Riphah International University

Thesis Proposal:

“AI for Multimodal Phishing Detection: Fusion of Email Text, Webpage Content, and Visual Features”

Superviser:

Dr. Jasim Saeed

By:

Talha Ali

SAP ID: 67484

Abstract

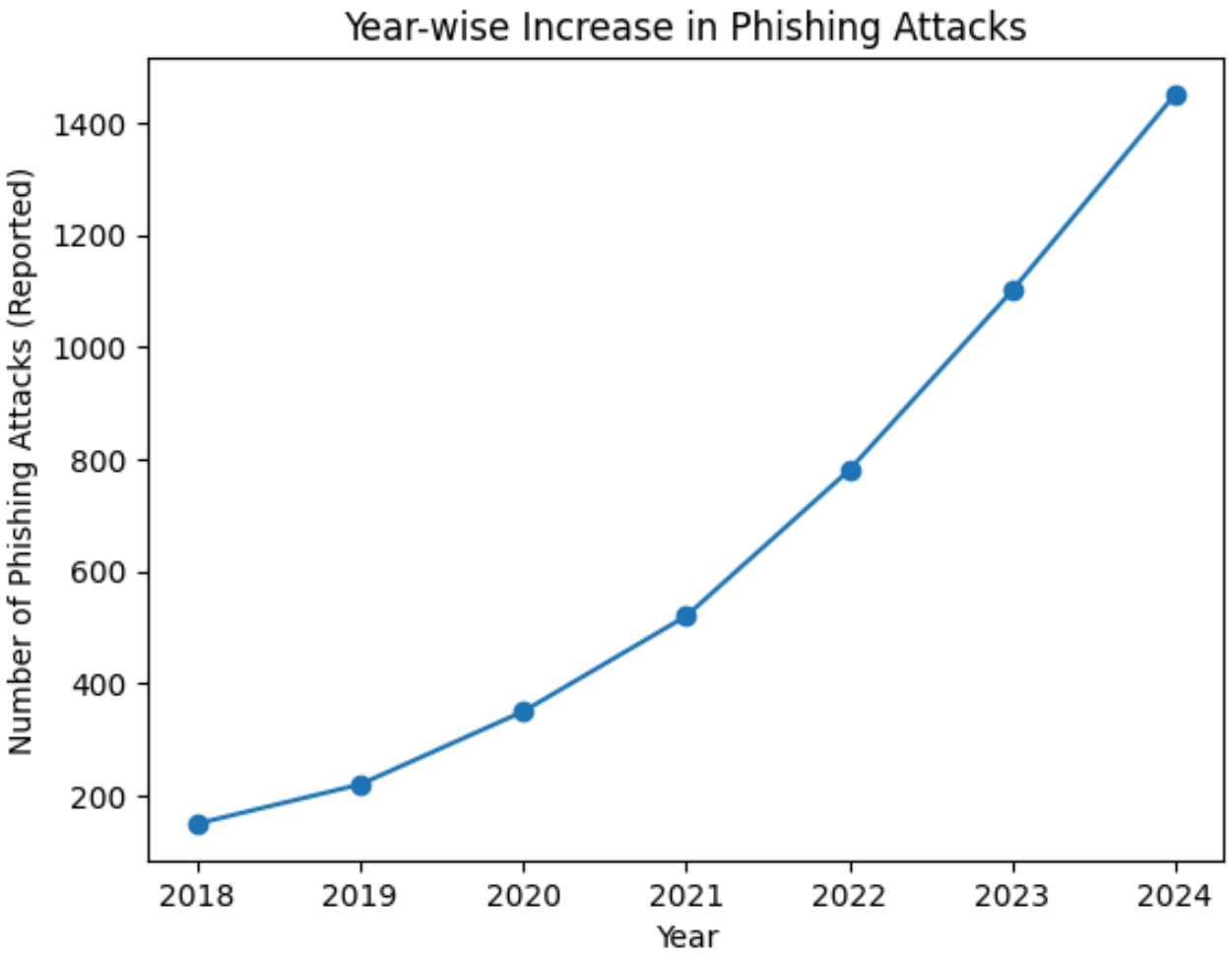
Phishing increasingly blends multiple modalities: deceptive emails link to malicious webpages that mimic real sites in terms of text, layout, and logos, and often embed persuasive images. Single-modality detectors (URL-only or text-only) miss many real-world attacks. This thesis proposes a multimodal detection framework that fuses email text features, webpage structural/HTML features, and visual embeddings from screenshots and images to improve detection performance and robustness. The approach trains modality-specific encoders (a transformer for text, a graph/DOM-based model for HTML structure, and a CNN/ViT for images) and combines them with an attention-based fusion module. Evaluation will use curated multimodal phishing datasets and newly collected samples, and will measure classification performance (Precision, Recall, F1, AUC) and robustness to evasion (visual/logo manipulation, text paraphrasing, and URL obfuscation). Expected contributions are: (1) a reproducible multimodal pipeline for phishing detection; (2) an analysis of which modality or fusion strategy contributes most; and (3) a robustness evaluation under common adversarial manipulations. The code and dataset splits will be released for reproducibility.

