International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 05 | May 2026 www.irjet.net p-ISSN: 2395-0072

PhishShield: A Lightweight Multi-Layer Email Phishing Detection Framework for Small and Medium Enterprises

Ketan Sawant, **Sujal Raut**, **Manohar Kumawat**, **Sahil Valanju**

B.E. Students, Dept. of Computer Science and Engineering (IoT and CS BC),
A. C. Patil College of Engineering, Kharghar, Navi Mumbai University of Mumbai, India
Guide: Dr. Manjusha Deshmukh, Head of Department, Dept. of CSE (IoT and CS BC)

***Abstract* —** *Email phishing remains one of the most persistent and damaging cybersecurity threats globally. Attackers increasingly exploit breach data and AI-generated content to craft convincing messages that bypass conventional filters. Existing machine-learning solutions, although effective, demand substantial computational resources and infrastructure investment, rendering them impractical for small and medium enterprises (SMEs). This paper presents PhishShield, a lightweight, cost-effective, multi-layer email protection framework that combines cloud-based AI content analysis (Google Gemini), email authentication protocols (SPF, DKIM, DMARC), keyword-based heuristic detection, and URL reputation scoring via the VirusTotal API. The system operates as an SMTP proxy, intercepting emails in transit before delivery and assigning each a composite threat score on a 0–40 scale. Emails are then automatically delivered, quarantined, or dropped according to configurable thresholds. Evaluation on a dataset of 20 real-world emails seven legitimate and thirteen phishing demonstrated 100% detection accuracy with zero false positives and an average processing latency of 10.8 seconds per message. PhishShield provides a scalable, explainable, and easy-to-deploy security layer that enables SMEs to strengthen email defences without requiring specialised expertise or costly on-premise infrastructure.*

Key Words: Phishing Detection, Email Security, SMTP Proxy, Machine Learning, SPF/DKIM/DMARC, VirusTotal, Google Gemini, Explainable AI, Cybersecurity, SME

1. INTRODUCTION

Email phishing constitutes one of the foremost cyber security threats of the modern era. Phishing emails are fraudulent communications engineered to deceive recipients into disclosing sensitive credentials, financial details, or personal information, or to trigger the installation of malicious software. What distinguishes contemporary phishing from earlier spam campaigns is the level of precision and personalisation involved. Whereas traditional spam was broadcast indiscriminately to large mailing lists, today's attacks are carefully researched, contextually relevant, and frequently indistinguishable from legitimate correspondence.

According to cyber security reports, phishing incidents accounted for over 300,000 reported cyber events in the United States in 2024 alone, resulting in financial losses measured in billions of dollars. The problem is compounded by three converging trends. First, the rapid proliferation of large language models enables attackers to generate grammatically correct, professionally toned emails at scale, eliminating the typographic errors that once served as reliable red flags. Second, the shift from bulk campaigns to spear phishing highly targeted attacks personalised with information sourced from social media, corporate directories, or previous breaches substantially increases the probability that a recipient will trust and act upon a fraudulent message. Third, attackers continuously adapt to evade newly deployed defences, rendering static, rule-based filters obsolete within weeks of deployment.

SMEs are disproportionately exposed to these threats. Large organisations typically employ dedicated security operations centres, expensive enterprise email gateways, and teams of analysts. SMEs, by contrast, rarely possess the budget or expertise to deploy and maintain such infrastructure. The consequence is a growing asymmetry: sophisticated phishing tools available to any attacker at negligible cost, versus under-resourced defenders relying on basic spam filters.

This paper addresses that asymmetry by proposing PhishShield a multi-layer, SMTP-proxy-based framework that delivers enterprise-grade phishing detection at a fraction of the cost. By combining cloud AI, authentication protocols, heuristic

analysis, and real-time threat intelligence in a lightweight Python stack, PhishShield is designed to be deployable by any organisation regardless of its technical capacity.

