A Transformer Model Approach for Robust Phishing Detection

by Glorina Pendie

M.S. Cybersecurity Operations & Leadership, August 2020, University of San Diego
Master's in Legal Studies, May 2018, Washington University School of Law, St. Louis MO
B.A. in Political Science & International Relations, May 2016, St. Catherine University

A Praxis was submitted to

The Faculty of
The School of Engineering and Applied Science
of The George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Engineering

May 19, 2024

Praxis directed by

Mehdi Hashemipour
Professorial Lecturer of Engineering Management and Systems Engineering

The School of Engineering and Applied Science of The George Washington University certifies that Glorina Bola Pendie has passed the Final Examination for the degree of Doctor of Engineering as of April 3, 2024. This is the final and approved form of the Praxis.

A Transformer Model Approach for Robust Phishing Detection

Glorina Bola Pendie

Praxis Research Committee:

Mehdi Hashemipour, Professorial Lecturer of Engineering and Applied Science, Praxis Director

John Fossaceca, Professorial Lecturer of Engineering and Applied Science, Committee Chair

Salah Hassan, Professorial Lecturer of Engineering and Applied Science, Committee Member

© Copyright 2024 by Glorina Bola Pendie
All rights reserved

Dedication

I want to dedicate this to my mother Florence Bassa. She has supported me throughout my educational journey, allowing me to continuously explore my interest. My grandparents were uneducated and having the opportunity to further my education to this level is where I pull my courage, strength, and tenacity to always try. I would not be where I am today without acknowledging my immigrant roots from Cameroon, and the unwavering support from my mother.

Acknowledgements

I want to begin by giving thanks to God the Almighty for the blessing that is my life. The challenges experienced in life are not often battles we can overcome ourselves, but possible with the help and support of others. In my life, God has always given me grace and strength when the mere possibilities seemed unattainable.

I would also like to give a major thank you to my advisor, Dr. Mehdi Hashemipour. His direction, encouragement, and mentorship through the completion of this praxis was everything I needed. Thank you, professor Dr. Mehdi.

Furthermore, to my support system: family, husband, friends please accept my sincerest gratitude for your steadfast support and patience. Lastly, I want to thank my colleagues of this program for the endless collaboration shared throughout this program.

Abstract of Praxis

A Transformer Model Approach for Robust Phishing Detection

This research uses a quantitative cross-sectional study to examine conventional models and traditional machine learning algorithm models in detecting email phishing.

The escalating threat of phishing, exemplified by the staggering volume of 3.4 billion daily phishing emails in 2022 (Baki & Verma, 2022), poses a significant risk to individuals and businesses, resulting in an alarming $8.8 billion in losses according to the Federal Trade Commission. In response, various deep learning models have emerged to enhance threat detection. However, a critical research gap exists in exploring the benefits of generative pre-training transformer models in email phishing detection. Catching these emails is crucial, but existing methods have limitations. Traditional models are effective on their own but using Natural Language Processing (NLP) methods such as transformers (BERT, GPT-2) in ensemble models may be even better. Characterizing and investigating the performance of email phishing detection ensemble models with transformers is an essential exploration when combating email phishing attacks.

In response to the negative impacts of email phishing, deep learning-based approaches developed to improve phishing email detection (Mughaid et al., 2022). Among, deep learning-based approaches, LSTM and GRU offer “long-term dependencies” and the “ability to handle sequential data” (Do et al., 2022; Halgaš et al., 2020; Sun et al., 2021). Multimodel approaches have also been introduced as a means of mitigating limitations identified in different deep-learning models (e.g. Roy et al., 2022). However, research on ensembled model approach using generative pre-training

transformers-based models is needed to understand the characteristics of each approach and their associated outcomes. Research limitations do not explore the possibility of ensemble transformer models.

Table of Contents

Dedication ..... **iv**

Acknowledgements ..... **v**

Abstract of Praxis ..... **vi**

List of Figures ..... **xi**

List of Tables ..... **xii**

List of Acronyms ..... **xiii**

Chapter 1 —Introduction **1**

1.1 Background 1

1.2 Research Motivation 3

1.3 Problem Statement 7

1.4 Thesis Statement 8

1.5 Research Objectives 9

1.6 Research Questions and Hypotheses 10

1.6.1 Research Questions 10

1.6.2 Research Hypotheses 10

1.7 Nature of the Study 11

1.8 Scope of Research 12

1.9 Research Limitations 13

1.10 Organization of Praxis 13

Chapter 2 —Literature Review **15**

2.1 Introduction 15

2.2 Literature Review Search Strategy 16
2.3 Organization of the Remainder of the Chapter 16
2.4 Conceptual Framework 17
2.5 Review of Relevant Literature 21
2.5.1 Email Phishing 21
2.5.2 Machine Learning and Phishing Threats 26
2.5.3 Deep Learning Neural Networks 30
2.5.4 Long Short-Term Memory Networks 32
2.5.5 Bidirectional Encoder Representations from Transformers 34
2.5.6 Generative Pre-trained Transformer 34
2.6 Summary and Conclusion 35
Chapter 3 —Methodology 37
3.1 Introduction 37
3.2 Research Design and Rationale 37
3.3 Methodology 40
3.3.1 Population 41
3.3.2 Sampling Strategy 42
3.3.3 Procedure for Recruitment, Participation and Data Collection 42
3.3.4 Data Collection 43
3.3.5 Data Analysis Plan 46
3.4 Threats to Validity 48
3.5 Ethical Procedures 49
3.6 Summary 50

Chapter 4 —Results 51

4.1 Introduction 51

4.2 Results 52

4.3 Base Models 55

4.4 Statistical Test – Analysis of Model Performance Metrics 61

4.4.1 Training Loss 61

4.4.2 Training Accuracy 62

4.4.3 Validation Loss 63

4.4.4 Validation Accuracy 64

4.5 Summary 66

Chapter 5 —Discussion and Conclusions 67

5.1 Discussion 67

5.2 Conclusions 68

5.3 Contributions to Body of Knowledge 68

5.4 5.4 Recommendations for Future Research 69

References 70

Appendix A 90

List of Figures

Figure 1. Research Methodology 46

Figure 2. Phishing and Spam Emails Count 53

Figure 3. Common words in Phishing Emails 54

Figure 4. Common Words in Spam Emails 54

Figure 5. SHAP Conventional Model 59

Figure 6. SHAP BERT Model 59

Figure 7. SHAP GPT-2 Model 60

List of Tables

Table 1. Random Forest Classifier Performance ReportI 55

Table 2. Logistic Liner Regression Classification Report 56

Table 3. Conventional Models 57

Table 4. Transformer-Based Models 57

Table 5. Generative Pre-Training(GPT-2) Transformers-Based Models 58

Table 6. Summary of Research Questions and Hypothesis 60

List of Acronyms

ANOVA	Analysis of Variance
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
DL	Deep Learning
DHS	Department of Homeland Security
DoD	Department of Defense
GPT-2	Generative Pretrained Transformers
GRU	Gated Recurrent Unit
HSD	Honestly Significant Difference
FTC	Federal Trade Commission
LLM	Large Language Model
LSTM	Long Short-Term Memory
ML	Machine Learning
NLP	Natural Language Processing
PII	Personal Identifiable Information
RF	Random Forest
RNN	Recurrent Neural Network
TAM	Technology Acceptance Model

Chapter 1—Introduction

1.1 Background

Humans account for the most significant security threat we face. Social Engineering is using manipulative or deceptive tactics to gain confidential information from a person that could potentially be used for fraudulent behavior. For example, Uber experienced a data breach in 2022 where the "attacker managed to gain access to the organization's internal IT systems, email dashboard, Slack server, endpoints, Windows domain and Amazon Web Services console" (Keary, 2022) through social engineering tactics. Data can be leveraged to find trends that explain human behavior relative to technology and social engineering. Social engineering depends on manipulating human emotions, which is controversial. Unsuspecting users are frequently duped into clicking on email links, unaware that attackers are waiting to deceive them into disclosing personal information such as birth dates, social security numbers, and bank account information, allowing the attackers to infect the user's computer by downloading malware via attachments clicked by the unsuspecting user. These approaches are known as email phishing. Kevin Mitnick gained notoriety in the 1980s and 1990s for his high-profile hacking activities, which often involved exploiting human vulnerabilities through social engineering tactics such as pretexting (creating false identities or scenarios to manipulate targets) and impersonation focused on creating the Social Engineering Attack Framework. The need to address these assaults is a chronic issue that researchers have worked to solve and will continue to do so until the risk is eliminated. Email has had a significant impact on the way we communicate with each other in the contemporary world, both personally and

professionally. Inopportunely, email is often exploited for phishing attacks (Jampen et al., 2020). “Phishing attacks have recently become among the most prominent social engineering attacks faced by businesses, governments, and public internet users” (Mughaid et al., 2022). According to Lain et al. (2022), email phishing is a type of social engineering attack focused on drawing sensitive data such as bank information, bank account numbers, and credit card details via a fraudulent solicitation on email, where the attacker masquerades as a reputable individual or trustworthy businessperson. It is a common cyber-attack that tricks people into clicking on malicious links or providing personal information (Lain et al., 2022).

Phishing has multiple adverse impacts on individuals and organizations, such as loss of money, disruption of operational activities, damage to reputation, and loss of intellectual property (Baki & Verma, 2022; Kwak et al., 2020). The FTC (2023) reported that in 2022, email phishing led to over $8.8 billion in loss for clients and businesses, an increase of over 30% from 2021. Email phishing attacks should thus be tracked, identified, and eliminated in response to the negative impacts. The proposed Praxis project will recommend a transformer model approach that improves the performance of email phishing detection.

The escalating number of email phishing attacks has resulted in the development of multiple techniques for detecting these attacks. Email phishing detection based on machine learning has gained incredible interest and attention among schools in the cybersecurity field (Do et al., 2022; Mughaid et al., 2022). Multiple recent studies have been completed on implementing machine learning in diverse solutions to detect the

rapidly increasing phishing threats (Do et al., 2022; Jampen et al., 2020; Mughaid et al., 2022). Deep learning algorithms, which comprise a subset of machine learning, have lately emerged as a possible alternative to conventional machine learning models (Do et al., 2022; Roy et al., 2022). Nonetheless, limited studies have currently comprehensively focused on adopting deep learning models in email phishing recognition, their merits and demerits, present issues, and future study guidelines to address the prevailing issues (Do et al., 2022). According to Jena et al. (2021), deep learning neural networks are applied to mimic the behavior of the human brain, although far from matching its capability to learn from enormous quantities of data. Deep learning drives multiple artificial intelligence (AI) services and applications, enhancing automation and the performance of physical and analytical functions with minimal human involvement (Jena et al., 2021). Accordingly, deep learning models are superior to other email phishing-detecting approaches, including those based on machine learning.

Given the growing threat of email phishing attempts, there is an urgent need for a transformer-based model approach for robust phishing detection. This study aims to fill the existing gap in understanding the effectiveness of existing models and their implications for mitigating losses resulting from successful email phishing attacks, utilizing quantitative data to inform the identification of the most effective approaches under varying conditions.

1.2 Research Motivation

Email phishing has detrimental impacts on individuals and organizations. Examples of its adverse impacts include loss of money, disruption of operational activities, damage

to reputation, and loss of intellectual property (Baki & Verma, 2022). Research and statistics show that email phishing attacks are steadily increasing, making individuals and organizations lose personal information, assets, and valuables to unsanctioned persons (Do et al., 2022; (FTC, 2023), Roy et al., 2022). The key motivation of this Praxis project is to help eradicate this growing threat through detection. The core outcome measure will be improved performance in email phishing detection by adopting a transformer model approach for robust phishing detection. The achievement of this core outcome will be measured by evaluating and comparing the impact of applying ensemble models.

Due to the growing threat, an efficient and precise approach is necessary for identifying and detecting malicious emails to deter the acknowledged adverse impacts of phishing on individuals and entities. According to Mughaid et al. (2022), more effective phishing detection technology is required to curb the threat of phishing emails, which have been increasing at an alarming rate recently. The other primary motivation is to leverage new technologies to help dissuade the growing threat of email phishing. The swiftly emerging deep learning algorithms, such as LSTM (Long Short-Term Memory) and CNN (Convolutional Neural Network), have great potential in email phishing detection.

According to Li et al. (2022), “deep learning algorithms have been applied” in the real world in financial services, healthcare, customer service, and law enforcement sectors. Of interest “is the” application “of” these “deep learning algorithms in law enforcement” (Catal et al., 2022). As described by Catal et al. (2022), deep learning algorithms are used in law enforcement to evaluate and learn from transactional

information to recognize precarious trends that signal probable criminal or fraudulent activity. Qin et al. (2020) add that deep "learning applications such as computer vision and speech recognition" enhance the effectiveness and efficiency of investigative evaluation by extracting trends and evidence from documents, images, video recordings, and sound, enabling law enforcement agencies to appraise large quantities of data more quickly and accurately.

Traditional machine learning models include algorithms such as logistic regression based on conventional statistical approaches. These methods do not typically involve deep learning architectures like transformers. "Machine learning (ML) is a branch of artificial intelligence (AI) that employs the method of data mining to discover new or existing patterns (or features) from a dataset, which is then used for classification" (Clarke et al., 2022). "Random forest (RF) is an ensemble learning classification and regression method suitable for handling problems involving grouping of data into classes" (Nar & Arslankaya, 2022). The algorithm was developed by Breiman and Cutler (2007). "In RF, prediction is achieved using decision trees" (Akinyelu & Adewumi, 2014).

> *During the training phase, several decision trees are constructed (as defined by the programmer), which are then used for the class prediction; this is achieved by considering the voted classes of all the individual trees, and the class with the highest vote is considered to be the output (Akinyelu & Adewumi, 2014).*

Advanced deep learning models are transformer-based models. Examples of transform-based models would be Generative Pretrained Transformer Model 2 (GPT-2), OpenAI, and Bidirectional Encoder Representations from Transformers (BERT) models founded by Google. The BERT model "uses encoders in a transformer as a sub-structure to pre-training models for NLP tasks" (Devlin et al., 2019). "BERT," a transformer-based deep learning model, "is designed to pre-train deep bidirectional representations from the unlabeled text by joint conditioning on both left and right context in all layers" (Melnyk & Feld, 2022). "The bi-directionality capability of BERT allows it to read input from right to left or left to right simultaneously" (Otieno, Namin, & Jones, 2023). Another advantage of BERT is that the model does not need to be retrained; it can be fine-tuned with "one additional output layer to create state-of-the-art models for a wide range of tasks such as" (Devlin et al., 2019). Pre-processed training data was utilized using BERT to create input masks, type IDs, and word IDs for categorizing emails for phishing and non-phishing.

OpenAI launched GPT-2, a language model and transformer trained on 1.5 billion parameters (Better Language Models and Their Implications). It can perform various NLP tasks such as text generation, summarization, and more (Khan et al., 2021). Providing text to the model will predict the following word (Khan et al., 2021). "GPT-2 is the successor of GPT-2 (released previously), which is trained on 10X data and parameters" (Khan et al., 2021). "This text-generating model can be fine-tuned to gain the necessary results, so there exists a threat of using GPT -2 for malicious use" (Khan et al., 2021). "It has many versions depending on the size of the parameters it has been

trained on; the small and medium versions are published, while the other versions are not currently available publicly" (Khan et al., 2021). This research aims to show the benefit of using the ensemble GPT-2 model approach for robust phishing detection.

1.3 Problem Statement

The problem to be addressed through this Praxis project is that email phishing attacks are gradually increasing, making public internet users, businesses, and governments lose confidential information, assets, and valuables to unsanctioned persons (Do et al., 2022; Mughaid et al., 2022; Roy et al., 2022; Shahbaznezhad et al., 2021). Phishing is a common cyber-attack. Mughaid et al. (2022) state that "phishing attacks have become the most prominent social engineering attacks faced by public internet users, governments, and businesses." In 2022, an estimated "3.4 billion phishing emails were sent daily" (Baki & Verma, 2022). Sun et al. (2021) add that with the increasingly complex phishing campaigns in recent years, phishing emails lure individuals and organizations using more authentic-appearing perspectives, making more people vulnerable to this cyber-attack.