Introduction

With the rapid growth of online banking, e-commerce, and cloud services, phishing has become one of the most common and damaging cyber threats. Modern phishing attacks no longer rely only on suspicious URLs or poorly written emails; instead, they carefully imitate trusted brands, clone login pages, and use convincing visual elements to deceive users. Traditional defenses such as blacklists, URL filters, and simple rule-based systems struggle to detect these advanced and constantly evolving attacks.

Machine learning and deep learning–based approaches have improved detection by treating phishing as a classification problem, using features from URLs, HTML code, or email text. However, most of these models focus on a single source of information and ignore the fact that a real phishing attack is a chain of steps, starting from a crafted email and ending on a spoofed web page. This thesis proposes an AI-based multimodal framework that jointly analyzes email text, webpage structure/content, and visual information, aiming to achieve more robust and reliable phishing detection than existing unimodal methods.

Growth of Phishing Attacks Over the Years



Problem Statement

Most current phishing detection systems are unimodal and analyze only one aspect of an attack, such as the URL, the email text, or the webpage HTML. As a result, they fail to fully model the complete attack chain, where the attacker uses a deceptive email to lure the victim to a visually spoofed landing page. These models are also highly vulnerable to adversarial changes like paraphrased text, obfuscated URLs, and modified logos or layouts.

There is limited research on how to effectively fuse multiple modalities email content, URL/HTML features, and webpage visuals within a single AI-based framework and how different fusion strategies affect robustness. Furthermore, there is a lack of standardized multimodal datasets that contain synchronized email–webpage pairs with visual information, which makes fair comparison and reproducible research difficult. Therefore, the central problem of this thesis is to design and evaluate a multimodal phishing detection framework that improves performance and robustness compared to state-of-the-art unimodal detectors.

Literature Review

Phishing detection has evolved from simple blacklist- and rule-based filters to data-driven approaches that apply machine learning (ML) and deep learning (DL) across URLs, HTML, email content and user behavior. Early work mainly optimized single-channel classifiers, but recent systematic reviews show a huge growth in AI-based phishing research spanning websites, emails, smishing and social media, and also highlight issues like data leakage, non-standard evaluation and concept drift in deployed systems [1–4], [10]. Wilk-Jakubowski et al. conduct a large-scale bibliometric and systematic review of ML and neural network methods in phishing detection from 2017–2024, mapping research trends across URL, webpage, email and message-based attacks, and stressing the need for more realistic datasets, cross-channel evaluation and robust models that can generalize beyond curated benchmarks [10]. Similarly, Alazaidah et al. systematically compare twenty-four classifiers and four feature-selection strategies on two website-phishing datasets; their results show that RandomForest, FilteredClassifier and J48, combined with InfoGain-based feature selection, consistently achieve the best trade-off between accuracy, precision and recall, confirming that careful algorithm and feature-engineering choices remain crucial in practice [14].

Beyond traditional single-view models, recent work increasingly exploits multi-dimensional and multi-modal information. Belfedhal and Belfedhal propose a multi-modal deep learning framework that jointly leverages URL, HTML content and script-level features to detect malicious webpages, demonstrating that fusing heterogeneous signals improves robustness over purely URL- or content-based baselines [11]. Murhej and Nallasivan go further and design a multimodal phishing-detection framework that integrates SMS, email and URL data; their architecture combines an EM-BERT encoder with SPCA-based EAI-SC-LSTM to capture both textual semantics and user behavior, achieving very high detection performance across multiple channels while still relying mainly on text-like inputs [12]. In parallel, Kim et al. target voice-based phishing and develop a multimodal system that fuses KoBERT-based text embeddings with CNN–BiLSTM voice features using an 8:2 weighted fusion scheme; experiments on a real-world Korean phishing dataset show that the multimodal model reaches an F1-score near 0.994 and accuracy close to 0.999, dramatically reducing false positives compared with voice-only classification [13].

