IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, June 2026
Impact Factor: 8.2

Heuristic-Based Detection of Phishing URLS: A Practical Approach for Enhanced Cyber Awareness

Khushabu Ramasare Yadav
Student, Department of Computer Science
Vidyavardhini's College of Engineering and Technology, Vasai, India

Abstract: *Phishing remains one of the most effective social engineering attacks, leading to significant financial and data losses globally. This paper proposes a lightweight, rule-based (heuristic) detection system that analyses URL structures to identify potential threats. By focusing on specific 'red flags' such as IP addresses, suspicious characters, and domain length, the proposed system provides an accessible method for real-time phishing detection suitable for non-technical users. The methodology demonstrates that proactive analysis of URL features can significantly reduce the window of vulnerability compared to traditional blacklisting methods.*

Keywords: Phishing Detection, Cyber security, Heuristic Analysis, URL Security

I. INTRODUCTION

The rapid digitalization of services has made the internet an indispensable part of modern life. However, this growth has been mirrored by a rise in cybercrime, specifically phishing. Phishing is a form of social engineering where attackers deceive users into revealing sensitive information by masquerading as a trustworthy entity in electronic communication.

Traditional security measures like blacklisting are often reactive; they rely on a database of known malicious sites which are only updated after an attack has been reported. This research explores a proactive approach using heuristics. Heuristics involve identifying patterns and characteristics common to phishing URLs, allowing the system to flag new, previously unseen threats in real-time.

II. LITERATURE REVIEW

1. Mechanisms of Heuristic-Based Detection

Heuristic detection operates by analysing a URL's structure and behaviour against predefined patterns associated with malicious activity. Research categorizes these features into several key areas: URL-Based Features: This includes analysing the length of the URL, the presence of special characters (e.g., "@", "-", "."), and the use of subdomains or IP addresses instead of domain names (Akinyelu, 2019; SPIE, 2025).

Domain Metadata: Heuristics often evaluate the domain age, registration details, and DNS records. Phishing sites typically use recently registered domains to avoid reputation filters (SPIE, 2025).

Technical Integrity: The presence or absence of HTTPS, the validity of SSL certificates, and the use of URL redirection are critical heuristic indicators (RJ Wave, 2026).

2. Practical Advantages and Limitations

Recent studies highlight a "practicality trade-off" between heuristic methods and more complex machine learning (ML) models:

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-36742

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 6, June 2026 Impact Factor: 8.2

Feature	Heuristic Rule-Based	Machine Learning (ML)
Response Time	Extremely fast; suitable for real time browsing (ITM, 2025).	Slower due to feature extraction and model inference.
Interpretability	High; easy to explain why a URL was flagged (ITM, 2025).	Low; often acts as a "black box" (Akinyelu, 2019).
Maintenance	Low cost; ideal for resource constrained environments (ITM, 2025).	High; requires massive datasets and frequent retraining (RJ Wave, 2026).
Accuracy	Prone to false positives and bypasses by slight URL modifications.	Higher accuracy and better at finding "hidden" patterns (Akinyelu, 2019).

III. HYBRID APPROACHES FOR CYBER AWARENESS

To enhance cyber awareness in practical settings—such as mobile devices or small-scale business networks—literature suggests a hybrid approach. By combining heuristic rules with lightweight ML models (like Gradient Boosting or Extreme Learning Machines), systems can achieve high precision without the heavy computational toll of deep learning (SPIE, 2025; RJ Wave, 2026).

Furthermore, heuristic systems serve as a "first line of defence," providing immediate feedback to users. This "immediate response capability" is vital for educational purposes, as it helps users identify suspicious patterns in real-time, thereby fostering long-term cyber vigilance (ITM, 2025).

III. PROBLEM DEFINATION

The Threat Landscape

Modern phishing attacks have evolved beyond simple credential harvesting into sophisticated, multi-staged social engineering campaigns. Traditional defence mechanisms rely heavily on Static Blacklisting—databases of known malicious URLs. However, these systems are inherently reactive; they cannot detect "Zero-Day" phishing sites that are created, utilized, and dismantled within hours, long before they are flagged by security vendors.

The Technical Gap

While Machine Learning (ML) and Deep Learning (DL) offer high accuracy in detecting these novel threats, they present significant hurdles for practical, real-world deployment:

Computational Overhead: Complex models often require high processing power and memory, making them unsuitable for mobile devices or low-power edge gateways.

Latency: The "time-to-verdict" in ML models can delay page loading, leading to a poor user experience or users bypassing security protocols entirely.

The "Black Box" Problem: Most advanced models lack transparency, providing a "block" or "allow" decision without explaining why a URL was deemed suspicious, which fails to educate the end-user.

IV. OBJECTIVE / SCOPE

The primary goals of this study are to:

- Evaluate Response Latency: Measure and compare the speed of heuristic rules against the slower feature extraction and model inference times of ML.
- Assess Detection Accuracy: Determine the effectiveness of ML in identifying "hidden" patterns compared to rule-based systems that are prone to bypasses via URL modifications.
- Analyse Operational Overhead: Quantify the maintenance costs associated with the frequent retraining and massive datasets required by ML models.
- Examine System Transparency: Investigate the "black box" nature of ML models versus the high interpretability and ease of explanation found in heuristic methodologies.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, June 2026
Impact Factor: 8.2

The scope of this research is constrained to the following parameters:

- Detection Methodologies: The study is limited to a comparative analysis of Heuristic Rule-Based systems and Machine Learning (ML) models.
- Performance Metrics: Analysis is restricted to four key areas: Response Time, Interpretability, Maintenance requirements, and Accuracy.
- Environment Suitability: The research evaluates these systems specifically for their application in real-time browsing and resource-constrained environments.
- Threat Type: The focus is primarily on the detection and classification of phishing URLs.

