**DEVELOPING AI TOOLS FOR DETECTING CYBER THREATS IN MULTIMODAL EMAIL ATTACHMENTS**

**BY**

**SALISU OLUWATOYIN SAMUEL**

**21/8788**

**A FINAL YEAR PROJECT SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE, COLLEGE OF PURE AND APPLIED SCIENCES, CALEB UNIVERSITY, IMOTA, LAGOS.**

**IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF BACHELOR OF SCIENCE (B.SC) DEGREE IN CYBERSECURITY.**

**JUNE 2025**

# DECLARATION

I, **SALISU OLUWATOYIN SAMUEL** do hereby declare that this project is entirely my work and composition. The work embodied in this project has not been submitted in candidature for any degree and is not concurrently being submitted for any other degree.

All references made to works of other persons have been duly acknowledged.

Signature: …………………………………….

Date: ………….................................................

# CERTIFICATION

This is to certify that the research work presented in this document entitled “**DEVELOPING AI**

**TOOLS FOR DETECTING CYBER THREATS IN MULTIMODAL EMAIL**

**ATTACHMENTS**” was conducted under my supervision and guidance. The content of this research represents the original work of the author and meets the academic standards required for such a study. I attest to the authenticity and integrity of the research findings and conclusion presented in this document.

**Prof. Moses Aregbesola \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(Project Supervisor) (Signature & Date)

**Dr. Adeniyi Akanni \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(Head of Department) (Signature & Date)

**Prof. Adetola Adesanya \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(Dean of COPAS) (Signature & Date)

**External Examiner \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(Signature & Date)

**DEDICATION**

This project is dedicated first and foremost to God Almighty, whose grace, wisdom, and strength have sustained me through every stage of this journey.

To my parents and siblings, your love, sacrifices, and unwavering support have been the foundation of my perseverance. Thank you for believing in me even when I doubted myself.

Finally, to every student who continues to chase excellence in the face of challenges, may this work be a reminder that persistence truly pays off.

**ACKNOWLEDGEMENT**

First and foremost, I return all glory to God Almighty, the giver of wisdom, strength, and grace. His presence has been my anchor, and without Him, this work would not have been possible.

I express my sincere appreciation to my supervisor, Prof. Aregbesola, for his unwavering support, constructive guidance, and insightful feedback. Your mentorship greatly shaped the direction and quality of this research, and I am truly grateful.

My heartfelt thanks go to all the lecturers in the Department of Cybersecurity. Your teachings, advice, and dedication laid the academic foundation that made this project achievable.

To my friends and colleagues, thank you for the collaboration, shared knowledge, and words of encouragement, and for being a source of strength through the challenges. The moments we shared made this experience both bearable and unforgettable.

To my family, thank you for your constant prayers, patience, love, and support. Your belief in me was the fuel that kept me moving forward, especially during the most difficult times.

Finally, to everyone who stood by me, offered a kind word, or silently cheered me on, thank you.

This project is more than a requirement; it is a reflection of growth, gratitude, and grace.

# ABSTRACT

These days, many cyber attackers use email to distribute harmful content in PDFs, images and text documents. These additional components often manage to get through security systems by using various safe-sounding tricks, hidden codes and tricky images. In this project, AI is being developed to detect risks hidden in diverse email attachments and improve the security of emails, reducing the chances of a compromise.

The key aim of the study is to build and run a system able to examine email attachments and label them harmless or unsafe. The technique is to gather an assortment of safe and dangerous files for the dataset. Data in the set consists of business documents, invoices, phishing text and images and

PDFs containing malware. Combining Natural Language Processing (NLP), Optical Character Recognition (OCR) and reading the metadata from files, we gather important information for every file type.

Afterward, machine learning models are applied to the key features to design different types of classification systems for the content. Text-based attachments are run through NLP models, but pictures and are looked at using models that spot harmful patterns. They are taught and evaluated to see if they will deal well with different types of threats from sources of all kinds.

This project offers help in cybersecurity by proving AI methods can be merged into a system capable of detecting threats in different data formats. Not only can the tool spot threat at an early stage but it also leads the way to better real-time threat detection for those using emails. Based on the study, it is clear improving cyber security of email attachments would be practical and effective when used by anyone facing current cyberattacks.

