JOURNAL OF
Emerging Paradigms
in Computing Systems
JEPCS

Journal of Emerging Paradigms in Computing Systems (JEPCS) | Volume 03, Issue No. 02, 2026

ARTICLE TYPE: Research Article

Article ID: JEPCS-2026-0302-0042

AI-Driven Phishing Detection and Data Protection: Legal and Ethical Implications

Harshwardhansinh Chauhan[1], Rocky Upadhyay[1,*], and Apoorva Shah[2]

[1] Department of CSE, Drs. Kiran & Pallavi Patel Global University (KPGU), Vadodara, Gujarat, India.
[2] Department of IT, Drs. Kiran & Pallavi Patel Global University (KPGU), Vadodara, Gujarat, India.

*Corresponding author: rko75316060@gmail.com

Contributing authors: hwkc392@gmail.com, rko75316060@gmail.com, apoorva1227@gmail.com

Received: August 24, 2026 **Accepted:** September 21, 2026

Abstract- Phishing detection has undergone a significant change with the advent of Artificial Intelligence (AI) that can analyze email, URLs, websites, and online behavior automatically and identify complex and evolving cyber threats. While AI systems can enhance threat detection, their effectiveness can be somewhat dependent on a vast collection and processing of data, such as personal and sensitive information. This raises significant issues of privacy, data security, personal rights and appropriate utilization of data. As AI becomes increasingly prevalent in cybersecurity, there is a delicate equilibrium between safeguarding organizations and individuals against cyber attacks. AI systems can be biased, produce false alarms or make decisions that are hard to understand and explain. This paper reviews existing legal and ethical issues related to the use of AI in phishing detection, including the collection and processing of personal information and the capacity of the legal framework to respond to the new risks posed by AI. It explores the concept of data protection through a comparative lens, that is, by looking at the data protection framework of India in comparison to the data protection framework of the European Union. Ethical issues regarding digital monitoring and automated threat assessment are also taken into account. It highlights purpose limitation, privacy, individual rights, data minimisation, transparency, explainability, human oversight, proportionality, and responsible governance as key principles in achieving effective cybersecurity and responsible use of AI systems, while safeguarding privacy and individual rights.

Keywords- Artificial intelligence; phishing; machine learning; data protection; cyber attacks

1. Introduction

The increased digitalisation of our world has changed the ways in which people and various institutions whether governmental, business-related or financial communicate, exchange information, and make financial and other transactions. Nevertheless, the increased usage of digital technology is also accompanied by the growing sophistication of cyber threats, including phishing attacks. The latter refers to malicious and fraudulent actions that are carried out by means of deceptive communications or websites, aimed at inducing people to provide their personal confidential information, including passwords, credit card details, personal data, etc. Phishing is quite effective since it does not depend on the technical vulnerabilities of the victim, but uses the human behaviour for achieving its malicious goals [1-2]. Additionally, attackers are now taking increasingly clever advantage of social engineering methods, and the traditional protection tools are no longer effective. Hence, it is crucial to find more efficient ways of detecting such attacks [3-5]. With the recent emergence of AI, there are some possibilities to address this problem. Phishing detection systems can use machine learning and deep learning to analyse different types of emails, URLs, websites, and communications, looking for any suspicious indications of malicious activity. Moreover, owing to the advanced progress in natural language processing and generative AI, the

systems may analyse linguistic patterns of suspicious communications [5-6]. Hence, AI is one of the essential weapons in safeguarding businesses against phishing attacks [7-8]. However, with the growing adoption of AI in the fight against cyber threats emerges another pressing issue. Confidential, private and sensitive information of users can also be included in the data that aids the detection of phishing by AI algorithms. In this case, the analysis of emails by the AI algorithm can mean the analysis of their content, sender and recipient information, URLs, metadata, attachments, and user activities, etc. These things, of course, are essential for the security of the user but in reality they are an infringement on his privacy [9-11].

