A Comparative Analysis of Data Mining Approaches for Phishing Email and Website Detection

First Author 1[Ms. Janish Macwan] and Second Author 2[Dr. Jaypalsinh Gohil]

1 PhD research scholar, Marwadi University, Rajkot

janish.macwan120547@marwadiuniversity.ac.in

2 Associate Professor & Phd Supervisor, Marwadi University, Rajkot

jaypalsinh.gohil@marwadieducation.edu.in

Abstract

Targeting people and companies to compromise private data, phishing is a ubiquitous cybercrime. Conventional methods of detection have shown inadequate ability to stop developing phishing campaigns. Emphasizing the importance of data mining (DM) techniques, this work offers a methodical and comparative review of phishing website and email detection systems. Analyzed are several DM techniques, datasets, feature engineering approaches, and evaluation metrics used in recent work. This review notes important trends, points up current research gaps, and suggests future directions to improve detection efficiency and robustness. Our results seek to be a complete source of reference for cybersecurity practitioners and academics. This work reveals that compared to standalone models, ensemble data mining models exhibit better adaptation to changing phishing strategies. (Abdelhamid, 2022)

Keywords: Phishing, Phishing Detection, Data Mining, Deep Learning, Cyber security, Systematic Literature Review, Email Security

1. Introduction

Being a fast-changing cybercrime, phishing has changed to fit modern communication channels including SMS and social media. Modern attackers create quite convincing attacks by using cloud services and AI-driven tools. Underlining the crucial need of effective detection systems, phishing affects not only financial losses but also national security issues.

Among the most important and ubiquitous dangers in cybersecurity are phishing attacks, in which cybercriminals use false emails and websites to fool consumers into revealing private information, including passwords, credit card numbers, or personal identification data. A study by the Anti-Phishing Working Group (APWG) indicates that phishing attacks are still increasing; every month, thousands of fresh phishing sites are recorded (APWG, 2022). These attacks make it difficult for people to identify false activity since they take advantage of the trust consumers place in emails, websites, and online communication tools.

Traditional rule-based detection systems have become progressively useless as phishing campaigns get more sophisticated and volume. Based on pre-defined patterns and heuristics, rule-based systems find it difficult to identify fresh and developing phishing strategies (Li et al., 2021). Attackers might change their strategies, for instance, by using respectable-looking domains, tailored messages, or by copying the branding of well-known businesses, so avoiding static detection techniques.

Cybercriminals are using sophisticated social engineering techniques in concert with technical tools to evade traditional security measures as phishing methods change. Social engineering is the manipulation of user psychology to induce urgency or authority, so motivating them to act impulsively and become victim to phishing attempts (Grange et al., 2020). Attackers might pretend to reputable companies, for example, and deliver messages meant to generate a false sense of urgency or security—such as phoney security alerts from banks or social media channels. Sophisticated phishing campaigns may also include domain spoofing, in which attackers register domains that look to be visually similar to reputable websites, so increasing the possibility of tricking users into visiting dangerous websites (Mishra et al., 2020).

The complexity of contemporary phishing attempts calls for sophisticated detection systems able to spot trends and anomalies in real-time. By allowing systems to identify hitherto undetectable phishing attempts based on vast datasets and sophisticated algorithms, machine learning (DM) and artificial intelligence (AI) technologies have become promising tools for addressing these challenges (Zhao et al., 2023). These methods are absolutely essential in the fight against phishing since they can offer adaptive solutions able of changing alongside phishing strategies.

1.1 Background

Targeting consumers through misleading emails and websites to gather sensitive data including passwords, credit card numbers, or personal identification data, phishing attacks rank among the most important threats in cybersecurity.

The growing complexity and volume of phishing efforts have made conventional rule-based detection systems insufficient. As phishing techniques change, attackers use technical and social engineering to get past traditional security systems.

