

MSc in Cybersecurity and Digital Forensics

Research Methods and Professional Development
(ACFA7004)

Explainable Detection of AI-Enhanced Phishing Emails and URLs

Student Name	Md Obaydul Haque
Student Number	2435442
Programme	MSc Computing
Assessment	Research Proposal

This work focuses on cybersecurity, machine learning, natural language processing, and explainable AI.

Table of Contents

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

1. Introduction

1.1 Context and Background

Phishing is still one of the most common cyber threats because it targets people's decisions, not just technical weaknesses. Today's phishing attacks are no longer just badly written emails with spelling errors. Attackers now use large language models (LLMs) and other generative AI tools to create messages that look real, are tailored to the recipient, and are hard to tell apart from genuine communication [1], [2]. Malicious URLs are also still a major way for attackers to steal credentials, spread malware, and hijack sessions. So, defending against phishing is not just about checking the message text. It also requires spotting suspicious URL patterns, domain names, and other warning signs [3], [4].

In cybersecurity research, phishing detection has moved from rule-based filtering to using machine learning, deep learning, and transformer-based natural language processing models. Recent studies show that classical machine learning models like logistic regression, support vector machines, and random forests can perform well on benchmark datasets. Deep learning and transformer models can capture more complex features and handle more subtle phishing content [5]–[8]. Still, two main challenges remain. First, phishing messages that are rewritten or generated by AI can make traditional detectors less effective, especially if they were trained on older or less varied datasets [2], [9]. Second, highly accurate models are often hard to interpret, which makes it difficult to build trust, ensure transparency, and use them in real security operations [10], [11].

This proposal brings together cyber security, machine learning, NLP, and explainable AI (XAI). It addresses the need for phishing detection systems that are both accurate and easy to understand. In security operations, false positives waste analysts' time, while false negatives can put organizations at risk. A model that explains why an email or URL is suspicious can help analysts validate threats more quickly, improve awareness training, and make automated decisions more reliable [10], [12].

1.2 Research Problem

This research focuses on how current phishing detection systems struggle to stay robust and understandable when dealing with AI-generated phishing emails and harmful URLs. Many studies show high accuracy, but most models are tested on static datasets and do not pay enough attention to new LLM-rephrased phishing content, cross-format detection, or clear explanations of how decisions are made [2], [9], [11]. As generative AI makes it easier to create convincing phishing content quickly, the difference between real-world phishing tactics and older training data is growing [1], [2].

A second issue is that many state-of-the-art detection another problem is that many advanced detection models work like black boxes. In businesses, analysts and users need more than just a yes-or-no answer; they want clear reasons, such as unusual words, urgent language, mismatched sender and content, strange URL tokens, or odd domains. If there is no explanation, users might not trust correct results or might ignore mistakes, which lowers the system's usefulness. So, the main research problem is the need for a detection framework that can spot AI-enhanced phishing emails and URLs and also explain the reasons behind each decision.

1.3 Research Aim and Objectives

This research aims to design and test a machine learning framework that can explain how it detects AI-enhanced phishing emails and URLs.

1. Review recent studies on phishing detection, AI-driven social engineering, transformer based text classification, and explainable AI in cybersecurity.
2. Compare how well baseline machine learning models and transformer-based models classify phishing emails and URLs.
3. To compare the performance of baseline machine learning models and transformer based models for phishing email and URL classification.
4. Use explainability tools like SHAP and LIME to find out which features and content patterns influence the model's predictions.
5. Evaluate the framework using metrics such as accuracy, precision, recall, F1-score, ROC AUC, and confusion matrices, and discuss how useful it is for real world cybersecurity.

1.4 Significance of the Research

This research is important for three main reasons. First, it addresses a current cybersecurity problem: AI tools are making phishing easier to create, more convincing, and harder to detect [1], [2]. Second, it supports the development of trustworthy AI by promoting security systems that are transparent, accountable, and easy to understand, rather than being black boxes [10], [12]. Third, the results are helpful for organizations, universities, and small businesses that depend on email and need reliable automated tools to sort messages and raise awareness. The study also provides a clear way to evaluate these tools, which future MSc projects can use to build larger prototypes or test them in real-world situations.

2. Literature Review

This literature review is organized by theme to highlight how the research area has evolved and to identify the main gap in current knowledge. Most of the review covers studies from the past five years, with some reference to key earlier work on trustworthy AI and phishing detection. The review focuses on four main themes: the changing phishing threat landscape with generative AI, machine learning and deep learning methods for email phishing detection, phishing detection based on URLs and websites, and the importance of explainability in practical security analytics.

