

NAME :- VAISHNAVI PATEL

ID :- 2503493

**Module: Research Methods and
Professional Development**

Detecting AI-Generated Phishing Emails Using Machine Learning

Research Proposal Presentation



What This Presentation Covers

01

Introduction & Background

How phishing has changed with generative AI

02

Research Problem

Why current detectors fail against AI-generated phishing

03

Aim & Objectives

What this study sets out to achieve

04

Literature Review

What existing research tells us — and where it falls short

05

Methodology

Models, datasets, and evaluation plan

06

Expected Outcomes

Anticipated results and what they mean in practice

07

Timeline & Conclusion

Project roadmap and summary

01

Introduction & Background

Introduction & Background

The Phishing Threat

Phishing tricks users into handing over credentials or financial details. It remains the leading initial attack vector in data breaches worldwide — consistently outranking malware and vulnerability exploits [2].

>80%

of recent phishing emails show evidence of AI-generated content [1]

The AI Shift

Traditional phishing was easy to spot: bad grammar, generic greetings, suspicious links. GPT-4 and tools like WormGPT now produce polished, personalised emails at scale [1], [3]. Conventional filters were not built for this.



APWG reports rising phishing volumes year-on-year — generative AI toolkits are accelerating that trend [1]

Research Problem

Why existing detection systems are failing:

Rule-Based Filters Are Outdated

Built to flag misspellings, blacklisted URLs, and formulaic sentence patterns. AI-generated emails bypass all of these cleanly.

ML Models Trained on Human-Written Phishing

Most training datasets contain only human-crafted phishing [3]. Models trained on these don't generalise to AI content.

Polymorphic Campaigns

Generative AI produces a unique variant per recipient [1], [3]. Signature-based detection is entirely defeated — there's nothing to match against.

Email Platforms Struggle Too

Gmail and Outlook show reduced effectiveness filtering AI-generated phishing compared to traditional phishing [2].

Aim & Objectives

AIM

To create and evaluate a machine learning system that can detect AI-generated phishing emails more accurately than existing methods, helping improve protection against modern phishing attacks.

O1

Conduct a literature review on phishing detection, focusing on AI-generated content.

O2

Curate a dataset combining legitimate emails, traditional phishing, and AI-generated phishing samples.

O3

Implement and compare Random Forest, SVM, LSTM, and BERT classification models.

O4

Evaluate models using accuracy, precision, recall, and F1-score.

O5

Discuss practical implications, limitations, and future directions.

02

Literature Review

Literature Review: Key Findings

Traditional ML

- Naive Bayes, SVM, Random Forest widely used
- NLP-enhanced features improve accuracy
- SVM reaches ~97.6% on standard datasets [4]
- All struggle with polished AI-generated text

Deep Learning & Transformers

- CNN, LSTM, BERT outperform classical ML
- BERT's bidirectional context makes it the best performer [5]
- Hybrid CNN-BiGRU also shows strong results
- Explainable AI (XAI) integration is emerging

AI-Generated Phishing Detection

- Eze & Shamir [3] built 865 AI-phishing samples — pioneering work
- Stylometric analysis is limited against prompted AI text
- AI phishing is harder to detect than traditional phishing

Gaps in the Literature

- Most datasets lack AI-generated phishing samples
- No systematic comparison of ML vs BERT on AI phishing [5]
- Cross-study comparison difficult — varied datasets

03

Research Design & Methodology

Research Design & Methods

Design Science Research (DSR) — iterative design, build, test, refine [6]

1

Data Collection

Kaggle Phishing Dataset [7] + Eze & Shamir AI-phishing corpus [3] (865 samples) + Enron legitimate emails [8]

2

Preprocessing

Tokenisation, stop-word removal, TF-IDF for classical models, BERT tokeniser for transformer input

3

Model Training

Random Forest, SVM (classical ML) | LSTM (deep learning) | BERT (transformer) — all under identical conditions

4

Evaluation

Model performance will be evaluated using accuracy, precision, recall, F1-score, confusion matrices, and statistical significance testing.

ML Models & Datasets

Models Compared

Model	Type	Rationale
Random Forest	Classical ML	Strong baseline; widely used
SVM	Classical ML	Top classical performer ~97.6% [4]
LSTM	Deep Learning	Captures sequential text patterns
BERT	Transformer	State-of-the-art NLP; best in class

Datasets Used

Kaggle Phishing Email Dataset [7]

Balanced phishing and legitimate emails — primary source

Eze & Shamir AI-Phishing Corpus [3]

865 AI-generated phishing emails from DeepAI platform

Enron Email Dataset [8]

Legitimate email samples for class balance and diversity

Ethics: All datasets are publicly available with no personal data involved. Any AI-generated samples produced for this study are for research purposes only, under UWTSD institutional ethics approval.

04

Expected Outcomes & Significance

Expected Outcomes & Significance

Anticipated Results



BERT expected to achieve the highest detection accuracy for AI-generated phishing, consistent with published benchmarks [5].



Training on a mixed dataset (traditional + AI phishing) should improve robustness over single-source training [3].



LSTM expected to perform competitively but may miss the subtler patterns in AI-generated text.



Random Forest and SVM will provide useful classical baselines for comparison [4].

Why This Matters

Academic

First systematic ML vs BERT comparison on datasets with explicit AI-generated phishing samples.

Practical

Informs next-generation email filters better equipped to handle AI-crafted attacks.

Societal

Protects individuals, SMEs, healthcare, and educational institutions from increasingly sophisticated threats.

Limitations: Dataset representativeness · Offline experimental setting · BERT computational cost · Focus on email text only (not headers or URLs)

Project Timeline (May – July 2026)



Key References (IEEE Format)

[1]

F. Alharbi et al., "In-Depth Analysis of Phishing Email Detection: Evaluating the Performance of ML and Deep Learning Models," *Applied Sciences*, vol. 15, no. 6, p. 3396, 2025.

[2]

Verizon, "Data Breach Investigations Report (DBIR)," Verizon, 2024. [Online]. Available: <https://www.verizon.com/business/resources/reports/dbir/>

[3]

C. S. Eze and L. Shamir, "Analysis and Prevention of AI-Based Phishing Email Attacks," *Electronics*, vol. 13, no. 10, p. 1839, 2024.

[4]

S. Atawneh and H. Aljehani, "Phishing Email Detection Model Using Deep Learning," *Electronics*, vol. 12, no. 20, p. 4261, 2023.

[5]

R. Meléndez et al., "Comparative Investigation of Traditional ML Models and Transformer Models for Phishing Email Detection," *Electronics*, vol. 13, p. 4877, 2024.

[6]

A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design Science in Information Systems Research," *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004.

[7]

Kaggle, "Phishing Email Dataset," Kaggle, 2023. [Online]. Available: <https://www.kaggle.com>

[8]

B. Klimt and Y. Yang, "The Enron Corpus: A New Dataset for Email Classification Research," in *Proc. European Conf. on Machine Learning (ECML)*, 2004.

Thank You

“Improving Phishing Detection in the Age of AI”