MSc Data and Artificial Intelligence

Module: Research Methods and Professional Development

Module Code: ACFA7004

Assessment Component 1: Research Proposal

Title : A Study of Phishing Detection Using AI and User-Behavior Signals

Student Name: MD. Mehedi Hasan Mixon

Student ID: 2436061

Table of Contents

1. Introduction 2
1.1 Context and Background 2
1.2 Research Problem 3
1.3 Research Aim and Objectives 3
1.4 Significance of the Research 4
2. Literature Review 4
3. Research Design and Methodology 5
3.1 Research Design 6
3.2 Research Methods and Materials 6
3.3 Research Materials 6
3.4 Ethical Considerations 7
4. Expected Outcome 7
5. Limitations 7
6. Timeline 8
7. References 8
8. Appendices 9

1. Introduction

1.1 Context and Background

One of the most common and destructive types of cyberattacks is still phishing. Because it takes advantage of human judgment, which defenders still find difficult to completely secure, it continues to have an impact on people, companies, and public sector organizations. Phishing relies on more than only technological flaws, in contrast to many other types of assaults. It operates by tricking users into opening infected attachments, exchanging credentials, clicking malicious links, or believing false requests. As a result, the security issue is both behavioral and technical. Phishing activity is still very high, according to recent threat reports; in the fourth quarter of 2024 alone, the Anti-Phishing Working Group reported 989,123 phishing incidents. This demonstrates that phishing is evolving and continuing on a large scale rather than decreasing. At the same time, phishing's technical environment is evolving. Conventional phishing emails were frequently simple to identify due to misspellings, shoddy layout, or dubious-looking links. That is becoming less trustworthy. Attackers may now produce more polished, customized, and convincing phishing content more easily because to the proliferation of generative AI. According to recent study, generative AI makes phishing assaults more convincing, realistic, and large-scale, making it more difficult for consumers to discern them from genuine correspondence.

To respond to this threat, researchers have increasingly turned to artificial intelligence and machine learning for phishing detection. AI-based phishing detection systems can analyse large volumes of URLs, email content, metadata, and webpage features far more quickly than manual inspection or rule-based filters. Recent reviews show that machine learning and deep learning approaches have significantly improved phishing detection performance by learning complex patterns from technical artefacts. However, most of these systems are still designed mainly around the content of the message rather than the behavior of the person receiving it.

This creates an important research opportunity. Phishing does not succeed only because a message looks malicious. It succeeds because a user reacts in a certain way under pressure, trust, distraction, urgency, or habit. If phishing is partly a behavioural problem, then detection may improve when behavioral indicators are considered alongside technical features. This proposal therefore explores whether a phishing detection approach that combines AI with user-behavior signals can provide a more realistic and effective defensive model. This aligns with the brief’s expectation that the proposal should clearly define the research area, current developments, and the problem the study aims to address.

1.2 Research Problem

Although phishing detection has advanced significantly, an important gap remains in the literature. A large body of work focuses on technical detection using machine learning or deep learning models trained on message-based features such as URLs, lexical patterns, email content, and metadata [3], [4]. A separate body of research examines why users still

fall for phishing attacks, highlighting factors such as security awareness, intuition, knowledge, attention, and decision style [5]–[7]. The problem is that these two areas are often studied independently.

As a result, many phishing detection approaches may be incomplete. They can identify suspicious technical artefacts, but they often overlook behavioural signals that influence whether a phishing attempt is likely to succeed. This matters because phishing does not operate in a vacuum. Even a technically sophisticated message may fail if the recipient behaves cautiously, while a moderately convincing message may succeed if the recipient acts quickly or inattentively [6], [7].

The research problem addressed by this proposal is therefore that current AI-based phishing detection approaches rely heavily on technical features and do not adequately integrate user-behaviour signals, which may limit their effectiveness and practical realism.

