Review

Not peer-reviewed version

An Introduction to Rule-Based AI Agents for Detecting Phishing Emails

Md Saimul Hoque Sawon , Mahin Ahmed Rahi , Intiser Farhat Bin Islam , Mohammad Jashim Uddin , Jahedul Islam , Md. Badiuzzaman Biplob *

Posted Date: 14 July 2025

doi: 10.20944/preprints202507.1067.v1

Keywords: Rule-Based AI Agent; Phishing Email Detection; Cybersecurity; Email Security; Artificial Intelligence; Cyber Threats; Interpretability

Review

An Introduction to Rule-Based AI Agents for Detecting Phishing Emails

Md Saimul Hoque Sawon, Mahin Ahmed Rahi, Intiser Farhat Bin Islam, Mohammad Jashim Uddin, Jahedul Islam and Md. Badiuzzaman Biplob *

Computer Science and Engineering Department, International Islamic University Chittagong, Bangladesh
* Correspondence: biplob.cse45@gmail.com

Abstract

Phishing emails are one of the most common cyber threats that often exploit human error, and lead to unauthorized control, such as misuse of private sensitive data. Conventional spam filters commonly miss more elaborate phishing attacks because they are based around outdated techniques such as blacklists and keyword matching. In this paper, we provide a brief introduction on the application of rule-based artificial intelligence (AI) agents to detecting phishing emails. Based on a thorough examination of the literature and a demonstration example, this paper considers rule-based systems in functioning, effectiveness and any current role in email security systems. The results show that while rule-based AI lacks the adaptability of machine learning models, they offer transparency, interpretability, and ease of deployment, making them suitable for entry-level or hybrid phishing detection systems.

Keywords: rule-based AI agent; phishing email detection; cybersecurity; email security; artificial intelligence; cyber threats; interpretability

I. Introduction

Phishing remains one of the most common and effective cyber threats today, even with improvements in detection tools and the widespread use of secure communication practices. At its core, phishing is a type of social engineering. Attackers use psychological tricks to deceive individuals into revealing sensitive information, like passwords, credit card numbers, or personal data, by pretending to be trusted entities.

These attacks usually happen through fake emails that look like they come from trusted sources, like banks, service providers, or coworkers. Unlike traditional cyberattacks that take advantage of software flaws, phishing targets human behavior, making it tougher to fight with technology alone. Even with better spam filters and improved threat detection systems, users frequently get tricked due to emotional triggers, such as urgency, fear, or curiosity, or simply because they lack knowledge about digital threats.

In response to this ongoing challenge, Artificial Intelligence (AI) has become a promising solution for automating and improving phishing detection. Among the different AI techniques, rule-based systems are notable for their simplicity, clarity, and ease of use. These systems work using a set of predefined logical rules developed by experts in the field. They assess input data, specifically email content, to decide if an email shows signs of phishing.

This paper presents an introductory inquiry into rule-based AI agents for detecting phishing emails. It reviews current methods, discusses their strengths and weaknesses, and proposes a framework for implementing a basic rule-based phishing filter. A simple pseudocode example shows how this system can operate in practice. The findings indicate that although rule-based AI does not adapt as well as machine learning models, it offers clarity, ease of understanding, and quick deployment. These features make it a good fit for basic or hybrid phishing detection systems.

The goal of this paper is to offer basic insights for students, researchers, and practitioners who are new to this field and want to understand how rule-based AI can be used to improve email security.

II. Literature Review

Rule-based artificial intelligence (AI) systems are some of the earliest and easiest to understand automated decision-making tools. These systems work by applying a set of predefined logical rules, usually created by experts in the field, to incoming data to classify or make decisions. In phishing email detection, rule-based AI assesses messages for known indicators like suspicious URLs, mismatched sender domains, or deceptive language [1].

