IJARETY

International Journal of Advanced Research in Education and TechnologY (IJARETY)

Volume 12, Issue 6, November-December 2025

Impact Factor: 8.152

www.ijarety.in editor.ijarety@gmail.com

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 6, November-December 2025 ||

DOI:10.15680/IJARETY.2025.1206012

Detection of Phishing Websites using Data Mining Techniques: A Review

Fasseela Mol KJ, Shejin Mathulla Thomas, Smitha C Thomas

PG Student, Department of Computer Science and Engineering, APJ Abdul Kalam Technological University, Kerala, India

Assistant Professor, Department of Computer Science and Engineering, APJ Abdul Kalam Technological University, Kerala, India

Professor, Department of Computer Science and Engineering, APJ Abdul Kalam Technological University, Kerala, India

ABSTRACT: Phishing has emerged as one of the most widespread cyber threats targeting millions of users worldwide. Modern phishing attacks imitate legitimate websites to trick victims into revealing sensitive information such as passwords, banking details, and personal identity data. This review article examines existing techniques used for phishing detection, including email-based filtering, visual similarity analysis, fuzzy logic models, and machine-learning-based approaches. Special focus is given to data-mining techniques and the RIPPER rule-based classifier, which have demonstrated promising accuracy in detecting newly generated phishing URLs with no previous history. The article presents a comparative analysis of major methods, highlights challenges in zero-day phishing detection, and outlines future research directions for building more robust phishing detection systems.

KEYWORDS: Phishing Detection, Data Mining, Machine Learning, Fuzzy Logic, RIPPER Algorithm, Cybersecurity, URL Analysis

I. INTRODUCTION

Phishing attacks continue to rise globally as attackers exploit user trust by creating fake websites and deceptive email messages. These attacks often mimic banks, online shopping platforms, and financial service providers. With the increasing sophistication of phishing campaigns, detection has become challenging, especially for newly generated URLs that do not appear in traditional blacklists.

Traditional defense mechanisms, such as browser filters, anti-spam systems, and visual verification methods, are effective only against known phishing pages. However, these approaches frequently fail when facing zero-day phishing attacks. As a result, researchers have turned towards data mining and machine learning models that analyze URL features, domain attributes, and email content to identify suspicious behaviour.

This review provides an analytical overview of major phishing detection techniques and evaluates the role of data mining—particularly the RIPPER algorithm—in building adaptive and intelligent detection systems.

II. OVERVIEW OF PHISHING TECHNIQUES

Phishing techniques generally fall into several categories:

2.1 Email-Based Attacks

Attackers send fake emails that contain malicious links or attachments. These emails often resemble official communication from trusted organizations.

2.2 Website Spoofing

Fake websites visually replicate legitimate login pages to steal credentials. Attackers manipulate URLs, use similar domain names, or hide malicious redirects.

2.3 DNS Hijacking and Domain Manipulation

Cybercriminals alter DNS entries to redirect legitimate web traffic to fraudulent pages.

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 6, November-December 2025 ||

DOI:10.15680/IJARETY.2025.1206012

2.4 Social Engineering Techniques

Attackers use fear, urgency, or reward-based messages to manipulate user behaviour.
Understanding these techniques is essential for designing effective detection models.

III. EXISTING ANTI-PHISHING SOLUTIONS

3.1 Blacklist and Heuristic Filters

Browsers maintain lists of known phishing sites, but blacklists cannot detect new phishing URLs until they are reported.

3.2 Visual Similarity Detection

Techniques compare webpage screenshots or layout structures to detect spoofing. However, this method fails when attackers make subtle modifications.

3.3 Two-Factor Authentication (2FA)

Although 2FA protects accounts, it cannot prevent attackers from stealing other sensitive information such as credit card numbers.

3.4 Browser Toolbars and Security Plugins

Toolbars provide warnings, but studies show that many users ignore alerts, and plugins often fail to detect advanced spoofing attacks.

3.5 Machine Learning Approaches

Machine learning algorithms classify URLs or emails using extracted features. Algorithms such as decision trees, neural networks, Naive Bayes, and SVMs have been widely studied.

