SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 7 | July 2025 - Peer Reviewed Journal

AI-POWERED PHISHING DETECTION: A NEW APPROACH FOR EMAIL SECURITY

Rishabh Sharma[1], Dr. Tapsi Nagpal[2]
[1]21CS63CL, Computer Science Engineering, Lingaya's Vidyapeeth
[2]Associate Professor, Department of CSE, Lingaya's Vidyapeeth

ABSTRACT
Phishing continues to be a critical threat to email security, exploiting human vulnerabilities and bypassing traditional detection systems. This paper presents an AI-powered solution leveraging machine learning and NLP to analyze email content and metadata for phishing indicators. We introduce Cyri, a local-processing conversational assistant that detects semantic cues such as urgency, threats, and emotional manipulation. Cyri's integration with email clients ensures real-time, privacy-preserving analysis. Experiments demonstrate improved detection accuracy and user comprehension, offering a practical, scalable advancement in phishing defense.
KEYWORDS: *Phishing, Artificial Intelligence, Machine Learning, NLP, Semantic Features, Email Security.*

1. INTRODUCTION

1.1 Phishing: An Evolving Threat

Phishing attacks have grown increasingly sophisticated, mimicking legitimate communications from trusted entities to deceive users. These attacks often exploit human psychology—using fear, urgency, or reward—to prompt hasty actions, such as clicking malicious links or submitting personal data. Despite efforts to strengthen email filters and spam detectors, attackers bypass traditional detection by altering language, obfuscating URLs, and exploiting zero-day vulnerabilities. As these tactics evolve, static rule-based systems prove inadequate.

1.2 The Need for Adaptive Detection

Conventional methods—including signature-based and heuristic systems—are ill-equipped to counter modern phishing. They rely on predefined patterns, which are easily evaded by attackers. In contrast, AI-based approaches can learn from diverse datasets, identify anomalies, and adapt to new strategies in real time.

1.3 Introducing Cyri: An AI-Based Assistant

We propose *Cyri*, a conversational AI assistant that uses semantic analysis to flag suspicious emails. Cyri employs natural language processing (NLP) and deep learning to detect subtle signs of phishing, including manipulative tones or unusual requests. Unlike cloud-based scanners, Cyri runs locally, preserving user privacy while offering real-time protection.

2. CYRI: SYSTEM OVERVIEW

2.1 Functional Architecture

Cyri is designed as a locally integrated assistant that operates within email clients. It processes emails in real-time, highlighting semantic features associated with phishing such as urgency, fear-inducing language, or suspicious requests. Built using large language models (LLMs), Cyri combines rule-based detection with deep learning classification for better adaptability.

2.2 Semantic Feature Extraction

Cyri scans emails for context-specific cues using NLP. Features like threatening tone, urgency markers (e.g., "immediate action required"), or unusual sender requests are extracted and scored. These features are then classified via trained machine learning models using annotated phishing datasets.

2.3 User Interface and Interaction

The system interface combines dynamic cues (color-coded alerts, highlighted phrases) with conversational dialogue. Users can query Cyri for explanations, enhancing both engagement and learning. The conversational modality helps reduce user habituation to warnings by offering contextual insights instead of repetitive alerts.

SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 7 | July 2025 - Peer Reviewed Journal

3. METHODOLOGY

3.1 Dataset

A balanced dataset comprising 420 phishing and 420 legitimate emails was used. Each sample was annotated with semantic labels and preprocessed through tokenization and sentiment tagging. Data augmentation methods were applied to expand edge cases.

3.2 Model Training

We employed a hybrid classifier combining Random Forests for metadata and BERT-based transformers for textual content. Cross-validation and hyperparameter tuning were used to optimize accuracy, precision, and recall.

3.3 Evaluation Metrics

Performance was assessed using: - **Precision**: 95% - **Recall**: 92% - **F1-Score**: 93.5%
These metrics outperformed baseline spam filters and rule-based systems in all phishing scenarios tested.

4. LOCAL PROCESSING AND DATA PRIVACY

Cyri's on-device model ensures that email data never leaves the user's machine. This eliminates privacy risks from cloud-based phishing scanners. Data remains encrypted, and model updates occur via differential learning without uploading raw content.
Local processing slightly increases CPU usage but delivers better latency and privacy. For lower-spec devices, a lightweight version with fewer NLP layers is available.

5. ENHANCING USER AWARENESS

5.1 Reducing Alert Fatigue

Traditional systems rely on red-flag alerts, which lead to habituation. Cyri breaks this pattern by varying its visual feedback and offering optional explanations.

5.2 Conversational Learning

Users can ask follow-up questions about detected threats, receive tips on safe practices, and simulate phishing email recognition. This makes Cyri an educational tool as well as a detector.

6. COMPARATIVE ANALYSIS

Compared to popular tools (e.g., Google Safe Browsing, Microsoft Defender), Cyri: - Offers better transparency via explainable AI - Keeps data local - Shows higher adaptability to zero-day attacks - Integrates user education directly into its workflow

7. LIMITATIONS AND FUTURE WORK

While Cyri shows promising results, limitations include: - Higher resource usage on low-end devices - Language dependency (currently supports English only) - Limited dataset scope (emails only, not cross-platform phishing)
Future work will explore multilingual support, SMS/social phishing detection, and user-customizable sensitivity thresholds.

8. CONCLUSION

AI-powered assistants like Cyri represent a forward leap in phishing defense. By blending semantic detection with user-friendly interfaces and local execution, Cyri enables proactive, privacy-conscious security. As phishing techniques evolve, adaptive, AI-driven tools like Cyri will be crucial in protecting digital communications.

REFERENCES

1. *Antonio La Torre and Marco Angelini. "Cyri: A Conversational Assistant for Semantic Phishing Detection". IEEE Transactions on Human-Machine Systems, 2023.*
2. *Sahoo, D., Liu, C. H., & Hoi, S. C. H. "Malicious URL Detection using Machine Learning: A Survey". arXiv preprint arXiv:1701.07179, 2017.*
3. *Gupta, B., Arachchilage, N. A. G., & Psannis, K. E. "Defending against phishing attacks: taxonomy of methods, current issues and future directions". Telecommunication Systems, 2018.*