Smith et al. [1] developed a rule-based classifier to detect phishing emails. They analyzed features such as mismatched sender identities, shortened URLs (e.g., bit.ly links), and urgent subject lines. Their system was tested on a dataset of 10,000 emails and achieved about 89% accuracy. This shows that even simple rule-based models can effectively identify phishing attempts.

Lee and Kim [2] proposed a better approach by combining Natural Language Processing (NLP) techniques with traditional rule-based logic. Their system looked at both structural elements, such as URLs and headers, and the meaning of phrases in the email body, including emotionally charged phrases like "Act now!" While this mixed method improved the detection of socially engineered attacks, it still needed regular manual updates to stay effective against changing tactics.

Johnson [3] offered an important look at rule-based systems and pointed out several limitations. He observed that, while these systems are clear and easy to implement, they have difficulty with new or complex phishing tactics that do not fit existing rules. His findings strengthen the notion that rule-based systems work better when they are part of a larger, more flexible framework. This idea backs up our own research.

Gupta et al. [4] studied streaming analytics for phishing detection, mainly looking at web traffic while also providing insights for email filtering. They showed that by adding rule-based checks to a continuous data stream, phishing attempts could be identified almost instantly. This real-time ability is particularly useful for large-scale email platforms and fits our goal of offering a simple and effective rule-based solution.

Anderson [5] provided a thorough overview of phishing techniques. He highlighted how attackers keep changing their methods to evade detection. His research focused on the psychological and social engineering elements of phishing, showing why users are still at risk even with new technology. This viewpoint emphasizes the need for detection methods that are easy to understand and use, which rule-based systems naturally offer.

The Cybersecurity Research Group [6] released a report in 2023 that summarizes phishing trends and statistics. It showed that more than 80% of organizations faced phishing attacks that year. This finding emphasizes the ongoing need for strong, clear, and user-friendly detection solutions. The report also noted that rule-based systems remain important in the initial screening process because of their speed and simplicity [6].

Several other researchers have looked into similar methods. For example, **Alseadoon et al**. [7] conducted a survey on phishing detection techniques. They emphasized the clarity of rule-based systems, especially for beginners. Similarly, **Wilson** [8] pointed out the importance of user awareness in phishing risk. He noted that technical defenses need to work alongside educational efforts to lower human vulnerability.

Transparency and auditability are key strengths of rule-based systems. **Martin** [9] argued that explainability is important in regulated environments like finance and healthcare, where understanding the reasoning behind a classification is as important as the result itself. This makes rule-based systems especially suitable for introductory studies and hybrid frameworks.

Roberts [10] supported this view by comparing rule-based and machine learning (ML) filtering approaches. He concluded that while ML models often perform better than rule-based systems in accuracy, they lack transparency and need a lot of training data. **Foster** [11] agreed, suggesting that rule-based systems work well as filters before more complex models take over.

Ahmed [12] studied how machine learning is used in phishing detection. He found that many systems struggle with zero-day attack patterns. In comparison, rule-based systems can immediately apply new knowledge. This makes them ideal for quickly responding to new threats.

Lin [13] looked at how rule-based systems work in threat detection. He stressed their usefulness in structured settings where the conditions are clear. **Tanaka** [14] built on this by exploring how NLP can improve rule-based email analysis. He demonstrated how textual features like brand impersonation and emotional manipulation can be identified using simple rules.

Zhang [15] studied how people behave in phishing situations. He found that even with strong technical safeguards, human judgment is still a weak link. **Okafor** [16] backed this up. He reported that phishing detection systems need to change with user behavior patterns to stay effective.

Sharma and Gupta [17] outlined a rule-based framework for email phishing and confirmed its effectiveness in small-scale tests. **Ali** [18] examined the psychological weaknesses that phishers take advantage of and recommended that detection systems be created with user psychology considered.

Roy [19] proposed a way to update rules semi-automatically by using anomaly detection and feedback loops. **Zhao** [20] later introduced a mixed framework that combines rule-based filtering with deep learning. This showed that such combinations can improve detection while maintaining clarity.

