

University of Wales Trinity Saint David

MSc Cybersecurity and Digital Forensics

ACFA7004 – Research Methods and Professional Development

Research Proposal

Detecting AI-Generated Phishing Emails Using Machine Learning

Student Name: Vaishnavi Riteshkumar Patel

Enrolment Number: 2503493

Lecturer: Dr. Suresh Shanmugasunadaram

Submission Date: 1st May 2026

Table of Contents

1. Introduction.....	3
1.1 Context and Background.....	3
1.2 Research Problem	4
1.3 Research Aim and Objectives.....	5
2. Significance of the Research.....	5
3. Literature Review.....	6
3.1 Introduction to the Literature Review	6
3.2 Traditional Machine Learning Approaches	6
3.3 Deep Learning and Transformer-Based Methods.....	7
3.4 Detection of AI-Generated Phishing Content.....	8
3.5 Key Contributions and Gaps.....	8
3.6 Relation to This Study	9
3.7 Literature Summary	9
4. Research Design and Methodology	10
4.1 Research Design.....	10
4.2 Research Methods.....	10
4.3 Research Materials.....	11
5. Expected Outcomes	12
6. Timeline	13
7. References.....	14

1. Introduction

1.1 Context and Background

Cybersecurity is one of the biggest challenges in today's digital world. As organisations and individuals depend more on electronic communication, cyber threats have become more common and complex. One of the most serious and widespread threats is phishing. Phishing is a type of cyberattack, usually carried out through email, where attackers Scam persons into Sharing sensitive information such as usernames, Credentials, financial data, or personal data [1]. Despite awareness campaigns and security training, phishing remains highly effective and causes many data breaches worldwide [2].

The fast growth of generative artificial intelligence (AI), particularly large language models (LLMs) like GPT-4, has made phishing attacks more advanced. Traditional phishing emails were often easy to spot because of spelling mistakes, poor grammar, and generic messages [3]. In contrast, AI-generated phishing emails are more convincing because they are grammatically correct, personalised, and relevant to the target. This makes them much harder for users and traditional detection systems to identify [4]. Recent reports show that over 80% of phishing emails contain signs of AI-generated content, highlighting the seriousness of this threat [5].

Machine learning (ML) has shown strong potential in detecting phishing threats. By training models on large datasets of both legitimate and phishing emails, researchers can classify suspicious messages with high accuracy [6]. Different techniques have been used, including traditional approaches like Random Forest and Support Vector Machines (SVM), as well as advanced deep learning models such as CNN, LSTM, and transformer-based models like BERT [7][8]. These methods are effective because they can recognise hidden patterns in data instead of relying only on fixed rule-based filters.

Despite these advancements, most research has focused on detecting traditional human-written phishing emails. Less attention has been given to AI-generated phishing emails, which are harder to detect because they often do not contain spelling mistakes, grammar errors, or other common warning signs. This creates an important research gap and motivates this study, which intentions to develop and evaluate a machine learning-based framework for detecting AI-generated phishing emails.

1.2 Research Problem

This research addresses the limited ability of current phishing detection systems to recognise AI-generated phishing emails. Traditional spam and phishing filters are mainly designed to detect clear warning signs such as spelling mistakes, suspicious links, blacklisted domains, and repeated language patterns [9]. However, these methods often fail with AI-generated emails because they are usually well-written, relevant, and more varied in structure.

Even advanced machine learning-based detection systems struggle with AI-generated phishing emails because these messages often look very similar to legitimate emails. Studies show that major email platforms like Gmail and Outlook are less effective at detecting AI-generated phishing emails than traditional phishing attempts [10]. This problem becomes worse because generative AI can create unique phishing emails for each recipient, making signature-based detection methods much less effective [5].

Many existing phishing detection studies mainly use datasets of human-written phishing emails, which means the models may not work well on AI-generated phishing emails [3]. There is also a absence of big public datasets covering AI-generated phishing emails, making it difficult for researchers to build and test reliable detection models. This shows a clear need for research focused on detecting AI-generated phishing emails using both traditional machine learning methods and advanced transformer-based models on datasets that include AI-generated samples.

1.3 Research Aim and Objectives

Aim: This research aims to design and evaluate a machine learning-based framework for detecting AI-generated phishing emails and improving detection accuracy compared to existing methods.

Objectives:

1. To review existing research on phishing email detection, with a focus on methods for identifying AI-generated content.
2. To collect and prepare a dataset of legitimate emails, traditional phishing emails, and AI-generated phishing emails for training and testing.