These recent studies confirm that (i) ML/DL-based phishing detection is mature for individual channels such as websites and emails, and (ii) multi-modal fusion across complementary signals (URL + HTML + script, or text + audio) yields clear gains in robustness and generalization [10–14]. However, the literature also reveals several gaps: most models are still trained and evaluated on narrow, single-source datasets; multi-modal frameworks often target only one attack vector (e.g., either websites or voice calls) rather than providing a unified solution across different phishing media; and many works do not explicitly address real-world deployment issues such as adaptive adversaries, evolving attack patterns and explainability for end-users [10–13]. These limitations motivate the development of an integrated phishing-detection approach that can combine multiple feature types and channels while maintaining high accuracy, interpretability and practicality for real-world environments.

Objectives

1. Design modality-specific encoders for email text, webpage HTML/URL, and visual screenshots.

2. Create an attention-based fusion module and compare fusion strategies (early, mid, late).

3. Collect/curate a multimodal dataset (email + landing page screenshot + URL + labels) and standardize splits.

4. Evaluate detection performance against state-of-the-art uni-modal baselines and measure robustness to common evasions.

5. Produce reproducible code, results, and a discussion on deployment tradeoffs (latency, storage, privacy).

Methodology

- Data: Combine public phishing email/website datasets + crawl landing pages to obtain screenshots; augment with synthetic examples for robustness tests.

- Models:
    - Text: fine-tuned transformer (BERT / DistilBERT) on email body + subject + sender embedding.

    - HTML/URL: transformer over URL tokens + DOM graph encoder (GCN/GAT) for structural cues.

    - Visual: CNN or Vision Transformer on webpage screenshots / embedded images (logo detection).

    - Fusion: Attention-weighted concatenation or cross-modal transformer; early, mid, and late fusion strategies will be systematically compared, with late fusion using cross-modal attention selected as the primary approach because recent multimodal surveys and KnowPhish results demonstrate superior robustness when one or more modalities are attacked or missing [8], [9].

    - Other (alternative models not adopted): Prior work has explored additional architectures such as CNN–BiLSTM audio models for voice phishing [13], large language model–knowledge graph (LLM–KG) pipelines [9], CSOA-based feature optimization, and the EAI-SC-LSTM model for multimodal text streams [12]. These models are not adopted in this thesis, which instead focuses on modular transformers, CNN/ViT, and GCN/GAT encoders with a simpler fusion and optimization pipeline.

- Training: multi-task objective (primary: phishing/non-phishing; auxiliary: brand/logo match, suspiciousness score).

- Robustness: test paraphrase/text obfuscation, logo removal/replacement, homoglyphs/homograph URL attacks.

- Evaluation: Precision, Recall, F1, AUC-ROC, and robustness metrics (attack success rate, drop in F1 under attack).

- Ethics & safety: anonymize any real user data; run all web crawling in isolated VMs and follow institutional policies.

Datasets & Tools

- Datasets: CIC-Phishing, other public phishing URL datasets, Enron and other benign email corpora, and the TR-OP and KnowPhish datasets with their brand-knowledge extensions.

- Tools and frameworks: Python, PyTorch/TensorFlow, Hugging Face Transformers, OpenCV/PIL for image processing, Selenium/requests with a headless browser for webpage rendering and screenshots, and semgrep/BeautifulSoup for HTML parsing and feature extraction.