1.1 Phishing Taxonomy

Phishing manifests across multiple channels and targeting strategies. Email phishing represents the most prevalent form, in which fraudulent messages impersonate trusted institutions to harvest credentials or install malware. Spear phishing refines this approach with individualised content derived from open-source intelligence. Whaling targets senior executives, typically to authorise fraudulent financial transfers a variant known as Business Email Compromise (BEC). Smishing and vishing extend the same social-engineering principles to SMS and telephone channels respectively. PhishShield targets the email vector, which remains the primary attack surface for organisational compromise.

1.2 Phishing Attack Lifecycle

A typical phishing campaign proceeds through a structured lifecycle: (1) target selection an individual or organisation is identified; (2) reconnaissance publicly available information is gathered to personalise the message; (3) message crafting a convincing email is composed, often assisted by AI tools; (4) delivery the message is dispatched, frequently from spoofed or compromised domains; (5) exploitation the victim clicks a malicious link or provides credentials; (6) data exfiltration stolen information is leveraged for fraud, further intrusion, or resale on dark-web markets. Effective defences must intervene before step five, ideally at the point of delivery.

2. LITERATURE SURVEY

Phishing detection research has evolved from simple blacklist-based approaches to sophisticated ensemble learning systems. Abu-Nimeh et al. [1] conducted a foundational comparative study of machine-learning techniques including naive Bayes, random forests, and support vector machines and demonstrated that learned models outperform static rules for email classification. Ma et al. [2] extended this work to URL-level analysis, showing that features derived from domain registration and lexical URL structure can reliably distinguish malicious from benign links.

Verma and Hossain [3] applied semantic feature selection to phishing email text, demonstrating that contextual meaning captures deceptive intent more effectively than surface-level keyword frequency. Sahingoz et al. [4] evaluated seven machine learning algorithms on a large URL dataset, with random forests achieving the highest classification accuracy. Fette et al. [5] proposed the PILFER system, which extracted ten features from email headers and bodies to train an SVM classifier, achieving 99.5% detection with a low false-positive rate.

Deep learning approaches have further improved detection capability. Convolutional and recurrent neural networks, including LSTM architectures, can capture sequential and contextual patterns in email text that elude shallower models. However, these gains come with substantially higher computational requirements and the need for large, continuously updated labelled datasets barriers that SMEs cannot readily overcome.

Authentication-protocol-based approaches using SPF, DKIM, and DMARC [7, 8] provide orthogonal protection by verifying sender domain authority at the DNS level. These protocols are effective against domain spoofing but cannot detect phishing originating from compromised legitimate accounts or newly registered lookalike domains that pass authentication checks.

Hybrid, multi-layer frameworks that combine content analysis, URL inspection, and authentication have demonstrated superior accuracy compared to single-method approaches [3, 4]. Nevertheless, most published systems require significant infrastructure and continuous retraining cycles, limiting their applicability to resource-constrained environments. The emergence of cloud-based AI APIs, such as those provided by Google and OpenAI, presents a promising direction: powerful natural-language understanding accessible via simple HTTP calls, without local model hosting. Academic evaluation of such APIs in the phishing-detection context remains limited, representing a gap that this work seeks to address.

2.1 Identified Research Gaps

Based on the survey above, the following gaps are evident: (i) static and rule-based approaches fail to generalise to novel phishing patterns; (ii) deep-learning models impose computational and maintenance burdens unsuitable for SMEs; (iii) existing datasets under-represent breach-based, AI-generated phishing; (iv) multi-layer integration remains complex to deploy in practice; and (v) cloud AI APIs are under-studied in this domain. PhishShield is designed to address all five gaps through a lightweight, modular architecture.

3. SYSTEM DESIGN AND ARCHITECTURE

PhishShield is architected as an SMTP proxy an intermediary positioned between external senders and the organisation's mail server. All incoming SMTP traffic is routed through the proxy on port 1025, where each message is subject to multi-stage analysis before a delivery decision is issued. This placement requires no modification to existing mail clients or user workflows, reducing deployment friction.

3.1 Multi-Layer Detection Pipeline

Upon receipt, the proxy parses each message to extract the sender address, recipient details, subject line, plain-text and HTML body, embedded URLs, and attachment metadata. This structured representation is simultaneously forwarded to four independent detection engines, each of which returns a sub-score on a 0–10 scale. The four engines are described below and summarised in Table 1.