The challenge with these common attacks is that, according to the FTC (2023), email phishing resulted in over $8.8 billion in loss for consumers and businesses in 2022, an increase of over 30% from 2021. Email phishing attacks should, therefore, be tracked, detected, and eliminated in response to such negative impacts. Conventional mechanisms of detecting email phishing attacks have yet to be effective and efficient enough to help eradicate this cybercrime. An efficient and accurate approach is necessary for identifying and detecting malicious emails to deter the acknowledged adverse impacts of phishing on individuals and organizations. Mughaid et al. (2022) identified that more effective

phishing detection technology is required to curb the threat of phishing emails, which have been increasing at an alarming rate recently.

1.4 Thesis Statement

There is a need to leverage the capabilities of transformer models in combating the persistent threat of email phishing. This approach enhances accuracy and adaptability through in-depth analysis of email content, structure, and sender information compared to traditional methods, thus addressing this growing security threat more effectively.

There is an undeniable need to refine existing methodologies by incorporating open-source, moderately scaled Large Language Model (LLM) architectures, such as GPT-2LSTM-CNN, to synthesize textual data effectively. This is crucial for rectifying the imbalances commonly observed in cybersecurity datasets and enhancing text embeddings to optimize the performance of Machine Learning (ML) models when classifying email phishing.

The primary objective of this research focuses on the exploration and capabilities of GPT-2LSTM-CNN models in generating language and synthesizing data to address dataset imbalances. Moreover, the research aims to leverage the long-range dependency modeling of LSTMs and the local feature extraction capabilities of CNNs to optimize text embeddings derived from the dataset. These objectives remain relevant across both multi-class and binary datasets, providing adaptability to various classification tasks. To evaluate these objectives effectively, the research meticulously examines the performance of GPT-2LSTM-CNN models when coupled with ML classifiers, systematically

analyzing model classification performance across diverse experimental factors and levels.

An integral aspect of this research is its practical applicability, especially in environments constrained by computational resources or limited access to proprietary APIs. The methodology involves two main steps, supported by the development of specialized Python programs tailored to facilitate the achievement of research objectives.

This Praxis conducted this research by addressing specific challenges encountered in text classification for cybersecurity applications. By leveraging the long-range dependency modeling of LSTMs and the local feature extraction capabilities of CNNs within the GPT-2LSTM-CNN architecture, the research aims to synthesize imbalanced and unstructured textual datasets effectively. Additionally, this research explores the potential of GPT-2LSTM-CNN models in optimizing text embeddings, thereby enhancing the performance of ML classifiers.

1.5 Research Objectives

Enhance GPT-2 based Model Performance for Robust Email Phishing Detection by focusing on improving the ability of GPT-2 based models to detect phishing emails. The specific approach to investigate the use of fine-tuned GPT-2 models, would be testing these following ensemble models: (GPT-2, GPT-2+LSTM, GPT-2+CNN, GPT-2+LSTM+CNN). This allows for realistic synthetization of phishing emails to address imbalanced datasets. The transformer model would be able to extract informative text features for improved classification using machine learning models. Lastly, the evaluation of this transformer model will develop and apply a novel metric that considers

the risk associated with misclassifying different email types (e.g., prioritizing accurate detection of phishing emails).

1.6 Research Questions and Hypotheses

The project will be guided by the following research questions and hypotheses.

1.6.1 Research Questions

The following research questions and hypothesis guided this study utilizing quantitative methods and a cross-sectional design.

RQ1. What is the performance of base model algorithms (random forest classifier and logistic linear regression) in email phishing detection?

RQ2. What is the performance of the proposed models (linear layers' models, transformer-based models, generative pre-training transformer models) in email phishing detection?

RQ3. How does the performance of base models compare to the proposed models?

RQ4. What are the key factors influencing the effectiveness of different machine learning models in email phishing detection?

1.6.2 Research Hypotheses

$H1_0$. There is no statistically significant difference in performance between the random forest classifier and logistic regression algorithms for detecting email phishing.

$H1_a$. The performance of the random forest classifier and logistic regression algorithms for detecting email phishing is significantly different.

$H2_0$. There is no significant relationship between the performance of proposed models (linear layers' models, transformer-based models, generative pre-training transformer models) and email phishing detection.

$H2_a$. Proposed models (linear layers' models, transformer-based models, generative pre-training transformer models) perform differently for email phishing detection.

$H3_0$. There is no statistically significant difference in the performance of base models (random forest classifier and logistic regression) and proposed models (linear layers' models, transformer-based models, generative pre-training transformer models) in email phishing detection.

$H3_a$. There is a statistically significant difference in the performance of base models (random forest classifier and logistic regression) and proposed models (linear layers' models, transformer-based models, generative pre-training transformer models) in email phishing detection.

$H4_0$. There are different factors that significantly impact the effectiveness of various machine learning models used for email phishing detection.

$H4_a$. Various factors such as feature sets, model architectures, and hyperparameters significantly influence the effectiveness of different machine learning models utilized for email phishing detection.

1.7 Nature of the Study

Quantitative research relies on numerical data and statistical analysis, which provide objective and measurable results. According to Taherdoost (2022), quantitative methods use empirical and descriptive statements to give insight and meaning to the phenomenon

being studied. Li et al. (2020) noted that quantitative methods focus on quantifying casual relationships of the studied variables. By quantifying the performance of the algorithms using appropriate metrics, the study can provide clear and precise evaluations of their effectiveness in detecting email phishing. This objectivity enhances the credibility and reliability of the study's findings. Quantitative research aims to ascertain the relationship between a population's independent and dependent variables (Kandel, 2020). This research employs a quantitative approach, utilizing a structural equation model as the analytical framework.

1.8 Scope of Research

The scope of this research is centered around evaluating and comparing various machine learning approaches for email phishing detection. Specifically, the study will investigate the performance of:

1. Base Models: Random Forest Classifier and Logistic Regression Model
2. Conventional Linear layers Models: LSTM, CNN, LSTM+CNN
3. Transformer-based models: BERT, BERT+LSTM, BERT+CNN, BERT+LSTM+CNN
4. Generative pre-training transformers-based models: GPT-2, GPT-2+LSTM, GPT-2+CNN, GPT-2+LSTM+CNN

Additionally, the study will integrate SHAP (SHapley Additive exPlanations) to conventional linear layer models, transformer-based models, and generative pre-training transformer-based models to interpret their predictions and understand the factors influencing model decisions.

The research will focus on assessing the performance, accuracy, and interpretability of these machine learning models in detecting email phishing attempts. It will involve comprehensive experimentation and analysis to evaluate the strengths and limitations of each model architecture. The goal is to provide a transformer model approach effective for email phishing detection, contributing to the development of more robust cybersecurity solutions.

1.9 Research Limitations

The main research restriction encountered was limited resources. A massive amount of processing power was required to fine-tune the transformer models. RAM limit warnings were received while using Google Colab. Google Colab Pro+ was utilized, which included 500 computing units. It still took over 24 minutes to execute each transformer model with lower epochs. Furthermore, even with these resources, data needed to be limited to handle the batch size to prevent running out of GPU RAM.

1.10 Organization of Praxis

Email has become a very important form of communication in the modern world, serving as a channel for private and business conversations. The problem to be addressed through this Praxis project is that email phishing attacks are gradually increasing, making public internet users, businesses, and governments lose confidential information, assets, and valuables to unsanctioned persons (Do et al., 2022; Mughaid et al., 2022; Roy et al., 2022; Shahbaznezhad et al., 2021). This Praxis project is categorized into five chapters for enhanced flow and dissemination of information. Chapter 1 covers the research background, research motivation and problem, project thesis, research objectives,

research questions and hypotheses, the scope of the project, and research limitations. Chapter 2 presents the conceptual framework. The conceptual framework covers the key concepts of the research in providing a transformer model approach for email phishing. Chapter 3 describes the research methodology, design, and procedures to be applied to achieve the specified project objectives. The proposed project will draw evidence quantitative cross-sectional study. Chapter 4 will detail the findings from the data analysis conducted. These findings will be categorized based on the stated research questions and hypotheses. Chapter 5 will present a discussion of the findings conforming to the research approaches specified in the previous chapter. Moreover, Chapter 5 will include the project's main contributions and conclusions, recommendations, and direction for future research.

Chapter 2—Literature Review

2.1 Introduction

This Praxis project aims to address the increasing threat of email phishing attacks, which have resulted in the loss of confidential information, assets, and valuables for public internet users, businesses, and governments (Do et al., 2022; Mughaid et al., 2022; Roy et al., 2022; Shahbaznezhad et al., 2021). Phishing attacks are a prevalent cybercrime that deceives individuals into clicking on malicious links or divulging personal information. Recent studies have highlighted that phishing attacks have become one of the most prominent social engineering threats faced by various entities, including businesses, governments, and public internet users (Barlow et al., 2020; Hijji & Alam, 2021; Mughaid et al., 2022). Around 3.4 billion phishing emails were estimated to be sent daily in 2022 alone (Baki & Verma, 2022; Nifakos et al., 2021).

Phishing campaigns have become more sophisticated over the years, employing deceptive techniques that make it increasingly challenging for individuals and organizations to identify fraudulent emails (Sun et al., 2021). As a result, more people are falling victim to these cyber-attacks. The repercussions of email phishing are significant, with the Federal Trade Commission reporting that in 2022, consumers and businesses suffered losses exceeding $8.8 billion, representing a 30% increase from the previous year (FTC, 2023). To mitigate these detrimental effects, developing effective and efficient mechanisms for detecting and eliminating email phishing attacks is crucial.

Traditional methods of detecting phishing emails have proven insufficient in combating this type of cybercrime. Therefore, there is a pressing need for more advanced

and accurate approaches to identify and detect malicious emails, thereby safeguarding individuals and organizations from the harmful impacts of phishing (Mughaid et al., 2022).

2.2 Literature Review Search Strategy

Several databases were used to search for relevant literature on the research topic. The databases used to locate the literature included PlosOne, PubMed, Scielo, Scopus, Web of Science, and Google Scholar OECD Library, IEEE, and ProQuest Engineering Database. The keywords that were used to search databases included: *email phishing, long short-term memory networks, CNN, GPT-2, BERT, phishing attacks, social engineering attacks, cyber-attacks, deep learning, deep learning-centered approaches, cybersecurity, and deep learning algorithms*. Once relevant articles were identified, the studies were further refined to include only articles published between 2019 and 2023. After the older articles were filtered out, the remaining articles were further refined to only include ones that had the full text.

2.3 Organization of the Remainder of the Chapter

The remainder of this chapter will cover various topics concerning the study. The conceptual framework used to guide this study will be discussed first. The framework covers the concepts of deep learning algorithms, LSTM, GRU, and multimodel approaches and how these approaches are connected to phishing emails. The literature review portion of the chapter will cover topics and studies related to email phishing, machine learning and phishing threats, and networks related to deep learning neural

networks. The chapter will conclude with a summary of the literature reviewed in this chapter.

2.4 Conceptual Framework

This research leverages various machine learning models, along with interpretability techniques (SHAP), to evaluate their effectiveness in detecting phishing emails. An ensemble transformer model approach is used, incorporating different metrics and statistical comparisons is employed to identify the most suitable models for this task. By acknowledging the limitations and potential challenges, the research aims to contribute to the development of robust email phishing detection systems.

In the realm of information systems and technology adoption, the Technology Acceptance Model (TAM) is a commonly used paradigm that explains how people accept and use technology. Although TAM is commonly used to study how users react to technology, it may also be modified to study how algorithms for email phishing detection are received and how well they work (Lemay et al., 2020). Fred Davis created the idea in 1989 with the intention of measuring system usage and forecasting. Perceived usefulness and perceived ease of use are the two key components that make up the TAM.

Deep learning algorithms are a subset of machine learning algorithms designed to mimic the functioning of the human brain's neural networks (Shrestha & Mahmood, 2019). These algorithms are particularly effective in handling complex and large-scale data, making them suitable for tackling sophisticated cyber threats like email phishing (Adebowale et al., 2023). LSTM networks are a type of recurrent neural network (RNN) that can effectively model and capture long-term dependencies in sequential data

(Staudemeyer & Morris, 2019). In the context of email phishing detection, LSTM networks can analyze the sequential nature of email content, including the order and context of words or characters, to identify patterns and indicators of phishing attempts (Rahman & Ullah, 2020). GRU is another type of recurrent neural network that, similar to LSTM, can handle sequential data effectively (Mani et al., 2023). GRU networks can also capture long-term dependencies with a simpler architecture than LSTM (Brindha et al., 2023). GRU networks have been applied in email phishing detection tasks to analyze and process the sequential patterns of phishing emails. Multimodal approaches involve integrating multiple data sources to improve a model's accuracy and robustness (Ghose et al., 2019). In the context of email phishing detection, multimodal approaches combine different features or representations of emails, such as text content, metadata, header information, or visual elements, to create a more comprehensive and holistic understanding of the email content. Integrating multiple modalities enhances the model's ability to detect phishing emails accurately. Email phishing refers to the fraudulent practice of sending deceptive emails to trick recipients into revealing sensitive information, clicking on malicious links, or taking actions that can compromise their security (Gangavarapu et al., 2020). Email phishing attacks exploit human vulnerabilities and social engineering techniques to deceive individuals or organizations (Alhogail & Alsabih, 2021).

The conceptual framework outlined in this research explores the comparative analysis of various machine learning models for email phishing detection, incorporating

interpretability techniques for better understanding. Here is a breakdown of the key elements:

1. Model Categories:

- **Base Models:**
    - **Random Forest Classifier:** Makes predictions based on a majority vote from an ensemble of decision trees, offering robustness to noise and overfitting.
    - **Logistic Regression Model:** Estimates the probability of an email belonging to the phishing class based on a linear relationship between features and the outcome.
- **Conventional Models:**
    - **Linear Layers with LSTMs and CNNs:**
        - **LSTMs (Long Short-Term Memory):** Capture long-term dependencies within sequential data like email text, crucial for understanding context.
        - **CNNs (Convolutional Neural Networks):** Extract local features from the data, potentially identifying relevant keywords or patterns.
    - **LSTM + CNN Combination:** Leverages the strengths of both architectures, potentially improving feature extraction and sequence analysis for better classification.
- **Transformer-based Models:**

- **BERT (Bidirectional Encoder Representations from Transformers):** Pre-trained model capable of understanding the deeper meaning and relationships within text, offering superior context analysis compared to traditional models.
- **BERT with LSTM/CNN Combinations:** Integrates the contextual understanding of BERT with the feature extraction capabilities of LSTMs or CNNs, potentially leading to more robust phishing detection.

- **Generative Pre-training Transformer Models:**
  - **GPT-2 (Generative Pre-trained Transformer):** Similar to BERT but trained with a focus on generating realistic text.
  - **GPT-2 with LSTM/CNN Combinations:** Aims to leverage the strengths of GPT-2 in text generation and feature extraction capabilities of LSTMs/CNNs for potential data augmentation and improved phishing detection.

2. Interpretability (SHAP):

- **SHAP (SHapley Additive exPlanations):** This technique is applied to **CNN, BERT and GPT-2** to:
  - **Gain insights into model predictions:** Understand the reasoning behind each model's classification (phishing vs. legitimate).
  - **Identify influential features:** Determine the specific email features (e.g., sender address, keywords, suspicious links) that each model relies on most heavily for making the classification.

2.5 Review of Relevant Literature

2.5.1 Email Phishing

Email phishing is an insidious and ever-present threat in the digital landscape, posing severe risks to individuals, businesses, and organizations (Lin et al., 2019). This cyber-attack involves attackers employing sophisticated tactics to trick unsuspecting individuals into revealing sensitive and confidential information (Kumar et al., 2020). The information these malicious actors seek often includes usernames, passwords, credit card details, and other personal data that can be exploited for various nefarious purposes, such as identity theft, financial fraud, and unauthorized access to sensitive accounts (Valecha et al., 2021). Email phishing is a prevalent cyber-attack technique that relies on social engineering tactics to deceive individuals into providing confidential information. The findings from this research can potentially enhance the protection of individuals and organizations against the detrimental consequences of falling victim to phishing scams.