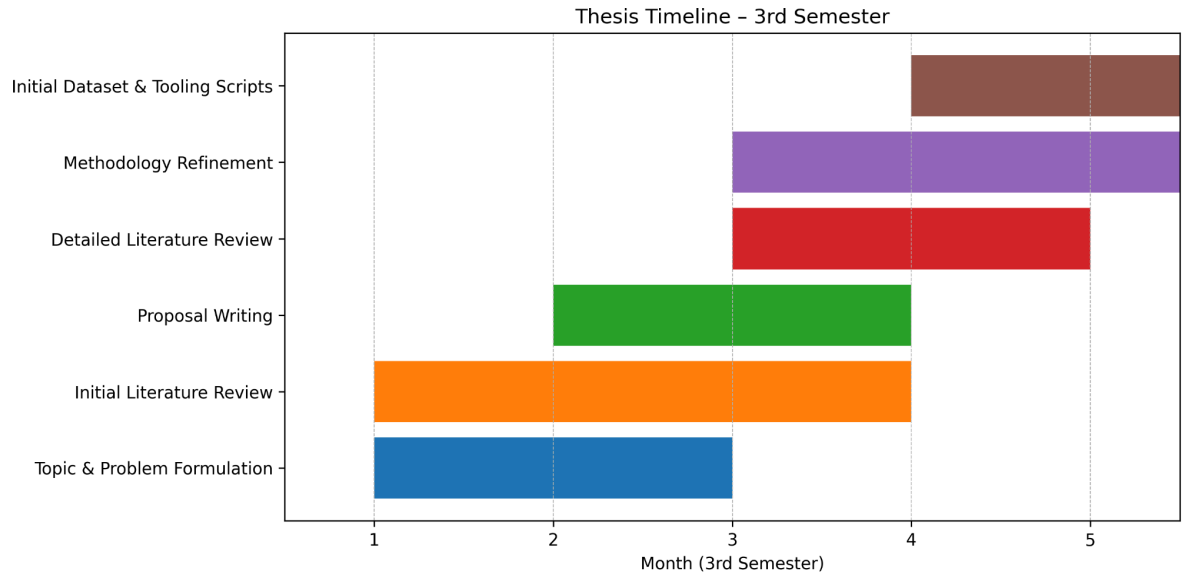
Timeline

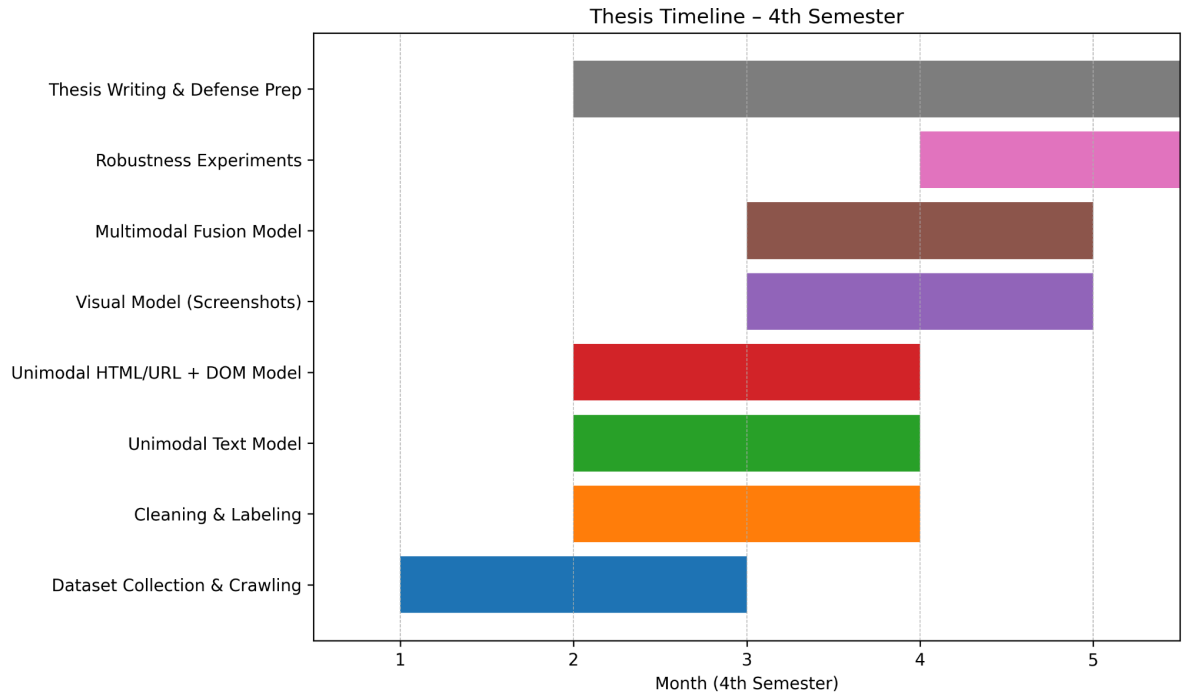
The thesis work spans the second half of the 3rd semester and the full 4th semester of the MSCS program (approximately ten months in total).