Table 1: PhishShield Detection Modules

Detection Module	Weight	Function
AI Content Analysis (Gemini)	High	Semantic understanding of email intent, social engineering cues, and impersonation detection using cloud-based LLM
Email Authentication (SPF/DKIM/DMARC)	Medium	DNS-level verification of sender domain authority to detect spoofed identities
Keyword/Pattern Matching	Medium	Heuristic scanning for urgency phrases, suspicious requests, and obfuscation patterns
URL Reputation (VirusTotal)	High	Static and dynamic link analysis against a global threat intelligence database

The AI Content Analysis Engine submits the email body and subject to the Google Gemini API with a carefully engineered system prompt that instructs the model to evaluate semantic intent, identify social-engineering tactics (urgency creation, authority impersonation, fear induction), and return a structured JSON response containing a numeric score and a naturallanguage explanation. This component operation alises the principle of Explainable AI (XAI): every automated decision is accompanied by a human-readable rationale that administrators can audit and challenge.

The Email Authentication Engine performs DNS lookups to evaluate SPF records, verifies DKIM cryptographic signatures, and enforces DMARC policy. Failure of any check elevates the sub-score proportionally. This module reliably intercepts domain-spoofing and sender-forgery attacks but is acknowledged to be ineffective against phishing originating from compromised legitimate accounts.

The Keyword and Pattern Matching Engine applies a curated library of regular expressions to the email content, targeting urgency markers ('act immediately', 'verify within 24 hours'), impersonation phrases, and obfuscation techniques such as character substitution. While straightforward, this component provides fast, deterministic detection of known attack patterns and serves as an important counterweight when the AI engine is deceived by professionally worded text.

The URL Reputation Engine extracts all hyperlinks from the message, applies heuristic checks (suspicious TLDs, typosquatting patterns, excessive subdomains), and submits each URL to the VirusTotal API for evaluation against over 70 commercial

antivirus and threat-intelligence engines. The aggregated results are converted into a sub-score reflecting the proportion of positive malicious detections.

3.2 Composite Scoring and Decision Engine

Sub-scores from all four engines are combined using a weighted summation to produce a final composite threat score on a 040 scale. Weights are assigned to reflect the reliability and coverage of each module, with the AI and URL engines carrying higher weights due to their broader detection surface. The decision engine maps composite scores to three dispositions using configurable thresholds: (i) scores below 40 email delivered normally; (ii) scores between 40 and 70 email quarantined for administrator review; (iii) scores above 70 email dropped and logged. These thresholds are adjustable via the administrative dashboard to accommodate varying organisational risk tolerances.

3.3 Technology Stack and Deployment

The backend is implemented in Python 3.10 on a Linux server, utilising the Flask web framework for API request handling and inter-component communication. Email metadata, detection scores, and administrative actions are persisted in a SQLite database, chosen for its zero-configuration deployment profile and low resource footprint. The overall architecture supports deployment on local servers, virtual machines, and cloud platforms with equal ease, providing organisations with full flexibility over their hosting arrangements. The administrative dashboard is a responsive web application offering real-time threat monitoring, configurable thresholds, whitelist and blacklist management, and downloadable audit logs.

4. RESULTS AND ANALYSIS

PhishShield was evaluated on a dataset of 20 emails comprising seven legitimate messages and thirteen phishing samples. The phishing samples were drawn from multiple attack categories, including standard credential harvesting, Business Email Compromise attempts, prize-baiting spam, and breach-based campaigns reconstructed from real 2025 security incidents involving Qantas, Microsoft, and Google. The results are summarised in Table 2.

Table 2: Evaluation Results Summary

Metric	Value	Metric	Value
Test dataset	20 emails	Phishing detected	13 / 13 (100%)
Legitimate emails	7 / 7 (100%)	False positives	0
Avg. processing time	10.8 s / email	Score range used	0 – 40 (composite)

All thirteen phishing emails were correctly identified and either quarantined or dropped. All seven legitimate emails were delivered without disruption, yielding a false-positive rate of zero. The highest composite score (24.0 / 40, after normalisation) was assigned to a Qantas breach-based phishing email, which leveraged authentic brand language and contextual alignment with a widely reported security incident, illustrating the sophistication of contemporary attacks.