One of the primary techniques employed in email phishing involves impersonating trusted and well-known entities, banks, social media platforms, online retailers, or government agencies (Basit et al., 2020). Attackers create deceptive emails that resemble legitimate communication from these reputable sources (Magdy et al., 2022). They use carefully crafted language, logos, and even spoofed email addresses to make the messages appear authentic (Parsons et al., 2019). As a result, recipients are often lured into a false sense of security, making it increasingly challenging to distinguish these phishing attempts from genuine communications (Zuraiq & Alkasassbeh, 2019). As a

result, individuals and organizations are constantly at risk of falling victim to these scams, leading to financial losses, data breaches, and other detrimental consequences.

Extensive research has shed light on the severity of the email phishing threat and its detrimental impact on online security. Researchers, in various studies (Butavicius et al., 2022), have emphasized the importance of understanding the nuances of phishing attacks and developing effective countermeasures to protect users and organizations from falling victim to these scams. The deceptive nature of phishing emails has become increasingly sophisticated, exploiting psychological and cognitive vulnerabilities to increase their success rates (Kwak et al., 2020). Scholars Ferreira & Teles (2019) have analyzed the characteristics of these fraudulent emails, highlighting their compelling appearances and psychological manipulations that can deceive even cautious and well-informed users. For instance, some phishing emails instill a sense of urgency, pressuring recipients to act quickly and bypass their critical thinking.

Furthermore, cybercriminals often leverage social engineering techniques to enhance the credibility of their phishing attempts (McAlaney & Hills, 2020). By personalizing the messages or tailoring them to specific demographics, attackers increase the likelihood of victims falling for their ploys. The use of data obtained from previous breaches or publicly available information makes these phishing attempts even more convincing and difficult to identify.

Phishing attacks, fueled by cunning social engineering tactics, have become one of the most formidable threats to cybersecurity. These deceptive schemes have garnered considerable attention from researchers, who have sounded the alarm about their

widespread prevalence and devastating impact on the security of digital environments (Bountakas & Xenakis, 2023). As cybercriminals become increasingly adept at manipulating human psychology, organizations face a significant challenge in safeguarding their sensitive information from falling into the wrong hands.

An intriguing aspect of the phishing threat lies in the pivotal role of employees' behavior in an organization's cybersecurity posture. Studies have extensively explored the influence of employees' responses to phishing emails, revealing that their actions can either fortify the company's defenses or create vulnerabilities that cyber attackers can exploit (Buckley et al., 2023; Jampen et al., 2020; Weaver et al., 2021). Buckley et al. (2023) explored “the relationship between sociodemographics, cyber security training, phishing email typology, information processing factors, and risky and secure email response behaviors." The study's findings revealed "that the likelihood of clicking on a link in a phishing email was lower for participants who had greater faith in their intuition and paid more attention to the sender's email address" (Buckley et al, 2023). Buckley et al. (2023) concluded that, "the likelihood of reporting a phishing email was" more remarkable "for participants" with more "elaborative processing to evaluate the email than" for "those who used less elaboration." The findings by Buckley et al. (2023) have also been confirmed by Lain et al. (2022), who found similar results in their study. Weaver et al. (2021) "examined the effectiveness of a short online phishing training program to help users identify phishing emails." The study results revealed "that the training group's sensitivity" to "correctly classifying emails as legitimate or phishing

increased by 1.14", and "the control group's sensitivity increased by only 0.48" (Weaver et al., 2021).

The inability to identify phishing emails among individuals is a concern. The two studies by Buckley et al. (2023) and Weaver et al. (2021) shed light on a concerning issue in cybersecurity: the difficulty many individuals face in accurately identifying phishing emails. However, both studies also bring encouraging news, suggesting that individuals can significantly enhance their effectiveness in detecting such deceptive messages with the right training programs. As a result, it becomes paramount for organizations to cultivate a culture of cyber awareness and prioritize comprehensive training initiatives to empower their employees in recognizing and managing phishing attempts effectively. The findings from Buckley et al. (2023) and Weaver et al. (2021) underscore the significance of addressing the human element in cybersecurity, as cybercriminals continue to exploit psychological and cognitive vulnerabilities to deceive their targets. Phishing emails have become increasingly sophisticated, and even vigilant individuals can sometimes fall victim to their persuasive tactics. This makes it imperative for organizations to take proactive measures to fortify their defense lines and arm their employees with the knowledge and skills to defend against such threats.

In cybersecurity, researchers have been diligently working to shed light on the challenges organizations face concerning their knowledge of social engineering attacks. Aldawood and Skinner (2019) and Blancaflor et al. (2022) are among those who have contributed valuable insights through their studies, mainly focusing on the human detection aspects of phishing email detection. Blancaflor et al. (2022) conducted a study

to gauge the knowledge level of individuals working from home regarding the threats posed by phishing attacks. Their findings indicated that most respondents acknowledged the importance of understanding phishing attacks and keyloggers. Initially, only 35 percent of the participants strongly agreed that understanding phishing was essential. However, after demonstrating the workings and potential impact of keyloggers on their systems, this percentage rose to 60 percent. Only 10 percent of the respondents believed that knowledge about phishing and keyloggers would not significantly impact the safety of their PCs or their work-from-home jobs. This study highlights the value of providing employees comprehensive knowledge about phishing and related threats to bolster cybersecurity measures.

In today's digital landscape, the threat to privacy and safeguarding knowledge are crucial. With the increasing sophistication of phishing attacks, it is imperative to develop robust detection methods to protect organization and individuals against attackers of these malicious activities. Email phishing poses a significant threat as it targets users through deceptive emails, attempting to trick them into revealing sensitive information or downloading malware. Similarly, the research conducted by Aldawood and Skinner (2019) aimed to explore the challenges organizations face in safeguarding human knowledge against social engineering attacks. The researchers examined existing literature to support their arguments and analyzed contemporary cybersecurity approaches. The results revealed that organizations still faced vulnerabilities despite employing cutting-edge cybersecurity measures and well-trained personnel. Hackers

successfully executed malicious acts, including stealing sensitive information from organizations.

Preventing phishing attacks has become increasingly crucial as cybercriminals develop sophisticated techniques to deceive unsuspecting individuals. The studies conducted by Aldawood and Skinner (2019) and Blancaflor et al. (2022) highlight the importance of not solely relying on human behavior to prevent phishing attacks. Singh et al. (2019) also conducted a study to investigate whether the frequency of experiencing phishing emails influenced the core complex detection decisions in phishing. "Results revealed that participants receiving a higher frequency of phishing emails had a higher hit rate but also higher false alarm rate at detecting phishing emails post-training compared to participants encountering lower frequency levels during training" (Singh et al., 2019). While educating employees is crucial, exploring technological alternatives to enhance threat detection capabilities is equally vital. By combining practical human training with robust technological solutions, organizations can fortify their defenses against social engineering attacks, ultimately contributing to a more resilient and secure digital environment. As research continues, a comprehensive and adaptive approach to cybersecurity will remain crucial in safeguarding valuable information and resources from cyber threats.

2.5.2 Machine Learning and Phishing Threats

Phishing attacks, characterized by their deceptive nature and increasing prevalence, have recently become a significant concern (Mughaid et al., 2022). Combating this escalating threat requires a pressing need for more effective phishing detection

technology (Rafat et al., 2022). Artificial Intelligence (AI) methods have been embraced in this domain for approximately a decade, offering promising potential in identifying and thwarting phishing attacks (Soon et al., 2020). As the volume of emails received daily grows, automated management of phishing emails has become crucial in today's context (Alauthman, 2020). Though valuable, traditional lookup systems and fraud cue-based methods are limited (Swain et al., 2022). This has driven researchers to develop advanced deep representation-based methods capable of learning deep fraud cues to bolster the capacity of anti-phishing mechanisms (Chai et al., 2021). In the ongoing battle against phishing, companies adopt various software-based solutions to detect malicious URLs effectively. These solutions encompass a range of cutting-edge technologies, including image processing, natural language processing, machine learning (ML), and AI (Indrasiri et al., 2021). Curbing phishing attacks is crucial in today's digital age, as they pose a significant threat to individuals and organizations. By analyzing the effectiveness of different deep learning algorithms in detecting email phishing attempts, this study seeks to contribute to developing more robust and accurate anti-phishing measures.

By leveraging the power of these technologies, organizations aim to strengthen their defenses and stay ahead of increasingly sophisticated phishing attacks. One such study was conducted by Aljofey et al. (2020), who introduced an efficient solution for phishing detection based on the URL of websites utilizing a fast deep learning-based model. What made the model introduced by Aljofey et al. (2020) different from traditional models is that the model did not require accessing the content of the target website or relying on third-party services. Adebowale et al. (2019) also developed a deep learning phased

phishing detection solution that used the Universal Resource Locator and website content. The detection system had a classification accuracy of 93.28 percent. Both studies by Aljofey et al. (2020) and Adebowale et al. (2019) provided evidence of utilizing machine learning and AI to detect phishing emails. With the increasing sophistication of phishing techniques, exploring advanced machine-learning approaches that can effectively identify and mitigate these malicious activities is essential.

The academic community has also proposed diverse solutions to counter the threat of web phishing. Researchers have explored many methodologies, each designed to tackle the challenges posed by phishing attacks (Adebowale et al., 2019). These proposed approaches range from innovative AI-driven systems to ingenious techniques to disrupt phishing campaigns and safeguard users. In email filtering, machine learning has emerged as a favored approach to classify whether an email is legitimate or unwanted (Butt et al., 2022). Leveraging machine learning algorithms, such as classifiers, enables efficient and automated identification of potentially malicious emails (Sethi et al., 2021). These algorithms continually learn from data and adapt their detection capabilities to stay resilient against evolving phishing tactics.

Cultivating the use of natural language processing techniques to detect phishing. Researchers such as Shahrivari et al. (2020) and Salloum et al. (2022) have compared the results of multiple machine-learning methods for predicting phishing websites and using natural language processing techniques to detect phishing. Shahrivari et al. (2020) evaluated twelve different classifiers using a dataset of 6,157 legitimate websites and 4,898 phishing websites. The classifiers examined included Logistic Regression,

Decision Tree, Support Vector Machine, AdaBoost, Random Forest, Neural Networks, KNN, Gradient Boosting, and XGBoost. The results revealed that the ensemble classifiers, specifically Random Forest and XGBoost, demonstrated excellent computation duration and accuracy performance. Salloum et al. (2022) synthesized research on using "natural language processing techniques for" detecting "phishing, except one that sheds light on" using "natural language processing "techniques for classification and training purposes while exploring a few alternatives. The findings" revealed "that the main research area in phishing detection studies is feature extraction and selection, followed by methods for classifying and optimizing the detection of phishing emails." The studies by Salloum et al. (2022) and Shahrivari et al. (2020) shed light on the direction research has been going on phishing emails and detection. These studies also highlight "that more research needs to be conducted" (Yeo et al., 2024) on effective ways to detect these emails other than human detection.

The research and efforts directed toward phishing detection are ongoing, driven by the ever-increasing threats posed by cybercriminals. The development of robust AI-powered systems and advanced machine-learning models continues to be at the forefront of the battle against phishing attacks (Mansoor et al., 2021). As technology evolves and improves, the hope is to create a safer digital environment where individuals and organizations can confidently navigate their online communications, free from the fear "of falling victim to phishing scams" (Truong et al., 2023).

2.5.3 Deep Learning Neural Networks

Deep learning, a subset of machine learning, is a category of algorithms inspired by the structure and functioning of the human brain (Bashar, 2019; Dilhara, 2021). Given the detrimental effects of email phishing, researchers have turned to deep learning-based algorithms to enhance the detection of phishing emails (Mughaid et al., 2022). While deep learning detection is effective in most instances, these phishing detection applications do not apply to mobile devices (Haynes et al., 2021). An essential characteristic of any deep learning model is its ability to generalize well over testing data, ensuring its effectiveness beyond the training dataset (Kaddoura et al., 2020). Machine learning algorithms, including neural networks, have previously attempted to detect spam messages and differentiate standard messages or ham from SMS messages (Chandra & Khatri, 2019). However, deep learning algorithms, a subset of machine learning, have emerged as a potential alternative to conventional models due to their superior performance in specific domains (Do et al., 2022; Roy et al., 2022). Jena et al. (2021) describe that deep learning neural networks attempt to mimic the human brain's behavior, though they do not match their full capabilities. These networks can learn from vast quantities of data, enabling them to excel in various applications (Liu et al., 2021). Indeed, deep learning drives multiple artificial intelligence (AI) services and applications, enhancing automation and the performance of physical and analytical functions with minimal to no human involvement (Jena et al., 2021). These algorithms have shown promising results in various domains, such as image recognition, natural language processing, and speech recognition. In the context of email phishing detection, (e.g.,

Jemili and Korbaa (2023)) propose that deep learning algorithms can improve the accuracy and efficiency of detecting malicious emails. By comparing the performance of LSTM, GRU, and multimodal deep learning algorithms, this study aims to identify the most effective approach for phishing detection and contribute to developing more robust cybersecurity measures.

A variety of deep learning algorithms exist, each designed for specific tasks. “Deep learning algorithms fall into the following four main categories: convolutional neural networks, recurrent neural networks, generative adversarial networks, and reinforcement learning networks” (Hao et al., 2023). Apart from the realm of email phishing detection, deep learning algorithms also find applications in law enforcement. Catal et al. (2022) explain how these algorithms analyze and learn from transactional information, helping law enforcement recognize suspicious trends that may indicate potential criminal or fraudulent activity. Additionally, deep learning applications such as computer vision and speech recognition have significantly enhanced the effectiveness and efficiency of investigative evaluation. By extracting trends and evidence from various sources, such as documents, images, video recordings, and sound, these algorithms enable law enforcement agencies to process vast amounts of data more quickly and accurately (Catal et al., 2022). Brindha et al. (2023) proposed a deep learning model to differentiate between legitimate and phishing emails effectively and concluded in their experimental findings that the accuracy of their model was 99.72 percent.

2.5.4 Long Short-Term Memory Networks

Long Short-Term Memory (LSTM) networks have demonstrated remarkable success in time series forecasting (Liu et al., 2023). Eckhardt and Bagui (2021) developed a classification system for identifying phishing emails using deep neural networks. The researchers specifically compared two types of deep neural networks: long short-term memory (LSTM), a recurrent neural network architecture, and convolutional neural networks (CNN). The study's findings indicated that the LSTM model achieved the highest accuracy in classifying phishing emails, with an impressive accuracy rate of 98.32% and an ROC-Score (Receiver Operating Characteristic) of 96.57%. However, the CNN model also performed remarkably well, with a high accuracy rate of 96.52% and an ROC-Score of 96.16%. Both models demonstrated excellent classification capabilities, showcasing the potential of deep neural networks in detecting and distinguishing phishing emails. As a modified Recurrent Neural Network (RNN) family type, LSTM architecture presents unique advantages, making it particularly effective in handling sequential data (Isik et al., 2020). This has led to the widespread adoption of LSTMs, and they have proven to operate effectively even on large-scale problems, as observed by Sheneamer (2021). Rahman and Ullah (2020) proposed "a new model for detecting spam messages based on the sentiment analysis of the textual data of the email body." Rahman and Ullah (2020) incorporated "Word-Embeddings and Bidirectional LSTM network to analyze texts'" sentimental and sequential properties. Rahman and Ullah (2020) also sped "up the training time and extracted higher-level text features for Bi-LSTM network using Convolution Neural Network." The model achieved an "improved performance of

accuracy of about 98-99%" (Rahman & Ullah, 2020). According to Rahman and Ullah (2020), "Apart from this, we demonstrate our model outperforms not only some popular machine learning classifiers but also state-of-the-art approaches for detecting spam messages and hence, proves the superiority by itself" (Rahman & Ullah, 2020).

One significant advantage of LSTM networks over the Gated Recurrent Unit (GRU) lies in their superior accuracy when dealing with datasets containing longer sequences. As mentioned by Do et al. (2022), this characteristic makes LSTMs a preferred choice for tasks that involve extended temporal dependencies, where GRUs might struggle to maintain precision. This strength has been harnessed in various applications, including email content analysis. Chandra and Khatri (2019) point out that LSTM networks can effectively analyze the sequential nature of email content to capture relevant patterns. By considering multiple aspects of the emails, LSTM-based models can better discern phishing attempts and reduce false positives, thereby improving the overall effectiveness of phishing detection systems. LSTM networks are a powerful tool in text classification and sequential data analysis, particularly in handling long sequences. Their unique architecture and ability to capture temporal dependencies have led to their comprehensive utilization in various applications, including email content analysis for phishing detection. As researchers continue to explore and refine LSTM-based techniques, their impact on improving the accuracy and efficiency of data analysis tasks is expected to grow further.