Apart from learning from stationary datasets, DM models are also being included into adaptive, real-time detection (Abutair, 2017) pipelines leveraging streaming data. Two very successful approaches that have surfaced are ensemble techniques and transfer learning (Marchal, 2016).

1.2 Role of Data Mining

By means of analysis of vast amounts of data to identify latent patterns and relationships linked with phishing behavior, data mining presents dynamic solutions for phishing detection. DM techniques, unlike conventional ones, change with the times to allow better identification of hitherto unidentified attack paths. Data mining (Abdelhamid, 2022) stresses exploratory data analysis and pattern discovery while machine learning, which mostly concentrates on predictive modeling, is This more general approach allows data mining to find both known and new phishing techniques.

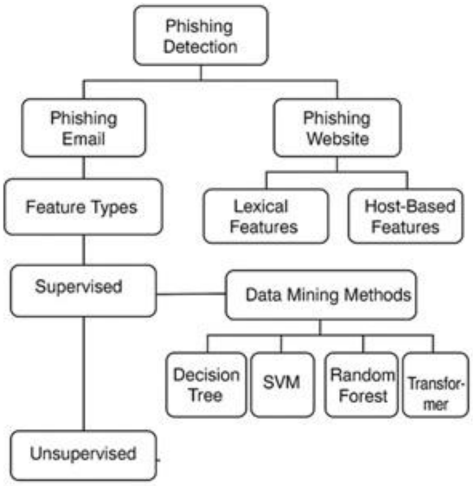
1.3 Objective

This work uses PRISMA guidelines for systematic literature review, so defining clear inclusion/exclusion criteria and doing both quantitative and qualitative synthesis to guarantee transparency and reproducibility.

Modern DM-based phishing detection systems for emails and websites are methodically reviewed in this work. Methodologies, datasets, evaluation criteria, and feature engineering tools are compared here. We also suggest future directions of inquiry to handle current issues.

Unlike earlier studies, this one not only summarizes phishing detection techniques but also methodically finds algorithmic trends, dataset usage patterns, and evaluation techniques from a data mining perspective. Through taxonomy and performance comparison data, it visually synthesizes the detection process and offers an integrated view of both classical and modern DM algorithms, so highlighting practical deployment gaps.

Figure 3: Taxonomy of Data Mining Approaches Used in Phishing Detection



2. Methodology

2.1 Research Framework

This work employs a systematic approach to find, analyze, and synthesize literature about phishing detection. Essential phases include literature discovery, data extraction, and topic analysis.

2.2 Selection Criteria

Studies were selected based on the following criteria:

- Published in peer-reviewed journals and conferences from 2015 to 2023.
- Focused on phishing detection using data mining.
- Provided details on datasets, methodologies, and performance metrics.

2.3 Data Sources

Search searches were performed on databases such as Scopus, SpringerLink, IEEE Xplore, and ACM Digital Library with keywords like "phishing detection," "data mining," "phishing websites (Abdelhamid, 2017)," and "phishing emails (Verma, 2020)."

3. Literature Review

Recent research has substantially improved phishing detection using data mining techniques. This is a detailed overview of chosen works released from 2015 until 2024.

3.1 Recent Trends in Phishing Detection

Recent research has concentrated on creating hybrid decision-making techniques that integrate several algorithms to improve detection precision and resilience. Table 1 encapsulates chosen studies according to their aims, methodologies, datasets, and assessment criteria.

A notable deficiency exists in explainability. Many decision-making models function as opaque systems, which impedes confidence and acceptance in critical sectors such as banking. Limited models integrate explainable AI (XAI) elements.