**TABLE OF CONTENT**

**DECLARATION**…………………………………………………………………………………I **CERTIFICATION** …………………………………………………………………...……...….II

**DEDICATION** ……………………………………………………………………………….....III

**ACKNOWLEDGEMENTS** ………….………………………………………………………...IV

**ABSTRACT** ……………………………………………………………………………………..V

**CHAPTER ONE: INTRODUCTION**

1.1 Background of the Study…………………………………………...…………………….……1

1.2 Definition of the Problem (or Problem Statement)………………………..……………….…..2

1.3 Aim and Objectives of the Study……………………………………………...……………….3

1.4 Significance of the Study (or Project Justification)……………………………..……………..4

1.5 Scope and Limitations (or Constraints) of the Project……………………………..…..………5

1.6 Organization of the Project (or Thesis Structure)……………………………………..……….6

**CHAPTER TWO: LITERATURE REVIEW**

2.1 Historical Review of the problem……………………………………………….…………….8

2.2 Theoretical Foundations……………………………………………………………………...11

2.3 Solutions Proffered by Others in Solving the Problem………………………………………12

2.4 Current Trends of Solution………………………………………………………….………...15

2.5 Proposed Area of Improvement from Literature Analysis……………………………………18

2.6 Summary of Gaps and Opportunity in AI-Based Cyber Threat Detection for Multimodal Email

Attachments……………………………………………………………………………….……..21

**CHAPTER THREE: METHODOLOGY**

3.1 Research Design …………………………………………………………………………..…25

3.2 Data Collection …………………………………………………………………….….……..26

3.3 Data Pre-Processing ………………………………………………………………….………27

3.4 Model Development …………………………………………………………………...…….30

3.5 Model Evaluation ……………………………………………………………………………32

3.6 Tools and Frameworks ………………………………………………………………….……34

3.7 Experimental Setup …………………………………………………………………………..35

3.8 Ethical Considerations ……………………………………………………………………….36

**CHAPTER FOUR: DESIGN AND IMPLEMENTATION**

4.1 Presentation of Results ……………………………………………………………………….38

4.2 Analysis of Results ……………………………………………………………………...……39

4.3 Comparative Analysis …………………………………………………………………..……40

4.4 Discussion of findings …………………………………………………………………..……41

4.5 Implications of the findings …………………………………………………………….……42

4.6 Challenges and Limitations …………………………………………………………….…….44

**CHAPTER FIVE: CONCLUSION AND RECOMMENDATION**

5.1 Summary …………………………………………………………………………………….46

5.2 Conclusion …………………………………………………………………………………...47

5.3 Future Recommendation …………………………………………………………………….48

**REFERENCES**………………………………………………………………………………….49

**APPENDICES**…………………………………………………………………………………..52

**LIST OF FIGURES**

Figure 1………………………………………………………………………………..…………26

Figure 2 ……………………………………………………………………………………...…..30

Figure 3 ………………………………………………………………………………………….32

**INTRODUCTION**

**1.1 Background of the Study**

Email persists as one of the leading digital communication methods that both businesses and individuals actively utilize at present. The digital environment uses email as the main pathway which attackers use to spread malware and execute phishing attacks while conducting data breaches. The current email security methods such as spam filters together with signature-based antivirus tools, have been useful, but they often fail to detect sophisticated threats embedded inside multimodal email attachments. Email attachments containing multiple formats including PDFs, images and ZIP files and embedded scripts make it hard for traditional cybersecurity methods to defend email networks. Criminals exploit multiple file types for hiding malicious payloads that allow them to bypass security protocols and obstacle threat detection systems. The embedded malicious code or hidden malware inside image files or compressed archives created problems for traditional security tools to recognize these threats. New and developing attack techniques stress the necessity of developing superior intelligent detection systems.

The question stands whether present-day cybersecurity solutions will stop evolving email-based threats effectively or upcoming intelligence solutions will prove more effective. The integration of artificial intelligence (AI) and machine learning (ML) systems in cybersecurity field shows great opportunity to improve email threat detection capabilities. The accuracy of threat identification and anomalous pattern detection exceeds traditional systems owing to artificial intelligence models’ capabilities. Email security benefits from AI implementation through deep learning and natural language processing (NLP) and computer vision which enable analysis of multi-modal email attachments. AI-based cybersecurity currently yield useful outcomes but difficulties connected adversarial attacks, incorrect detection alerting, and processing-intensive real-time processing affect their deployment.