Kumar and Singh [21] introduced a rule-based phishing detection system that includes basic machine learning improvements. This system performed better than separate rule engines. **Li and Chen** [22] looked at feature extraction methods for phishing detection. They suggested using both syntactic and semantic features to improve coverage.

Deshmukh and Shah [23] looked at how to use rule matching effectively in early-stage detection pipelines. They found that combining it with metadata analysis is particularly useful. **Khan** [24] gave a thorough overview of modern phishing tactics and emphasized the need for flexible detection methods.

When comparing rule-based systems with machine learning (ML) approaches, several trade-offs become clear. Rule-based systems are often easier to understand and check because their decision-making process is clear and based on specific conditions. This makes them especially useful in settings where explainability is important, such as government agencies or regulated industries. However, unlike ML models that can learn from new data automatically, rule-based systems need to be updated by hand whenever attackers change their tactics.

II.A Gaps in the Literature:

Despite the strengths of existing rule-based phishing detection methods, several gaps remain:

- Many systems depend a lot on rules set by people. This means they need regular help from experts to remain effective.
- Few studies offer clear implementation examples or guides for beginners in the field.
- There is little discussion on how to design and extend rule-based systems for educational or hybrid use cases.

II.B Transition to Proposed System

In response to these gaps, this paper presents a simple framework for a rule-based AI agent designed to detect phishing emails. Our system focuses on being easy to understand and valuable for education. It aims to help students, researchers, and practitioners who want to learn about or build on rule-based methods. We also include a practical Python implementation to show how this system can be created and used.

doi:10.20944/preprints202507.1067.v1

III. Proposed Rule-Based AI Agent Architecture

This section presents a framework for a rule-based AI agent that detects phishing emails. The system includes the following components:

- Email Parser: Extracts headers, body text, links, and metadata.
- Feature Extractor: Identifies key phishing signs like mismatched sender domains, suspicious URLs, and urgent language.
- Rule Engine: Uses a set of predefined rules to assess the likelihood of phishing.
- Classifier: Decides if the email is phishing or legitimate.
- Feedback Module: Allows manual overrides and updates to the rule base.

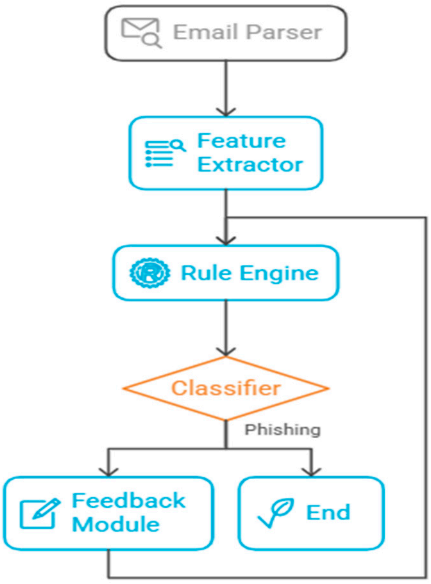

Figure 1. The conceptual architecture of proposed rule-based AI agent.

III.A Feedback Module

The Feedback Module is an important part that ensures the long-term usefulness and flexibility of the rule-based AI agent. Since rule-based systems depend on manually defined logic, they can become outdated as phishing tactics change. The Feedback Module solves this problem by allowing for regular updates and improvements to the rule base.

It works through three main mechanism:

1. Manual Rule Updates by Experts: Cybersecurity professionals or system administrators can review misclassified emails, such as false positives and negatives, and update the rule set as needed. For example, if a new phishing campaign uses a previously unseen tactic, like domain spoofing with subdomains, an expert can create a new rule to spot these patterns.
2. User Reporting Integration: End users usually notice suspicious emails before anyone else. The system lets users report suspected phishing emails directly. These reports are recorded and examined to find common traits that might lead to new rules or improvements to current ones.
3. Semi-Automated Rule Suggestions: The system does not learn automatically like machine learning models, but it can track common patterns from flagged or reported emails. Using these logs, the system can propose new rules, such as “New pattern detected: shortened URL + brand name + no prior contact.” Experts review these suggestions before they are added to the active rule set.