During the 3rd semester, the focus is on topic selection, problem formulation, proposal writing, an initial and detailed literature review, and methodology refinement, as well as preparing scripts for data collection and tooling.

In the 4th semester, the emphasis shifts to dataset collection and preprocessing, implementation and training of unimodal and multimodal models, robustness experiments, and final thesis writing and defense preparation.

A detailed Gantt chart is provided to illustrate the planned schedule across both semesters.





References

[1] A. Odeh, I. Keshta, and E. Abdelfattah, “Machine learning techniques for detection of website phishing: A review for promises and challenges,” in *Proc. IEEE 11th Annu. Comput. Commun. Workshop Conf. (CCWC)*, Las Vegas, NV, USA, Jan. 2021, pp. 351–358, doi: 10.1109/CCWC51732.2021.9375997.

[2] R. Zieni, L. Massari, and M. C. Calzarossa, “Phishing or not phishing? A survey on the detection of phishing websites,” *IEEE Access*, vol. 11, pp. 18499–18518, 2023, doi: 10.1109/ACCESS.2023.3247135.

[3] P. A. Bhavani, C. Madhumitha, P. S. Likhitha, and C. P. Sai, “Phishing websites detection using machine learning,” SSRN Electron. J., 2022, doi: 10.2139/ssrn.4208185.

[4] K. Barik, S. Misra, and R. Mohan, “Web-based phishing URL detection model using deep learning optimization techniques,” *Int. J. Data Sci. Anal.*, 2025, doi: 10.1007/s41060-025-00728-9.

[5] L. Tang and Q. H. Mahmoud, “A deep learning-based framework for phishing website detection,” *IEEE Access*, vol. 10, pp. 1509–1519, 2022, doi: 10.1109/ACCESS.2021.3137636.

[6] T. Kim, N. Park, J. Hong, and S.-W. Kim, “Phishing URL detection: A network-based approach robust to evasion,” in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS ’22)*, Los Angeles, CA, USA, Nov. 2022, pp. 1–14, doi: 10.1145/3548606.3560615.

[7] L. R. Kalabarige, R. S. Rao, A. Abraham, and L. A. Gabralla, “Multilayer stacked ensemble learning model to detect phishing websites,” *IEEE Access*, vol. 10, pp. 79543–79556, 2022, doi: 10.1109/ACCESS.2022.3194672.

[8] W. Li, S. Manickam, Y. W. Chong, M. A. Razak, and S. S. H. Shah, “A state-of-the-art review on phishing website detection techniques,” *IEEE Access*, vol. 12, pp. 173496–173521, 2024, doi: 10.1109/ACCESS.2024.3514972.

[9] Y. Li, C. Huang, S. Deng, M. L. Lock, T. Cao, N. Oo, H. W. Lim, and B. Hooi, “KnowPhish: Large language models meet multimodal knowledge graphs for enhancing reference-based phishing detection,” in *Proc. 33rd USENIX Secur. Symp. (USENIX Security ’24)*, Philadelphia, PA, USA, Aug. 2024, pp. 1–19.

[10] J. L. Wilk-Jakubowski, Ł. Pawlik, G. Wilk-Jakubowski, and A. Sikora, “Machine Learning and Neural Network Methods in Phishing Detection: A Systematic Review (2017–2024),” *Electronics*, vol. 14, no. 18, article 3744, 2025.

[11] A. E. Belfedhal and M. A. Belfedhal, “Multi-Modal Deep Learning for Effective Malicious Webpage Detection,” *Revue d’Intelligence Artificielle*, 2023.

[12] A. Murhej and K. Nallasivan, “A Multimodal Framework for Detecting Phishing Attacks Using EM-BERT and SPCA-Based EAI-SC-LSTM,” *Frontiers in Communications and Networks*, vol. 6, article 1587654, 2025.

[13] J. Kim *et al.*, “A Multimodal Voice Phishing Detection System Integrating Text and Audio Analysis,” *Applied Sciences*, vol. 15, article 11170, 2025.

[14] R. Alazaidah, A. Al-Shaikh, M. R. Al-Mousa, H. Khafajah, G. Samara, M. Alzyoud, N. Al-Shanableh, and S. Almatarneh, “Website Phishing Detection Using Machine Learning Techniques,” *Journal of Statistics Applications & Probability*, vol. 13, no. 1, pp. 119–129, 2024.