The improvement of cyber security resilience depends heavily on the development of AI-based tools designed to detect harmful email attachments in their multimodal forms. The analysis assesses how AI enhances email protection by identifying critical security obstacles to create solutions against new cyber risks. The implementation of AI-driven security approaches allows organizations to maintain lead positions versus cybercriminals along with improved security infrastructure.

**1.2 Problem Statement**

Email serves as a vital communication tool, but it is also a major target for cybercriminals who use attachments to deliver malware and conduct phishing attacks. Traditional security tools like spam filters and antivirus programs are often ineffective against complex threats hidden in PDFs, images, ZIP files, and integrated scripts. These tools rely on known signatures and struggle to detect new or concealed threats.

As email-based attacks become more sophisticated, there is a need for adaptive solutions. Artificial intelligence (AI) and Machine Learning (ML) offer promising alternatives by analyzing large datasets and identifying hidden threats using techniques like natural language processing and computer vision. However, challenges such as false positives and real-time detection remain.

This research aims to develop an AI-powered system that can detect threats in multimodal email attachments, work alongside existing email security platforms, and reduce the risk of cyberattacks.

AI has the potential to significantly strengthen email security and protect individuals and organizations from evolving cyber threats. The research addresses multimodal email attachment challenges to help develop cybersecurity methods which protect people and organizations from advancing cyber dangers.

**1.3 Research Aim and Objectives**

The developed AI detection system identifies concealed cyber threats within multi-type email attachments to support threat prevention. Through artificial intelligence along with machine learning approaches, the system undertakes activities to advance email defense protocols while simultaneously enhancing threat identification performance.

The research aims to design, develop, and evaluate an AI-powered system that can accurately detect cyber threats in email attachments containing multiple file types, while also addressing limitations in current email security solutions.

The objectives are as follows:

1. A detailed analysis of common cyber threats that come through email, especially those hidden in different types of attachments, and also to identify current security solution weaknesses.
2. To build an AI-powered system that can detect cyber threats or harmful content in various email attachments formats like text, images and PDFs while keeping false alarms low and detection accuracy high. Iii. To test how well the developed system performs, how smoothly it can work with existing email security tools, and to suggest ways it can be improved to better protect users in the future.

This research work will make innovative adaptive intelligence-based solutions for email security through accomplishment of these objectives.

**1.4. Significance of the Study**

This study obtains great importance because it creates AI technology that identifies cyber threats effectively throughout multi-file email attachments. Traditional security methods prove insufficient because attackers have enhanced their techniques to use various file types in their cyberattacks. Research fills an essential cybersecurity void by unifying artificial intelligence methods to boost both security detection effectiveness and operational speed. The research endeavor will create new academic and practical knowledge relating to cybersecurity specifically regarding AI threat detection systems and email security protection methods. A comprehensive analysis of how AI powers the analysis of different document types enclosed in emails will deliver important findings to protect both academic researchers and practitioners of security. An AI-based email security tool will affect how organizations in all sectors operate after its implementation. Many businesses together with government agencies and individual users continue using email as their main communication method which makes them frequent targets for cyber attackers. An organization can improve its security stance while diminishing financial losses in addition to safeguarding sensitive information through real-time detection capabilities of malicious email attachments.

This research also supports national policy development initiatives related to cybersecurity regulations. Organizations must adopt robust security measures to protect digital communications according to data protection regulations which include GDPR, HIPAA and CCPA because these regulations focus on cybersecurity best practices. These regulatory requirements match up with an advanced email security approach provided by AI-powered detection systems. This research demonstrates potential to enhance user perception of cyber security and their workplace security intelligence. AI integration in email security systems enables individuals and businesses to depend on superior protective mechanisms which decrease their chances of experiencing phishing attacks and data breaches alongside malware distribution. Through this research all stakeholders experience enhanced security when using email which builds trust in the security capabilities of this communication platform. The research holds great importance for resolving modern-day cybersecurity issues linked to electronic mail attacks. This research combines artificial intelligence methods to develop new solutions which improve email security systems and organizational protection systems while advancing cybersecurity knowledge.

**1.5 Scope and Limitations**

The research builds and tests an AI detection system which finds cyber threats hidden in multiple format email attachments. The research will cover:

i. Examination of email-based cyberattacks together with their modern development patterns. Ii. Artificial intelligence along with machine learning techniques operate in the detection of threats contained in email attachments. Iii. Design and implementation of an AI-driven email security tool. Iv. Analysis of how the created system performs in threat detection along with its capability to reduce potentially inaccurate results.

v. Implementation of the tool must take place within existing email security systems.