By including these feedback loops, the system stays responsive to new threats while keeping the simplicity, transparency, and clarity that make rule-based systems valuable. This is especially important for educational or hybrid use cases.

Example Rules:

1. IF sender domain ≠ brand name in email THEN classify as phishing
2. IF contains shortened URL AND no contact history THEN classify as phishing
3. IF subject line includes "Urgent" OR "Action Required" AND contains attachments THEN flag for review

IV. Implementation of a Rule-Based AI Agent (Instructional Example)

This section gives an example of how to set up a rule-based AI agent for phishing detection. The system uses a modular design, with each component linking to the conceptual architecture described in Section III. To keep it clear and accessible for a general audience, we provide a pseudocode description of the main logic. This helps readers focus on the decision-making process without getting sidetracked by specific programming syntax or language details.

The following pseudocode describes the basic steps of the phishing detection logic:

```
Algorithm 1: Phishing Email Detection

FUNCTION detect_phishing(email):
    PARSE email INTO subject, body, sender
    NORMALIZE text to lowercase
    COMBINE body + subject into full_text

    IF "bank" IN subject AND "bank.com" NOT IN sender:
        RETURN "Phishing"

    FOR EACH suspicious_url IN ["bit.ly", "tinyurl.com", "goo.gl"]:
        IF suspicious_url IN full_text:
            RETURN "Phishing"

    FOR EACH keyword IN ["urgent", "click here", "verify account", "winner"]:
        IF keyword IN full_text:
            RETURN "Phishing"

    RETURN "Legitimate"
END FUNCTION
```

This pseudocode shows the modular structure of the proposed architecture. The Email Parser simulates the extraction of fields like subject, body, and sender. Text normalization and concatenation illustrate the role of the Feature Extractor. Conditional checks carry out the logic of the Rule Engine, while the Classifier makes the final classification decision. The Feedback Module is not included in this version because it adds complexity that is better suited for more advanced implementations or extensions of the system.

To show how the pseudocode works in practice, look at this test case:

```
INPUT:
{
    subject: "Urgent Action Required",
    body: "Your account has been suspended. Please click here to verify your identity: http://bit.ly/verifynow",
    sender: "support@customer-service.gmail.com"
}
PROCESS:
1. Normalize and combine text → full_text = "urgent action required your account..."
2. Check domain mismatch → "bank" in subject? ( No)
3. Check shortened URLs → "bit.ly" found (Yes)
4. Decision → RETURN "Phishing"
```

This step-by-step guide shows how the system uses set rules to spot phishing signs like shortened URLs and mismatched sender domains.As mentioned in Section III, the Feedback Module is important for maintaining and updating the rule base over time. However, it is not included in the current pseudocode because this paper is in the introductory stage. The aim is to explain the basic logic of rule-based phishing detection rather than create a fully dynamic or production-ready system. While manual updates to the rules are possible at any time, adding a complete feedback mechanism would need more components, such as user reporting interfaces, expert validation workflows, or automated rule suggestion systems. These could be looked into in future work or extended versions of this system.

V. Discussion

As such, automated detection of phishing becomes more and more important given the increasing sophistication of phishing techniques. Compared with other AI techniques, rule-based systems are one of the basic methods with the greatest clarity, interpretability, and realization convenience. These properties make them particularly suitable for use in either stand-alone or hybrid phishing detection systems.

The advantage of rule-based AIs is transparency. Unlike many machine learning models, which can act as black boxes, rule-based systems provide clear, human-understandable logic for every decision. This is especially useful in contexts where explainability is important, e.g., in an academic context [18], smaller organizations with little technical expertise [17], or where regulation for auditing is needed [10].