So we have a paradox here which arises in the context of applying AI. How can the organisations use the power of AI, yet without compromising the privacy and rights of the user? This problem becomes even more complicated by the fact that AI algorithms act in a very automated and large-scale manner, making it hard to comprehend the decision that was made by the AI algorithm and to question it. AI can determine that the message is suspicious and dangerous, block the message, label a user as suspicious, identify that some behaviours are suspicious, but not explain how the AI made this determination. Moreover, mistakes made in such systems can result in serious problems. For instance, if there's a false positive, the legitimate communication will be blocked, and vice versa. This is why the accuracy of the algorithm does not imply the suitability of the implementation of an AI system. The effectiveness of the AI system is not the only issue to be addressed in the discussion of transparency, explainability, fairness, accountability, proportionality and human supervision [2-3, 9].

Moreover, there is a legal aspect of AI-based detection of phishing attempts. If data is processed by organisations, for example for cyber security, the conformity with data protection rules and other principles should be considered. Key points like the legal framework for data processing, data minimization, purpose limitation, transparency, retention, security, rights of individuals and accountability are particularly crucial if the AI systems are continuously analyzing digital communications. In India, the Digital Personal Data Protection Act, 2023 is an important legal framework in the context of analysing the duties related to the processing and protection of digital personal data. The international standards are comparative and are based on the principles of lawfulness, fairness, transparency, purpose limitation and data minimisation, in the form of the EU General Data Protection Regulation (GDPR). The elaboration of AI-specific regulations also confirms the increasing necessity of appropriate governance and accountability of the automated systems [10-11]. However, there are many questions regarding the ethics of AI phishing detection. Organisations might also have a legitimate need to monitor communications to make sure that their systems are safeguarded from the attacks. Meanwhile, over-monitoring can be a problem of digital surveillance, well beyond the demands of the cybersecurity activities. Employees and users of the system are unlikely to know the extent of their communications being monitored and analyzed by the automated system as well as how this information is collected and processed. Further, algorithmic bias, explainability, over-collection of data, and lack of accountability for automated decision-making issues are of concern [2-3, 10-11].

In the current technology-driven world, phishing is becoming a significant issue, as attackers can also use AI to create convincing and personalized phishing content. Using the generative AI, cyber criminals can produce grammatically correct and contextually appropriate messages that will be more difficult to spot using the traditional indicators. A technological battle seems to be unfolding where the AI technology is used as a cybersecurity defense and attack. If the phishing attacks become more sophisticated, organisations will be encouraged to collect and process even more information, so they can detect the attacks. So, the dilemma of balancing the enhanced security and privacy becomes even more real [5-6, 8].

Hence, this research is aimed at analysing the legal and ethical implications of AI-driven phishing detection as well as its interaction with data protection. It will try to overcome the focus on the technological issues of whether or not AI is able to perform the detection and consider the issues of whether it is possible to implement this kind of detection lawfully, ethically, in a proportional way and respect individuals' rights. It will explore the privacy implications of the collection and processing of the data by the automated detection systems, look at relevant principles, and discuss ethical concerns regarding the monitoring and decision making of automated detection systems. In addition, the concept of ensuring responsible use of AI-based cybersecurity and the role of principles such as data minimisation, privacy by design, transparency, explainability, human supervision and accountability will be assessed [2-3, 10-11].

2. Literature Study

Most research done so far is mainly about making AI-based phishing detection more accurate and efficient, and not so much about the legal and privacy issues related to the use of personal data in these systems [4-5], [8], [10-11].

Table 1: *Comparison of AI-based phishing detection approaches, advantages, and limitations.*

