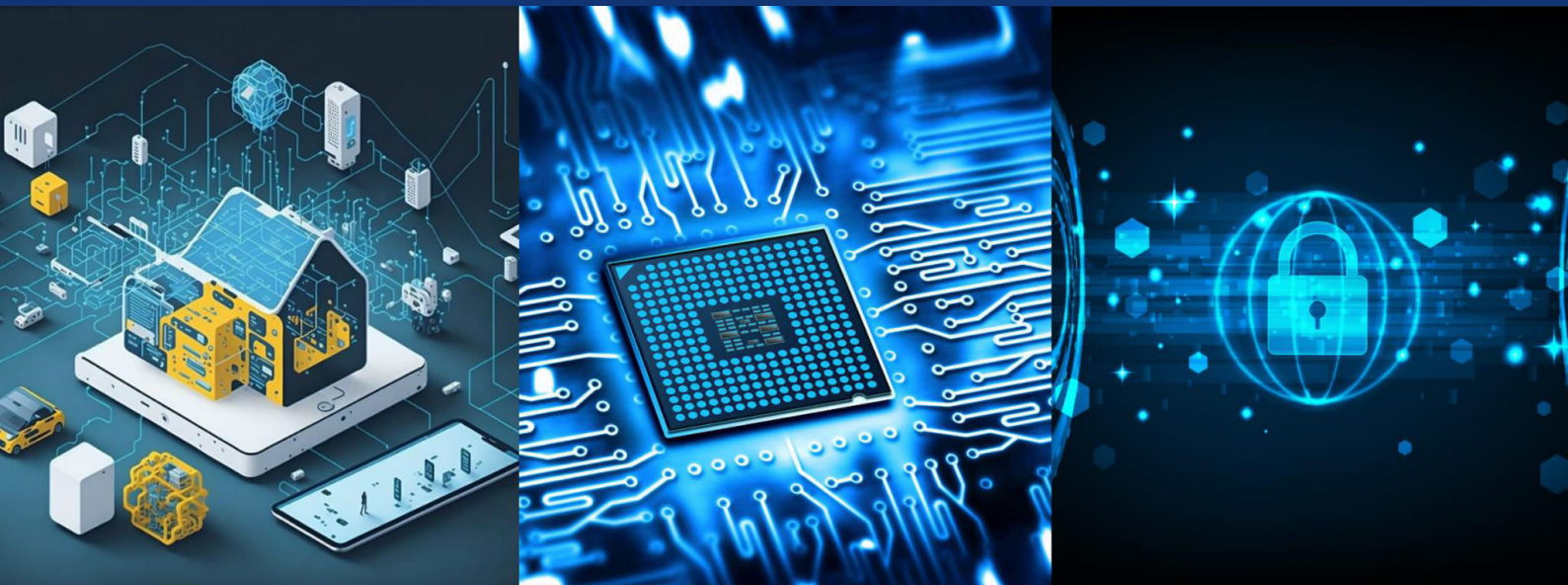
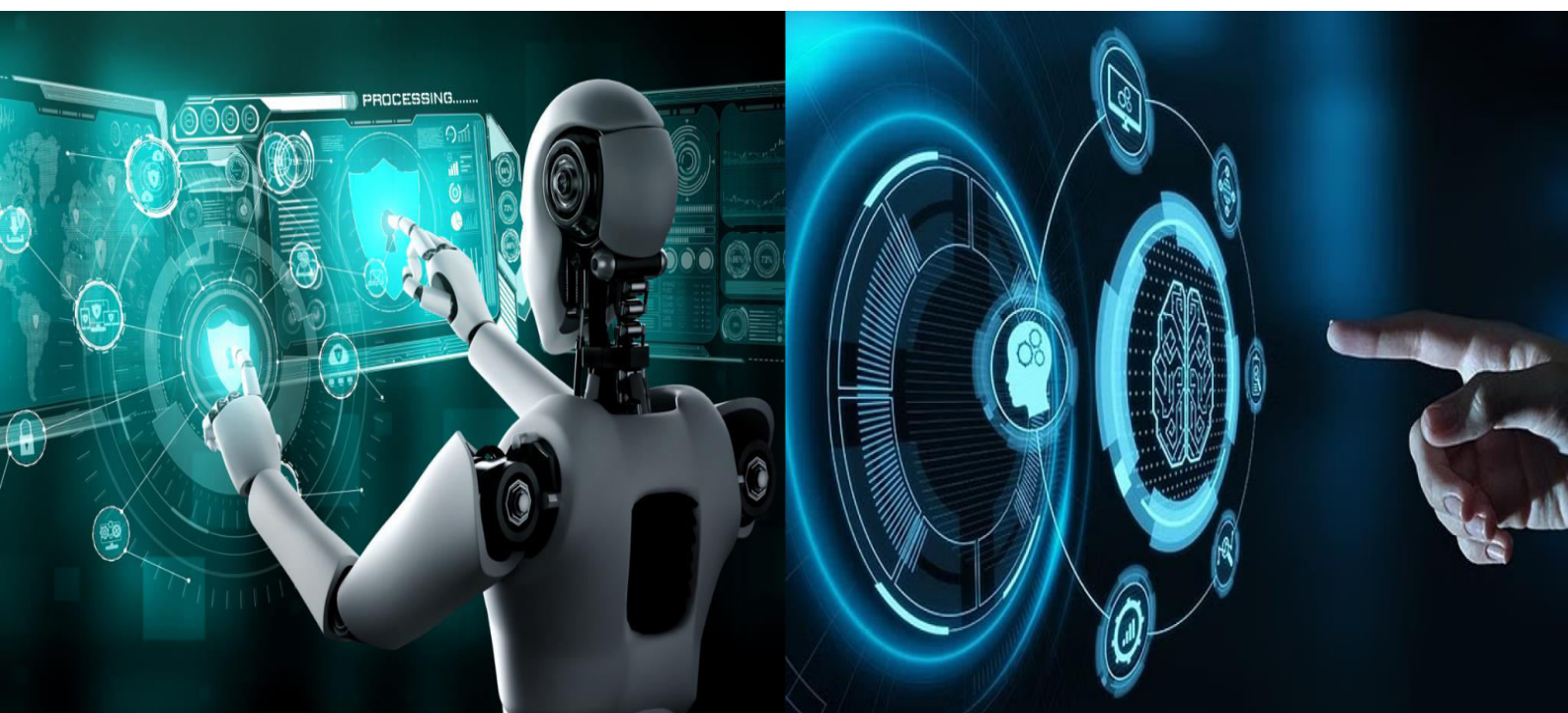