IV. DATA MINING AND FUZZY LOGIC IN PHISHING DETECTION

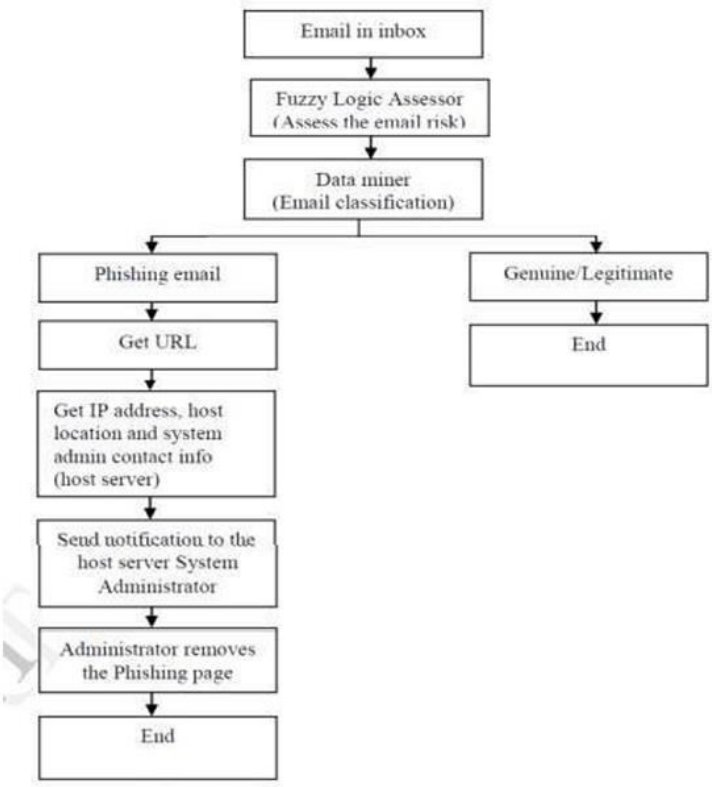

Fig 1 Overall approach

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 6, November-December 2025 ||

DOI:10.15680/IJARETY.2025.1206012

4.1 Importance of Data Mining

Data mining identifies patterns hidden within large datasets. In phishing detection, it analyses:

- URL structure
- Domain registration details
- Redirect patterns
- Page features
- Email content characteristics

4.2 Role of Fuzzy Logic

Fuzzy logic handles uncertainties in phishing detection. Many phishing features—such as word choice, link appearance, or tone—are not strictly true or false. Fuzzy models interpret these "linguistic indicators" more effectively than binary systems.

4.3 Combined Fuzzy–Mining Models

Hybrid models combine fuzzy rules with data-driven classifiers to improve detection accuracy, especially for ambiguous or borderline phishing URLs.

Classification approach	Category/Criteria	Component	Layer
Non Content Based Approach	URL	IP URL	1
		Redirect URL	
		Non Matching URL	
		Crawler URL	
		Long URL address	
		URL prefix/suffix	
Content Based Approach	Email Message	Spelling errors	2
		Keywords	
		Embedded Links	

V. RIPPER ALGORITHM FOR PHISHING DETECTION

The RIPPER (Repeated Incremental Pruning to Produce Error Reduction) algorithm is a rule-based classifier widely used in data mining.

Advantages of the RIPPER Algorithm

- Produces human-readable rules
- Detects hidden relations between features
- Effective with noisy data
- Adapts well to new phishing patterns

How RIPPER Works

RIPPER processes data in three main phases:

1. **Rule Building** – Generates rules from positive examples.
2. **Rule Optimization** – Removes redundant conditions and selects the best rules using minimum description length (DL).
3. **Rule Pruning/Deletion** – Eliminates low-performing rules to reduce overfitting.

Studies show that RIPPER can achieve **over 85% accuracy** when trained with URL-based features.