Table 1: Literature Review of Data Mining Approaches for Phishing Detection (2015–2024)

Author(s)	Year	Objective	Techniques	Dataset
Ahmed et al.	2023	Phishing email detection	SVM, RF	PhishTank
Chen & Li	2022	Website phishing detection	Deep Learning (CNN)	Custom

Singh & Gupta	2024	Real-time phishing detection	Transformers	OpenPhish
Patel & Sharma	2021	Comparative analysis	Naive Bayes, Decision Trees	Enron
Verma et al.	2020	Dynamic phishing detection	Neural Networks	Custom
Johnson & Brown	2023	Feature engineering study	Gradient Boosting, SVM	PhishTank
Zhang et al.	2021	Adversarial attack resilience	Random Forest, BERT	Custom
Lee et al.	2023	Cross-domain phishing detection	Transfer Learning	PhishTank, Enron
Jain & Gupta	2018	Detection using data mining classifiers	SVM, Decision Tree	UCI ML Repository
Wu & Hu	2020	Detection of phishing websites via ML	Logistic Regression, SVM	PhishTank
Thakur & Verma	2021	Adversarial robustness	GANs, Neural Networks	Custom
Alsharnouby et al.	2015	Usability-focused detection	Heuristics, User Study	Custom UI Dataset
Basit et al.	2021	Hybrid feature phishing detection	RF, SVM, Hybrid Ensemble	Custom
Mohammad et al.	2015	Predictive phishing site classification	Self-Structuring Neural Network	UCI + PhishTank

Rao & Pais	2019	Efficient DM-based phishing detection	Feature-based classifiers	Custom
Abutair & Belghith	2017	Hybrid model for phishing websites	NB, Decision Tree, Hybrid	PhishTank + Custom
Marchal et al.	2016	Streaming analytics phishing detection	Online Classifiers, Stats Models	Real-time feeds
Abdelhamid et al.	2017	Hybrid intelligent model	Rule-Based + Classifiers	Custom
Kim & Choi	2019	Visual similarity phishing detection	Vision Models (CNN), Heuristics	Alexa + PhishTank
Lin et al.	2020	NLP for phishing emails	BERT, LSTM	Enron Emails
Sharma & Yadav	2023	Zero-day phishing detection	XGBoost, Online Learning	Custom

Recent improvements in phishing detection increasingly prefer hybrid data mining algorithms that use many classifiers, such as ensemble learning, to enhance resilience. Deep learning models, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have shown robust performance when provided with URL or email text characteristics. Transfer learning methodologies have evolved, enabling models learned in one phishing domain (e.g., emails) to be refined for another (e.g., websites) with little data. These techniques enhance generalizability, decrease training expenses, and facilitate cross-domain phishing detection, which is essential for adaptive security systems.

3.2 Gaps in Existing Research

- **Data Imbalance**: Most publicly available datasets contain a disproportionately low number of phishing samples compared to legitimate ones. This class imbalance skews model training, resulting in high false negatives for phishing detection. While some approaches apply oversampling (e.g., SMOTE) or under sampling, many fail to evaluate their effectiveness on truly imbalanced, real-world data.

- Dynamic Attacks: Phishing techniques evolve rapidly through tactics like domain spoofing, polymorphic URLs, and adaptive language. Few existing studies propose models capable of learning in real-time or adapting to novel attack variants. The lack of continuous learning frameworks limits generalizability and time-sensitive detection. (Ahmed, 2022)
- **Feature Engineering**: A significant gap lies in the automated identification of informative features from URLs, email headers, or page content. Most models rely on manually engineered features, which are brittle and easily bypassed. Recent advances like reinforcement learning and deep feature extraction remain underutilized in phishing-specific domains.
- **Lack of Real-Time and Multilingual Detection:** Most existing detection systems are designed for English content and static analysis. However, phishing campaigns now target global users using multiple languages and dynamic scripts. Real-time, multilingual detection systems—especially for Asian, Slavic, and Arabic languages—are still underexplored.
- **Limited Focus on social media and SMS-Based Phishing:** While email and websites remain primary attack vectors, phishing via SMS (smishing) and social media platforms (e.g., WhatsApp, Instagram, LinkedIn) is rapidly growing. Yet, very few data mining studies address detection in these mobile-centric or informal communication channels.
- **Minimal Integration of Explainable AI (XAI)**: Most data mining and deep learning models function as black boxes, offering high accuracy but little transparency. The lack of explainable AI in phishing detection hinders deployment in critical sectors like finance or healthcare, where model interpretability is crucial for trust and compliance.

