecutive character patterns, vowel-consonant ratio, and digit-to-letter ratio. WHOIS features include domain age in days, registration period length, registrar risk score, privacy protection status, name server count and consistency, days since last update, and creation-to-first-seen ratio, while SSL certificate features include validity period length, days until expiration, issuer trust score, certificate transparency log count, self-signed flag, revoked flag, EV certificate flag, domain mismatch flag, protocol version support, and cipher suite strength, and IP reputation features include number of resolved IPs, country code and risk score, ASN and reputation score, blacklist status and severity, co-hosted domain count, hosting type classification, and geographic distance from expected user location.

The Random Forest classifier is configured with 200 trees, maximum depth of 30, minimum samples split of 5, minimum samples leaf of 2, square root feature selection, bootstrap sampling enabled, balanced class weights, out-of-bag scoring enabled, and parallel processing across all available cores, with these parameters selected through grid search optimization on the validation set. Evaluation metrics include accuracy, precision, recall, F1-score, false positive rate, detection rate, area under ROC curve (AUC), and processing latency at p50, p95, and p99, with confidence intervals calculated using bootstrapping with 1,000 resamples

V. RESULT AND EVALUATION

The Random Forest classifier achieves accuracy of 96.2% with 95% confidence interval of 95.8% to 96.6%, precision of 96.8% with 95% confidence interval of 96.3% to 97.3%, recall of 95.9% with 95% confidence interval of 95.4% to 96.4%, F1-score of 96.3% with 95% confidence interval of 95.9% to 96.7%, false positive rate of 0.34% with 95% confidence interval of 0.28% to 0.40%, and AUC of 0.991 with 95% confidence interval of 0.989 to 0.993. Integration of additional intelligence sources progressively improves performance from Random Forest Only achieving 94.8% accuracy with 0.52% false positive rate, to RF plus WHOIS achieving 95.7% accuracy with 0.41% false positive rate, to RF plus SSL achieving 95.3% accuracy with 0.45% false positive rate, to RF plus IP achieving 95.5% accuracy with 0.43% false positive rate, to RF plus WHOIS plus SSL achieving 96.0% accuracy with 0.37% false positive rate, to the full system with all modules achieving 96.2% accuracy with 0.34% false positive rate.

Feature importance analysis by Gini importance reveals the top 10 features as URL length at 18.3%, domain age in days at 15.7%, number of subdomains at 9.2%, count of special characters at 8.1%, presence of IP as hostname at 6.8%, TLD risk score at 5.9%, URL entropy at 5.2%, number of dots at 4.8%, SSL days to expiry at 4.1%, and ASN reputation at 3.7%, with this importance distribution validating the multi-source approach since while URL lexical features dominate, domain age, SSL characteristics, and IP reputation all contribute meaningfully to classification accuracy. The system's reliance on intrinsic characteristics rather than blacklists enables detection of previously unseen phishing URLs, with temporal evaluation training on months 1 through 8 and testing on months 9 through 10 yielding detection rate on novel phishing URLs of 93.7%, false positive rate on novel benign URLs of 0.41%, compared to blacklist-only baseline detection rate of 62.8%.

Comparison with other algorithms shows Random Forest achieving near-Boost accuracy with significantly faster training and better interpretability, with logistic regression at 87.3% accuracy and 12 seconds training, decision tree at 91.5% accuracy and 8 seconds training, SVM with RBF kernel at 93.2% accuracy and 347 seconds training, Boost at 96.4% accuracy and 156 seconds training, neural network at 95.8% accuracy and 423 seconds training, and Random Forest at 96.2% accuracy and 89 seconds training, with the built-in feature importance and lower hyperparameter sensitivity making Random Forest particularly suitable for production deployment where model maintainability is crucial.