2.5.5 Bidirectional Encoder Representations from Transformers

"Bidirectional Encoder Representations (BERT's) design consists of many encoders layered on top of one another in a feed-forward manner, where the output from one encoder is fed as the input for the following" (Islam & Zhang, 2024). Islam & Zhang (2024) highlight that fact that, "data used for training was composed of plain Wikipedia text corpus and Google's BookCorpus." BERT's design consists of many encoders layered on top of one another in a feed-forward manner, where the output from one encoder is fed as the input for the following one. "It is a bidirectional model which helps in analyzing the text in both directions of the target word such that it can predict the previous text as well as the next text based on the surrounding words" (Mittal et al., 2022). "Due to its training on a huge dataset of Wikipedia articles, it is mighty and has been used in the industry for various tasks like sentiment analysis, text prediction, chatbots, auto-completion of queries and email, etc." (Mittal et al., 2022). "Using BERT as transfer learning proves to be a powerful tool in predicting the outcome of the email texts malicious or legitimate" (Mittal et al., 2022). One major limitation of BERT is that "While complex BERT models obtain impressive results, the full-sized models are computationally expensive and memory intensive" (Lee et al. 2020).

2.5.6 Generative Pre-trained Transformer

GPT-2 is an NLP. NLP focuses on the interaction between computers and human language. GPT-2 excels at this since it is trained on large volumes of text data. GPT-2 employs deep learning techniques, notably a transformer architecture, which is extremely

effective in NLP applications. These techniques allow GPT-2 to understand the relationships between words and sentences, enabling it to process and generate human language.GPT-2 discards the encoder blocks reducing it to a decoder stack. It provides traditional language model functionality, allowing to predict the next word in a sequence given the history (Klein & Nabi, 2019).

2.6 Summary and Conclusion

Phishing emails are a growing threat, causing significant financial losses and data breaches for individuals and organizations. These deceptive emails aim to trick recipients into revealing sensitive information or clicking on malicious links. Traditional methods for detecting phishing emails are becoming insufficient as attackers develop more sophisticated tactics. This research aims to explore the effectiveness of various machine learning models in email phishing detection, with the goal of improving detection accuracy and offering a more robust defense against phishing scams.

The study investigates a range of machine learning models, including random forest classifiers, logistic regression, LSTMs, CNNs, and transformer-based models like BERT. Additionally, the research incorporates interpretability techniques (SHAP) to gain insights into model predictions and identify the features each model relies on for classification. This multi-faceted approach allows for a comprehensive evaluation of different models' strengths and weaknesses in the context of phishing email detection.

The research places a particular emphasis on deep learning algorithms, considering their effectiveness in handling complex data like email content. LSTM networks, known for their ability to capture long-term dependencies in sequential data, are explored for

their potential in analyzing the sequential nature of email text and identifying patterns indicative of phishing attempts. The study acknowledges the challenges associated with deep learning models, such as the need for large amounts of data for training and potential limitations in interpretability.

By comparing the performance of various machine learning models and incorporating interpretability techniques, this research seeks to contribute to the development of more robust and accurate email phishing detection systems. The findings from this study can empower organizations and individuals with improved defenses against phishing scams, ultimately fostering a safer digital environment.

Chapter 3—Methodology

3.1 Introduction

The purpose of this quantitative cross-sectional study on the Praxis project is to assess how well base models and proposed models algorithms perform in email phishing detection. The goal of the study is to compare how well these algorithms perform in identifying email phishing and to pinpoint the elements and interactions that affect their performance. This methodology section comprises the research design and rationale as well as the methods chosen. It also includes procedures for recruitment, participation, and data collection. A data analysis plan, threats to validity, ethical considerations, and a summary will also be part of this section.

3.2 Research Design and Rationale

This study's key variables included the independent variables (base models and traditional algorithms models) and the dependent variable (email phishing detection performance). Ideally, the independent variables represented the different machine learning algorithm models used to detect email phishing attempts. Particularly, base models included random forest classifier and logistic linear regression while proposed models algorithm models included the linear layers' models, transformer-based models, and generative pre-training transformer (GPT) models. The dependent variable, or the performance of the machine learning models in email phishing detection was assessed to understand how well they could classify emails as phishing or non-phishing primarily using the validation accuracy metrics. Validation accuracy represents the percentage of

correctly predicted outcomes (or classifications) by the model on the validation set (Manjunath et al., 2022). Other metrics used include precision, recall, F1-score, and ROC-AUC score, and confusion matrix.

The study's research design chosen was a quantitative cross-sectional study design. According to Setia (2016), cross-sectional study design enables the researcher to measure the outcomes and exposures of the participants. It means the design allows observation and measurement of different exposures or factors influencing specific outcomes of research interest (Brown et al., 2021). This study was focused on assessing the performance of different machine learning models in email detection. It was further concerned with exploring key factors influencing the effectiveness of different machine learning models in email phishing detection, thereby aligning with the chosen design.

This design is particularly suitable for this study as it allows for data collection at a specific point in time, providing a snapshot of the participant's experiences with email phishing detection. By using this design, the researcher was able to gather information on the performance base models, and traditional algorithms models comprehensively (Brown et al., 2021; Yang et al., 2023). Additionally, the cross-sectional design allows for the examination of multiple variables simultaneously, enabling a more holistic understanding of the effectiveness of the algorithms (Yang et al., 2023). This study benefited from the quantitative cross-sectional research design because it examined more than one variable while assessing how well base models and proposed models algorithms perform in email phishing detection.

Quantitative cross-sectional research design is analytical and descriptive design making it suitable for this study (Brown et al., 2021). Different algorithms' performances were systematically analyzed and compared to give a better understanding of how well perform. This ability to yield analytical and descriptive results was important for this study to achieve its goals. Since this design allows the computation of large data, it was suitable to utilize in manipulating the large datasets required to build machine learning algorithms (Yang et al., 2023). It worked well for representing a variety of characteristics or variables that machine learning algorithms can use as input. The model can learn complex patterns and relationships within the dataset with this thorough data representation.

Other research designs were not selected in this study due to inherent weakness in answering this study's research questions compared to quantitative cross-sectional research design. For instance, an experimental design was not adopted since the focus of this study was not on establishing cause and effect but rather on assessing the performance of different algorithms in email phishing detection (Siedlecki, 2020; Stockemer et al., 2019). Equally, quantitative longitudinal research was not selected since it would require data collection from the same participants for a long time which could mean more time to collect data (Siedlecki, 2020). However, the quantitative cross-sectional research design utilized data collected and stored in different datasets for a long time.

3.3 Methodology

A quantitative methodology was selected for this study. Quantitative methods are suitable when carrying a research involving numerical analysis (Schoonenboom, 2023; Schutt, 2019; Stockemer et al., 2019). According to Mohajan (2020), quantitative methods utilize clearly defined numerical data to conduct examinations that are quantifiable. The use of a quantitative research approach saves time because of the various statistical tools that are at the disposal of researchers. According to Stockemer et al. (2019), the numerical data and statistical analysis, which provide impartial and quantifiable conclusions, are the foundation of quantitative research. Taherdoost (2022) quantitative research allows the opportunity for accurate generalization without the possibility of guessing as to the outcome of an event or events.

This research was conducted using a large data set of numerical values and the results regarding the performance of different machine learning algorithms were aimed to be presented in numbers. For instance, performance metrics such as validation accuracy, precision, recall, F1-score, ROC-AUC score, and confusion matrix could all be presented numerically (Manjunath et al., 2022). Still data analysis methods adopted included utilizing building machine learning models using Python programming which allowed statistical manipulation of data to produce the desired results that could answer the research questions adequately (Ziogas et al., 2021). Quantitative methods of research allowed for testing of the hypothesis formulated to help answer the research questions (Schoonenboom, 2023; Schutt, 2019). Quantitative methods were complementary to the

quantitative cross-sectional research design chosen further elucidating its suitability (Stockemer et al., 2019).

Qualitative methods and mixed research methods were not suitable for this study. For example, qualitative methods could not explain the performance of base models and proposed models models in phishing detection objectively. Qualitative methods deal with subjective meanings, experiences, and interpretations of individuals or groups which was not the focus of this study (Creswell & Creswell, 2022; Schoonenboom, 2023). This type of methodology was also unsuitable for results generalization due small samples they utilized (Schoonenboom, 2023). Testing the hypothesis would not be possible if a qualitative approach was adopted. Mixed methods could also be used in this study, but it could not adequately answer all the research questions considering some parts would receive objective answers while others could get subjective answers (Creswell & Creswell, 2022). It would also not support hypothesis testing.

3.3.1 Population

The target population was organizations from various industries and sectors that have experienced phishing attacks between 1998 and 2022. The dataset utilized in this study was found online and therefore, was assumed to have represented diverse factors such as organization's sizes, industries, and geographical locations. These criteria ensure that the selected organizations represent diverse entities susceptible to phishing attacks. For this study, the research objectives and the availability of suitable secondary data were critical. Therefore, all the 11 datasets available from the public website were considered for download and analysis. They include CEAS_08.csv, Enron.csv, Ling.csv, Nazario.csv,

Nazario_5.csv, Nigerian_5.csv, Nigerian_Fraud.csv, SpamAssasin.csv, TREC_05.csv, TREC_06.csv, and TREC_07.csv. The inclusion included requirements such as the availability of labeled or annotated email data that accurately identifies phishing emails.

3.3.2 Sampling Strategy

A purposive sampling strategy and simple random sampling were employed to select the datasets for analysis. According to Hakizimana et al. (2019), a two-stage cluster sampling design would ensure that estimates are representative. The purposive sampling ensured only datasets with more than two features were selected. This approach was chosen because it allowed for a more efficient and cost-effective way of gathering relevant data. The researcher was keen on the availability of labeled or annotated email data that accurately identifies phishing emails. Selecting a random sample of 3 datasets from those that met the inclusion criteria ensured relevant and manageable data was utilized.

3.3.3 Procedure for Recruitment, Participation and Data Collection

The procedure for getting access to the dataset was downloading publicly available datasets from https://zenodo.org/records/8339691. Out of the 11 datasets, 2 contained only 2 features. These datasets were Ling and Enron datasets, and they were excluded. The data collection process for the PRAXIS project involved studying variables from the dataset system using a Python program to assess the data contents in each column. Varoquaux and Cheplygina (2022) argued that in developing machine learning models, large, labeled datasets are required. However, they recommended availability of such data

be a key factor in choosing the dataset. Also, a size ranging from a few thousand to millions of data points is ideal for machine learning model development and testing (Varoquaux & Cheplygina, 2022). The 3 randomly selected datasets were to be merged before the resulting dataset was divided into training, testing, and validation segments.

3.3.4 Data Collection

Instrumentation and Operationalization of Constructs-

Target data used in this study included Nazario, SpamAssasin, and TREC_07 datasets. Nominal values of Phishing (1) and Spam (0) were assigned to columns with the target variables. Independent variables were organized before DataFrames were formed using Pandas in Python (Ziogas et al., 2021). Data cleaning was conducted before the actual analysis. Data was labeled *phishing* or *spam* to differentiate phishing emails from other spam emails in the data set. The resulting data set was saved as a clean comma-separated value file (CSV) saved as *Final_Clean_Data.csv* for further analysis.

The operation definitions as per this study are:

Base models: These are the fundamental algorithms that serve as building blocks for more complex models (Kumar et al., 2020). In this study, they included random forest classifier and logistic linear regression.

Ensemble models: Are machine learning techniques that combine multiple individual models to improve the overall performance and predictive accuracy (Hao et al., 2020; Sarker, 2021). According to Hao et al. (2020), ensemble models aggregating the predictions of multiple base models to generate a final prediction. In this study, they

include random forest classifier since it combines multiple decision trees to make predictions.

Generative Pre-Training(GPT) Transformers-Based (Proposed Models): These transformer-based models are pre-trained on vast amounts of text data to understand and generate human-like text (Sarker, 2021; Vaswani et al., 2017). The models included the GPT, GPT+LSTM, GPT+CNN, GPT+LSTM+CNN (GPT2Classifier, GPT2CNNClassifier, PT2LstmClassifier, and GPT2CNNLstmClassifier).

Individual algorithm models: Are single models or standalone models that rely on a single algorithm and do not incorporate predictions from other models (Kumar et al., 2022). They make predictions based solely on the learned parameters from the training data without any ensemble techniques (Hao et al., 2020).

Linear layers models: These are the kind of foundation model that machine learning applications frequently utilize for classification jobs. These models are made up of one or more linear layers, each of which applies an activation function and a linear transformation to the input characteristics (Sarker, 2021). In this study, they included conventional models including the Mlp4LayerClassifier, Mlp8LayerClassifier, LstmClassifier, CnnClassifier, and CnnLstmClassifier.

Logistic regression: It is a statistical model used for binary classification tasks, where the outcome variable is categorical and has two classes (Sarker, 2021). It predicts the probability of occurrence of an event by fitting data to a logistic function. (Manjunath et al., 2022; Sarker, 2021).

Random forest classifier: It is a machine learning model that is utilized for ensemble classification and regression tasks combining multiple decision trees to make predictions (Hao et al., 2020; Sarker, 2021).

Traditional machine learning models: These are fundamental algorithms which include supervised learning (classification and regression), unsupervised learning (clustering, dimensionality reduction, and association rule learning), and reinforcement learning (Sarker, 2021). In this study, they included the linear layers' models, transformer-based models, and generative pre-training transformer (GPT) models.

Transformer-based models: They include a class of deep learning models that rely on a transformer architecture, which was introduced by Vaswani et al. (2017). These models included BERT, BERT+LSTM, BERT+CNN, BERT+LSTM+CNN (BertClassifier, BertLstmClassifier, BertCNNClassifier, and BertCNNLstmClassifier). Figure 1 below shows the research methodology used for this research.

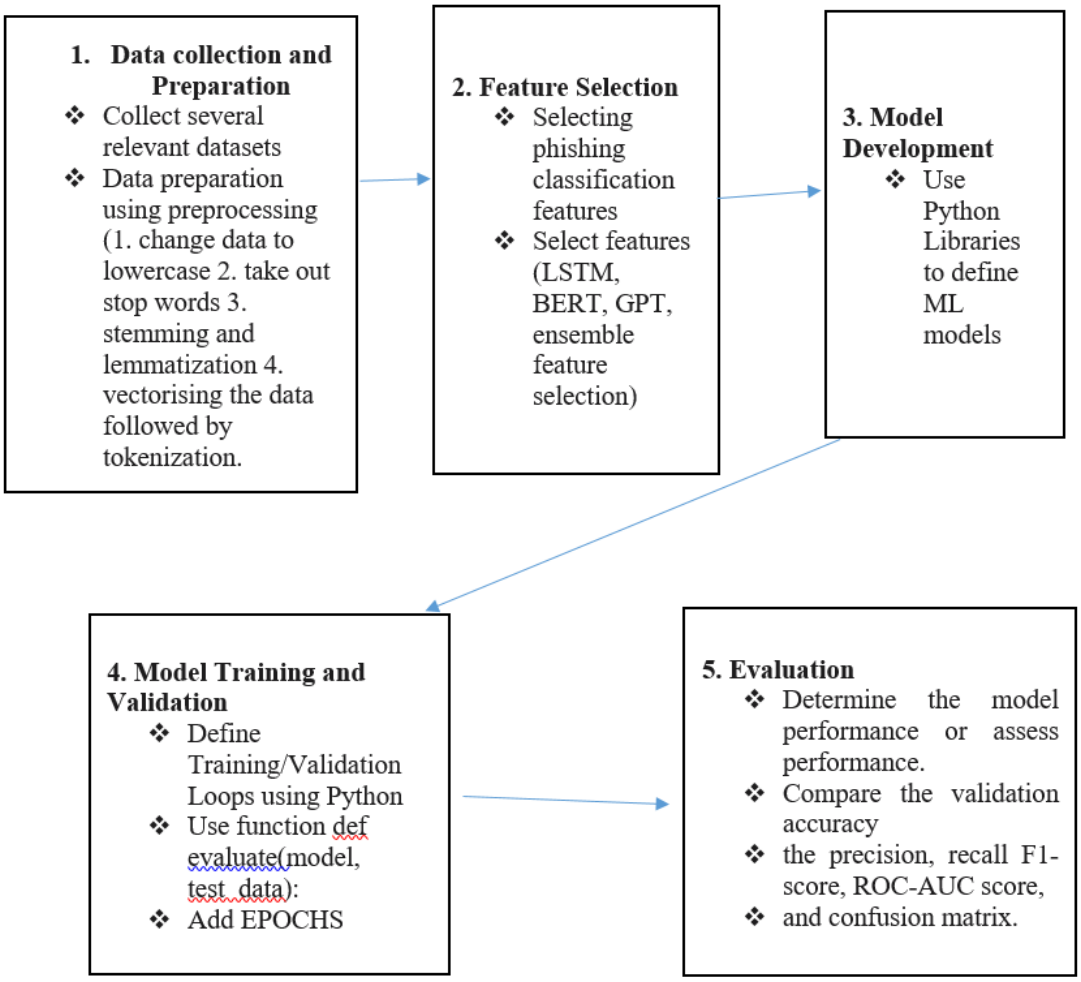

Figure 1. Research Methodology

Note: Research methodology showing model evaluation from data collection to data analysis

3.3.5 Data Analysis Plan

This quantitative cross-sectional study adopted Python program codes utilizing powerful inbuilt libraries such as numpy and scipy to manipulate data (Ziogas et al., 2021). First data was cleaned, and datasets were merged to form one large dataset saved

in a comma-separated value (CSV) file. Eliminating the problematic data could fix it in the following methods. Leys et al. (2019) assert that outliers may affect the connection between variables. They suggest three strategies for handling outliers. Keeping the outliers, eliminating the outliers, or recoding the outliers are all options. As a result, any missing or inaccurate data dropped. All the model types were implemented in Python after the data wrangling process to maintain only relevant data that would help assess model performances. Next, the dataset was divided into training and testing sets, ensuring that both sets contained a balanced distribution of legitimate and phishing emails. The analysis plan involved assessing the performance of the models using validated accuracy figures since the dataset was balanced. No further analysis was conducted such as t-test or other statistical tests because the validated accuracy values were sufficient to compare how different models performed in email phishing detection.