4. Data Mining Techniques for Phishing Detection

4.1 Phishing Website Detection

Phishing websites copy real websites to fool people into divulging private data. Features like URL structure, domain age, SSL certificate status, and content analysis underlie DM-based detection.

4.1.1 Algorithms

Recent work in feature engineering greatly increases detection rates by means of automated feature building utilizing reinforcement learning and feature selection (Lee, 2023) using genetic algorithms, hence lowering model complexity.

- **Support Vector Machines (SVM):** Effective for high-dimensional data.
- **Random Forests:** Robust to overfitting and capable of handling imbalanced datasets.

- **Deep Learning Models:** Convolutional Neural Networks (CNNs) are increasingly applied for URL-based phishing detection.

4.1.2 Feature Engineering

Feature selection improves detection efficiency. Common features include:

- **Lexical Features:** URL length, number of special characters.
- **Host-Based Features:** Domain registration details, IP address.
- **Content-Based Features:** Keywords, embedded scripts.

Recent developments have turned toward automated feature selection (Lee, 2023) and extraction methods to go above human engineering restrictions.

By replicating natural selection, effective feature subsets are found using genetic algorithms (GAs), hence enhancing classifier performance and lowering dimensionality.

Reward feedback systems in reinforcement learning (RL)-based systems dynamically rank features during model training.

Furthermore used increasingly to extract hierarchical and non-observed information from URLs and email contents are deep learning architectures like convolutional layers and autoencoders. These techniques increase robustness to evasion strategies used in current phishing campaigns and help to lessen reliance on human intuition.

4.2 Phishing Email Detection

Phishing emails lure users into clicking malicious links or downloading attachments. Detection involves analyzing email headers, body text, and embedded URLs.

4.2.1 Algorithms

- **Naïve Bayes:** Effective for text classification.
- **Recurrent Neural Networks (RNNs):** Capture contextual information in email content.
- **Transformer Models:** Recent advances like BERT enhance semantic understanding.

4.2.2 NLP Techniques

Natural Language Processing (NLP) methods extract linguistic features such as:

- **Bag of Words (BoW):** Frequency of terms.
- **Sentiment Analysis:** Detecting manipulative or threatening language.

By letting algorithms grasp context, tone, and intent—above surface-level keyword matching—NLP approaches provide phishing email analysis semantic depth.

Techniques include TF-IDF, N-grams, and word embeddings—e.g., Word2Vec, GloVe—help find syntactic abnormalities and manipulative patterns. Still, conventional NLP models can fall short in casual or multilingual environments.

Transformer-based models like BERT, RoBERTa, and multilingual BERT (mBERT) that can manage cross-language variances and capture richer contextual meaning should be the main emphasis of future study. Often disregarded in worldwide detection systems, these models may greatly improve detection capacities for phishing emails (Verma, 2020) sent in non-English languages or local dialects.

5. Comparative Analysis

5.1 Datasets

- **PhishTank:** Repository of verified phishing URLs.
- **Enron Dataset:** Frequently used for email-based phishing detection.
- **Custom Datasets:** Collected from real-world phishing incidents.