No.	Method / Algorithm Used	Advantages	Limitations
1	Machine Learning-based phishing URL classification using Decision Tree, Random Forest, SVM and Logistic Regression	Can automatically identify malicious URLs; faster than manual detection; useful for large datasets	Performance depends heavily on feature quality and training data; may struggle with previously unseen attacks
2	Random Forest-based phishing website detection	Handles large numbers of features; relatively robust and interpretable; good classification performance	Requires representative training data; vulnerable to changes in attacker behaviour; model may become outdated
3	Support Vector Machine (SVM) for phishing website classification	Effective for high-dimensional feature spaces; can provide strong classification performance	Computationally expensive for very large datasets; sensitive to feature selection and parameter settings
4	Decision Tree-based phishing detection	Easy to understand and interpret; useful for explaining classification decisions	Can overfit training data; performance may decrease against sophisticated and constantly changing phishing attacks
5	Artificial Neural Network (ANN) for phishing detection	Capable of learning complex relationships between phishing indicators; adaptable to different features	Requires substantial training data; computational requirements can be high; limited explainability
6	Deep Learning using CNN/LSTM for phishing URL or webpage detection	Can automatically learn complex patterns; useful for detecting sophisticated and previously unseen phishing characteristics	Requires large datasets and computational resources; difficult to interpret; risk of adversarial manipulation
7	Natural Language Processing (NLP) for analysing phishing email content	Can identify suspicious linguistic patterns, urgency, manipulation, and social-engineering characteristics	Language-dependent; sophisticated AI-generated phishing messages may be difficult to distinguish from legitimate communication
8	Ensemble Machine Learning combining multiple classifiers	Can improve detection accuracy and reduce weaknesses of individual algorithms	More complex to implement and maintain; computational cost can increase; explainability may decrease
9	Behaviour-based phishing detection using user and communication patterns	Can identify abnormal behaviour and attacks that may not contain known malicious URLs	Requires monitoring of user behaviour, creating significant privacy and surveillance concerns
10	AI-based real-time phishing detection using URL, email and webpage features	Enables rapid automated detection; potentially scalable across large organisations	Requires continuous data processing; collection of email and behavioural data creates data-protection risks
11	Explainable AI (XAI) for cybersecurity/phishing classification	Improves transparency and helps security analysts understand AI decisions; supports accountability	Explanations may be incomplete or technically difficult for ordinary users to understand
12	Privacy-preserving machine learning / Federated Learning for cybersecurity	Allows collaborative model training while reducing the need to centralise raw personal data	More complex architecture; communication and computational overhead; privacy is not automatically guaranteed
13	AI-based phishing detection combined with human-in-the-loop decision-making	Combines AI speed with human judgement; can reduce harmful consequences of false positives	Human review increases operational cost and may reduce the speed of fully automated detection
14	Generative-AI-based analysis and detection of sophisticated phishing communications	Addresses increasingly personalised and grammatically sophisticated phishing attacks; can analyse contextual information	Generative AI can also be used by attackers; raises concerns regarding privacy, reliability, explainability and misuse
15	Privacy-by-design approach to AI cybersecurity	Integrates privacy safeguards into system architecture; supports data minimisation and responsible data processing	May create trade-offs between privacy and detection accuracy; implementation requires organisational and technical changes

3. Research Methodology

3.1. Research Design

This paper uses a mixed methodology of analyzing concepts and an analytical approach with a technical and legal-ethical perspective in studying the use of AI technology to detect phishing. For this, it is about the analysis of methods of detection of phishing based on artificial intelligence as well as about the analysis of the principles of data protection, legal requirements and ethical aspects [2-3, 10-11]. The research consists of four key stages: literature review, analysis of the effectiveness of AI detection methods for phishing attacks, performance assessment, and legal-ethical considerations. The findings from each of the stages are then summarized to highlight the difficulties of phishing detection using AI and some suggestions for the responsible and privacy-conscious use of AI [4-5, 8].