Theme 1: Theme 1: AI-enhanced phishing is becoming a serious threat. Recent studies and threat reports show that generative AI makes phishing attempts more convincing, better targeted, and easier to scale. The NCSC reports that AI will likely increase both the number and impact of cyber intrusions by improving the quality of social engineering content [1]. ENISA also notes that AI supported phishing and social engineering are growing threats [2]. Research shows that phishing emails generated or rewritten by large language models can be as effective as those written by experts and can also make traditional filtering systems less effective [9], [13]. These findings suggest that detection research needs to keep pace with how attackers use AI rather than relying on older phishing patterns.

Theme 2: Researchers use machine learning, deep learning, and transformer-based methods to detect phishing emails. Earlier studies found that feature-based machine learning models work well when trained on well labeled phishing datasets [5]. More recent research compares traditional machine learning with deep learning and transformer models, and often finds that transformer-based models better capture the context of email language [6], [7], [8]. Distil BERT and other lightweight transformer models are appealing because they provide strong language modeling while using less computing power than larger models [7]. However, the literature highlights a benchmarking problem: very high accuracy may depend on narrow datasets or similarities between training and test data, so performance can drop when tested on more realistic or reworded phishing emails [9], [14].

Theme 3: Detecting phishing URLs and websites remains important because attackers often use links to target victims. Recent studies show that both neural and feature based models can detect phishing URLs using lexical, structural, and host-based features [3], [4]. However, cross dataset stability is a concern. Some features that seem highly predictive in one dataset may not work as well in another. This suggests that URL detection should not rely on a single dataset or feature type, and explainability can help show whether a model is learning real patterns or just details specific to a dataset.

Theme 4: Explainability and trustworthy AI are important in cybersecurity. NIST lists explainable and interpretable AI as key parts of trustworthy AI, along with reliability, security, resilience, and accountability [10], [12]. In phishing research, recent papers have started using SHAP, LIME, and interpretable feature selection to build user trust and make models more transparent [11], [15]. Explainability is especially useful in cyber defense because analysts need reasons for automated decisions, and feature level insights can help with training and policy development. However, the literature is scattered. Some papers focus on email text, others on URLs, and a few examine AI-enhanced phishing across both. This clearly shows an opportunity for the proposed study.

Overall, the literature shows strong progress in phishing detection but also highlights three main gaps. First, there is little evaluation focused on AI-enhanced phishing and attacks rewritten by large language models. Second, many studies focus more on predictive performance than on how understandable the models are in practice. Third, email and URL evidence are often studied separately instead of together as indicators of phishing. This research aims to address these gaps by combining the current threat context, explainable modeling, and a practical comparative approach.

Theme	What the literature shows	Gap for this study
AI-enhanced phishing	LLMs make phishing more fluent, personalized, and scalable.	Need an evaluation focused on AI-rephrased / AI-generated phishing content.
Email detection models	ML and transformer models can achieve high accuracy.	Need comparison across realistic, current datasets and supporting explanations.
URL detection	Lexical and neural approaches are effective for malicious links.	Need interpretable feature analysis and a discussion of cross-dataset robustness.
Explainable AI	SHAP/LIME improves transparency and trust.	Need integration into a practical phishing framework with cybersecurity relevance.

3. Research Design and Methodology

3.1 Research Design

This study will use an experimental design within a design science research framework. Design science fits because the goal is to create and test a practical tool: an explainable phishing detection framework. The experimental part is needed to measure how well the model works, how robust it is, and how easy it is to interpret, all under controlled conditions. This method meets the assignment's requirement to justify the research approach and connect the methods to the study's aim.

The study will move through several stages: preparing the dataset, building baseline models, developing advanced models, generating explanations, and comparing results. This comparative approach fits the research goals, which include not just detecting phishing, but also finding out which models work best and which give the clearest explanations. The results will cover both technical performance and analysis of why the models make their decisions.

3.2 Research Methods and Materials

Methodological approach: This study will use a quantitative approach because the main research questions focus on measurable classification results. Using quantitative experiments allows for direct comparison between models and datasets with objective metrics like precision, recall, F1-score, and ROC-AUC. The research will mainly involve developing models and testing them on public datasets, along with feature analysis and explanation methods. If needed, a small set of AI-rephrased phishing messages may be created manually for evaluation, following an ethical, offline process that does not involve real people.