1.3 Research Aim and Objectives

The purpose of this study is to determine whether phishing detection effectiveness can be enhanced by combining artificial intelligence with user behavior signals in comparison to traditional technical feature-based models. The objectives of the study are:

1. To review recent literature on AI-based phishing detection and user susceptibility to phishing.
2. To identify behavioural indicators that are relevant to phishing risk.
3. To develop a baseline phishing detection model using technical features only.
4. To develop a second model that combines technical features with user-behaviour signals.
5. To compare the performance of both models using appropriate evaluation metrics.
6. To assess the interpretability, practical value, and limitations of the combined approach.

1.4 Significance of the Research

This study is important from an academic and practical standpoint. From an academic standpoint, it fills a glaring vacuum by combining two frequently distinct fields of study: human-factor research on phishing vulnerability and AI-based phishing detection [3], [5]. This could lead to a more comprehensive understanding of phishing as a socio-technical danger. Practically speaking, the results may help create phishing defense systems that are more sophisticated and cognizant of human behavior. Organizations may be able to improve email security, awareness training, and even risk-based alerting by taking into account the user's behavioral context in addition to the message's suspiciousness if behavioral signals prove beneficial. In operational settings, where phishing is still one of the most prevalent attack methods, this could be especially helpful [1].

2. Literature Review

The purpose of this literature review is to examine the existing body of research related to phishing detection, user-behaviour factors, and the role of artificial intelligence in both cybersecurity defence and phishing evolution. The review is organised thematically because that structure best highlights the relationship between technical detection studies and behaviour-focused studies.

One major theme in the literature is the growth of AI-based phishing detection. Kavya and Sumathi provide a recent review of phishing detection methodologies and show that machine learning and deep learning techniques are now central to modern anti-phishing research [3]. Their review explains that AI models can detect phishing by analysing URLs, email content, metadata, and structural features, often outperforming traditional rule-based systems because they adapt better to changing phishing patterns. Similarly, Kyaw et al. offer a systematic review of deep learning techniques for phishing detection and note that deep learning models are especially useful in identifying complex patterns in phishing emails and websites [4]. These studies make it clear that AI has become a major pillar of phishing detection research.

However, this technical progress does not fully solve the human side of the problem. Fan et al. investigated phishing susceptibility using explainable artificial intelligence and found that human-related and demographic factors can provide meaningful insight into why certain users are more vulnerable to phishing [5]. This study is important because it moves beyond simple technical detection and examines how behavioural characteristics contribute to phishing susceptibility. It also demonstrates the value of explainability, showing that useful models should not only produce accurate results but also reveal why those results occur.

Another key theme in the literature is phishing as a behavioural and cognitive problem. Buckley et al. found that employee phishing behaviour is influenced by intuition, elaboration, attention, and email typology [6]. Their work suggests that users often do not rely purely on rational analysis when assessing an email. Instead, decisions may be shaped by limited attention, habitual behaviour, or fast judgement. Sturman et al. similarly showed that phishing detection is associated with security awareness, knowledge, and decision style [7]. Their findings suggest that more reflective decision-making and stronger awareness can reduce vulnerability to phishing. Together, these studies show that phishing success is influenced not only by attacker sophistication but also by how users process information.

A related issue is whether training alone is enough to solve the problem. Marshall et al. conducted a scoping review of email phishing training and found that although training is widely used, the evidence base remains mixed and uneven [8]. This is a valuable insight because it suggests that awareness programmes, while useful, may not be sufficient on their own. Users may still make poor decisions under realistic conditions, especially when messages are persuasive or time pressure is involved. This reinforces the case for developing systems that support users rather than relying entirely on them.

A more recent development is the role of generative AI in strengthening phishing attacks. Jabir et al. argue that generative AI has made phishing more persuasive and scalable by allowing attackers to create realistic, personalised, and fluent phishing messages more easily [2]. This has important implications for defence. If phishing messages become more polished and human-like, older warning signs such as grammatical errors or awkward wording become less reliable. This means that defenders may need more adaptive detection systems that take account of both technical patterns and human behaviour.