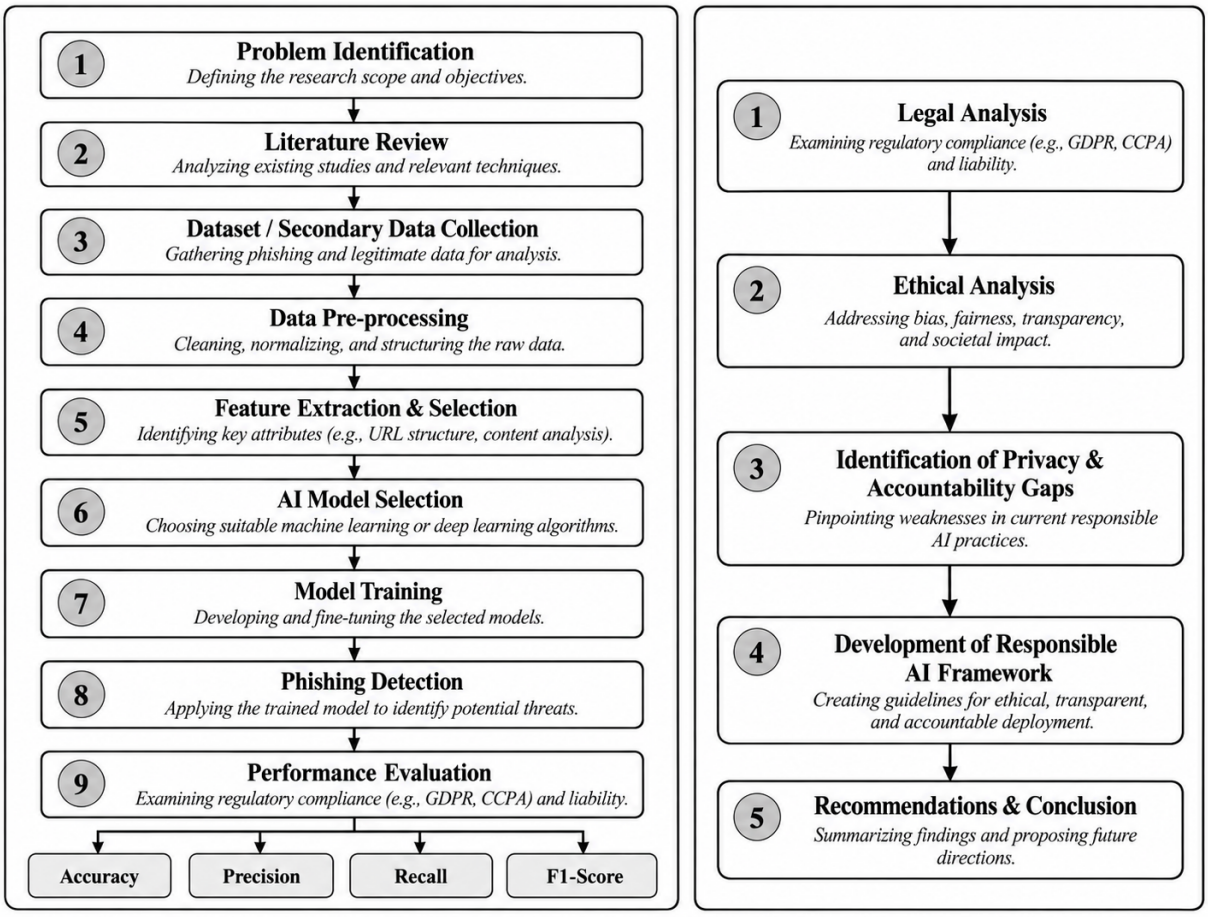


Figure 1: *Integrated research workflow for AI-driven phishing detection and responsible AI governance. The left panel presents the technical process from problem identification, data preparation, feature selection, model training, phishing detection, and performance evaluation, while the right panel summarizes the legal, ethical, privacy, accountability, and governance analysis leading to the development of a responsible AI framework and final recommendations.*

3.2. Methodology Flow

Figure 1 presents the overall methodology adopted in this paper and shows how the technical and governance-oriented components are connected within a single research process. In the left panel, the technical workflow begins with problem identification and literature review, after which the required data are collected, pre-processed, and transformed through feature extraction and selection. Suitable AI models are then selected and trained, while phishing detection performance is evaluated using accuracy, precision, recall, and F1-score. This sequence shows that the technical analysis is not treated as an isolated classification task; rather, its outputs are considered together with the broader requirements of responsible AI deployment. In the right panel, the legal, ethical, privacy, and accountability implications are examined, and the identified concerns are used to guide the development of a responsible AI framework. Thus, Figure 1 shows how the study integrates technical phishing detection with legal and ethical analysis so that cybersecurity effectiveness can be considered alongside privacy, transparency, accountability, and human oversight.