3. To grow and compare different machine learning and deep learning models, including Random Forest, SVM, LSTM, and BERT, for phishing email classification.
4. To evaluate the models using correctness, precision, memory, and F1-score to recognise the most operative method for detecting AI-generated phishing emails.
5. To analyse the practical implications, limitations, and possible future improvements of the proposed detection framework.

2. Significance of the Research

This research is important both practically and academically. From a practical view, phishing remains one of the most common causes of cybersecurity incidents. The Anti-Phishing Working Group has reported a yearly increase in phishing attacks, and generative AI has made this problem even worse [2]. If detection systems cannot keep up, organisations face higher risks of data breaches, financial loss, and damage to their reputation. A strong machine learning-based system for detecting AI-generated phishing emails could deliver better protection for organisations of all sizes.

From an academic perspective, this study contributes to an important and growing area of cybersecurity research. Although many studies have explored phishing detection using machine learning, fewer have focused on detecting AI-generated phishing emails [3][11]. By comparing traditional machine learning models by transformer-based models on datasets that include AI-generated emails, this research will provide useful evidence for future studies. The findings may also help improve commercial email filtering systems and connect academic research with real-world applications.

The social importance of this research is also significant. Phishing attacks affect not only large organisations but also individuals, small businesses, healthcare providers, and educational institutions. Successful phishing attacks can cause identity theft, financial loss, and the exposure of sensitive personal, financial, or medical information. By improving the detection of AI-generated phishing emails, this research helps create safer and more secure digital communication for everyone.

3. Literature Review

3.1 Introduction to the Literature Review

The purpose of this literature examination is to scrutinize current research on phishing email detection, with a focus on machine learning and deep learning techniques. It mainly covers studies published in the last five years because phishing methods and detection technologies are changing quickly. The review is organised by topic, starting with traditional machine learning methods, followed by deep learning and transformer-based approaches, and finally studies on detecting AI-generated phishing emails. The purpose is to identify key findings, common methods, and research gaps that this study will address.

3.2 Traditional Machine Learning Approaches

Many studies have discovered the use of traditional machine learning algorithms for phishing email detection. These methods usually extract features from email content, such as word frequency, keywords, URL structures, and header information, and use them as input for supervised learning models. Common algorithms used include Naive Bayes, Decision Trees, Random Forest, Support Vector Machines (SVM), and Logistic Regression [6][12].

Alhogail and Alsabih [13] used machine learning and natural language processing (NLP) to improve phishing email detection by extracting language features from email content. Their results showed that NLP-based features improved detection compared to traditional rule-based systems. Similarly, Fares et al. [14] developed a phishing detection framework combining machine learning and deep learning methods, achieving strong results on standard email datasets. Their study also highlighted the importance of effective feature selection for improving accuracy.

A 2024 study compared Random Forest, Support Vector Machine (SVM), and XGBoost using a phishing email dataset, and the SVM model achieved the highest accuracy of about 97.6% [15]. This supports earlier research showing that SVM remains a strong method for text classification when combined with good feature engineering. However, the study also found that these models may struggle with more advanced phishing emails, especially those without clear language or structural warning signs.

A major weakness of traditional machine learning methods is their reliance on manually created features. As phishing emails become more advanced, especially AI-generated ones, these handcrafted features become less effective at identifying malicious messages. This has led to a

shift toward deep learning methods, which can automatically learn complex patterns directly from raw text data.

3.3 Deep Learning and Transformer-Based Methods

Deep learning methods have gained strong attention in phishing email detection because they can identify complex patterns in text without needing much manual feature engineering. Models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Bidirectional LSTM (Bi-LSTM) have been used for this task and have shown promising results [7][10].

Atawneh and Aljehani [7] developed a phishing email detection framework using deep learning and compared CNN, RNN, LSTM, and BERT models on public datasets. Their study found that all deep learning models performed better than traditional machine learning methods, but BERT achieved the best overall results. This was because BERT could better understand word meaning and context using bidirectional attention. Since BERT is pre-trained on large text datasets and then fine-tuned for phishing detection, it was able to work well across different writing styles and email structures.

Altwaijry et al. [16] compared several deep learning models for phishing email detection, including CNN combined with LSTM, Bi-LSTM, GRU, and Bi-GRU. Their results showed that the hybrid CNN-BiGRU model achieved the highest accuracy. This suggests that combining CNN's ability to extract local features with GRU's strength in handling sequence patterns can better detect phishing emails. These findings show that hybrid deep learning models are a promising approach for improving phishing detection.