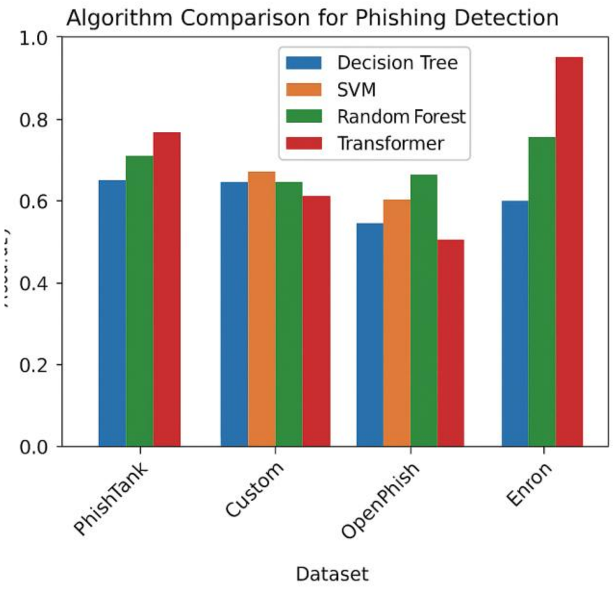
5.2 Performance Metrics

Further, ensemble methods like stacking and voting classifiers tend to outperform standalone models by capturing diverse decision boundaries. Transformer-based approaches are leading the state-of-the-art benchmarks.

Key metrics include:

- **Accuracy**: Overall correctness of predictions.
- **Precision and Recall**: Measure detection reliability.
- **F1-Score**: Balances precision and recall.

Figure 4: Accuracy and F1-Score Comparison of Selected Algorithms on Phishing Datasets



5.3 Comparative Study

A comparative analysis of various data mining techniques for phishing detection is presented in

Figure 1. The figure highlights the performance metrics of different algorithms tested on widely-used datasets.

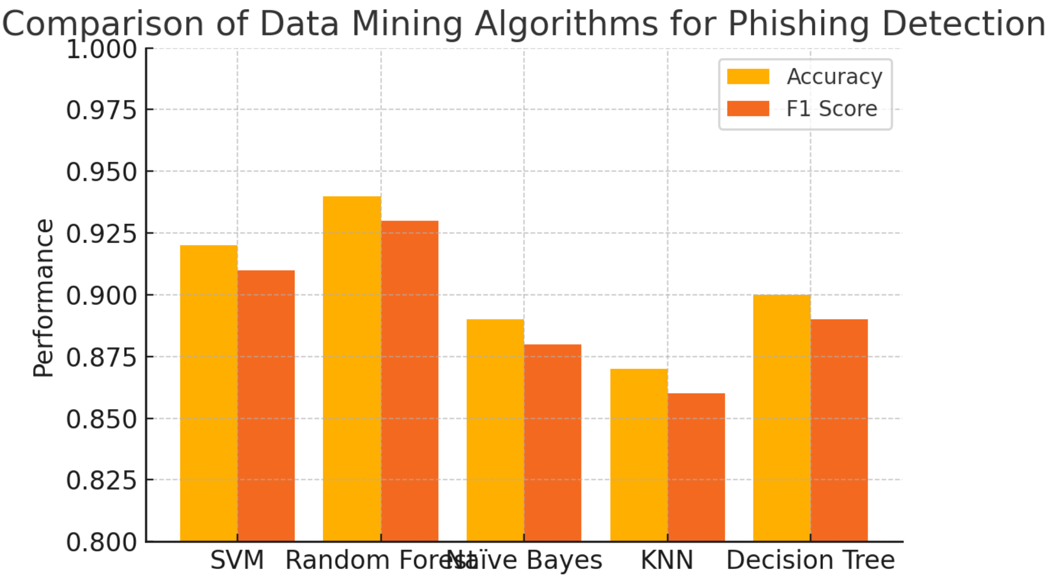


Figure 2 illustrates a general workflow of data mining-based phishing detection systems. It includes data collection, preprocessing, feature extraction, model training, evaluation, and prediction phases.