International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)





International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Role of AI in Detecting Online Scams and Fake Messages

Ms. Pranita Dipak Daund, Mr. Swaraj Malve

Department of Master of Computer Applications, JCEI's Jaihind Institute of Management and Research, Vadgaon Sahani, Maharashtra, India

Assistant Professor, Department of Master of Computer Application, JCEI's Jaihind Institute of Management and Research, Vadgaon Sahani, Maharashtra, India

ABSTRACT: This paper examines how artificial intelligence could help people recognise scam messages before they respond to them. Fraudulent messages may arrive by email, SMS, social media or chat apps. They often pretend to come from a bank, delivery service, employer or someone the recipient knows, then ask for money, account details, passwords or a click on a link. AI tools can review several clues together, including the words used in a message, the web address it contains and available sender information. Natural language processing can help analyse the text, while machine-learning models can learn patterns from examples labelled as legitimate or suspicious. The paper reviews published work on these approaches and outlines a possible detection framework with risk levels and plain-language explanations. It is a literature-based proposal, not an experiment: no new model was trained, and no detection-accuracy result is claimed. AI may help flag suspicious messages, but it cannot guarantee that every scam will be found or that every unflagged message is safe. Users still need to verify important requests through trusted channels.

KEYWORDS: Artificial intelligence, online scams, fake messages, phishing detection, machine learning, natural language processing, cybersecurity.