A notable contribution comes from Uddin and Sarker [17], who developed an explainable transformer-based phishing email detection model using BERT. Their model achieved high accuracy and also used Explainable AI (XAI) techniques to make its decisions more transparent. This is important in real-world cybersecurity, where it is not enough to only label an email as phishing; analysts also need to understand why a decision was made so they can make better judgments and improve detection systems.

The existing literature shows a clear trend that transformer-based models, especially BERT, are currently the most effective approach for phishing email classification. Their main strength is their ability to understand words in context rather than as isolated terms. This contextual

understanding gives them an advantage over traditional machine learning methods and earlier deep learning models [1][8].

3.4 Detection of AI-Generated Phishing Content

With the rise of generative AI tools such as ChatGPT, DeepAI, and WormGPT, a new type of phishing threat has emerged: fully AI-generated phishing emails. Unlike traditional phishing emails, these messages often lack common language errors that were previously used as detection signals. As a result, they are much harder to detect using conventional methods [3][4].

Eze and Shamir [3] conducted an early study on detecting and mitigating AI-generated phishing emails. They created a dataset of 865 AI-generated phishing emails using the DeepAI platform and tested several machine learning models on a dataset that also included legitimate emails from the Enron corpus and Nigerian scam emails. Their results showed that while machine learning models could identify AI-generated phishing emails with reasonable accuracy, the task was harder than detecting traditional phishing. This was especially true when the AI-generated emails closely matched the tone, structure, and professionalism of real business communication.

A 2025 study by Opara, Modesti, and Golightly [18] investigated the use of spam filters and stylometric techniques to detect AI-generated phishing emails. Stylometric analysis examines writing style features such as sentence length, vocabulary diversity, and sentence structure. The study found that while these features can help distinguish emails to some extent, their effectiveness decreases when AI-generated text is prompted to imitate a specific writing style. These results highlight how adaptable generative AI can be and show the need for detection methods that go beyond simple linguistic features.

Another recent study proposed a dual-layer detection framework that combines supervised learning with unsupervised methods to identify phishing emails [11]. The supervised model was trained on labelled data to classify emails as either legitimate or malicious, while the unsupervised component used clustering techniques to detect unusual patterns that may indicate new phishing attacks. This hybrid approach is especially useful for AI-generated phishing, where tactics change quickly and fixed models may not adapt well to new variations.

3.5 Key Contributions and Gaps

The reviewed literature highlights several important findings. First, both machine learning and deep learning methods are effective for phishing email detection, with transformer-based models like BERT showing the best performance overall [1][8][17]. Second, the growing threat of AI-generated phishing is increasingly being recognised, and initial studies have started to explore ways to detect it [3][11][18]. Third, there is increasing agreement that explainability and interpretability are essential for deploying these systems in real-world cybersecurity settings [17].

However, several important gaps still exist in the literature. Most studies use datasets that mainly contain human-written phishing emails, with only a few including AI-generated samples in both training and testing [3]. In addition, datasets differ greatly in size, structure, and quality, making it difficult to compare results across studies or draw consistent conclusions [1]. There is also limited research that directly compares traditional machine learning, deep learning, and transformer-based models specifically for detecting AI-generated phishing emails, rather than general phishing detection. Finally, although explainability is widely recognised as important, only a small number of studies have incorporated Explainable AI (XAI) into their models [17].

3.6 Relation to This Study

This study aims to address the key gaps identified in the literature. It focuses specifically on detecting AI-generated phishing emails and goes beyond studies that rely only on traditional phishing datasets by including AI-generated samples in both training and testing. This allows for a more realistic evaluation of model performance against this emerging threat. The research also compares traditional machine learning models (Random Forest and SVM) with transformer-based models (BERT) using a mixed dataset, providing empirical evidence that is currently limited. In addition, it considers practical deployment factors such as computational cost and the trade-off between accuracy and interpretability.

3.7 Literature Summary

In summary, the literature shows that phishing email detection has developed significantly, moving from simple rule-based systems to advanced deep learning methods. While traditional machine learning approaches are still effective in some cases, they are becoming less reliable

against AI-generated phishing emails, which often lack typical warning signs. Transformer-based models, especially BERT, currently show the best performance, but they have not been widely tested on AI-generated phishing content. Therefore, detecting AI-generated phishing emails remains an important and emerging research area that needs further study. This research addresses this gap by developing and evaluating a machine learning-based framework designed specifically for this problem.