4.1 Ensemble Resilience Against AI-Generated Phishing

A particularly instructive test case involved a Chase Bank impersonation email constructed with professional, grammatically flawless language and a plausible transaction narrative. When evaluated in isolation, the DistilBERT-based NLP su-component a finetuned transformer model used to benchmark against PhishShield's AI engine assigned this email a low-threat score, incorrectly classifying it as legitimate. However, the URL analysis module identified the sender domain as a suspicious .tk toplevel domain, and the keyword engine flagged urgency phrasing. Their combined contribution raised the composite score to 95, triggering an automatic drop decision. This outcome demonstrates the core architectural benefit of ensemble design: individual modules may be deceived, but multi-engine agreement is considerably harder to evade.

4.2 Sandbox Testing Environment

PhishShield includes a Threat Analysis Sandbox a manual submission interface that enables security analysts to test arbitrary email content against the detection pipeline without risk of delivery. The sandbox returns a full analysis report including the composite score, per-module sub-scores, the AI engine's natural-language explanation, and identified keywords and URLs. This facility supports analyst training, incident response investigations, and regression testing after configuration changes.

5. SECURITY ANALYSIS

5.1 Attack Resistance Profile

Standard phishing emails characterised by suspicious sender domains, urgency language, and malicious links are reliably intercepted by multiple engines simultaneously, producing high composite scores. Breach-based phishing, which exploits realworld events to achieve contextual credibility, was correctly detected in all test instances, with the semantic AI engine providing the most discriminating signal. SPF/DKIM/DMARC validation provides robust resistance against domain-spoofing attacks, elevating threat scores for any email whose authenticated identity diverges from the claimed sender.

The system's most significant vulnerability lies in zero-day phishing campaigns that employ entirely novel language patterns and newly registered, unindexed malicious domains. For such attacks, the keyword engine and VirusTotal lookup provide limited coverage, and the AI engine must bear the primary detection burden. A second known limitation involves phishing content hosted on trusted platforms Google Drive, Dropbox, GitHub whose domains are whitelisted by most threat-intelligence services. Contextual URL analysis beyond domain reputation is required to close this gap and is identified as a priority for future development.

5.2 Privacy and Compliance Considerations

Operating as a full-content SMTP proxy, PhishShield necessarily processes the complete content of every email passing through the organisation. Two aspects require particular attention. First, submission of email content to external cloud APIs (Google Gemini and VirusTotal) constitutes a transfer of potentially sensitive data outside the organisation's network boundary. Organisations subject to regulations such as GDPR, HIPAA, or India's Digital Personal Data Protection Act must assess whether this transfer is permissible and, if required, execute appropriate data-processing agreements with the relevant service providers. Second, SQLite storage of email metadata creates a persistent record that demands a clear retention policy and access controls, including multi-factor authentication for dashboard access. Future versions should incorporate database-at-rest encryption to mitigate the impact of server compromise.

6. ADVANTAGES AND LIMITATIONS

6.1 Key Advantages

PhishShield offers several notable strengths. Its ensemble architecture four independent engines contributing to a single composite decision provides resilience against the evasion techniques that defeat single-method approaches. Cloud based AI integration via the Gemini API delivers advanced semantic understanding without local model training or GPU infrastructure. The three-tier decision system (deliver, quarantine, drop) preserves legitimate mail while providing a safety buffer for ambiguous cases, substantially reducing the operational cost of false positives. Every decision is logged and explained, satisfying the Explainable AI requirements increasingly mandated by enterprise governance frameworks. The system achieved 100% detection accuracy across a carefully curated test set that included breach-based attacks representative of the current threat landscape.

6.2 Current Limitations

Several limitations bound the present implementation. The evaluation dataset of 20 emails, while carefully selected and diverse, is insufficient to characterise performance at production scale. Processing latency of 10.8 seconds per email may introduce unacceptable delivery delays in high-volume environments and warrants optimisation through asynchronous

pipeline processing and API call batching. Attachment scanning a critical attack vector is not yet implemented. The keyword database requires regular curation to remain effective against evolving vocabulary. Finally, dependence on external APIs introduces both privacy considerations and availability risk: service outages or quota exhaustion would degrade detection capability.