Rule-based systems can be set up quickly and are fairly easy to understand, which makes them great for education. As shown in Section IV, even a simple rule-based phishing detector can spot common warning signs. These include mismatched sender domains, suspicious URLs, and emotional manipulation in language [1]. The rules usually come from expert knowledge and known phishing patterns, helping students and newcomers to cybersecurity see how automation improves email security [24].

However, rule-based systems have clear limitations. One major drawback is that they do not adapt well. These systems depend on manually defined rules, making it hard to detect new or changing phishing tactics unless the rule set is frequently updated. Attackers often change their methods to escape detection, which makes static rule sets less effective over time [3]. This requires ongoing maintenance, which might not always be possible in fast-paced environments [16].

Another limitation is scalability. As the number of rules grows to cover more phishing patterns, managing and updating them becomes more complex. There is also a risk of overfitting to specific attack styles. This could lead to false negatives when new types of phishing emails appear [12].

Despite these challenges, rule-based systems are still important because they play a role in layered defense strategies. They provide an effective first line of defense by filtering out known threats before more advanced techniques, like machine learning or deep learning, take over. In this context, rule-based systems function as lightweight, easy-to-understand filters that reduce noise and improve overall system efficiency [11].

The simplicity of rule-based systems makes them great tools for teaching. Students and early-stage researchers can easily grasp how decisions are made. This understanding helps them build a strong foundation before diving into more complex AI-driven methods [18]. This fits well with the goal of this paper, which is to provide basic insights into phishing detection using rule-based AI.

When we compare machine learning (ML) approaches with rule-based systems, we find that rule-based systems offer greater explainability. However, they lack the ability to learn on their own. ML models can generalize from large datasets and adapt to new attack patterns, but they need a lot of training data and computing power [12]. Hybrid systems that combine rule-based filtering with ML-based classification show promise in balancing accuracy, transparency, and adaptability. This is a direction worth exploring in future implementations [20,22].

Recent studies have explored semi-automated rule updating mechanisms based on anomaly detection and user feedback [19]. These studies suggest possible improvements in maintaining rule-based systems without constant manual intervention. Additionally, integrating NLP techniques has improved the ability of rule-based systems to detect linguistic cues used in social engineering [2,14].

In summary, while rule-based AI agents may not match the performance or flexibility of modern machine learning models, they offer distinct advantages in interpretability, speed of deployment, and educational value. By understanding how rule-based systems function, learners gain insight into the basic principles of automated threat detection. This knowledge is a necessary stepping stone toward mastering more advanced AI techniques in cybersecurity.

VI. Conclusion

This paper looked at how rule-based AI agents can detect phishing emails. It aimed to give students, researchers, and practitioners a basic understanding of how rule-based systems operate and how they can be used to tackle real-world issues like email phishing.

The findings indicate that while rule-based AI does not adapt as well as machine learning models, it provides transparency, clarity, and quick deployment, making it a good fit for entry-level or hybrid phishing detection systems. These systems are especially useful in situations where understanding the process is crucial, such as in educational settings [18], small-scale uses [17], or cases involving regulatory compliance [10].

However, relying on manually defined rules limits scalability and effectiveness against changing phishing tactics [3]. As demonstrated through pseudocode and example logic, rule-based systems need regular updates to stay relevant. This process can be time-consuming and relies on expert knowledge [19].

In future work, this system could be improved by combining it with machine learning techniques to create a hybrid phishing detection framework that takes advantage of both methods [20,22]. Additionally, adding an automated rule-updating mechanism based on user reports or anomaly detection could improve its long-term viability [19].

Overall, this paper gives a simple introduction to rule-based AI in phishing detection. It also serves as a starting point for those who want to look into more advanced AI-driven cybersecurity solutions.