Data sources: Only public datasets will be used to prevent privacy and consent problems. Likely sources are well-known phishing email collections such as Nazario, Enron-based ham/phishing sets, CEAS, and other cleaned datasets. Phishing URL data may be taken from repositories like Phish Tank, UCI datasets, Kaggle benchmarks, or similar academic sources [3], [4], [6].

The final dataset will have phishing emails, legitimate emails, phishing URLs, and safe URLs. During cleaning, duplicates, obvious leakage fields, and malformed entries will be removed. Email text will be normalized, tokenized, and any metadata that could reveal labels will be taken out. URL strings will be turned into features such as length, token count, entropy, number of subdomains, suspicious keywords, and character patterns.

Model development: For email detection, the baseline models are TF-IDF with Logistic Regression or Linear SVM. Tree-based models such as Random Forests or XGBoost may also be used for comparison. For more advanced text analysis, a fine-tuned transformer such as Distil BERT will be used because it provides strong semantic understanding while keeping computational needs moderate [7]. For URL detection, the baseline approach uses lexical feature engineering with Logistic Regression, Random Forest, or Gradient Boosting. Depending on the data size and practicality, an advanced neural or sequence-based model may also be included [3], [4].

Explainability and evaluation: We will use SHAP to analyze how features contribute both globally and locally, especially in tree-based and linear models. LIME might help us interpret specific phishing emails or URLs. For transformer models, we will consider attention visualization or token-level contribution methods when possible.

The evaluation will use train, validation, and test splits, with stratification to keep class balance. We may also use k-fold cross-validation for the baseline models. To measure performance, we will look at accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices. Since phishing detection is a security issue, recall and false negatives are especially important. Precision is also important because it affects how much work analysts have to do.

The study will also check if the explanations match well-known phishing cues found in the literature, like urgent language, requests for credentials, fake domain names, and impersonation tactics.

Objective	Method	Output
Review current evidence	Conduct a thematic literature review on phishing, AI-driven social engineering, and explainable AI (XAI).	Clearer definition of the research gap and the criteria used for evaluation
Prepare data	Gather, clean, and label public datasets containing emails and URLs.	Description of the dataset used in experiments and the steps taken to prepare the data
Build baseline models	Apply TF-IDF features with models such as Logistic Regression, SVM, and tree-based algorithms.	Results from testing traditional methods as benchmarks
Build advanced models	Fine-tune DistilBERT and use a chosen model for analyzing URLs.	Comparison of performance results with baseline methods
Explain results	Use SHAP, LIME, and specific token or feature analysis to interpret model results.	Evidence supporting decisions that can be easily understood by people

Software and hardware: For this project, I will use Python, Jupyter Notebook, pandas, scikit learn, Hugging Face Transformers, SHAP, LIME, and visualization tools such as matplotlib and seaborn. I will use Git for version control. Most classical models should run on standard hardware, but fine-tuning transformers may need a GPU. If so, I can use Google Colab, university labs, or cloud notebooks. To keep the project suitable for an MSc, I will select efficient models and avoid making things more complex than necessary.

Ethical considerations: This proposal relies only on public, secondary datasets, so the ethical risk is low. There will be no real phishing campaigns, no use of live malicious URLs, and no targeting of personal email accounts. If AI-rephrased samples are needed for evaluation, they will be created only for offline defensive research and will not be shared outside the study. When working with datasets, personal identifiers will not be stored unless necessary, and any examples used in the dissertation or presentation will be sanitized. An ethics form will still be submitted, as required by the module brief, since all students must include one, even for low risk projects.

4. Expected Outcome

Because this is still the proposal stage, there are no final empirical results yet, but several outcomes are expected. First, transformer-based email models will probably outperform simple bag-of-words baselines on nuanced phishing texts, especially for AI-rephrased examples, since they capture context more effectively [7], [9]. Second, models that use lexical and structural features of URLs should still be useful, especially when combined with strong preprocessing and well-chosen features [3], [4]. Third, explainability tools should help show whether models rely on genuine phishing indicators or merely on unstable patterns in the dataset. This is important because a high-accuracy model that learns the wrong cues may not work well in real-world situations.

This research aims to develop a reliable, effective, and transparent phishing-detection framework. The framework could help with email sorting, cyber awareness training, and analyst review. Some possible limitations include dataset bias, class imbalance, domain drift, limited access to real enterprise email data, and challenges in measuring the quality of the explanations. Another limitation is that the study will likely focus on static lexical or host-based URL features rather than analyzing full dynamic webpage behavior, due to scope and safety constraints. These issues will be discussed in the evaluation and recommendations sections of the final project.