I. INTRODUCTION

People now use digital messages to manage banking, shopping, work and everyday communication. Scammers take advantage of these habits by sending messages about account problems, parcel deliveries, job offers or unexpected prizes. The message may insist that the recipient act immediately, open a link, share login details or transfer money. The pressure to respond quickly can make it harder to check whether the sender is genuine.

Not every scam contains spelling mistakes or strange formatting. Some messages copy the style of a genuine organisation and use a convincing explanation to encourage the recipient to click a link or open an attachment. For this reason, checking grammar alone is not a reliable way to judge a message. The US Federal Trade Commission warns that phishing messages often use a story or urgent request to persuade people to click or open something that may put their information or devices at risk [1].

AI provides another layer of screening. A model can learn from labelled examples of genuine and fraudulent messages, while NLP methods examine the words and context used in a message. A separate URL check can look for features such as an unexpected domain or a link that does not match the organisation named in the text. Combining several clues may be more useful than relying on a single suspicious word. Reviews of phishing research describe a range of approaches, from traditional filters and manually selected features to machine learning and deep learning [2–4].

Detection systems also make mistakes. Scammers can change their wording, use new domains or adapt their methods, while genuine messages may contain words that appear suspicious. A system can therefore miss a scam or incorrectly flag a legitimate message. Any warning should explain the main reasons for concern and avoid promising certainty. The following sections review the main approaches and set out a framework that could be implemented and tested in future work.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

II. PROBLEM STATEMENT

Users may receive scam attempts mixed in with genuine messages, and the difference is not always obvious. Blocklists and fixed rules can catch known threats, but they may not recognise a new website or a message rewritten to look different. If a filter is too strict, it may also mark a real message as fraudulent. The central challenge is to identify likely scams without creating too many false alarms.

This paper considers how AI can combine message text, link features and available sender information to flag suspicious messages. It also considers two practical requirements: limiting incorrect warnings and protecting private information. A useful system should explain why it raised an alert so that users can decide what to check next.

III. OBJECTIVES

1. To explain the role of AI and machine learning in detecting online scams and fake messages.
2. To review how NLP, URL analysis and classification models can be used for suspicious-message detection.
3. To identify key challenges, including changing scam patterns, false positives, privacy and language variation.
4. To propose a conceptual AI-assisted framework that gives users a risk level and understandable reasons for a warning.

IV. LITERATURE REVIEW

4.1 NLP and machine learning for phishing messages

A literature survey by [2] examined the use of NLP and machine-learning methods for phishing email detection. It describes how message text can be transformed into features that a model uses to distinguish phishing from legitimate communication. Common steps include cleaning the text, tokenisation, feature extraction and classification

4.2 Recent phishing-detection methods

Kavya and Sumathi reviewed recent phishing-detection methods, including supervised machine learning, deep learning and hybrid approaches [3]. Their review discusses the value of combining information such as URL structure, webpage content and domain characteristics. It also identifies recurring problems such as dataset bias, changes in attacker behaviour and difficulties in reproducing reported results. These concerns suggest that a model should be evaluated on data that differs from its training examples, rather than relying only on a single test split.

4.3 Explainable AI in cybersecurity

Yan and colleagues surveyed explainable machine learning in cybersecurity [4]. They emphasise that users and security professionals often need to understand why a model has made a prediction, not simply receive a label. For a fake-message detector, an explanation could identify a suspicious link, an urgent request for credentials or a mismatch between the claimed organisation and the sender domain. Such explanations can help users assess a warning, although the explanation itself must not expose sensitive information or create a false sense of certainty.

4.4 Human behaviour and phishing difficulty

The National Institute of Standards and Technology (NIST) studies phishing in real-world contexts and highlights that detection difficulty depends on both the message and the recipient's situation [5]. This is important because a technically suspicious message may appear convincing when it refers to a current delivery, account problem or work task. AI can assist screening, but user awareness and safe reporting procedures remain necessary.

4.5 Evolving threats and adversarial behaviour

NIST's work on adversarial machine learning explains that AI systems can be targeted or manipulated, and that risks need to be considered throughout the model lifecycle [6]. In scam detection, attackers may change wording, use URL shorteners, rotate domains or craft messages to evade known patterns. Systems therefore require monitoring, updated datasets and a process for reviewing incorrect classifications.