4. Research Design and Methodology

4.1 Research Design

This research uses a quantitative experimental design and follows a Design Science Research (DSR) approach. DSR is well suited for studies that involve designing and evaluating computational artefacts, in this case a machine learning-based framework for phishing email detection [19]. The approach is based on iterative cycles of design, development, and evaluation, which fits the aims of this study, including building machine learning models, testing them on real datasets, and improving them based on performance results.

The experimental design involves training several classification models on a labelled email dataset, evaluating their performance using standard metrics, and comparing the results to determine the most effective method for detecting AI-generated phishing emails. This approach is suitable because the research questions are quantitative, focusing on model performance, detection accuracy, and differences across email categories. An experimental design therefore provides a clear and structured way to address these objectives.

4.2 Research Methods

4.2.1 Overview of Methods

This study uses a quantitative methodology, with supervised machine learning as the main approach for classification. Supervised learning is suitable for this task because it can be framed as a binary or multi-class problem, where the model classifies emails as legitimate, traditional phishing, or AI-generated phishing. The approach relies on labelled training data, sourced from publicly available email datasets and supplemented with AI-generated phishing examples.

4.2.2 Research Strategy

The research strategy is based on controlled experiments. Several machine learning and deep learning models will be implemented and trained under the same conditions, using identical data splits and preprocessing steps to ensure a fair comparison. The models evaluated include Random Forest, Support Vector Machine (SVM), LSTM, and BERT. Random Forest and SVM represent traditional machine learning methods commonly used in phishing detection [6][15]. LSTM is included as a deep learning model that can capture sequence patterns in text [7]. BERT is selected as a transformer-based NLP model that has shown strong performance in recent phishing detection studies [8][17].

4.2.3 System Architecture

The proposed system architecture consists of four main stages. The first stage is data collection and preparation, where emails are gathered from publicly available datasets and AI-generated phishing samples are added. The second stage is preprocessing, which includes cleaning the text, tokenisation, and feature extraction depending on the model. For traditional machine learning models, TF-IDF features are used, while BERT uses its built-in tokenizer to generate embeddings. The third stage is model training and optimisation, where each model is trained on the dataset and tuned using cross-validation. The final stage is evaluation and comparison, where the models are tested on a separate test set and compared using standard performance metrics.

4.2.4 Data Analysis and Evaluation

The performance of each model will be evaluated using accuracy, precision, recall, and F1-score. These metrics are commonly used in phishing detection research and provide a comprehensive assessment of model performance [1][7]. Accuracy measures the overall proportion of correct predictions, precision indicates how many predicted positives are actually correct, recall shows how many actual positives are correctly identified, and F1-score provides a balanced measure of precision and recall. Confusion matrices will also be used to give a detailed view of classification outcomes across different email classes. In addition, statistical significance tests will be applied where appropriate to ensure that performance differences are not due to chance.

4.3 Research Materials

4.3.1 Data Sources and Collection

The primary data for this research will be collected from publicly available email datasets. The main source will be the Kaggle Phishing Email Dataset, which includes a balanced set of phishing and legitimate emails [20]. This dataset is widely used in recent research and provides a reliable basis for training and evaluating the models.

To include AI-generated phishing content, the dataset created by Eze and Shamir [3], which contains 865 AI-generated phishing emails generated using the DeepAI platform, will be used. If further samples are needed, additional AI-generated emails will be produced in a controlled and ethical way using publicly available language model APIs, with safeguards to ensure responsible use and prevent misuse.

The Enron Email Dataset [21] will also be used to provide legitimate email samples, helping to ensure class balance and improve dataset diversity.

These datasets are selected due to their widespread use in research, ease of access, and relevance to the objectives of this study. Combining multiple data sources also allows the models to learn from a wider variety of email writing styles, reducing the risk of overfitting to a single dataset and improving overall generalisability.

4.3.2 Ethical Considerations

Several ethical considerations are taken into account in this study. All datasets used are publicly available and do not contain personally identifiable information (PII), ensuring that individual privacy is protected. Any AI-generated phishing emails created for the research will be produced strictly for academic purposes under controlled conditions and will not be shared or used maliciously. The study will follow the ethical guidelines of the University of Wales Trinity Saint David, and ethical approval will be obtained before data collection begins.