Taken together, the literature shows strong progress in two parallel areas. One area focuses on AI-based technical detection, while the other explores human susceptibility and behavioural risk [3]–[8]. The main gap is that relatively few studies combine both perspectives into one practical phishing detection framework. This study is positioned directly within that gap. It proposes that phishing detection may become more accurate, explainable, and realistic when user-behaviour signals are analysed alongside technical indicators.

3. Research Design and Methodology

3.1 Research Design

A quantitative, experimental research design will be used for this investigation. Because it entails creating and assessing an artifact in the form of an AI-based phishing detection model, the entire strategy is design-oriented. Two models—one based only on technical phishing traits and the other combining technical features with behavioral indicators—will be compared in this study.

Because the primary goal is to determine whether adding user-behavior signals enhances phishing detection performance, this approach is suitable. A quantitative design facilitates an organized assessment of whether the suggested method provides a significant improvement over current technical-feature-based models and permits objective comparison using established performance criteria [3], [4].

3.2 Research Methods and Materials

The study will use a comparative modelling strategy. First, a baseline model will be developed using conventional technical phishing features. These may include lexical properties of URLs, domain-based characteristics, message structure, email metadata, and content-based indicators depending on the dataset selected. Second, a combined model will be developed using both technical features and user-behaviour signals. The behavioural features may include awareness score, response time, sender-checking behaviour, confidence level, click tendency, and attention-related indicators. These variables are chosen because prior research suggests that awareness, intuition, attention, and decision style influence phishing outcomes [5]–[7]. The study will likely use machine learning models such as Random Forest and XGBoost because they perform well on

structured tabular data and offer useful interpretability. If the data is suitable, a deep learning model may also be included for comparison, especially in light of recent interest in deep learning for phishing detection [4]. Explainability tools such as SHAP will be used to understand which features contribute most to model decisions. Evaluation will be based on standard classification metrics including accuracy, precision, recall, F1-score, and ROC-AUC. Precision and recall are especially important in phishing detection because false negatives may allow successful attacks, while false positives can overwhelm users and reduce trust in the system.

3.3 Research Materials

The publicly accessible phishing datasets utilized in earlier phishing detection experiments will provide technical data. Phishing email datasets, malicious URL datasets, and associated open cybersecurity repositories are a few examples. Because they enable controlled model construction and repeatable evaluation, these datasets are suitable [3], [4]. Three methods are available for gathering the behavioral data. The first option combines a phishing-classification job with a brief user survey. In the second, participants evaluate sample emails in a controlled phishing simulation, and their answers are recorded. The third option is to create a synthetic proof-of-concept behavioral dataset based on factors found in the literature [5]–[7] if live data gathering proves to be unfeasible. The ultimate decision will be based on ethical acceptability and practicality. Python, Jupyter Notebook, pandas, scikit-learn, XGBoost, and SHAP will all be a part of the software ecosystem. It is also possible to use PyTorch or TensorFlow if deep learning is involved. A typical laptop can meet the minimal hardware requirements for classical machine learning models. Cloud-based solutions like Google Colab can facilitate more computing if necessary.

3.4 Ethical Considerations

When it comes to behavior-related data, ethics are particularly crucial. Informed consent, anonymization of responses, safe data storage, and a clear description of the study's objective are all necessary if participants are used. No needless personal information will be gathered. If a phishing simulation is employed, it will be regulated, have a restricted scope, and be made to prevent injury or discomfort. These actions are required because user-behavior data should be handled carefully and used responsibly, even if it is not extremely sensitive.

4. Expected Outcome

Several consequences are expected, but no actual results are available at this point of the proposal. First, given the advancements recently demonstrated in AI-based phishing detection research, it is anticipated that the technical-only model will offer a strong performance baseline [3], [4]. Second, by adding behavioral cues that more accurately reflect user vulnerability, it is anticipated that the combined model will demonstrate some improvement in phishing detection.

An additional anticipated result is that the integrated model might provide improved interpretability. A stronger case for human-aware detection systems may be made if specific behavioral characteristics consistently influence predictions. The study may nevertheless provide important information regarding which behavioral characteristics are worth incorporating and which are less helpful in reality, even if performance gains are modest.