4.6 Research gap

Existing literature covers phishing emails, URL-based detection, machine-learning classifiers and explainability. A practical gap remains in combining these signals into a user-facing workflow that works across common message



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

formats and explains its warning in plain language. Another challenge is that strong results on a benchmark dataset do not automatically establish reliable performance on new scams, different languages or real user messages. This paper addresses that design gap by proposing a conceptual framework that combines multiple signals, gives a reason for the risk label and keeps human verification in the process. The framework is a proposal for future implementation, not a tested system.

V. SYSTEM REQUIREMENTS

5.1 Functional requirements

- Allow a user to paste message text or submit a suspicious URL for analysis.
- Clean and process the text without unnecessarily storing personal details.
- Extract message, sender and URL features where the information is available.
- Classify the input into risk categories such as Low, Medium or High risk.
- Show the main signals that contributed to the warning.
- Allow the user to mark a result as incorrect or report a suspicious message.
- Provide basic safety advice, such as not sharing OTPs or passwords and verifying the sender through an official channel.

5.2 Non-functional requirements

- Privacy: minimise collection and retention of message content and personal identifiers.
- Usability: display clear language, readable warnings and accessible controls.
- Performance: return results quickly enough for practical use.
- Reliability: evaluate false positives and false negatives, not accuracy alone.
- Maintainability: support updates to threat data and periodic model evaluation.
- Security: protect submitted data, logs and model endpoints against unauthorised access.

5.3 Possible technologies

Component	Possible option	Purpose
Interface	HTML, CSS, JavaScript or a simple Python web UI	Accept a message or URL and display the result
Text processing	Python with NLP preprocessing	Prepare text for feature extraction
Baseline classifier	TF-IDF with Logistic Regression or Naive Bayes	Create an interpretable baseline model
URL analysis	URL parsing and domain-feature extraction	Identify suspicious URL characteristics
Evaluation	Precision, recall, F1-score, confusion matrix	Measure model performance and errors
Storage	Minimal, access-controlled database if required	Store only necessary feedback or model metadata

VI. PROPOSED SYSTEM

This paper proposes an AI-assisted online scam and fake-message detection framework. Its purpose is to help users recognise warning signs before they respond to a suspicious message. The system should be treated as a decision-support tool rather than a replacement for official verification.

6.1 Proposed workflow

1. Input: the user pastes a message or a URL into the tool. The interface warns users not to submit passwords, OTPs or other secrets.
2. Preprocessing: the system normalises text, identifies links and extracts relevant features while removing unnecessary personal information where possible.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

3. Feature analysis: NLP features are extracted from the message; URL and sender indicators are assessed when available. Examples include urgent wording, requests for credentials, suspicious domain patterns and mismatches between the sender identity and link destination.
4. Classification: a trained model estimates the likelihood that the message is suspicious. A rule-based safety layer can add warnings for high-risk indicators.
5. Risk explanation: the interface shows a Low, Medium or High risk label and explains the main reasons. It should avoid saying a message is definitely safe.
6. User action: the user is advised to verify the claim using a known official website, app or phone number, avoid clicking unexpected links and report suspected scams.
7. Feedback and review: incorrect results can be reviewed by an authorised maintainer. New data should be validated before being added to training.

6.2 Model development approach

A practical implementation could begin with a supervised baseline model. Labelled examples of legitimate and fraudulent messages would be cleaned and divided into training, validation and test sets. Text can be represented using TF-IDF features, while URL features can be represented separately and combined with text features. Logistic Regression or Naive Bayes would provide a useful baseline because they are relatively simple to train and interpret. More complex models, including transformer-based classifiers, could be compared only after a reliable baseline has been established.

To reduce data leakage, duplicate or near-duplicate messages should not appear across training and test sets. Evaluation should include messages collected from different periods or sources where possible. A model should be checked for performance on different message types and languages before any claim is made about general use. The paper does not report model accuracy because no model was trained or tested for this study.