V. PROPOSED METHODOLOGY

The system operates by extracting features directly from the URL string. Unlike deep-learning models that require heavy computational resources, this rule-based approach is lightweight and fast. These features are evaluated against a set of predefined rules derived from historical phishing data analysis. If a URL triggers a certain number of flags, it is categorized as "Suspicious" or "Phishing."

Key Heuristics Used:

IP Address Presence: Legitimate brands rarely use IP addresses (e.g., 192.168.1.1) as their public URL.
URL Length: Phishing URLs are often significantly longer than legitimate ones to hide the real domain.
"@" Symbol: The presence of an "@" symbol often leads the browser to ignore everything before it, a common trick to hide malicious domains.
Subdomain Analysis: Multiple subdomains (e.g., login.update.secure.bank.com) are used to confuse users.
HTTPS Protocol: Checking for a valid SSL certificate presence.

VI. ANALYSIS & KEY FINDINGS

Speed vs. Complexity Trade-off: The analysis shows a clear inverse relationship between detection depth and speed. While heuristic rules allow for "extremely fast" real-time browsing, the introduction of machine learning (ML) adds latency due to the heavy computational requirements of feature extraction.
The "Black Box" Barrier: A significant finding is the disparity in transparency. Heuristic systems provide high interpretability, making it "easy to explain" why a URL was flagged, whereas ML models often function as a "black box," complicating the audit process for security professionals.
Evasion Resilience: Findings indicate that heuristic methods are highly susceptible to "bypasses by slight URL modifications". ML demonstrates a superior ability to identify these "hidden" patterns, suggesting it is the more resilient choice against evolving adversarial tactics.

VII. LIMITATIONS & FUTURE SCOPE

Limitations

Heuristic Rule-Based Systems: Vulnerability to Evasion: These systems are highly prone to being bypassed by attackers who make slight modifications to a URL to avoid matching known patterns.
False Positives: The rigid nature of rule-based logic often leads to a higher rate of false positives, incorrectly flagging legitimate sites as threats.
Machine Learning (ML) Models: Resource Intensity: ML models require high maintenance costs, massive datasets, and frequent retraining to remain effective against new threats.
Operational Latency: Due to the time required for feature extraction and model inference, these systems generally have a slower response time than rule-based methods.
Lack of Transparency: These models often act as a "black box," providing low interpretability, which makes it difficult for security analysts to explain exactly why a specific URL was flagged.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, June 2026
Impact Factor: 8.2

Future Scope

Development of Hybrid Architectures: Future research could focus on merging these methodologies to create systems that use fast heuristic rules for initial screening and ML for deeper, high-accuracy analysis of "hidden" patterns.

Improving ML Interpretability: A key area for future development is moving away from "black box" models toward "Explainable AI" (XAI) to improve the interpretability of ML-based detection.

Optimizing Resource Efficiency: To address high maintenance costs, future work could investigate lightweight ML models or automated retraining pipelines specifically designed for resource-constrained environments.

Real-Time Inference Optimization: Research into faster feature extraction techniques could help bridge the gap in response times, making ML more suitable for real-time browsing applications.

VIII. CONCLUSION

The comparative analysis of phishing detection methodologies reveals a fundamental trade-off between operational efficiency and detection depth. While Heuristic Rule-Based systems offer superior response times and high interpretability, making them ideal for resource-constrained environments and real-time browsing, they remain inherently vulnerable to slight URL modifications.

In contrast, Machine Learning models provide the higher accuracy necessary to identify "hidden" patterns and sophisticated threats. However, these benefits come at the cost of high maintenance requirements, including the need for massive datasets and frequent retraining, alongside a "black box" nature that reduces transparency.

Ultimately, the choice between these methodologies is not binary. For a robust cyber security posture, organizations should consider a hybrid approach:

Layer 1: Utilize heuristic rules for an extremely fast first line of defense to catch known threats instantly.

Layer 2: Employ machine learning for deeper, secondary analysis to identify complex, evolving phishing patterns.

By balancing the interpretability of rules with the predictive power of ML, developers can create detection systems that are both fast and resilient against modern cyber threats.

REFERENCES

[1]. Akinyelu, A. A., & Adewumi, A. O. (2019). Classification of phishing email using random forest machine learning algorithm. Journal of Applied Mathematics.

[2]. Elgharbi, S. E., Ait Yahia, M., & Ouchani, S. (2024). Online phishing detection: A heuristic-based machine learning framework. 2024 13th Mediterranean Conference on Embedded Computing (MECO), 1–4. https://doi.org/10.1109/meco62516.2024.10577848

[3]. Jayaprakash, R., Natarajan, K., Daniel, J. A., Chinnappan, C. V., Giri, J., Qin, H., & Mallik, S. (2024). Heuristic machine learning approaches for identifying phishing threats across web and email platforms. Frontiers in Artificial Intelligence, 7, Article 1414122. https://doi.org/10.3389/frai.2024.1414122

[4]. Kandula, L. R. R., Lakshmi, T. J., Alla, K., & Chivukula, R. (2022). An intelligent prediction of phishing URLs using ML algorithms. International Journal of Safety and Security Engineering, 12(3), 381–386. https://doi.org/10.18280/ijsse.120312

[5]. Wave, R. J. (2026). The evolution of phishing datasets and maintenance in machine learning models.