The wider implication is that moving away from a strictly message-centered strategy may be advantageous for phishing defense. Future systems might take into account if a communication is dangerous for a specific user context in addition to questioning whether it is suspicious. That would be a more practical defense model that might also be more successful.

5. Limitations

The study is likely to be impacted by a number of limitations. The amount or realism of behavioral data may be constrained, particularly if recruiting participants is challenging. If utilized, synthetic data would aid in proof-of-concept research but fall short of accurately capturing human behavior. The depth of behavior-related data that can be gathered may also be limited by privacy considerations. Furthermore, user behavior can differ depending on jobs, contexts, and cultures, which could restrict generalizability. Lastly, the gain may be less than anticipated even if behavioral factors enhance detection. That would still be helpful to report since it would make the approach's practical limitations clear.

6. Timeline

A realistic timeline for the project is as follows:

- Month 1: Finalise research questions, proposal structure, and literature review plan.
- Month 2: Complete literature review and identify datasets, tools, and ethical requirements.
- Month 3: Design behavioural data collection approach and begin data preparation.
- Month 4: Develop baseline technical-feature model and prepare combined dataset.
- Month 5: Train and evaluate both models, then analyse and compare the results.
- Month 6: Interpret findings, write the final dissertation chapters, revise, and prepare for submission.

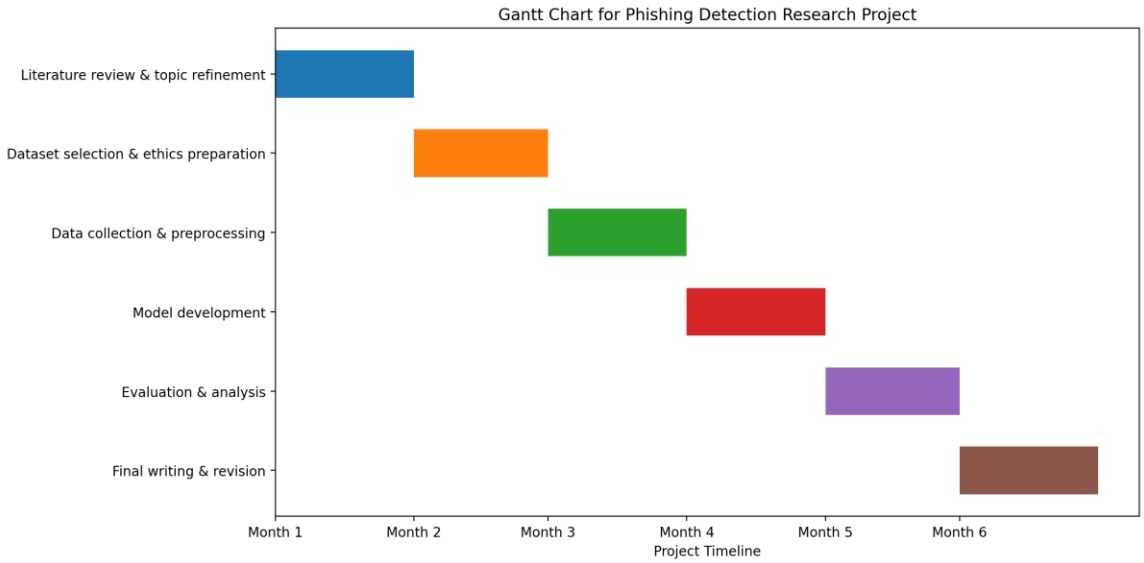

Fig. 1 – Project Timeline

7. References

[1] Anti-Phishing Working Group, “Phishing Activity Trends Report, 4th Quarter 2024,” APWG, Mar. 2025.

[2] R. Jabir, M.-T. Le, and T. T. Nguyen, “Phishing Attacks in the Age of Generative Artificial Intelligence: A Systematic Review of Human Factors,” AI, vol. 6, no. 8, Art. no. 174, 2025, doi: 10.3390/ai6080174.