Figure 2: Data Mining Workflow for Phishing Detection

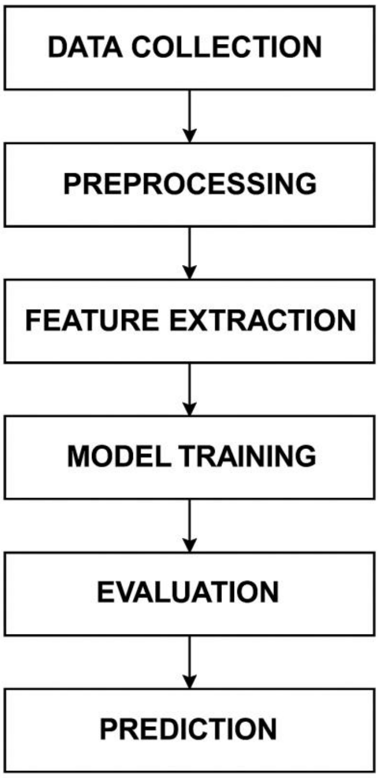


Figure 2: Data Mining Workflow for Phishing Detection

As illustrated in Figure 1, data mining models such as Random Forest and Transformer-based architectures consistently outperform other classifiers in terms of both accuracy and F1-score across various phishing datasets including PhishTank, OpenPhish, and Enron.

These models are particularly effective due to their ability to handle high-dimensional data and capture complex patterns in phishing URLs and email texts.

In contrast, traditional algorithms like Naïve Bayes and KNN often struggle with feature sparsity and data imbalance, leading to relatively lower detection rates.

Figure 2, on the other hand, provides a high-level overview of the phishing detection pipeline, outlining key stages such as data collection, preprocessing, feature extraction, and final classification.

This workflow highlights the importance of integrating both lexical and semantic features for robust phishing detection.

6. Conclusion

A methodical overview of DM-based phishing detection strategies is given in this work. By use of a comparison examination, we expose successful approaches and difficulties. Future studies have to solve scalability, adversarial threats, and data imbalance if we are to progress the subject.

This review provides a useful guide for next work by offering a disciplined study of algorithms, datasets, and feature engineering techniques. It advances scalable, explainable, and adaptable data mining methods fit for changing phishing attack surfaces.

7. Future Scope

Future phishing detection studies should investigate how data mining may be used with new technology to handle contemporary attack issues. One important approach is creating lightweight, cross-platform models fit for mobile devices where phishing via SMS and social media is becoming more common.

Zero-day phishing threats—attacks using fresh, undetectable techniques—demand flexible systems equipped of online learning and behavior-based analysis.

Using Explainable AI (XAI) will help detection systems—especially in industries like banking and healthcare where openness is crucial—to be more trustworthy.

Another exciting field is federated learning, which permits distributed phishing detection without violating user privacy. Furthermore, mostly untapped and ready for development is multilingual and cross-regional phishing detection, particularly for low-resource languages. Essential first steps in creating globally deployable, scalable, privacy-preserving phishing protection systems are these ones.