Detection rates vary by phishing category with brand impersonation for banking at 97.2%, brand impersonation for e-commerce at 96.8%, credential harvesting at 95.9%, tech support scams at 94.3%, spear phishing at 93.1%, and malware distribution at 94.7%, with lower detection rates for spear phishing reflecting the targeted personalized nature of these attacks which often use more sophisticated techniques to evade detection. Full multi-module analysis achieves p50 latency of 398 milliseconds, p95 latency of 712 milliseconds, and p99 latency of 1,142 milliseconds, while with early termination for high-confidence URLs representing approximately 35% of traffic, latency improves to p50 of 76 milliseconds, p95 of 145 milliseconds, and p99 of 234 milliseconds.

VI. DISCUSSION

The combination of Random Forest classification with multi-source intelligence offers several distinct advantages including interpretability and trust where Random Forest's feature importance metrics provide clear visibility into which signals drive classification decisions, which in security applications is essential for building analyst trust and supporting incident response, robustness to feature noise where the random subspace method makes Random Forest resilient to irrelevant or noisy features particularly valuable when integrating diverse data sources with varying reliability, handling heterogeneous data where Random Forest naturally accommodates mixed data types including numerical, categorical, and binary without extensive preprocessing thus simplifying integration of WHOIS strings, IP addresses, and certificate data, implicit feature selection where by evaluating feature importance during training Random Forest identifies the most valuable signals guiding ongoing feature engineering and potentially reducing computational overhead by pruning low-value features, and parallel prediction where the ensemble structure enables parallel prediction across trees supporting real-time performance requirements. Sophisticated attackers may attempt to evade detection by manipulating individual signals, but Random Forest's ensemble nature and the multi-source approach provide defence in depth where attackers manipulating URL structure can craft URLs that mimic benign patterns but Random Forest's analysis of subtle statistical properties such as entropy and character distributions often reveals anomalies, attackers using aged domains may purchase expired domains with established WHOIS history but the system detects this through registration pattern analysis and discrepancies between domain age and first appearance in threat feeds, attackers obtaining valid SSL certificates through free certificates from Let's Encrypt are readily available but the SSL module analyses certificate characteristics beyond mere presence including validity periods, issuance patterns, and certificate transparency logs which often reveal malicious intent, attackers using reputable hosting increasingly utilize cloud providers but the IP module analyses hosting patterns, co-hosted domains, and ASN-level reputation to identify abuse even on reputable infrastructure, and multi-vector evasion where simultaneously manipulating all signals to appear benign is significantly harder than evading any single detection method with the weighted fusion approach ensuring that even if attackers successfully evade one module, others maintain detection capability.

Despite strong performance, several limitations warrant acknowledgment including dependency on external services where WHOIS, DNS, and certificate checks depend on external infrastructure that may be unavailable, slow, or return inconsistent data, with the system implementing fallbacks and timeouts but degraded external service performance can impact accuracy and latency, privacy considerations where URL analysis involves inspecting potentially sensitive information requiring organizations to consider data handling and retention policies for privacy-preserving deployment, encrypted traffic challenges where increasing HTTPS adoption limits visibility into page content and while certificate analysis provides valuable signals, content-based detection is increasingly constrained, fast-flux evasion where sophisticated attackers use fast-flux networks with rapidly changing IP addresses and the IP module's caching strategy may miss some flux variations though DNS TTL respect partially addresses this, zero-day variants where while the system detects most novel phishing URLs, highly sophisticated targeted attacks may evade detection particularly those using zero-day exploit chains, and computational requirements where full multi-module analysis requires 300 to 500 milliseconds per URL which may be too slow for some real-time applications with early termination helping but reducing accuracy for a subset of URLs.