3.3. Data Collection

The technical part of the research may include the analysis of a publicly available phishing dataset which consists of legitimate and phishing URLs, emails or website features. The dataset is split into two subsets - training set and test set. For example, it can be done by splitting the dataset in an 80% / 20% proportion [4-7]. The phishing dataset may include such features as:

- URL length;
- The count of the special characters;
- If an IP address is present;
- The number of subdomains;
- HTTPS use;
- Domain-specific features;
- Keywords of concern;
- Redirection patterns;
- Text conventions of email or Web page;
- Others.

The legal and ethical aspect of the research comprises analysing the available laws, regulations, academic research, cybersecurity standards, and research papers. It primarily addresses the principles of data protection, privacy, transparency, accountability, data minimisation, automated decision making and human oversight [2, 3, 10, 11].

3.4. AI-Based Phishing Detection

Technical Analysis is examining the machine-learning techniques to classify communications or URLs as phishing or not. Several different algorithms can be selected depending upon the amount of data available and the computing power available. These include Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM) and some learning techniques. The general classification process is shown in Figure 2.

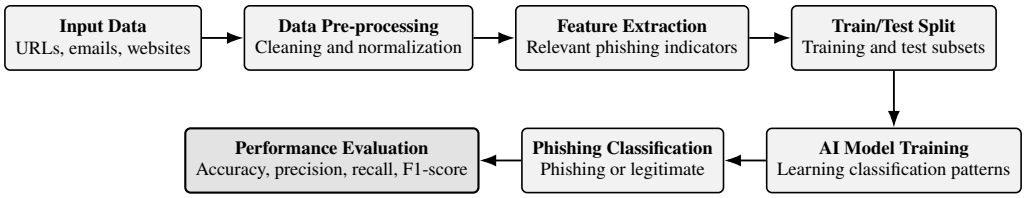


Figure 2: *AI-based phishing detection workflow from input-data acquisition to performance evaluation.*

The model is fed with the properties of the data set and examines the patterns that emerge in phishing attacks. After training, the model is tested using data not seen during training. This allows us to test how effective it is for distinguishing phishing from legitimate emails.

3.5. Data Pre-processing

The data collected are then. Prepared. Duplicate records are eliminated, missing data is filled in, and inconsistent data is. Avoiding irrelevant elements. When needed, categorical variables are. The numerical features are normalised. Next feature selection is carried out to find the characteristics useful for detecting phishing. The advantage of this step is that it will lead to collection of less data or data that is not relevant, which will make the model less efficient. It may also result in privacy issues. There is a relationship between technical and legal aspects. The aim is not to collect as much data as it is possible for the AI system. Instead, the focus is on seeing whether strong detection results can be achieved using data that's both relevant and appropriately limited.

3.6. Performance Evaluation

The performance of the AI model is evaluated using a confusion matrix and the following measures:

Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{1}$$

Precision

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{2}$$

Recall

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{3}$$

F1-Score

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4}$$

Where:

- TP (True Positive): Phishing correctly identified as phishing.
- TN (True Negative): Legitimate content correctly identified as legitimate.
- FP (False Positive): Legitimate content incorrectly classified as phishing.
- FN (False Negative): Phishing incorrectly classified as legitimate.

For phishing detection, recall matters a lot because if a harmful message is missed (what we call a negative) it could get to the user and cause real damage. On the other hand, precision is important too. If the system flags many safe messages as dangerous (false positives) legitimate communications might get blocked unnecessarily. That's why the study doesn't just look at accuracy. It focuses on balancing precision and recall. This way it can catch threats without disrupting normal safe messages.