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 6, November-December 2025 ||

DOI:10.15680/IJARETY.2025.1206012

ENDDO

Rule	IP URL	Redirect URL	Non Matching URL	Crawler URL	Long address URL	URL prefix/ suffix	Output
1	Low	Low	Low	Low	Low	Low	Genuine
2	Low	Low	Low	Low	Low	Moderate	Genuine
3	Moderate	Low	Moderate	Low	Low	Moderate	Suspicious
4	Low	Low	Low	Moderate	Moderate	Moderate	Suspicious
5	Moderate	Moderate	Moderate	High	High	High	Fraud
6	High	High	High	High	Moderate	Moderate	Fraud

Rule base for Layer 1

Rule	Spelling errors	Keywords	Embedded Links	Output
1	Low	Low	Moderate	Genuine
2	Low	Moderate	Moderate	Suspicious
3	High	High	High	Fraud
4	Low	Low	Low	Genuine
5	Moderate	Low	Moderate	Suspicious
6	High	Moderate	Moderate	Fraud

Rule base for Layer2

VI. COMPARATIVE ANALYSIS OF DETECTION METHODS

Method	Strengths	Limitations
• Blacklists	• Simple, widely used	• Fails on new URLs
• Visual Similarity	• Detects layout spoofing	• Easily bypassed
• Machine Learning	• High accuracy	• Needs training data
• Fuzzy Logic	• Handles uncertainty	• Requires expert rule design
• RIPPER	• Fast, rule-based	• Less effective with highly complex pages

| ISSN: 2394-2975 | www.ijarety.in | | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 6, November-December 2025 ||

DOI:10.15680/IJARETY.2025.1206012

Data mining techniques, especially rule-based classifiers like RIPPER, have shown excellent performance in detecting zero-day phishing pages when combined with fuzzy logic scoring.

VII. CHALLENGES IN PHISHING DETECTION

- Rapid generation of new malicious domains
- Short lifespan of phishing websites
- Attackers using HTTPS, making detection harder
- Use of homograph attacks (e.g., replacing letters with similar characters)
- Evasion through URL shortening services

VIII. FUTURE RESEARCH DIRECTIONS

Future models must focus on:

- Deep learning–based URL interpretation
- Real-time image analysis of webpages
- Detection of multilingual phishing campaigns
- Automated removal of phishing pages
- Behaviour-based browser plugins
- Cloud-based collective intelligence systems

These directions will help create next-generation phishing detection platforms.

IX. CONCLUSION

Phishing remains a persistent and evolving cyber threat. Traditional detection mechanisms are insufficient against newly created phishing URLs. Data mining techniques, fuzzy logic, and rule-based algorithms such as RIPPER offer a highly effective solution by analyzing structural and behavioural patterns. This review highlights the current state of phishing detection research, compares methodologies, and outlines future directions for improving the accuracy and adaptability of detection systems.

REFERENCES

1. Smadi, Sami, et al. "Detection of phishing emails using data mining algorithms." *2015 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*. IEEE, 2015.
2. Şentürk, Şerafettin, Elif Yerli, and İbrahim Soğukpınar. "Email phishing detection and prevention by using data mining techniques." *2017 International Conference on Computer Science and Engineering (UBMK)*. IEEE, 2017.
3. Abdelhamid, Neda, Aladdin Ayesh, and Fadi Thabtah. "Phishing detection based associative classification data mining." *Expert Systems with Applications* 41.13 (2014): 5948-5959.
4. Aburrous, Maher, et al. "Intelligent phishing detection system for e-banking using fuzzy data mining." *Expert systems with applications* 37.12 (2010): 7913-7921.
5. Pandey, Mayank, and Vadlamani Ravi. "Detecting phishing e-mails using text and data mining." *2012 IEEE International Conference on Computational Intelligence and Computing Research*. IEEE, 2012.
6. Ali, Mohd Mahmood, and Lakshmi Rajamani. "APD: ARM deceptive phishing detector system phishing detection in instant messengers using data mining approach." *International Conference on Computing and Communication Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.
7. Jibat, Dina, et al. "A systematic review: Detecting phishing websites using data mining models." *Intelligent and Converged Networks* 4.4 (2023): 326-341.
8. Aburrous, Maher Ragheb, et al. "Modelling intelligent phishing detection system for e-banking using fuzzy data mining." *2009 International Conference on CyberWorlds*. IEEE, 2009.

International Journal of Advanced Research in Education and Technology

ISSN: 2394-2975 Impact Factor: 8.152

www.ijarety.in editor.ijarety@gmail.com