6.3 Evaluation plan

Measure	Why it matters
Precision	Of messages flagged as suspicious, how many are actually suspicious?
Recall	Of the suspicious messages in the test set, how many does the model identify?
F1-score	Balances precision and recall in one measure.
False-positive rate	Shows how often genuine messages are incorrectly flagged.
Confusion matrix	Shows correct and incorrect predictions for each class.
Performance by group	Checks whether results vary across languages, message types or sources.

For a user-facing security tool, accuracy alone is insufficient. A model that labels almost every message as legitimate may have high accuracy on an imbalanced dataset while missing many scams. Precision and recall should therefore be considered together, and the threshold for a warning should reflect the intended use. High-risk warnings should be tested with users to ensure that the wording is understandable and does not create unnecessary panic.

VII. RESULTS AND SURVEY ANALYSIS

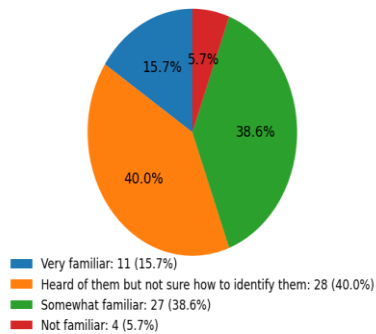
A Google Forms questionnaire received 70 responses. The charts below summarise selected single-choice questions from the response spreadsheet provided for this study. Percentages are calculated from the recorded responses for each question ($n = 70$); they describe this respondent group and should not be treated as representative of all internet users. The questionnaire results are descriptive and do not test the accuracy of an AI detection model.



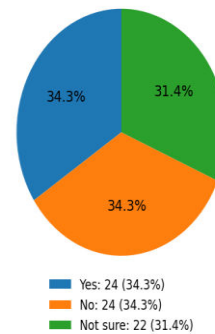
International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

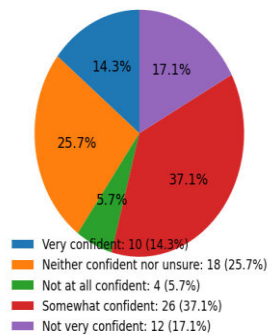
Q2. Familiarity with Online Scams



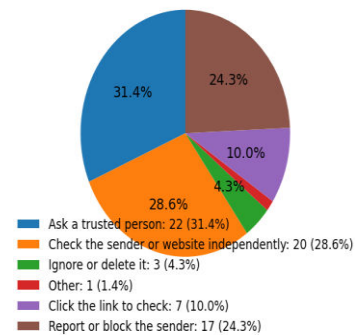
Q3. Experience Receiving Suspected Scam Messages



Q6. Confidence in Identifying Scam Messages



Q7. Usual Response to Suspicious Messages



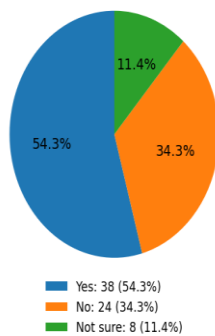


International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

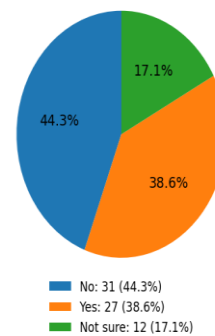
(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VIII. RESULTS AND SURVEY ANALYSIS (CONTINUED)

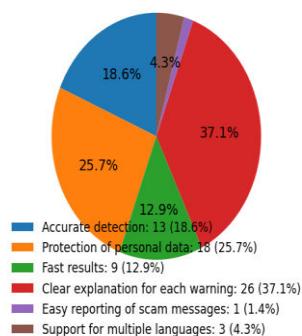
Q8. Awareness of AI-Based Scam Detection



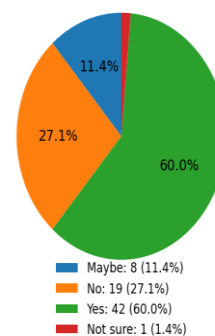
Q9. Use of AI-Based Detection Tools



Q12. Most Important AI Detection Feature



Q13. Perceived Benefit of AI Scam Detection



Brief observation: Responses show varied familiarity with scam messages and differing confidence levels. Respondents also identified privacy, missed new scams, and clear explanations as relevant considerations for AI-assisted detection. These descriptive results support the need for understandable warnings and privacy-conscious design, but they do not establish model performance or causal effects.