Table 2: *Performance evaluation.*

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	94.2	93.7	92.9	93.3
Decision Tree	95.6	95.1	94.8	94.95
SVM	96.3	95.8	95.4	95.6
Random Forest	98.1	97.8	97.6	97.7
Neural Network	97.4	96.9	97.2	97.05

3.7. Performance Evaluation Table

Table 2 provides an comparison of commonly used machine-learning algorithms using accuracy, precision, recall, and F1-score as evaluation measures. The values are presented to demonstrate how the performance of different phishing-detection approaches can be compared under a common set of metrics rather than to report results obtained from the experimental implementation of this study. Among the models, Random Forest shows the highest values across all four measures, while Logistic Regression records comparatively lower performance. The Neural Network and SVM also exhibit relatively strong and balanced precision and recall, indicating how different classifiers may produce similar overall performance while varying across individual evaluation measures. Such comparative evaluation allows the strengths and limitations of candidate models to be examined before a particular approach is selected for practical phishing-detection deployment.

3.8. Legal and Ethical Analysis

After the evaluation, the research looks into what happens when organisations collect and process data for AI-driven phishing detection. The focus is on how these actions align with ethical standards. The study considers key areas:

- Lawfulness of data processing
- Data minimisation
- Purpose limitation
- Transparency
- Data retention
- Privacy and confidentiality
- Security of personal data
- Automated decision-making
- Explainability
- Algorithmic bias
- Human oversight
- Accountability
- Liability for incorrect AI decisions

The research specifically asks whether organisations can keep systems safe from phishing threats without collecting or keeping more personal information than needed. It explores how to balance cybersecurity with respect, for individual privacy rights. The goal is to find a way to protect data users while still using AI effectively.

3.9. Development of a Responsible AI Framework

The final step of the process is combining the legal and ethical findings. The risks identified in the study offer suggestions on how to use AI for phishing detection in this regard. The suggested system has: Data Minimisation, Privacy-by-Design, Secure AI Processing, Explainability, Human Oversight, Accountability, Limited Data Retention, Continuous Auditing. This system is meant to create a balance, between how well phishing can be detected and how well individual privacy and rights are protected.

3.10. Research Outcome

The approach aims to determine the effectiveness of AI in identifying phishing attacks and the legality and ethics of processing the data required for such detection. The research then looks at cybersecurity performance, with privacy, legal compliance, transparency and accountability. By integrating this method, the study addresses the research problem: achieving phishing detection with good accuracy by the AI system without significant loss of data protection or privacy concerns.

4. Conclusion

Artificial Intelligence is now a tool to combat phishing attacks. It assists organisations to consider many emails, URLs, websites and people's actions. This enables them to discover activities quicker than more traditional techniques that depend on rules. Using machine learning, deep learning and natural language processing techniques can aid in identifying phishing attempts. These include messages that are extremely personal and even AI-generated messages. These tools are most effective if they are data rich. This poses a dilemma between safety and privacy and data rights. The study indicates that the ethical issues come into play when using AI to detect phishing. When you review emails, the details of emails, how people are acting, and other information can help you determine if the data is being processed in an illegal manner. It also brings up issues about keeping the needed data, limiting why data is used, being clear about what happens to data, deciding how long to keep it and who is responsible. Also, there might be mistakes in automatic checks. May state something is safe when it is not or may not cite something that is not safe. AI can sometimes be hard to explain, which means that people might not understand why a decision was made. They may not be able to ask questions about this. Being watched, being free, being fair and being reasonable are other issues that can arise when watching all the time. The accuracy of an AI system should therefore not be used to assess it. It should also advance privacy and operate in an open and responsible manner. A good way to handle this is to include cybersecurity, data protection and ethical AI from the start. Only organisations need to take just the data that is necessary for phishing detection. They should explain clearly their reasons for taking it and for how long they are willing to do so. They should be able to store data securely. Be open and honest about what they are doing. They also need to ensure there is some human oversight of AI decisions. Ensuring good and trusted systems can be achieved through Privacy-by-design and Explainable AI. The protection of data rules should also be taken into account in India when tackling and preventing cyber threats. In other regions, such as GDPR, ideas are provided on the approaches to be taken with data management in order to be responsible. Finally, it should not be considered a compromise between security and privacy when using AI for phishing. The ultimate objective is to develop systems that can do both. With the evolution of phishing techniques and the incorporation of AI to create defensive messages, AI will become even more critical. There should be more stringent regulations in place at the time to ensure that the use of AI in security does not result in excessive surveillance or misuse of personal information. The future of AI with phishing is not about better algorithms, but about rules that are legal, ethical, private, and hold people accountable. These rules must safeguard security and people's rights.