For base models, a full classification report was provided to assess performance. The assessment was the validation accuracy which shows the percentage of correctly predicted outcomes (or classifications) by the model on the validation set, the precision, recall, F1-score, ROC-AUC score, and confusion matrix. The confusion matrix confusion matrix diagonal elements tell the correct number of predictions. The first element tells the number of emails that were not phishing and were not identified as phishing -indicating correct prediction. The second element tells the number of emails that were actually phishing, and the model identified them as phishing, indicating the correct prediction. Traditional machine learning models, linear layers, transformer-based models, and GPT models' performances were assessed by the validation accuracy which showed the

percentage of correctly predicted outcomes (or classifications) by the model on the validation set.

3.4 Threats to Validity

During the study, the researcher used strict and open approaches to address validity concerns. Confounding variables were considered during data analysis to control for potential confounding effects. An adequate sample was selected to enhance statistical power. The results will be carefully interpreted and discussed in the context of email phishing detection and cybersecurity practices, minimizing the risk of misinterpretation (Mellinger & Hanson, 2022). The researchers employed various methodological approaches to address the potential threats to validity and ensure the robustness and credibility of the findings in this study.

Ecological validity is a potential threat to the generalizability of the findings in this study. Since the data for testing the algorithms were collected from a specific sample of email phishing attacks, it may not accurately represent the broader population of phishing attempts (Mellinger & Hanson, 2022). Additionally, simulated phishing emails may only partially capture the complexity and variability of real-world phishing scenarios. Therefore, caution should be exercised when interpreting the results of this study and generalizing them to real-world settings.

Another potential threat to validity is the issue of construct validity. Construct validity measures used in a study accurately represent the underlying constructs or concepts being studied (Mellinger & Hanson, 2022). In the context of email phishing detection, construct validity could be a concern if the measures used to assess the

performance of base model algorithms and traditional machine learning algorithms models are not truly reflective of their effectiveness in detecting phishing emails. This could occur if the chosen measures do not capture all relevant aspects of phishing detection or if they are biased toward certain types of phishing attacks. To mitigate this threat, careful measures towards evaluation, ensure they are comprehensive and unbiased. One crucial aspect to consider is the accuracy of the measures in capturing the true positive rate, which indicates the percentage of phishing emails correctly identified as such.

3.5 Ethical Procedures

While ethical procedures for this study will not deal with human subjects, ensuring research integrity, and upholding ethical standards in data analysis and reporting was observed. As this study involved using secondary data from published literature, direct interactions with human participants are not applicable, but respecting intellectual property rights through proper citation and acknowledgment is paramount (Page et al., 2022). The researcher handled the data with confidentiality, ensuring no personally identifiable information was disclosed. Adherence to research integrity principles, such as honesty, transparency, and unbiased reporting, was maintained throughout the study. The researcher diligently followed ethical guidelines and practices to maintain the credibility and integrity of the research process including seeking institutional review board approval before conducting the main study.

3.6 Summary

This section demonstrated how this quantitative cross-sectional study on the Praxis project to assess how well base models and proposed models perform in email phishing detection was conducted. This study's key variables included the independent variables (base models and traditional algorithms models) and the dependent variable (email phishing detection performance). The study's research design chosen was a quantitative cross-sectional study design and a quantitative methodology of data collection and analysis. The target population would consist of participants in organizations from various industries and sectors that have experienced phishing attacks. The data collection process for the Praxis project involved studying variables from the dataset system using a Python program to assess the data contents in each column. Target data used in this study included Nazario, SpamAssasin, and TREC_07 datasets. This quantitative cross-sectional study adopted Python program codes utilizing powerful inbuilt libraries such as NumPy and Scipy to manipulate data. Ecological and construct validity were the possible threats to the validity of this study. Ethical procedures including seeking IRB approval before conducting this study were followed. Chapter 4 will present the study's findings, providing a comprehensive analysis of the performance-based models and proposed models in email phishing detection.

Chapter 4—Results

4.1 Introduction

The purpose of this quantitative cross-sectional study on the Praxis project is to assess how well base models and proposed models perform in email phishing detection. The following research questions and hypotheses were used in this study. This section restates the research questions and hypothesis. It also describes data collection and presents the actual results obtained after the analysis and a summary.

Data Collection

Data used in this study was from 3 datasets: Nazario, SpamAssasin, and TREC_07. Data label columns in each data set were broken down into Phishing (1) and Spam (0). Organized the independent variables to *sender*, *receiver*, *date*, *subject*, *body*, *label*, and *url*. Data wrangling was performed on the three data sets and transformed into DataFrames. For instance, data columns *urls* were dropped because they were not useful in the analysis as part of data cleaning before the three DataFrames were joined to form the main DataFrame with 61,131 records. More data cleaning was conducted by dropping duplicates, joining similar columns, and dropping other unnecessary columns. Data was labeled *phishing* or *spam* to differentiate phishing emails from other spam emails in the data set. The resulting data set was saved as a clean comma-separated value file (CSV) saved as *Final_Clean_Data.csv* and further analyzed.

The file was divided into training, testing, and validation segments as follows: For train= (42462)70% of the dataset; Test= 30% of the dataset; Validation of 30% split to (9100)50% (9099)50%. Pre-processing involved tokenization and embedding. It entailed

merging some columns such as joining *subject* & *body* and leaving *sender email content url* (yes or no). There was hugging face BERT and hugging face GPT 2, and updating dataset is labels.

Data classification was performed via Python code. For any binary classification task, a neural network with at least one linear later was required. The rationale was to assess how just having a linear layer compares to having an LSTM classifier (lstm+linear) along with a CNN classifier (cnn+linear) and a classifier that has both LSTM and CNN layers followed by the classifying last linear layer. Also, adding transformers to this comparison with BERT and GPT was implemented. In each of these cases, a linear classification head was added followed by an LSTM classification head (contains a unidirectional LSTM layer followed by a linear layer), a CNN classification head (contains a 1D convolutional layer followed by a linear layer), and a custom head consisting of a unidirectional LSTM layer and a 1D CNN layer. In all of the custom classification heads the unidirectional LSTM layer appears first followed by the 1D convolutional layer and then the linear layer. The alternate combination could also be potentially explored.

4.2 Results

The flowing distribution was for 32,217 normal records and 28,444 phishing and spam records identified in the data set. Figure 2 below shows a bar graph of the distribution of normal records and phishing and spam email records from the data set.

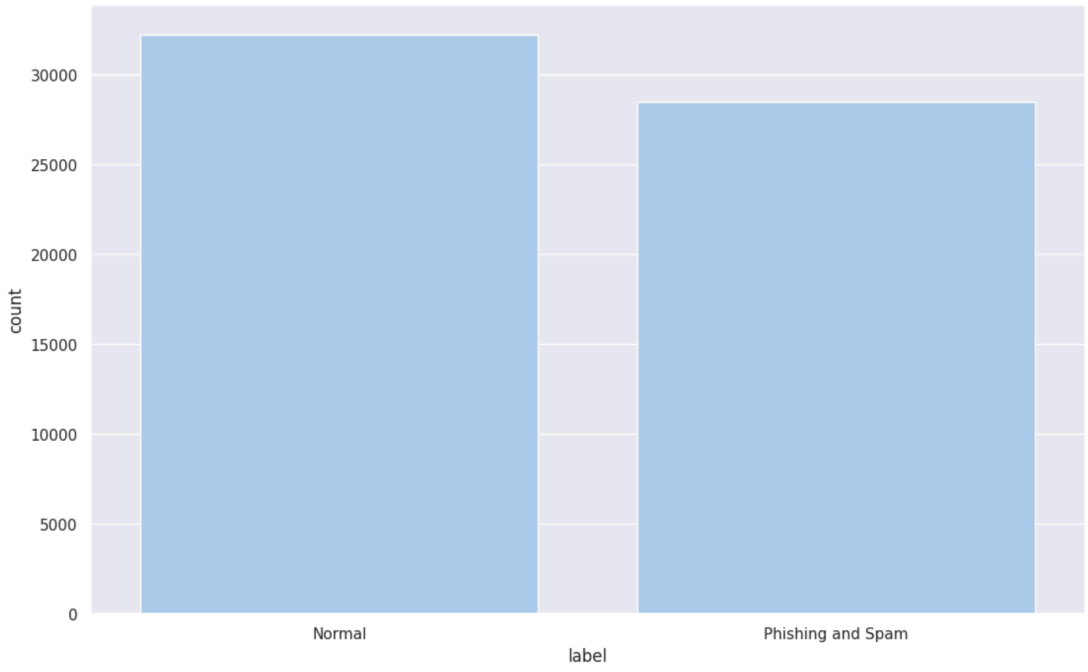

Figure 2. Normal and Phishing/Spam Email Count

Note. Records identified as either phishing or spam email from the data set indicating phishing emails are more common.

For phishing emails, the wordcloud below shows the top 100 most recurring words that are used in such emails.

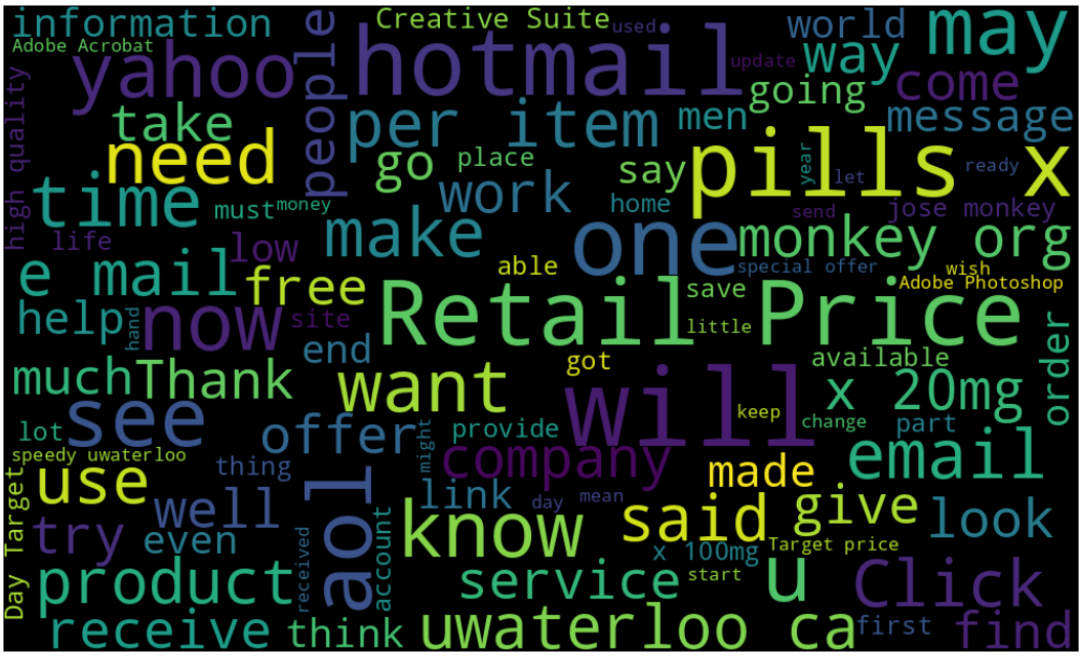

Figure 3. Common words in Phishing Emails

Note. Top 100 most recurring words that are used in phishing emails
For Spam emails, the wordcloud below shows the top 100 most recurring words that are used in such emails.

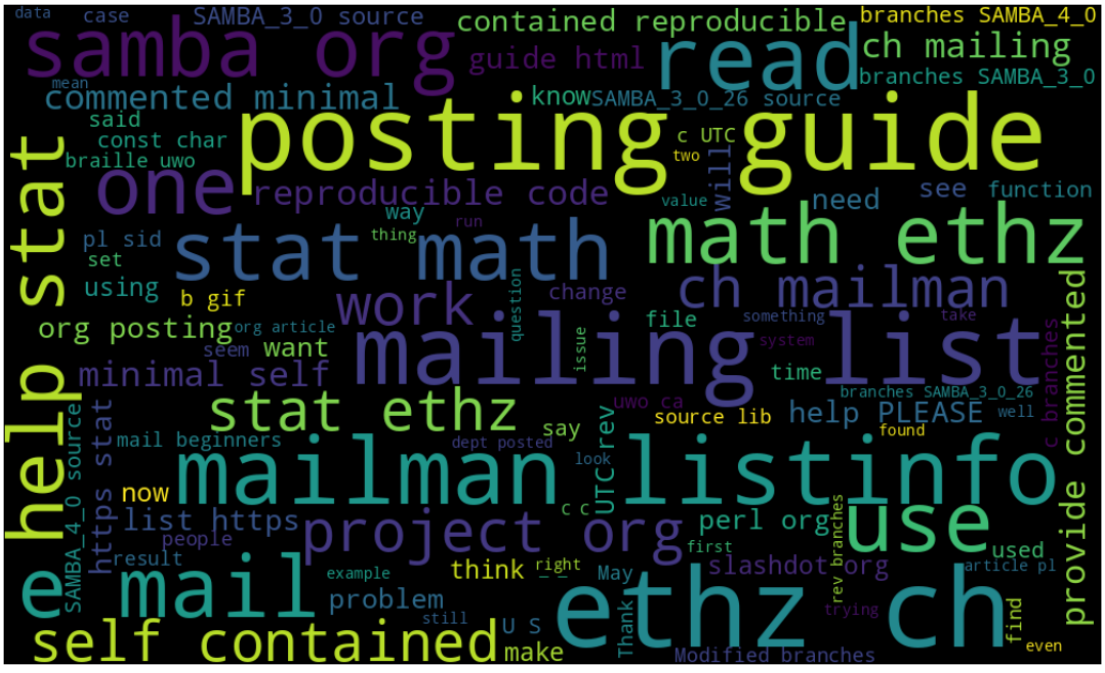

Figure 4. Common Words in Spam Emails

Note. Top 100 most recurring words that are used in spam emails

4.3 Base Models

The performances of major machine learning algorithms were assessed. The two common algorithms selected were random forest classifier and logistic linear regression. The results are presented in the table and in the form of a confusion matrix. Base Models: Random Forest Classifier and Logistic Regression Model achieved an average accuracy of 98%.