7. CONCLUSIONS

This paper presented PhishShield, a lightweight multilayer email phishing detection framework designed to address the security needs of SMEs without imposing the resource demands of enterprise-grade solutions. By integrating cloud-based AI semantic analysis, email authentication protocol validation, heuristic keyword matching, and real-time URL reputation scoring within an SMTP proxy architecture, PhishShield achieves comprehensive coverage across the major phishing attack categories in use today.

Experimental evaluation demonstrated 100% phishing detection and zero false positives on a dataset spanning standard, AIgenerated, spear, and breach-based phishing emails. The ensemble design proved particularly valuable against AIgenerated content: when individual modules were deceived, corroborating signals from complementary engines ensured correct final classification. The administrative dashboard and sandbox environment further reduce the operational burden on security teams by providing explainable, auditable decisions and a safe testing environment for analyst investigation.

Future work will focus on four directions: (i) evaluation on large-scale, publicly available phishing datasets to validate generalisation; (ii) asynchronous pipeline redesign to reduce per-email latency; (iii) integration of attachment scanning using sandboxed execution environments; and (iv) development of a privacy-preserving local AI inference option to eliminate dependency on external cloud services for organisations operating under strict data-residency requirements.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the guidance of Dr. Manjusha Deshmukh, Head of Department, Computer Science and Engineering (IoT and CS BC), A. C. Patil College of Engineering, whose expertise, feedback, and consistent encouragement were instrumental throughout this project. The authors also thank the faculty and staff of the Department of Computer Science and Engineering (IoT and CS BC) for their support, and the developers and maintainers of the open-source libraries and cloud APIs that made this work possible.

REFERENCES

[1] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, 'A Comparison of Machine Learning Techniques for Phishing Detection,' in Proc. Anti-Phishing Working Group eCrime Researchers Summit, Pittsburgh, PA, 2007, pp. 60–69.

[2] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, 'Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs,' in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, Paris, France, 2009, pp. 1245–1254.

[3] R. Verma and N. Hossain, 'Semantic Feature Selection for Text with Application to Phishing Email Detection,' in Proc. IEEE Int. Conf. Data Mining (ICDM), New Orleans, LA, 2017.

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

[5] I. Fette, N. Sadeh, and A. Tomasic, 'Learning to Detect Phishing Emails,' in Proc. 16th Int. World Wide Web Conference, Banff, Canada, 2007, pp. 649–656.

[6] P. Kumaraguru, Y. Rhee, A. Acquisti, et al., 'Getting Users to Pay Attention to Anti-Phishing Education,' in Proc. ACM CHI, Boston, MA, 2009.

[7] P. Resnick and P. Hoffman, 'Sender Policy Framework (SPF) for Authorizing Use of Domains in Email,' IETF RFC 4408, Apr. 2006.

[8] M. Kucherawy and E. Zwicky, 'Domain-based Message Authentication, Reporting, and Conformance (DMARC),' IETF RFC 7489, Mar. 2015.

[9] VirusTotal (Google LLC), 'VirusTotal API v3 Documentation for URL and File Analysis,' [Online]. Available: https://developers.virustotal.com/, 2024.

Biographies

Ketan Sawant is a final-year student in the B.E. Computer Science and Engineering (IoT and CS BC) programme at A. C. Patil College of Engineering, Navi Mumbai. His research interests include cybersecurity, network security, and full-stack web development. He served as the lead architect of the PhishShield backend and SMTP proxy implementation.

Sujal Raut is a final-year B.E. student in the same programme. His areas of interest span machine learning, data analysis, and software engineering. He led the integration of the AI content-analysis and URL reputation modules within PhishShield.

Manohar Kumawat is a final-year B.E. student with a focus on cloud computing, API integration, and systems programming. He was responsible for the email authentication engine and database management components of the PhishShield system.

Sahil Valanju is a final-year B.E. student whose interests encompass web development, data visualisation, and user experience design. He designed and implemented the PhishShield administrative dashboard and sandbox testing environment.