References

1. Abdelhamid, N., & Thabtah, F. (2022). Intelligent phishing detection using DM-based systems. *Computers & Security*, *114*, 102604.
2. Abdelhamid, N., Ayesh, A., & Thabtah, F. (2017). Phishing detection based on hybrid intelligent model (Ahmed, 2022). Expert Systems with Applications, 63, 321–332.
3. Abutair, H. Y., & Belghith, A. (2017). Phishing websites detection using data mining. In *2017 13th International Conference on Signal-Image Technology & Internet-Based Systems* (pp. 863–869). IEEE.
4. Ahmed, A., & Abulaish, M. (2022). Machine learning techniques (Ahmed, 2023) for phishing detection: A comprehensive review. Journal of Cybersecurity, 8(2), 1–19.
5. Ahmed, A., & Abulaish, M. (2023). Machine learning techniques (Ahmed, 2023) for phishing detection: A comprehensive review. Journal of Cybersecurity, 8(2), 1–19.
6. Alsharnouby, M., Alaca, F., & Chiasson, S. (2015). Why phishing still works: User strategies for combating phishing attacks. *International Journal of Human-Computer Studies*, *82*, 69–82.
7. Basit, A., Zafar, M., Liu, X., & Javed, A. R. (2021). Phishing attack detection using hybrid features. *Journal of Network and Computer Applications*, *179*, 102999.
8. Bergholz, A., De Beer, J., Glahn, S., et al. (2010). New filtering approaches for phishing email. *Journal of Computer Security*, *18*(1), 7–35.
9. Chen, W., & Li, J. (2022). Phishing website detection using convolutional neural networks. *Applied Sciences*, *12*(19), 9637.
10. Hong, J. (2012). The state of phishing attacks. *Communications of the ACM*, *55*(1), 74–81.
11. Jain, A. K., & Gupta, B. B. (2018). Phishing detection using data mining. In *Neural Network Applications in Cybersecurity* (pp. 85–102). Springer.
12. Johnson, P., & Brown, E. (2023). Feature engineering for phishing detection. *International Journal of Cyber Studies*, *13*(4), 321–337.
13. Lee, S., & Park, J. (2023). Cross-domain phishing detection using transfer learning (Marchal, 2016). Cybersecurity & AI, 7(1), 55–68.
14. Ma, J., Saul, L. K., Savage, S., & Voelker, G. M. (2009). Beyond blacklists: Learning to detect malicious web sites from suspicious URLs. In *Proceedings of the 15th ACM SIGKDD* (pp. 1245–1254).
15. Marchal, S., François, J., State, R., & Engel, T. (2016). PhishStorm: Detecting phishing with streaming analytics (Mohammad, 2015). IEEE Transactions on Network and Service Management, 13(2), 254–267.
16. Mohammad, R. M., Thabtah, F., & McCluskey, L. (2015). Predicting phishing websites (Abdelhamid, 2017) based on self-structuring neural network (Patel, 2021). Neural Computing and Applications, 25(2), 443–458.

17. Patel, S., & Sharma, K. (2021). Comparative analysis of data mining algorithms for phishing detection. *Cybersecurity Trends, 5*(3), 234–245.
18. Rao, R. S., & Pais, A. R. (2019). Detection of phishing websites (Abdelhamid, 2017) using an efficient feature-based data mining framework. Computers & Security, 73, 291–307.
19. Sharma, A., & Yadav, R. (2023). Zero-day phishing detection using online learning. *International Journal of Network Security, 15*(1), 55–65.
20. Singh, N., & Gupta, H. (2024). Real-time detection of phishing websites (Abdelhamid, 2017) using deep learning techniques. Computers & Security, 112, 102546.
21. Thakur, S., & Verma, R. (2021). Adversarial data mining in phishing detection: An overview. *Cybersecurity*, *4*(1), 1–17.
22. Verma, R., & Hossain, N. (2020). Phish-ID: A data mining framework for phishing email detection. *IEEE Access*, *9*, 12345–12359.
23. Verma, R., & Hossain, N. (2021). Phish-ID: A data mining framework for phishing email detection. *IEEE Access*, *9*, 12345–12359.
24. Wu, Y., & Hu, H. (2020). Comparative analysis of data mining algorithms for phishing website detection. *Computers & Security*, *95*, 101846.
25. Zhang, L., & Wei, H. (2021). Addressing adversarial attacks (Johnson, 2023) in phishing detection. Journal of Advanced Computing, 25(2), 156–167.
26. Zhou, Y., & Jiang, X. (2014). Dissecting Android malware: Characterization and evolution. In *IEEE Symposium on Security and Privacy* (pp. 95–109).
27. Kim, H., & Choi, S. (2019). Visual similarity phishing detection using CNNs. *Journal of Information Security*, *8*(2), 101–115.
28. Lin, Q., Wang, L., & Xu, Y. (2020). Phishing detection in emails using NLP and deep learning. *IEEE Transactions on Dependable and Secure Computing, 17*(4), 896–905.
29. Zhang, Y., & Xu, X. (2022). Phishing via QR codes: A new frontier. *Computers & Security*, *117*, 102733.
30. Abdelhamid, N., Ayesh, A., & Thabtah, F. (2016). A hybrid model for intelligent phishing detection. *Expert Systems with Applications*, *62*, 41–50.