Random Forest Classifier

Confusion matrix
array([[5675, 107],
[53, 6298]])

Here, the top-left value (5675) represents true negatives (non-phishing emails correctly classified), the top-right value (107) represents false positives (non-phishing emails incorrectly classified as phishing), the bottom-left value (53) represents false negatives (phishing emails incorrectly classified as non-phishing), and the bottom-right value (6298) represents true positives (phishing emails correctly classified). Random forest classifier (ensemble model) achieved a ROC-AUC performance score of 0.9865 (98.65%). Table 1 below, shows the results of the Random Forest Classifier performance report.

Table 1. Random Forest Classifier Performance Report

	precision	recall	f1-score	support
1	0.98	0.99	0.99	6351
0	0.99	0.98	0.99	5782
accuracy			0.99	12133

macro avg	0.99	0.99	0.99	12133
weighted avg	0.99	0.99	0.99	12133

F1-Score: 0.9867779636723653
ROC-AUC 0.9865745750675312

Logistic Liner Regression

Confusion matrix:
array([[5675, 107],
[53, 6298]])

The confusion matrix results show 5675 non-phishing emails were correctly classified as non-phishing (true negatives), 107 non-phishing emails were incorrectly classified as phishing (false positives), 53 phishing emails were incorrectly classified as non-phishing (false negatives), and 6298 phishing emails were correctly classified as phishing (true positives). The logistic linear regression model (non-ensemble model) achieved a ROC-AUC performance of 0.9852 (98.52%). Table 2 below, shows the Logistic Liner Regression Classification report.

Table 2. Logistic Liner Regression Classification Report

	precision	**recall**	**f1-score**	**support**
1	0.98	0.99	0.99	6351
0	0.99	0.98	0.99	5782
accuracy			0.99	12133
macro avg	0.99	0.99	0.99	12133
weighted avg	0.99	0.99	0.99	12133

F1-Score: 0.9853749616291607
ROC-AUC 0.9852284556489306

Traditional Models

These models include linear layers models, transformer-based models, generative pre-training transformer models, and SHAP.

Linear layers

Conventional models assessed included Mlp4LayerClassifier, Mlp8LayerClassifier, LstmClassifier, CnnClassifier, and CnnLstmClassifier. The following table shows the performance of the models. On average, the models achieved between 76.6% and 96.8% accuracy levels. It showed a fairly good accuracy but not the best. Table 3 below, shows the results from the tested conventional models.

Table 3. Conventional Models

Model	Train Loss	Train Accuracy	Val Loss	Val Accuracy
Mlp4LayerClassifier	0.012	0.779	0.017	0.896
Mlp8LayerClassifier	0.016	0.919	0.013	0.938
LstmClassifier	0.032	0.745	0.031	0.766
CnnClassifier	0.007	0.963	0.006	0.968
CnnLstmClassifier	0.027	0.791	0.011	0.945

Note. Conventional models with 1 EPOCHS and 3 EPOCHS average performance

Transformer-based models

These models included BERT, BERT+LSTM, BERT+CNN, BERT+LSTM+CNN. On average they achieved an accuracy level of between 50%-53%. It showed the worst performance. Table 4 below, shows the results from the transformer based models.

Table 4. Transformer-Based Models

Model	Train Loss	Train Accuracy	Val Loss	Val Accuracy
BertClassifier	0.043	0.515	0.043	0.533
BertLstmClassifier	0.043	0.531	0.043	0.535
BertCNNLstmClassifier	0.043	0.529	0.043	0.536
BertCNNClassifier	0.044	0.513	0.043	0.517

Note. Transformer-based models with 1 EPOCHS and 3 EPOCHS (Average) performance

Generative Pre-Training(GPT-2) Transformers Based Models

The models evaluated include GPT2Classifier, GPT2CNNClassifier, PT2LstmClassifier, and GPT2CNNLstmClassifier. The GPT, GPT+LSTM, GPT+CNN, and GPT+LSTM+CNN are the recommended models since they show the best performance of an average of 99% val accuracy. Table 5 below, shows the results of the GPT-2 transformer-based models.

Table 5. Generative Pre-Training(GPT-2) Transformers-Based Models

Model	Train Loss	Train Accuracy	Val Loss	Val Accuracy
GPT-2Classifier	0.002	0.995	0.003	0.992
GPT-2CNNClassifier	0.002	0.996	0.003	0.993
GPT-2LstmClassifier	0.002	0.995	0.002	0.994
GPT-2CNNLstmClassifier	0.002	0.994	0.003	0.993

Note. Generative Pre-Training(GPT) Transformer-based models with 1 EPOCHS and 3 EPOCHS (Average) performance

SHAP

For additional visuals, SHAP was added to traditional models, BERT, and GPT. The results showed CnnClassifier Train Loss: 0.007, Train Accuracy: 0.963, Val Loss: 0.005, Val Accuracy: 0.971.

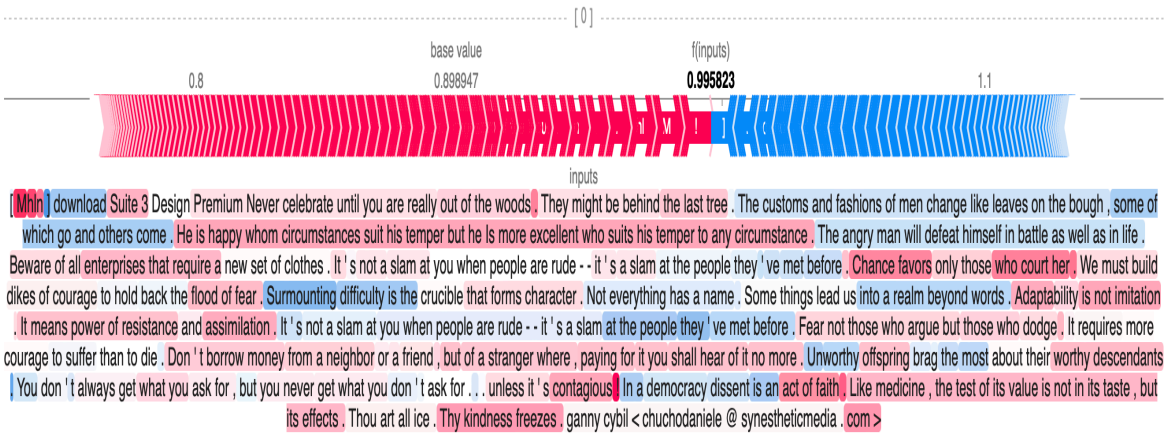

Figure 5. SHAP Conventional Model

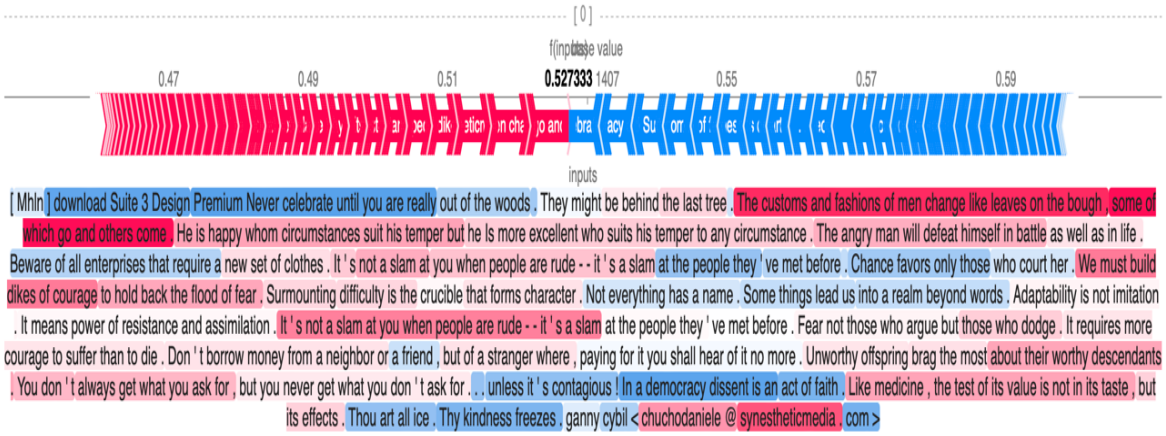

Figure 6. SHAP BERT Model

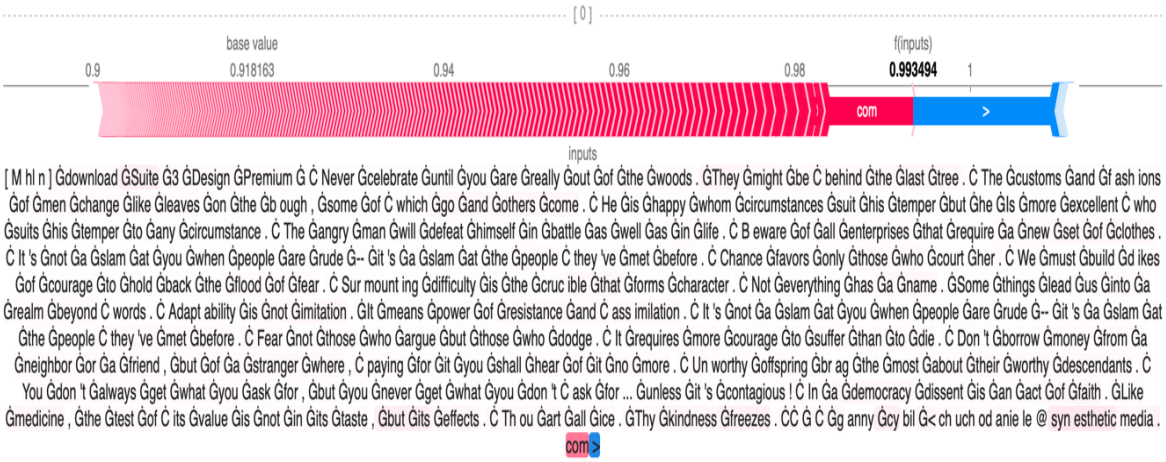

Figure 7. SHAP GPT-2 Model

Table 6. Summary of Research Questions and Hypothesis

Research Questions and Hypothesis	Results
RQ1. What is the performance of base model algorithms (random forest classifier and logistic linear regression) in email phishing detection?	Base Models (Random Forest Classifier and Logistic Regression Model) achieved an average accuracy of 98%. Random forest classifier (ensemble model) achieved a ROC-AUC performance score of 0.9865 (98.65%). The logistic linear regression model (non-ensemble model) achieved a ROC-AUC performance of 0.9852 (98.52%).
$H1_0$. There is no statistically significant difference in performance between the random forest classifier and logistic regression algorithms for detecting email phishing.	There is no statistically significant difference in performance (98.65% and 98.52%)
$H1_a$. The performance of the random forest classifier and logistic regression algorithms for detecting email phishing is significantly different.	Not true
RQ2. What is the performance of traditional machine learning algorithm models (linear layers' models, transformer-based models, generative pre-training transformer models) in email phishing detection?	Traditional ML algorithms can predict email phishing in varying degrees (performance: Linear layers models – between 76.6% and 96.8%, transformer-based models - 50% -53%, and GPT - 99% val accuracy).

Research Questions and Hypothesis	Results
$H2_0$. There is no significant relationship between the performance of proposed models (linear layers' models, transformer-based models, generative pre-training transformer models) and email phishing detection.	Not true.
$H2_a$: Proposed models (linear layers' models, transformer-based models, generative pre-training transformer models) perform differently for email phishing detection.	There is a significant relationship between traditional ML and email phishing detection (Performances vary significantly from 50% (worst) to 99% (best))
RQ3. How does the performance of base models compare to the proposed models?	Base models performed relatively better than traditional ML models
$H3_0$. There is no statistically significant difference in the performance of base models (random forest classifier and logistic regression) and proposed models (linear layers' models, transformer-based models, generative pre-training transformer models) in email phishing detection.	Not true
$H3_a$. There is a statistically significant difference in the performance of base models (random forest classifier and logistic regression) and proposed models (linear layers' models, transformer-based models, generative pre-training transformer models) in email phishing detection.	There is a statistical difference between the models with base models achieving an average of 98% while traditional ones achieved between 50% and 99%.
RQ4. What are the key factors influencing the effectiveness of different machine learning models in email phishing detection?	Different feature sets, model architectures, and hyperparameters could have influenced model effectiveness
$H4_0$. There are different factors that significantly impact the effectiveness of various machine learning models used for email phishing detection.	Not true
$H4_a$. Various factors such as feature sets, model architectures, and hyperparameters significantly influence the effectiveness of different machine learning models utilized for email phishing detection.	Different feature sets, model architectures, and hyperparameters could have influenced model effectiveness considering adding EPOCHS improved model performance

Note: Summary table of research questions and hypothesis with corresponding results obtained.

4.4 Statistical Test – Analysis of Model Performance Metrics

4.4.1 Training Loss

The analysis reveals significant differences in mean loss among the three groups.

The results of the statistical analysis indicate significant differences ($p < 0.05$) in the

mean training loss and accuracy across all three groups. Specifically, GPT-2 exhibits the lowest mean loss, indicating its superior performance in this regard. Thus, GPT-2 emerges as the preferred choice based on training loss.

Table 6. Train Loss

Group 1	Group 2	meandiff	p-adj	lower	upper	reject
BERT	GPT-2	-0.0412	0.0	-0.054	-0.0285	True
BERT	SNN	-0.0244	0.0007	-0.0366	-0.0123	True
GPT-2	SNN	0.0168	0.009	0.0047	0.0289	True

Note. F_onewayResult(statistic=39.58377118232471, pvalue=1.7740506544452924e-05)
Multiple Comparison of Means - Tukey HSD, FWER=0.05

4.4.2 Training Accuracy

Similar to the training loss findings, significant differences are observed in mean accuracy across the three groups. The statistical analysis reveals significant differences ($p < 0.05$) in both mean training loss and accuracy across the three groups. Notably, GPT-2 demonstrates the highest mean accuracy, underscoring its efficacy. Therefore, GPT-2 stands out as the optimal choice considering training accuracy.

Table 7. Train Accuracy

Group 1	Group 2	meandiff	p-adj	lower	upper	reject
BERT	GPT-2	0.473	0.0	0.3554	0.5906	True
BERT	SNN	0.3174	0.0	0.2059	0.4289	True
GPT-2	SNN	-0.1556	0.0085	-0.2671	-0.0441	True

Note. F_onewayResult(statistic=63.57143514501866, pvalue=2.061264988816717e-06)
Multiple Comparison of Means - Tukey HSD, FWER=0.05

4.4.3 Validation Loss

Analogous to the training metrics, there are discernible differences in mean loss among the three groups during validation. Statistically significant differences ($p < 0.05$) were detected in the mean validation loss, with GPT-2 models again exhibiting the lowest value. GPT-2 displays the lowest mean loss, solidifying its superiority in this aspect. Consequently, GPT-2 emerges as the preferred selection based on validation loss.

Table 8. Validation Loss

Group 1	Group 2	meandiff	p-adj	lower	upper	reject
BERT	GPT-2	-0.0402	0.0	-0.0519	-0.0286	True
BERT	SNN	-0.0274	0.0001	-0.0384	-0.0164	True
GPT-2	SNN	0.0128	0.0238	-0.0018	0.0239	True

Note. F_onewayResult(statistic=47.2700802461881, pvalue=8.009108691294093e-06)
Multiple Comparison of Means - Tukey HSD, FWER=0.05

4.4.4 Validation Accuracy

Although the statistical analysis indicates comparable accuracies between SNN and GPT models, GPT still exhibits a higher mean validation accuracy than SNN models. While the validation accuracy of SNN models was statistically comparable to GPT-2 models ($p > 0.05$), the latter still achieved a higher mean accuracy. Furthermore, considering GPT's superiority across other metrics, it remains the superior choice overall. Thus, it is conclusive that GPT-2 models represent the optimal choice.

Table 9. Validation Accuracy

Group 1	Group 2	meandiff	p-adj	lower	upper	reject
BERT	GPT-2	0.4628	0.0	0.3634	0.5621	True
BERT	SNN	0.3723	0.0	0.2781	0.4666	True
GPT-2	SNN	-0.0904	0.06	-0.1847	0.0039	False

Note. F_onewayResult(statistic=93.14606306492142, pvalue=3.431512652759489e-07)
Multiple Comparison of Means - Tukey HSD, FWER=0.05

In this research, analysis of variance (ANOVA) was selected as the statistical method, measuring three or more groups to see if at least one group's mean differs. It tests for significant differences between group means in a dataset. When you have

multiple groups and find a substantial effect with ANOVA, it doesn't tell you which groups are different from each other. This is where post-hoc tests come in.