[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, Art. no. 50, 2025, doi: 10.1007/s10462-024-11055-z.

[4] P. H. Kyaw, A. H. H. Htet, and T. Zin, “A Systematic Review of Deep Learning Techniques for Phishing Detection,” Electronics, vol. 13, no. 19, Art. no. 3823, 2024.

[5] Z. Fan, W. Li, K. B. Laskey, and K.-C. Chang, “Investigation of Phishing Susceptibility with Explainable Artificial Intelligence,” Future Internet, vol. 16, no. 1, Art. no. 31, 2024, doi: 10.3390/fi16010031.

[6] J. Buckley, D. Lottridge, J. G. Murphy, and P. M. Corballis, “Indicators of employee phishing email behaviours: Intuition, elaboration, attention, and email typology,” International Journal of Human-Computer Studies, vol. 172, Art. no. 102996, 2023, doi: 10.1016/j.ijhcs.2023.102996.

[7] D. Sturman, J. C. Auton, and B. W. Morrison, “Security awareness, decision style, knowledge, and phishing email detection: Moderated mediation analyses,” Computers & Security, vol. 148, Art. no. 104129, 2025, doi: 10.1016/j.cose.2024.104129.

[8] N. Marshall, D. Sturman, and J. C. Auton, “Exploring the evidence for email phishing training: A scoping review,” Computers & Security, vol. 139, Art. no. 103695, 2024, doi: 10.1016/j.cose.2023.103695.

[9] P. López-Aguilar, C. Urruela, E. Batista, J. Machin, and A. Solanas, “Phishing vulnerability and personality traits: Insights from a systematic review,” Computers in Human Behavior Reports, vol. 17, Art. no. 100784, 2025, doi: 10.1016/j.chbr.2025.100784.

[10] N. Ilany-Tzur and L. Fink, “Device and risk-avoidance behavior in the context of cybersecurity phishing attacks,” International Journal of Information Management, vol. 82, Art. no. 103050, 2025.

[11] U. Oner, O. Cetin, and E. Savas, “Human factors in phishing: Understanding susceptibility and resilience,” Information and Management, 2025.

[12] C. Tian, M. L. Jensen, and A. Durcikova, “Phishing susceptibility across industries: The differential impact of influence techniques,” Computers & Security, vol. 135, Art. no. 103487, 2023, doi: 10.1016/j.cose.2023.103487.

[13] K. Zhang, “Leveraging machine learning to proactively identify phishing campaigns before they strike,” Journal of Big Data, 2025.

8. Appendices

Appendix A: Literature Search Strategy

The literature for this proposal was gathered from Google Scholar, IEEE Xplore, ScienceDirect, SpringerLink, and MDPI. Main search terms included “phishing detection using AI,” “machine learning phishing detection,” “user behaviour and phishing susceptibility,” and “explainable AI in phishing.” The search focused mainly on recent studies published from 2020 onwards.

Appendix B: Proposed Features

The study will compare two phishing detection models. The first model will use technical features such as URL structure, email content, sender information, and metadata. The second model will combine these technical features with user-behaviour signals such as response time, confidence level, sender-checking behaviour, phishing awareness, and click tendency.

Appendix C: Draft Data Collection Plan

Technical data will be taken from public phishing datasets. Behavioural data may be collected through a short survey and a phishing email classification task, or through a controlled simulation if appropriate. If participant-based data collection is not feasible, a

synthetic behavioural dataset may be created based on variables identified in the literature.

Appendix D: Ethical Considerations Summary

The study will follow basic ethical principles including informed consent, voluntary participation, anonymity, and secure data storage. No unnecessary personal data will be collected. Any phishing simulation used in the study will be controlled and designed only for academic research purposes.

Appendix E: Project Timeline

Month 1: Literature review and topic refinement
Month 2: Dataset selection and ethics preparation
Month 3: Data collection and preprocessing
Month 4: Model development
Month 5: Evaluation and analysis
Month 6: Final writing and revision