5. Timeline

Table 3 shows a practical 12-week project timeline for a postgraduate research proposal and the following implementation planning. This schedule is based on the proposal being approved before implementation and assumes that ethics clearance will be simple and low risk.

Week(s)	Milestone / Task	Deliverable
1–2	Check the topic and scope with your supervisor, get their feedback, and finalize your research questions.	Proposal direction has been approved
3–4	Do a focused literature review and create an annotated reference library.	Literature matrix completed and research gap refined
5	Find the datasets you need and make a plan for cleaning and preparing the data.	Dataset has been selected and preprocessing notes are documented
6–7	Build the basic email and URL models, then run your first benchmark tests.	Initial benchmark results are available
8–9	Develop the advanced models and compare their results with the basic ones.	Comparative performance results have been compiled
10	Use SHAP or LIME to explain your models and look for the main signs of phishing.	Interpretability findings have been summarized
11	Bring together your evaluation, note any limitations, and suggest practical recommendations.	Draft of the analysis is ready
12	Finish your final write-up, prepare your presentation, and proofread everything.	Project pack is ready for submission

6. References

- [1] National Cyber Security Centre, “The near-term impact of AI on the cyber threat,” 2024. [Online]. Available: <https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat>
- [2] European Union Agency for Cybersecurity (ENISA), “ENISA Threat Landscape 2025,” 2025. [Online]. Available: <https://www.enisa.europa.eu/publications/enisa-threat-landscape-2025>
- [3] H. Ghalechyan, E. Israyelyan, A. Arakelyan, G. Hovhannisyan and A. Davtyan, “Phishing URL detection with neural networks: an empirical study,” *Scientific Reports*, vol. 14, 2024, Art. no. 25134.
- [4] F. Ait Messaoud et al., “An explainable feature selection framework for web phishing detection,” *Intelligent Systems with Applications*, vol. 23, 2024, Art. no. 200373.
- [5] A. Alsubhi et al., “Machine Learning Approach for Email Phishing Detection,” *Procedia Computer Science*, vol. 245, pp. 202–209, 2024.
- [6] A. Raza et al., “In-Depth Analysis of Phishing Email Detection: Evaluating the Performance of Machine Learning, Deep Learning, and LLM-Based Models,” *Applied Sciences*, vol. 15, no. 6, 2025, Art. no. 3396.
- [7] “An Explainable Transformer-based Model for Phishing Email Detection: A DistilBERT Approach,” *arXiv:2402.13871*, 2024.
- [8] S. Alshamrani et al., “Email Phishing Detection Using Machine Learning Approaches,” in *Proc. ICICIS*, Springer, 2025, pp. 63–79.
- [9] K. Afane, W. Wei, Y. Mao, J. Farooq and J. Chen, “Next-Generation Phishing: How LLM Agents Empower Cyber Attackers,” *arXiv:2411.13874*, 2024.
- [10] NIST, “Artificial Intelligence Risk Management Framework (AI RMF 1.0),” 2023. [Online]. Available: <https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10>
- [11] O. Aouad et al., “Novel interpretable and robust web-based AI platform for phishing email detection,” *Computers & Security*, vol. 145, 2024, Art. no. 104008.
- [12] NIST, “AI Risks and Trustworthiness,” 2025. [Online]. Available: <https://airc.nist.gov/airmf-resources/airmf/3-sec-characteristics/>
- [13] F. Heiding, S. Lermen, A. Kao, B. Schneier and A. Vishwanath, “Evaluating Large Language Models’ Capability to Launch Fully Automated Spear Phishing Campaigns,” *arXiv:2412.00586*, 2024.
- [14] E-PhishGen authors, “E-PhishGen: Unlocking Novel Research in Phishing Email Detection,” *arXiv:2509.01791*, 2025.
- [15] “EXPLICATE: Enhancing Phishing Detection through Explainable AI and LLM Integration,” *arXiv:2503.20796*, 2025.

7. Appendices

Appendix A: Ethical risk summary. This project is considered low risk since it uses public datasets, does not include real users, and does not share any harmful content. AI-generated examples will be created solely for offline testing and will remain within the project.

Appendix B: Key evaluation metrics include accuracy, precision, recall, F1-score, ROC-AUC, confusion matrix, false positive rate, false negative rate, and some observations on explanation consistency.

Appendix C: This chapter can be expanded into a complete MSc project by building on the current proposal. The next steps are to develop a prototype email triage dashboard, test it with various datasets, and assess how useful it is for analysts.