4.3.3 Software and Tools

The implementation will be developed using Python as the main programming language. The scikit-learn library will be used for traditional machine learning models, while TensorFlow and Keras will support the development of deep learning models. The Hugging Face Transformers library will be used to implement the BERT model. Natural language processing tasks such as

tokenisation, stop-word removal, and stemming will be handled using NLTK and spaCy. Data processing and visualisation will be carried out using Pandas, NumPy, Matplotlib, and Seaborn.

All experiments will be conducted on a GPU-enabled system to handle the computational requirements of deep learning and transformer-based models, with Google Colab Pro used as an additional resource if extra GPU capacity is needed.

4.3.4 Hardware Requirements

Due to the computational demands of training deep learning and transformer-based models, a system with a dedicated NVIDIA CUDA-enabled GPU is required. A minimum of 16 GB RAM is recommended, along with at least 8 GB of GPU VRAM to support efficient fine-tuning of BERT. As noted earlier, Google Colab Pro or a similar cloud computing platform will be used if local hardware resources are not sufficient.

5. Expected Outcomes

5.1 Anticipated Results

Based on trends reported in the literature, the BERT-based model is expected to achieve the highest detection accuracy, particularly for AI-generated phishing emails. This aligns with recent studies showing that transformer-based models outperform traditional machine learning and simpler deep learning approaches in text classification tasks [1][8][17]. The LSTM model is expected to perform at a moderate level but may struggle to capture subtle patterns in AI-generated text, which often lacks clear sequential structure. In comparison, Random Forest and SVM are expected to act as strong baseline models; however, they may be less effective when AI-generated phishing emails closely resemble legitimate messages in their feature patterns.

It is also expected that including both traditional phishing emails and AI-generated phishing samples in the training dataset will improve overall model robustness compared to using only traditional phishing data. This mixed-training approach should allow the models to learn a broader range of phishing patterns and reduce the risk of overfitting to features specific to human-written phishing emails.

5.2 Implications

If the expected results are confirmed, the study will have several important implications. From an academic perspective, it will provide empirical evidence on the performance of different machine learning approaches for detecting AI-generated phishing emails, addressing a research area where comparative studies are still limited. From a practical perspective, the findings could contribute to the development of more effective email filtering systems capable of detecting AI-generated threats. The results may also highlight the importance of including AI-generated samples in training datasets as a standard practice in future phishing detection research.

5.3 Limitations

Several limitations of this study should be acknowledged. First, the quality and representativeness of the AI-generated phishing emails may affect the validity of the results. If these samples do not fully reflect the diversity of real-world AI-generated phishing, the models may have limited generalisability beyond the experimental setting. Second, the computational requirements of transformer-based models may limit the extent of hyperparameter tuning and experimentation that can be carried out within the available timeframe. Third, the evaluation is conducted in an offline environment, which does not fully represent real-world deployment challenges such as adversarial adaptation, concept drift, and real-time detection requirements. Finally, the study focuses only on email text content and does not include other phishing indicators such as sender reputation, header information, or embedded URLs, which could further improve detection performance in practical systems.

6. Timeline

The following table presents the planned timeline for the completion of this research project, spanning approximately 10 weeks from May to August 2026.

Phase	Tasks	Duration	Weeks
Phase 1	Literature Review and Dataset Collection	2 weeks	Weeks 1–2
Phase 2	Data Preprocessing and Feature Engineering	1 weeks	Weeks 3

Phase 3	Model Implementation (RF, SVM)	1 weeks	Weeks 4
Phase 4	Model Implementation (LSTM, BERT)	2 weeks	Weeks 5–6
Phase 5	Model Evaluation and Comparison	1 weeks	Weeks 7
Phase 6	Analysis and Discussion of Results	2 weeks	Weeks 8–9
Phase 7	Report Writing and Final Submission	1 weeks	Weeks 10

Each phase includes built-in contingency time to accommodate unforeseen delays. Regular progress meetings with the supervisor will take place fortnightly to ensure the project remains on track and any issues are addressed promptly.