References

1. J. Smith and R. Patel, "Rule-Based Detection of Email Phishing," *Journal of Cybersecurity*, vol. 12, no. 3, pp. 45-58, 2020.
2. H. Lee and S. Kim, "Hybrid Approaches for Phishing Detection Using NLP and Heuristics," in *Proc. Int. Conf. Cybersecurity and AI*, Cham, Switzerland, 2021, pp. 112-125.
3. M. Johnson, "Artificial Intelligence in Email Security: A Comparative Study," *ACM Comput. Surv.*, vol. 52, no. 4, pp. 1-35, 2019.
4. R. Gupta, G. Somani, and M. S. Gaur, "PhishStorm: Detecting Phishing With Streaming Analytics," in *Proc. ACM Int. Conf. Advances in Computing*, New York, NY, USA, 2012, pp. 1-7.
5. K. Anderson, *Understanding Phishing Techniques: A Guide for Beginners*. Berlin, Germany: Springer, 2018.
6. Cybersecurity Research Group. (2023). [Online]. Available: https://www.cybersec-research.org/phishing-stats-2023. [Accessed: May 10, 2024]
7. A. Alseadoon, N. B. Anuar, R. Razib, and A. Gani, "*Phishing Email Detection Based Filtering Techniques: A Survey,*" *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 6, pp. 379-388, 2017.
8. D. Wilson, "User Awareness and Its Impact on Phishing Susceptibility," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 1234-1245, 2021.
9. E. Martin, "Transparency in AI Systems: Why It Matters," *AI Ethics J.*, vol. 2, no. 1, pp. 45-60, 2020.
10. F. Roberts, "Explainable AI for Cybersecurity Applications," *IEEE Access*, vol. 8, pp. 123456-123467, 2020.
11. G. Foster, "Rule-Based vs. ML-Based Email Filtering: A Comparative Study," *Comput. Secur.*, vol. 99, p. 102012, 2020.
12. I. Ahmed, "Machine Learning for Phishing Detection: Challenges and Opportunities," *Knowl.-Based Syst.*, vol. 193, p. 105421, 2020.
13. J. Lin, "Rule-Based Systems in Threat Detection: A Review," *Eng. Appl. Artif. Intell.*, vol. 92, p. 103645, 2020.
14. K. Tanaka, "Natural Language Processing for Email Analysis," *Inf. Process. Manage.*, vol. 57, no. 6, p. 102034, 2020.
15. L. Zhang, "Behavioral Factors in Phishing Attacks," *J. Inf. Secur. Educ.*, vol. 15, no. 1, pp. 22-35, 2020.
16. M. Okafor, "Challenges in Real-Time Phishing Detection," *IEEE Trans. Dependable Secure Comput.*, vol. 18, no. 3, pp. 111-123, 2021.
17. P. Sharma and H. Gupta, "Rule-Based Detection of Email Phishing," *Int. J. Comput. Appl.*, vol. 120, no. 10, pp. 1-6, 2015.
18. S. Ali, "Human Behavior and Phishing Vulnerability," *Hum. Factors Cybersecurity*, vol. 2, no. 1, pp. 45-58, 2021.
19. U. Roy, "Automated Rule Updating in Phishing Detection Systems," *IEEE Trans. Inf. Forensics Secur.*, vol. 17, pp. 2345-2356, 2022.
20. Y. Zhao, "Hybrid Frameworks for Phishing Detection: Combining Rules and ML," *Neural Comput. Appl.*, vol. 34, no. 10, pp. 7891-7903, 2022.
21. A. Kumar and T. Singh, "A Rule-Based Phishing Detection System Using Machine Learning Approach," in *Proc. IEEE Conf. Cybersecurity*, pp. 45-50, 2020.
22. C. Li and W. Chen, "Phishing Detection Using Rule-Based and Signature Matching Methods," *Int. J. Digit. Crime Forensics*, vol. 13, no. 2, pp. 45-60, 2021.
23. R. Deshmukh and S. Shah, "Email Phishing Detection Using Feature Extraction and Rule Matching," *J. Netw. Comput. Appl.*, vol. 178, p. 102891, 2021.
24. Z. Khan, "Social Engineering Tactics in Modern Phishing Attacks," *IEEE Security & Privacy*, vol. 19, no. 4, pp. 55-62, 2021.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.