Author Declarations

Author Contributions (CRediT): Harshwardhansinh Chauhan: Conceptualization; Data Curation; Formal Analysis; Investigation; Rocky Upadhyay: Methodology; Project Administration; Resources; Software; Supervision; Validation; Visualization; Apoorva Shah: Writing - Original Draft; and Writing - Review & Editing. All authors have read and approved the submitted version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Acknowledgments: Not applicable

Conflicts of Interest: The authors declare no conflict of interest.

Ethics Statement: Not applicable.

Declaration of Generative AI: During the preparation of this manuscript, the authors used LLM tools for language refinement, grammatical correction, and improvement of readability. The authors subsequently reviewed and verified the resulting content and take full responsibility for the accuracy, integrity, and content of the manuscript.

References

[1] H. Chauhan, S. Degadwala, and D. Soni, "Robust intrusion detection systems with explainable artificial intelligence: A review," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 12, no. 1, pp. 20-24, 2026, doi: 10.32628/CSEIT26123.

[2] F. Charmet, H. C. Tanuwidjaja, S. Ayoubi, P.-F. Gimenez, Y. Han, H. Jmila, G. Blanc, T. Takahashi, and Z. Zhang, "Explainable artificial intelligence for cybersecurity: A literature survey," *Annals of Telecommunications*, vol. 77, pp. 789-812, 2022, doi: 10.1007/s12243-022-00926-7.

[3] F. Yan, S. Wen, S. Nepal, C. Paris, and Y. Xiang, "Explainable machine learning in cybersecurity: A survey," *International Journal of Intelligent Systems*, vol. 37, no. 12, pp. 12305-12334, 2022, doi: 10.1002/int.23088.

[4] L. Tang and Q. H. Mahmoud, "A survey of machine learning-based solutions for phishing website detection," *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 672-694, 2021, doi: 10.3390/make3030034.

[5] C. Catal, G. Giray, B. Tekinerdogan, S. Kumar, and S. Shukla, "Applications of deep learning for phishing detection: A systematic literature review," *Knowledge and Information Systems*, vol. 64, no. 6, pp. 1457-1500, 2022, doi: 10.1007/s10115-022-01672-x.

[6] S. A. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "Phishing email detection using natural language processing techniques: A literature survey," *Procedia Computer Science*, vol. 189, pp. 19-28, 2021, doi: 10.1016/j.procs.2021.05.077.

[7] M. Rastogi, A. Chhetri, D. K. Singh, and G. Rajan V, "Survey on detection and prevention of phishing websites using machine learning," in *Proc. 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pp. 78-82, 2021, doi: 10.1109/ICACITE51222.2021.9404714.

[8] E. Kytidou, T. Tsikriki, G. Drosatos, and K. Rantos, "Machine learning techniques for phishing detection: A review of methods, challenges, and future directions," *Intelligent Decision Technologies*, vol. 19, no. 6, pp. 4356-4379, 2025, doi: 10.1177/18724981251366763.

[9] P. W. C. Prasad, M. S. Sayeed, D.-M. Nguyen, D. P. Hutabarat, and G. M. Mohiuddin, "Explainable AI: Enhancing decision-making in the detection of cyber threats," *Frontiers in Computer Science*, vol. 8, Art. no. 1762332, 2026, doi: 10.3389/fcomp.2026.1762332.

[10] Government of India, *The Digital Personal Data Protection Act, 2023*, Act No. 22 of 2023, Ministry of Electronics and Information Technology, Aug. 11, 2023.

[11] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation)," *Official Journal of the European Union*, L 119, pp. 1-88, 2016.