7. References

- [1] F. Alharbi et al., "In-Depth Analysis of Phishing Email Detection: Evaluating the Performance of Machine Learning and Deep Learning Models Across Multiple Datasets," *Applied Sciences*, vol. 15, no. 6, p. 3396, 2025.
- [2] Verizon, "2024 Data Breach Investigations Report," Verizon Business, 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] Q. Qi, Y. Luo, Y. Xu, W. Guo, and Y. Fang, "SpearBot: Leveraging Large Language Models in a Generative-Critique Framework for Spear-Phishing Email Generation," *arXiv preprint arXiv:2403.02985*, 2024.
- [5] KnowBe4, "2025 Phishing Threat Trends Report," KnowBe4, 2025. [Online]. Available: <https://www.knowbe4.com/>.
- [6] S. Bagui, D. Nandi, S. Bagui, and R. J. White, "Machine Learning and Deep Learning for Phishing Email Classification Using One-Hot Encoding," *Journal of Computer Science*, vol. 17, pp. 610–623, 2021.
- [7] S. Atawneh and H. Aljehani, "Phishing Email Detection Model Using Deep Learning," *Electronics*, vol. 12, no. 20, p. 4261, 2023.
- [8] R. Meléndez, M. Ptaszynski, and F. Masui, "Comparative Investigation of Traditional Machine-Learning Models and Transformer Models for Phishing Email Detection," *Electronics*, vol. 13, p. 4877, 2024.
- [9] A. Bergholz, J. De Beer, S. Glahn, M. F. Moens, G. Paaß, and S. Strobel, "New Filtering Approaches for Phishing Email," *Journal of Computer Security*, vol. 18, no. 1, pp. 7–35, 2010.
- [10] N. Altwaijry, I. Al-Turaiki, R. Alotaibi, and F. Alakeel, "Advancing Phishing Email Detection: A Comparative Study of Deep Learning Models," *Sensors*, vol. 24, no. 7, p. 2077, 2024.
- [11] M. Demir, R. Sheridan, M. Yildirim, and A. Gorur, "Machine Learning and Watermarking for Accurate Detection of AI-Generated Phishing Emails," *Electronics*, vol. 14, no. 13, p. 2611, 2025.
- [12] V. Shahrivari, M. M. Darabi, and M. Izadi, "Phishing Detection Using Machine Learning Techniques," *arXiv preprint arXiv:2009.11116*, 2020.
- [13] A. Alhogail and A. Alsabih, "Applying Machine Learning and Natural Language Processing to Detect Phishing Email," *Computers & Security*, vol. 110, p. 102414, 2021.
- [14] H. Fares, N. Mouakkal, Y. Baddi, and N. Hajraoui, "Robust Email Phishing Detection Using Machine Learning and Deep Learning Approach," *International Journal of Communication Networks and Information Security*, vol. 16, no. 1, pp. 19–32, 2024.
- [15] "Machine Learning Approach for Email Phishing Detection," *Procedia Computer Science*, 15th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN), 2024.
- [16] P. H. Kyaw, J. Gutierrez, and A. Ghobakhlou, "A Systematic Review of Deep Learning Techniques for Phishing Email Detection," *Electronics*, vol. 13, p. 3823, 2024.
- [17] M. A. Uddin and I. H. Sarker, "An Explainable Transformer-based Model for Phishing Email Detection: A Large Language Model Approach," *arXiv preprint arXiv:2402.13871*, 2024.

- [18] C. Opara, P. Modesti, and L. Golightly, "Evaluating Spam Filters and Stylometric Detection of AI-Generated Phishing Emails," *Expert Systems with Applications*, vol. 276, p. 127044, 2025.
- [19] 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.
- [20] S. Chakraborty, "Phishing Email Detection Dataset," Kaggle, 2023. [Online]. Available: <https://www.kaggle.com/datasets/subhajournal/phishingemails>.
- [21] B. Klimt and Y. Yang, "The Enron Corpus: A New Dataset for Email Classification Research," in *Proc. European Conference on Machine Learning*, 2004, pp. 217–226.
- [22] A. I. Champa, M. F. Rabbi, and M. F. Zibran, "Curated Datasets and Feature Analysis for Phishing Email Detection with Machine Learning," in *2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI)*, IEEE, 2024.
- [23] A. Gaurav, B. B. Gupta, A. Castiglione, S. Bansal, and K. T. Chui, "Optimized Deep Learning Based Phishing Email Detection Using BERT and Hill Climbing Algorithm," in *Computational Data and Social Networks (CSoNet 2024)*, *Lecture Notes in Computer Science*, vol. 15417, Springer, 2025.
- [24] Y. Fang, C. Zhang, C. Huang, L. Liu, and Y. Yang, "Phishing Email Detection Using Improved RCNN Model with Multilevel Vectors and Attention Mechanism," *IEEE Access*, vol. 7, pp. 56329–56340, 2019.
- [25] S. Jamal and H. Wimmer, "An Improved Transformer-based Model for Detecting Phishing, Spam, and Ham: A Large Language Model Approach," 2024.