Post-hoc tests are follow-up analyses conducted after an ANOVA if the initial results show statistical significance. These tests help determine exactly which group means are significantly different from each other. Post-hoc tests control for multiple comparisons to reduce the risk of committing Type I errors (incorrectly rejecting a true null hypothesis). There are several types of post-hoc tests, each with its assumptions and conditions of use. Some commonly used post hoc tests include:

Tukey's HSD (Honest Significant Difference): Compares all possible pairs of means and is considered one of the most robust post-hoc tests, maintaining a constant Type I error rate.

Bonferroni Correction: Adjusts the significance level (\(\alpha\)) based on the number of comparisons made. It's a conservative approach that reduces the risk of Type I errors but increases the risk of Type II errors (failing to reject a false null hypothesis).

Scheffé's Test: This test offers flexibility in hypothesis testing by allowing for comparing any set of means. It is beneficial to test hypotheses that must be specified before the data is examined.

ANOVA with post-hoc tests is a powerful statistical approach for comparing multiple group means and understanding their specific differences. While ANOVA determines if there is a significant effect, post-hoc tests identify where those effects are by comparing specific group pairs, helping researchers draw more detailed and accurate

conclusions from their data. While other statistical tests, like t-tests, were indeed possibilities, ANOVA was used because of the type of data analyzed.

4.5 Summary

This section has presented the results of the data analysis. The results have revealed that the performance of the models varies significantly. For instance, generative transformer-based models (GPT, GPT+LSTM, GPT+CNN, GPT+LSTM+CNN) are the recommended models since they show the best performance of an average of 99% val accuracy. On the other hand, BERT models which are transformer-based models achieved the worst performance of 50% -53% accuracy. Linear layers models achieved between 76.6% and 96.8% accuracy levels on average. Base Models that include Random Forest Classifier, and Logistic Regression Model achieved average accuracy of 98%.

Chapter 5—Discussion and Conclusions

5.1 Discussion

This chapter focused on analyzing the performance of various machine learning models for email phishing detection. The study encompassed a range of models, from traditional algorithms to advanced transformer-based and generative pre-training transformer models. Notably, the results of these models exhibited substantial variance, with generative transformer-based models emerging as the most effective, showcasing an average validation accuracy of 99%. This highlights the effectiveness of these models in leveraging pre-trained knowledge for email content analysis. Base models, Random Forest and Logistic Regression displayed average accuracy levels of 98%. Linear layer models (LSTM, CNN, LSTM+CNN) performed moderately well, achieving an average accuracy between 76.6% and 96.8%. Conversely, transformer-based models like BERT demonstrated lower accuracy levels, ranging from 50% to 53%. The comparative analysis between base models, traditional machine learning algorithms, and transformer-based models underscores the significance of leveraging advanced architectures for enhanced detection accuracy.

These findings suggest that generative pre-trained transformers hold immense potential for combating email phishing due to their superior ability to analyze the intricacies of email content, structure, and sender information. By capturing intricate patterns and nuances within email content, structure, and sender information, these models offer superior adaptability and accuracy compared to conventional methods. However, the inferior performance of BERT models suggests that not all transformer-

based architectures are equally effective in addressing the complexities of phishing email detection. This prompts further exploration into refining and optimizing transformer architectures tailored specifically for cybersecurity applications.

5.2 Conclusions

In conclusion, this study successfully evaluated the critical role of leveraging advanced machine learning models in combating the escalating threat of email phishing attacks. Through a comprehensive evaluation of various model architectures, this research unveiled generative transformer-based models (GPT-2 based), exhibit unparalleled performance, achieving an average validation accuracy of 99%. This reinforces the potential of leveraging pre-trained knowledge for in-depth email analysis. The research also validates the use of base models (Random Forest and Logistic Regression) and underscores the limitations of standard transformer models (BERT) in this context. While linear layer models (LSTM, CNN, LSTM+CNN) demonstrated moderate effectiveness, generative pre-trained transformers emerged as the superior choice. These findings pave the way for the development of robust email phishing detection systems utilizing the capabilities of generative pre-trained transformers. By enhancing the accuracy and adaptability of phishing email detection, these advanced models offer promising avenues for mitigating the detrimental impacts of cybercrime on individuals and organizations.

5.3 Contributions to Body of Knowledge

1. This research contributes to the advancement of cybersecurity by providing empirical evidence of the effectiveness of various machine learning models in

detecting email phishing attacks, with generative transformer-based models demonstrating superior performance.

2. Identified superior model for email phishing detection: This research pinpoints generative pre-trained transformer models (GPT-2 based) as the most effective approach for email phishing detection, achieving a 99% average accuracy.
3. Supported the use of transformer models for email security: This research contributes to the growing body of evidence supporting the effectiveness of transformer models for enhancing email security through advanced phishing detection capabilities.

5.4 Recommendations for Future Research

For future research endeavors, it is recommended to delve deeper into the optimization of transformer-based architectures tailored specifically for email phishing detection. Additionally, exploring ensemble techniques that combine the strengths of multiple model architectures could further enhance detection accuracy and robustness. Moreover, the integration with existing email security systems could be evaluated, exploring seamless integration of these models with existing email security infrastructure for deployment and impact assessment can be a valuable next step. By pursuing these avenues of research, we can leverage the power of generative pre-trained transformers to create a more secure email environment for individuals and organizations.

References

Adebowale, M. A., Lwin, K. T., & Hossain, M. A. (2019). *Deep learning with convolutional neural network and long short-term memory for phishing detection* [Conference presentation]. 2019 13th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), Island of Ulkulhas, Maldives. https://doi.org/10.1109/SKIMA47702.2019.8982427

Adebowale, M. A., Lwin, K. T., & Hossain, M. A. (2023). Intelligent phishing detection scheme using deep learning algorithms. *Journal of Enterprise Information Management*, *36*(3), 747-766. https://doi.org/10.1108/JEIM-01-2020-0036

Akinyelu, A. & Adewumi, A. (2014). Classification of phishing email using random forest machine learning technique. *Journal of Applied Mathematics, 2014,* Article ID 425731. http://dx.doi.org/10.1155/2014/425731

Alauthman, M. (2020). Botnet spam e-mail detection using deep recurrent neural network. *International Journal of Emerging Trends in Engineering Research*, *8*(5), 1979-1986. https://doi.org/10.30534/ijeter/2020/83852020

Aldawood, H. & Skinner, G. (2019). Reviewing cyber security social engineering training and awareness programs—Pitfalls and ongoing issues. *Future Internet*, *11*(3), 73. https://doi.org/10.3390/fi11030073

Alhogail, A., & Alsabih, A. (2021). Applying machine learning and natural language processing to detect phishing email. *Computers & Security*, *110*, 102414. https://doi.org/10.1016/j.cose.2021.102414

Aljofey, A., Jiang, Q., Qu, Q., Huang, M., & Niyigena, J. P. (2020). An effective phishing detection model based on character level convolutional neural network from URL. *Electronics*, *9*(9), 1514. https://doi.org/10.3390/electronics9091514

Baki, S., & Verma, R. M. (2022). Sixteen years of phishing user studies: What have we learned? *IEEE Transactions on Dependable and Secure Computing*, *20*(2), 1200-1212. https://doi.org/10.1109/TDSC.2022.3151103

Barlow, L., Bendiab, G., Shiaeles, S., & Savage, N. (2020). A novel approach to detect phishing attacks using binary visualisation and machine learning. In *2020 IEEE World Congress on Services (SERVICES)* Bejing, China, 177-182. https://doi.org/10.1109/SERVICES48979.2020.00046

Bashar, D. A. (2019). Survey on evolving deep learning neural network architectures. *Journal of Artificial Intelligence and Capsule Networks*, *1*(2), 73-82. https://doi.org/10.36548/jaicn.2019.2.003

Basit, A., Zafar, M., Javed, A. R., & Jalil, Z. (2020). *A novel ensemble machine learning method to detect phishing attack* [Conference presentation]. 2020 IEEE 23rd International Multitopic Conference (INMIC), Bahawalpur, Pakistan, 1-5. https://doi.org/10.1109/INMIC50486.2020.9318210

Blancaflor, E. B., Ayap, N. R. D., Dabu, A. C.., Reyes, A. M., Ventura, A. B. K., Trinidad, G. P. (2022). *Beat the bait: A case study on phishing attack.* [Conference presentation]. 3rd African International Conference on Industrial Engineering and Operations Management, Nsukka, Nigeria. https://doi.org/10.46254/af03.20220135

Bountakas, P., & Xenakis, C. (2023). HELPHED: Hybrid ensemble learning phishing email detection. *Journal of Network and Computer Applications*, *210*, 103545. https://doi.org/10.1016/j.jnca.2022.103545

Breiman, L. & Cutler, A. (2007). *Random forests; classification/clustering - description.* Berkeley.edu, Department of Statistics homepage. Retrieved from: http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.

Brindha, R., Nandagopal, S., Azath, Sathana, V., Joshi, G. P., & Kim, S. (2023). Intelligent deep learning based cybersecurity phishing email detection and classification. *Computers, Materials & Continua*, *74*(3). http://dx.doi.org/10.32604/cmc.2023.030784

Brown, J., McDonald, M., Besse, C., Manson, P., McDonald, R., Rohatinsky, N., & Singh, M. (2021). Nursing students' academic success factors: A quantitative cross-sectional study. *Nurse Educator*, *46*(2), E23-E27. https://doi.org/10.1097/NNE.0000000000000882

Buckley, J., Lottridge, D., Murphy, J. G., & Corballis, P. M. (2023). Indicators of employee phishing email behaviours: Intuition, elaboration, attention, and email typology. *International Journal of Human-Computer Studies*, *172*, 102996. https://doi.org/10.1016/j.ijhcs.2023.102996

Butavicius, M., Taib, R., & Han, S. J. (2022). Why people keep falling for phishing scams: The effects of time pressure and deception cues on the detection of phishing emails. *Computers & Security*, *123*, 102937. https://doi.org/10.1016/j.cose.2022.102937

Butt, U. A., Amin, R., Aldabbas, Mohan, S., Alouffi, B., & Ahmadian, A. (2023). Cloud-based email phishing attack using machine and deep learning algorithm. *Complex & Intelligent Systems*, *9*(3), 3043-3070. https://doi.org/10.1007/s40747-022-00760-3

Catal, C., Giray, G., Tekinerdogan, B., Kumar, S., & Shukla, S. (2022). Applications of deep learning for phishing detection: A systematic literature review. *Knowledge and Information Systems, 64*, 1457 - 1500. https://doi.org/10.1007/s10115-022-01672-x

Chai, Y., Zhou, Y. Li, W., & Jiang, Y. (2021). An explainable multi-modal hierarchical attention model for developing phishing threat intelligence. *IEEE Transactions on Dependable and Secure Computing*, *19*(2), 790-803. https://doi.org/10.1109/TDSC.2021.3119323

Chandra, A., & Khatri, S. K. (2019). *Spam SMS filtering using recurrent neural network and long short term memory.* [Conference presentation]. 2019 4th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India. https://doi.org/10.1109/ISCON47742.2019.9036269

Clarke, S. L., Parmesar, K., Saleem, M. A., & Ramanan, A. V. (2022). Future of machine learning in paediatrics. *Archives of Disease in Childhood*, *107*(3), 223-228.

Creswell, J. W., & Creswell, J. D. (2022). *Research design* (6th ed.). SAGE Publications.

Devlin, J. Chang ,M.-W., Lee, K., and Toutanova, K., (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding.* [Conference presentation]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dilhara, B. (2021). Phishing URL detection: A novel hybrid approach using long short-term memory and gated recurrent units. *International Journal of Computer Applications*, *975*, 8887.

Do, N. Q., Selamat, A., Krejcar, O., Herrera-Viedma, E., & Fujita, H. (2022). Deep learning for phishing detection: Taxonomy, current challenges, and future directions. *IEEE Access, 10*, 36429-36463. https://doi.org/10.1109/ACCESS.2022.3151903

Eckhardt, R., & Bagui, S. (2021). Convolutional neural networks and long short term memory for phishing email classification. *International Journal of Computer Science and Information Security*, *19*(5). https://doi.org/10.5281/zenodo.4898109

Federal Trade Commission. (2023, February 23). *New FTC data show consumers reported losing nearly $8.8 billion to scams in 2022*. [Press Release]. Federal Trade Commission. Retrieved from: https://www.ftc.gov/news-events/news/press-releases/2023/02/new-ftc-data-show-consumers-reported-losing-nearly-88-billion-scams-2022 (Accessed: 22 March 2024)

Ferreira, A., & Teles, S. (2019). Persuasion: How phishing emails can influence users and bypass security measures. *International Journal of Human-Computer Studies*, *125*, 19-31. https://doi.org/10.1016/j.ijhcs.2018.12.004

Gangavarapu, T., Jaidhar, C. D., & Chanduka, B. (2020). Applicability of machine learning in spam and phishing email filtering: Review and approaches. *Artificial Intelligence Review*, *53*, 5019-5081. https://doi.org/10.1007/s10462-020-09814-9

Ghose, N., Lazos, L., Rozenblit, J., & Breiger, R. (2019). *Multimodal graph analysis of cyber attacks* [Conference presentation]. 2019 Spring Simulation Conference (SpringSim), Tucson, AZ, USA. https://doi: 10.23919/SpringSim.2019.8732851

Hakizimana, D., Nisingizwe, M. P., Logan, J., & Wong, R. (2019). Identifying risk factors of anemia among women of reproductive age in Rwanda–A cross-sectional study using secondary data from the Rwanda demographic and health survey 2014/2015. *BMC Public Health*, *19*(1662), 1-11. https://doi.org/10.1186/s12889-019-8019-z

Halgaš, L., Agrafiotis, I., & Nurse, J. R. (2020). *Catching the phish: Detecting phishing attacks using recurrent neural networks (RNNs.)* [Conference paper]. In: You, I. (eds) Information Security Applications. WISA 2019. Lecture Notes in Computer Science, vol 11897. Springer, Cham. https://doi.org/10.1007/978-3-030-39303-8_17

Hao, T., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2020). Testing whether ensemble modelling is advantageous for maximising predictive performance of species distribution models. *Ecography*, *43*(4), 549-558. https://doi.org/10.1111/ecog.04890

Hao, J., Wang, X., Yang, S., & Gao, H. (2023). Intelligent Simulation Technology Based on RCS Imaging. *Applied Sciences*, *13*(18), 10119.

Haynes, K., Shirazi, H., & Ray, I. (2021). Lightweight URL-based phishing detection using natural language processing transformers for mobile devices. *Procedia Computer Science*, *191*, 127-134. https://doi.org/10.1016/j.procs.2021.07.040

Hijji, M., & Alam, G. (2021). A multivocal literature review on growing social engineering based cyber-attacks/threats during the COVID-19 pandemic: Challenges and prospective solutions. *IEEE Access*, *9*, 7152-7169. https://doi.org/10.1109/ACCESS.2020.3048839

Indrasiri, P. L., Halgamuge, M. N., & Mohammad, A. (2021). Robust ensemble machine learning model for filtering phishing URLs: Expandable random gradient stacked voting classifier (ERG-SVC). *IEEE Access*, *9*, 150142-150161. https://doi.org/10.1109/ACCESS.2021.3124628

Isik, S., Kurt, Z., Anagun, Y., & Ozkan, K. (2020). Recurrent neural networks for spam e-mail classification on an agglutinative language. *International Journal of Intelligent Systems and Applications in Engineering*, *8*(4), 221-227. https://doi.org/10.18201/ijisae.2020466316

Islam, M. S., & Zhang, L. (2024). A Review on BERT: Language Understanding for Different Types of NLP Task.