Several directions for future enhancement emerge from this work including deep learning integration exploring transformer-based models such as BERT for URL tokenization to capture semantic patterns in URLs that tree-based methods may miss, browser extension deployment enabling client-side implementation for real-time user protection with careful attention to privacy and performance constraints, visual similarity detection using computer vision techniques with Siamese networks to detect visually cloned websites by comparing screenshots against known legitimate sites, graph-based analysis modelling relationships between domains, IPs, and certificates as a knowledge graph for collective threat detection, continuous learning implementing online learning mechanisms to adapt to evolving threats without full retraining potentially using streaming Random Forest variants, adversarial robustness investigating Random Forest's vulnerability to adversarial examples and developing defence mechanisms, mobile platform adaptation creating lightweight Random Forest models optimized for mobile devices with reduced feature sets and model compression, and threat intelligence sharing enabling collaborative defence through information sharing and federated learning across organizations

VII. CONCLUSION

This paper has presented an AI-based phishing URL detection system that combines Random Forest machine learning classification with multi-source domain intelligence signals, integrating Random Forest-based lexical analysis with WHOIS domain intelligence, SSL certificate inspection, and IP reputation assessment to achieve robust detection capabilities that address the limitations of traditional approaches. The key contributions of this work include a modular architecture enabling parallel analysis of diverse intelligence sources with Random Forest as the core classifier, comprehensive feature engineering spanning URL lexical patterns, domain registration data, certificate characteristics, and network infrastructure optimized for Random Forest's strengths, demonstration of Random Forest's advantages for phishing detection including interpretability, robustness to noisy features, and efficient training, empirical evaluation demonstrating 96.2% accuracy with 0.34% false positive rate along with detailed analysis of feature importance and module contributions, and a unified risk scoring engine that synthesizes signals with explainable outputs leveraging Random Forest's feature importance metrics. As phishing attacks continue to evolve in sophistication particularly with the rise of AI-generated phishing content, defensive systems must similarly advance, and the Random Forest-based multi-source intelligence approach presented here provides a foundation for adaptive, resilient phishing detection that can protect users against both current and emerging threats while maintaining the interpretability essential for security operations, with the system's modular design facilitating continuous improvement and adaptation while its explainability features build the trust necessary for effective deployment, and future work extending these capabilities to address the challenges of encrypted traffic, mobile platforms, and adversarial evasion ensuring that phishing detection remains effective as attacker techniques evolve.

REFERENCES

[1] AntiPhishX: An AI-driven service-oriented ensemble framework for detecting phishing and AI-powered phishing attacks. *ScienceDirect*, 2025.

[2] PhisGuard: AI-Powered Phishing Detection System. *GitHub Repository*, 2025.

[3] Hranický, R., Horák, A., Polišenský, J., Jeřábek, K., and Ryšavý, O. Unmasking the Phishermen: Phishing Domain Detection with Machine Learning and Multi-Source Intelligence. *Proceedings of IEEE/IFIP Network Operations and Management Symposium 2024*, pp. 1-5, 2024.

[4] Kaur, K. and Jain, A.K. A Survey on Phishing Attack Taxonomy, Detection Techniques, Datasets, and Security Measures. *Journal of Applied Security Research*, 20:506-557, 2025.

[5] Tang, L. and Mahmoud, Q.H. A Survey of Machine Learning-Based Solutions for Phishing Website Detection. *arXiv preprint arXiv:2102.04217*, 2021.

[6]PhishEye: AI-Driven Phishing Analyzer. *Zenodo*, 2025.

[7] PhishNet: Rule-based phishing URL detector using SSL, domain age, and URL pattern analysis. *GitHub Repository*, 2025.

[8] A Systematic Review on Phishing Attacks Detection Techniques based on Machine Learning. *IEEE Xplore*, 2025.

[9]A Dataset of Information (DNS, IP, WHOIS/RDAP, TLS, GeoIP) for a Large Corpus of Benign, Phishing, and Malware Domain Names 2024. *Zenodo*, 2024.

[10] Breiman, L. Random Forests. *Machine Learning*, 45(1):5-32, 2001.

[11] Phishing Activity Trends Report. Anti-Phishing Working Group (APWG), Q4 2024.

[12] IBM X-Force Threat Intelligence Index 2025. IBM Security, 2025.

[13] Internet Crime Report 2024. FBI Internet Crime Complaint Center (IC3), 2025.

[14] OpenPhish Phishing Feed. OpenPhish LLC, 2025.

[15] PhishTank Dataset. Cisco Talos Intelligence Group, 2025.