IX. DISCUSSION

AI can review large numbers of messages and highlight patterns that deserve a closer look. Text analysis may identify requests for passwords or urgent payments, while URL checks can flag unexpected or misleading web addresses. Looking at several indicators together may provide more useful evidence than relying on a single keyword. Clear explanations can also help users understand the warning rather than simply accepting a label.

At the same time, AI has important limits. A legitimate message may include urgent language, and a fraudulent message may be carefully written without obvious warning words. A URL may look normal but lead to a malicious page, while a newly created scam domain may not yet appear in threat lists. Model performance may also decline as scam techniques change. For these reasons, the system should communicate uncertainty and allow users to verify important claims independently.

Privacy is another concern. Messages can contain names, phone numbers, account details or private conversations. A responsible implementation should process only the information needed for detection, avoid retaining full messages by



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

default, protect any feedback records and explain how data is handled. Human review is particularly important when a result could affect access to a financial account, employment communication or other important service.

The proposed framework also recognises that scammers can use AI to generate more convincing messages. This makes it unsafe to treat good grammar as proof of legitimacy. A stronger approach considers the whole context, including the requested action, destination URL, sender identity and whether the message can be verified through a trusted channel.

X. LIMITATIONS

This paper includes a descriptive analysis of 70 Google Forms responses, but it does not include a newly collected scam-message dataset, a trained detection model or experimental model-performance results. Therefore, it cannot claim a particular detection accuracy or prove that the proposed framework reduces scam losses. The reviewed literature also covers different datasets, attack types and evaluation methods, so their findings should not be treated as directly comparable. Further work should implement the framework, document the data sources and labels, evaluate it on unseen messages, and examine privacy, fairness and usability.

XI. CONCLUSION

AI can contribute to scam detection by checking message language, links and other available indicators quickly. NLP and machine-learning models may help screen messages, and an explanation can show users which features raised concern. These tools still have limits: they may miss new scams, and they may flag genuine messages by mistake.

The proposed framework combines text and URL analysis, a risk classification, plain-language explanations and safe next-step guidance. Future work should build a working prototype and evaluate it using labelled data, precision, recall, F1-score and false-positive rates. Testing across languages and message platforms would help establish whether the system generalises beyond its training data. Until such testing is completed, the framework should be understood as a research proposal rather than a validated product.

REFERENCES

- [1] Federal Trade Commission, "Phishing scams can be hard to spot," Consumer Advice, Dec. 2024. Available: <https://consumer.ftc.gov/consumer-alerts/2024/12/phishing-scams-can-be-hard-spot>
- [2] "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.
- [3] S. Kavya and D. Sumathi, "Staying ahead of phishers: a review of recent advances and emerging methodologies in phishing detection," *Artificial Intelligence Review*, vol. 58, article 50, 2025. doi: 10.1007/s10462-024-11055-z.
- [4] X. Yan et al., "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.
- [5] National Institute of Standards and Technology, "Human-Centered Cybersecurity: Phishing," NIST. Available: <https://csrc.nist.gov/Projects/human-centered-cybersecurity/research-areas/phishing>
- [6] A. Vassilev, A. Oprea, A. Fordyce, H. Andersen, and M. Davies, "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations," NIST AI 100-2, 2024. doi: 10.6028/NIST.AI.100-2.
- [7] E. Kytidou, T. Tsikriki, G. Drosatos, and K. Rantos, "Machine learning techniques for phishing detection: A review of methods, challenges, and future directions," *International Journal of Information Security*, 2025. doi: 10.1177/18724981251366763.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  ijircce@gmail.com



www.ijircce.com

Scan to save the contact details