Jampen, D., Gür, G., Sutter, T., & Tellenbach, B. (2020). Don't click: Towards an effective anti-phishing training. A comparative literature review. *Human-centric Computing and Information Sciences*, *10*(33), 1-41. https://doi.org/10.1186/s13673-020-00237-7

Jemili, F. & Korbaa, O. (2023). *Hybrid collaborative intrusion detection system based on blockchain & machine learning* [Preprint]. ResearchGate. http://dx.doi.org/10.21203/rs.3.rs-3262119/v1

Jena, B., Saxena, S., Nayak, G. K., Saba, L., Sharma, N., & Suri, J. S. (2021). Artificial intelligence-based hybrid deep learning models for image classification: The first narrative review. *Computers in Biology and Medicine*, *137*, 104803. https://doi.org/10.1016/j.compbiomed.2021.104803

Kaddoura, S., Alfandi, O., & Dahmani, N. (2020). *A spam email detection mechanism for English language text emails using deep learning approach.* [Conference presentation]. 2020 IEEE 29th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), Bayonne, France. https://doi.org/10.1109/WETICE49692.2020.00045

Kandel, B. (2020). Qualitative versus quantitative research. *Journal of Product Innovation Management, 32*(5), 658.

Keary, T. (2022, September 16). What Uber's data breach reveals about social engineering. *VentureBeat*. Retrieved from: https://venturebeat.com/security/what-ubers-data-breach-reveals-about-social-engineering/ (Accessed: 26 March 2024)

Khan, H., Alam, M., Al-Kuwari, S., & Faheem, Y. (2021). *Offensive AI: Unification of email generation through GPT-2 model with a game-theoretic approach for spear-phishing attacks.* [Online Conference]. Competitive Advantage in the Digital Economy (CADE 2021). https://doi: 10.1049/icp.2021.2422.

Klein, T., & Nabi, M. (2019). Learning to answer by learning to ask: Getting the best of GPT-2 and BERT worlds. *arXiv preprint arXiv:1911.02365*. https://doi.org/10.48550/arXiv.1911.02365

Kumar, A., Chatterjee, J. M., & Díaz, V. G. (2020). A novel hybrid approach of SVM combined with NLP and probabilistic neural network for email phishing. *International Journal of Electrical and Computer Engineering*, *10*(1), http://dx.doi.org/10.11591/ijece.v10i1.pp486-493

Kumar, M., Bajaj, K., Sharma, B., & Narang, S. (2022). A comparative performance assessment of optimized multilevel ensemble learning model with existing classifier models. *Big Data*, *10*(5), 371-387. https://doi.org/10.1089/big.2021.0257

Kwak, Y., Lee, S., Damiano, A., & Vishwanath, A. (2020). Why do users not report spear-phishing emails? *Telematics and Informatics*, *48*, 101343. https://doi.org/10.1016/j.tele.2020.101343

Lain, D., Kostiainen, K., & Čapkun, S. (2022, May). *Phishing in organizations: Findings from a large-scale and long-term study.* [Symposium]. 2022 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA. https://doi.org/10.1109/SP46214.2022.9833766

Lee, Y., Saxe, J., & Harang, R. (2020). CATBERT: Context-aware tiny BERT for detecting social engineering emails. *arXiv preprint arXiv:2010.03484*. https://doi.org/10.48550/arXiv.2010.03484

Lemay, D. J., Basnet, R. B., & Doleck, T. (2020). Examining the relationship between threat and coping appraisal in phishing detection among college students. *Journal of Internet Services and Information Security, 10*(1), 38-49. https://dx.doi.org/10.22667/JISIS.2020.02.29.038

Leys, C., Delacre, M., Mora, Y. L., Lakens, D., & Ley, C. (2019). How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. *International Review of Social Psychology, 32*(1). https://doi.org/10.5334/irsp.289

Li, T., Xu, H., Liu, Z., Dong, Z., Liu, Q., Li, J., Fan, S., & Sun, X. (2022). A spatiotemporal multi-feature extraction framework for opinion mining. *Neurocomputing, 490*, 337-346. https://doi.org/10.1016/j.neucom.2021.11.098

Li, Q., Cheng, M., Wang, J., & Sun, B. (2020). LSTM based phishing detection for big email data. *Institute of Electrical and Electronics Engineers* (*IEEE) Transactions on Big Data*, *8*(1), 278-288. https://doi.org/10.1109/TBDATA.2020.2978915

Lin, T., Capecci, D. E., Ellis, D. M., Rocha, H. A., Dommaraju, S., Oliveira, D. S., & Ebner, N. C. (2019). Susceptibility to spear-phishing emails: Effects of internet user demographics and email content. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *26*(5), 1-28. https://doi.org/10.1145/3336141

Liu, X., Zhang, X., & Baziar, A. (2023). Hybrid machine learning and modified teaching learning-based english optimization algorithm for smart city communication. *Sustainability*, *15*(15), 11535.

Liu, W., Qiu, J. L., Zheng, W. L., & Lu, B. L. (2021). Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, *14*(2), 715-729. https://doi.org/10.1109/TCDS.2021.3071170

Magdy, S., Abouelseoud, Y., & Mikhail, M. (2022). Efficient spam and phishing emails filtering based on deep learning. *Computer Networks*, *206*, 108826. https://doi.org/10.1016/j.comnet.2022.108826

Manjunath, C. R., Rathor, K., Kulkarni, N., Patil, P. P., Patil, M. S., & Singh, J. (2022). Cloud based DDOS attack detection using machine learning architectures: Understanding the potential for scientific applications. *International Journal of Intelligent Systems and Applications in Engineering*, *10*(2s), 268-271.

Mani, S., Gunasekaran, G., & Geetha, S. (2023). Email spam detection using gated recurrent neural network. *International Journal of Prograssive Research in Engineering Management and Science (IJPREMS)*, *3*(4), 90-99.

Mansoor, R. A. Z. A., Jayasinghe, N. D., & Muslam, M. M. A. (2021). *A comprehensive review on email spam classification using machine learning algorithms.* [Conference presentation]. 2021 International Conference on Information Networking (ICOIN), Jeju Island, South Korea. https://doi.org/10.1109/ICOIN50884.2021.9334020

McAlaney, J., & Hills, P. J. (2020). Understanding phishing email processing and perceived trustworthiness through eye tracking. *Frontiers in Psychology*, *11*, 1756. https://doi.org/10.3389/fpsyg.2020.01756

Mellinger, C. D., & Hanson, T. A. (2022). Considerations of ecological validity in cognitive translation and interpreting studies. *Translation, Cognition & Behavior*, *5*(1), 1-26. https://doi.org/10.1075/tcb.00061.mel

Melnyk, L. & Feld, L. (2022). Sentiment analysis and stance detection on German Youtube comments on gender diversity. *Journal of Computer-Assisted Linguistic Research, 6*, 59-86 https://doi.org/10.4995/jclr.2022.18224

Mittal, A., Engels, D. D., Kommanapalli, H., Sivaraman, R., & Chowdhury, T. (2022). Phishing detection using natural language processing and machine learning. *SMU Data Science Review*, *6*(2). https://scholar.smu.edu/datasciencereview/vol6/iss2/14

Mohajan, H. K. (2020). Quantitative research: A successful investigation in natural and social sciences. *Journal of Economic Development, Environment and People*, *9*(4), 50-79. https://doi.org/10.26458/jedep.v9i4.679

Mughaid, A., AlZu'bi, S., Hnaif, A., Taamneh, S., Alnajjar, A., & Elsoud, E. A. (2022). An intelligent cyber security phishing detection system using deep learning techniques. *Cluster Computing*, *25*(6), 3819-3828. https://doi.org/10.1007/s10586-022-03604-4

Nar, M., & Arslankaya, S. (2022). Passenger demand forecasting for railway systems. *Open Chemistry*, *20*(1), 105-119.

Nifakos, S., Chandramouli, K., Nikolaou, C. K., Papachristou, P., Koch, S., Panaousis, E., & Bonacina, S. (2021). Influence of human factors on cyber security within healthcare organisations: A systematic review. *Sensors*, *21*(15), 5119. https://doi.org/10.3390/s21155119

Otieno, D. O., Namin, A. S., & Jones, K. S. (2023, June). *The application of the BERT transformer model for phishing email classification.* [Conference presentation] 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC), Torina, Italy. doi: 10.1109/COMPSAC57700.2023.00198.

Parsons, K., Butavicius, M., Delfabbro, P., & Lillie, M. (2019). Predicting susceptibility to social influence in phishing emails. *International Journal of Human-Computer Studies*, *128*, 17-26. https://doi.org/10.1016/j.ijhcs.2019.02.007

Page, M. J., Moher, D., & McKenzie, J. E. (2022). Introduction to PRISMA 2020 and implications for research synthesis methodologists. *Research Synthesis Methods*, *13*(2), 156-163. https://doi.org/10.1002/jrsm.1535

Qin, H., Gong, R., Liu, X., Bai, X., Song, J., & Sebe, N. (2020). Binary neural networks: A survey. *Pattern Recognition*, *105*, 107281.

Rafat, K. F., Xin, Q., Javed, A. R., Jalil, Z., & Ahmad, R. Z. (2022). Evading obscure communication from spam emails. *Mathematcial Biosciences and Engineering*, *19*(2), 1926-1943. http://dx.doi.org/10.3934/mbe.2022091

Rahman, S. E., & Ullah, S. (2020). *Email spam detection using bidirectional long short term memory with convolutional neural network*. [Symposium]. 2020 IEEE Region 10 Symposium (TENSYMP*)*, Dhaka, Bangladesh. https://doi.org/10.1109/TENSYMP50017.2020.9230769

Roy, S. S., Awad, A. I., Amare, L. A., Erkihun, M. T., & Anas, M. (2022). Multimodal phishing URL detection using LSTM, bidirectional LSTM, and GRU models. *Future Internet*, *14*(11), 340. https://doi.org/10.3390/fi14110340

Salloum, S., Gaber, T., Vadera, S., & Shaalan, K. (2022). A systematic literature review on phishing email detection using natural language processing techniques. *IEEE Access*, *10*, 65703-65727. https://doi.org/10.1109/ACCESS.2022.3183083

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*, 160. https://doi.org/10.1007/s42979-021-00592-x

Schoonenboom, J. (2023). The fundamental difference between qualitative and quantitative data in mixed methods research. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research, 24*, 1. https://doi.org/10.17169/fqs-24.1.3986

Schutt, R. K. (2019). Quantitative methods. In G. Ritzer, W. W. Murphy (Eds.), *The Wiley Blackwell Companion to Sociology*, (2nd ed., pp. 39-56). John Wiley & Sons. https://doi.org/10.1002/9781119429333.ch3

Sethi, M., Chandra, S., Chaudhary, V., & Dahiya, Y. (2022). Spam email detection using machine learning and neural networks. [Conference paper]. In Shakya, S., Balas, VE., Kamolphiwong, S., Du, LK. (eds), *Sentimental Analysis and Deep Learning:Advances in Intelligent Systems and Computing, 1408.*. Springer, Singapore. https://doi.org/10.1007/978-981-16-5157-1_22

Setia, M. S. (2016). Methodology series module 3: Cross-sectional studies. *Indian Journal of Dermatology*, *61*(3), 261-264.

Shahbaznezhad, H., Kolini, F., & Rashidirad, M. (2021). Employees' behavior in phishing attacks: What individual, organizational, and technological factors matter? *Journal of Computer Information Systems*, *61*(6), 539-550. https://doi.org/10.1080/08874417.2020.1812134

Sheneamer, A. (2021). Comparison of deep and traditional learning methods for email spam filtering. *International Journal of Advanced Computer Science and Applications*, *12*(1). https://dx.doi.org/10.14569/IJACSA.2021.0120164

Shahrivari, V., Darabi, M. M., & Izadi, M. (2020). Phishing detection using machine learning techniques. *arXiv preprint arXiv:2009.11116*. https://doi.org/10.48550/arXiv.2009.11116

Shrestha, A., & Mahmood, A. (2019). Review of deep learning algorithms and architectures. *IEEE access*, *7*, 53040-53065. https://doi.org/10.1109/ACCESS.2019.2912200

Siedlecki, S. L. (2020). Understanding descriptive research designs and methods. *Clinical Nurse Specialist*, *34*(1), 8-12. https://doi.org/10.1097/NUR.0000000000000493

Singh, K., Aggarwal, P., Rajivan, P., & Gonzalez, C. (2019). Training to detect phishing emails: Effects of the frequency of experienced phishing emails. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 63*(1), 453-457. https://doi.org/10.1177/1071181319631355

Soon, G. K., Chiang, L. C., On, C. K., Rusli, N. M., & Fun, T. S. (2020). *Comparison of ensemble simple feedforward neural network and deep learning neural network on phishing detection.* [Conference paper]. In Alfred, R., Lim, Y., Haviluddin, H, On, C. (eds). Computational Science and Technology: Lecture Notes in Electrical Engineering, 63. Springer, Singapore. https://doi.org/10.1007/978-981-15-0058-9_57

Staudemeyer, R. C., & Morris, E. R. (2019). Understanding LSTM--a tutorial into long short-term memory recurrent neural networks. *arXiv preprint arXiv:1909.09586*. https://doi.org/10.48550/arXiv.1909.09586

Stockemer, D. (2019). *Quantitative methods for the social sciences* (Vol. 50, p. 181). Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-99118-4

Sun, Y., Chong, N., & Ochiai, H. (2022). *Federated Phish Bowl: LSTM-Based Decentralized Phishing Email Detection*. [Conference presentation]. IEEE International Conference on Systems, Man, and Cybernetics (SMC), Prague, Czech Republic. doi: 10.1109/SMC53654

Swain, D., Chillur, N., Kava, M., & Satapathy, S. (2022). Intelligent system for detecting email spam messages using GRU. In Tiwari, S., Trivedi, M.C., Kolhe, M.L., Singh, B.K. (Eds). *Advances in Data and Information Sciences. Lecture Notes in netwoks and systems, 522*.Springer, Singapore. https://doi.org/10.1007/978-981-19-5292-0_7

Taherdoost, H. (2022). What are different research approaches? Comprehensive review of qualitative, quantitative, and mixed method research, their applications, types, and limitations. *Journal of Management Science & Engineering Research*, *5*(1), 53-63. https://doi.org/10.30564/jmser.v5i1.4538

Truong, C. K., Do, P. H., & Le, T. D. (2023). A comparative analysis of email phishing detection methods: a deep learning perspective. In A A. A. El-Latif, M. A. Hannad, Y. Maleh, B. B. Gupta, & W. Mazurczyk (Eds.), *Artificial Intelligence for Biometrics and Cybersecurity: Technology and Applications.* https://doi.org/10.1049/PBSE020E_ch7I:

Valecha, R., Mandaokar, P., & Rao, H. R. (2021). Phishing email detection using persuasion cues. *IEEE Transactions on Dependable and Secure Computing*, *19*(2), 747-756. https://doi.org/10.1109/TDSC.2021.3118931

Varoquaux, G., & Cheplygina, V. (2022). Machine learning for medical imaging: Methodological failures and recommendations for the future. [Conference presentation]. *NPJ Digital Medicine*, *5*(1), 48. http://dx.doi.org/10.1038/s41746-022-00592-y

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). A*ttention is all you need.* [Poster presentation] 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Weaver, B. W., Braly, A. M., & Lane, D. M. (2021). Training users to identify phishing emails. *Journal of Educational Computing Research*, *59*(6), 1169-1183. https://doi.org/10.1177/0735633121992516

Yang, L., Li, X., Fang, L., Wu, L., Hu, H., & Cheng, Y. (2023). Comparative efficacy of different acupuncture-related therapies for primary tinnitus: A systematic review and network meta-analysis protocol. *Journal of Pain Research*,*16,* 2217-2225. https://doi: 10.2147/JPR.S414622

Yeo, D. J., Pollack, C., Conrad, B N., & Price, G. R. (2024). Functional and representational differences between bilateral inferior temporal numeral areas. *Cortex, 171*, 113-135 doi: 10.1016/j.cortex.2023.08.018

Ziogas, A. N., Ben-Nun, T., Schneider, T., & Hoefler, T. (2021, June). *NPBench: A benchmarking suite for high-performance NumPy*. [Conference presentation]. Proceedings of the ACM International Conference on Supercomputing, 63-74. https://doi.org/10.1145/3447818.3460360

Zuraiq, A. A., & Alkasassbeh, M. (2019). *Review: Phishing detection approaches*. [Conferene presentation] 2019 2nd International Conference on new Trends in Computing Sciences (ICTCS), Amman, Jordan. https://doi.org/10.1109/ICTCS.2019.8923069

Appendix A

Github Link: https://github.com/bolaglorina/Transformer_GWUPraxis