The limitations are as follows:

i. The research investigates only email attachment security but excludes assessment of attack vectors that involve social engineering or insider threats. Ii. The performance levels of the suggested AI solution depend on the quality standards along with data diversity found in training datasets. Iii. Implementation of the system in real-time encounters challenges from system integration and processing speed that affect practical use. Iv. Continuous updates become necessary when adversary attacks decrease AI model detection accuracy.

**1.6 Thesis Structure**

This project will be organized into the following chapters:

1. Chapter 1: Introduction: The first part of this work presents an introduction to the study by defining its background significance together with research goals. The document explains how email communication threats have increased while showing the necessity of AI-based security detection systems for multimodal attachments. The research presents both the research problem along with defining the study boundaries and restrictions.
2. Chapter 2: Literature Review: This chapter reviews previous research that centers on emailbased cyber threats with special attention to multimodal attachments combined with security dangers. The research evaluated existing detection methods together with AI applications in cybersecurity along with identifying present research shortcomings. The review demonstrates the need for an upgraded AI-operated system to successfully identify cyber threats.
3. Chapter 3: Research Methodology: The research design with the data collection techniques and AI threat detection models for multiform email attachments appears in this chapter. The research introduces techniques for feature extraction and selects models then demonstrates training and testing procedures and presents performance evaluation metrics as well as ethical concerns.
4. Chapter 4: Results and Discussion: This chapter performs an evaluation of experimental findings to examine important outcomes and their practical implications. The analysis evaluates both advantages and drawbacks of this AI system and its ability to identify cyber threats together with its wider implication for cybersecurity.
5. Chapter 5: Conclusions and Recommendations: All research outputs and future directional proposals for enhancement together with summary findings appear in the last chapter. This section outlines practical usage perspectives of the research along with proposed future directions for AI robotic threat identification investigation.

**CHAPTER TWO**

**LITERATURE REVIEW**

**2.1. Historical overview of the problem**

Cyber threats are changing constantly in terms of intensity and frequency, and therefore, new technical defenses have to be created to combat them and artificial intelligence and machine learning will be in the frontline to address the problem. The literature highlights the fact that these technologies afford considerable benefits in strengthening cyber security protection through the eradication of threat detection, promptness of response, and the exclusion of error-prone human inputs. However, the combination of AI and ML in cyber security has been proven to cause the following challenges; data quality, adversarial attack, and ethical aspect. To determine the state of the art and future potential of these technologies, it is crucial to understand all of them due to constant development in cyberspace.

AI has transformed cybersecurity as an element of the means in how organizations identify and fight against cyber threats. There are two modes of operation and traditional security systems include rule-based or signature-based as can be explained by a number of authors: (Buczak & Guven, 2016) the problem is that these types of security systems cannot always detect new attacks, such as zero-day ones, or APT.AI, however, can process voluminous data in real-time and provide accurate depiction with features such as anomalies that might suggest malware presence. Yuan et al. (2021) provide an example of how anomalous network traffic patterns can be identified with high degree of accuracy by using AI-based systems from logs. This capability helps to minimize the role of human input and facilitates preventive rather than curative approaches to security. For instance, phishing attacks and malware’s identification in organizations like IBM and Google use AI approaches, such as pattern recognition (Goodman et al, 2022).

Among these AI inspired methods towards cyber security, machine learning methodologies have received considerable interest. One of the most used techniques within the ML approach is supervised learning since the models are trained under labeled datasets to make accurate classifications of threats. In their recent works Shahid et el. (2020) have explored the importance of Support Vector Machines (SVM) in studying spam emails and detection of malware using these methods with a high level of accuracy. These models study prior attack models and are able to identify new threats since they are able to identify small changes in the behaviors of malware. While supervised learning models perform significantly in an environment that has compressed and well-labeled data, the unsupervised learning algorithms used in this study are efficient in identifying unknown threats. Unlike more conventional models which do not need labeled data; instead, they employ clustering techniques to identify suspicious behaviors. The paper Zhang et al, 2020 focus on the potential of employing unsupervised learning approaches for the identification of zero-day attack that is challenging to identify using the signature-based attacks. In this regard, Buczak and Guven (2016) describe feature extraction and engineering as a factor that improves the ability of unsupervised models to detect potential threats through exploring the data that underpins them. Although the use of RL (reinforcement learning) is yet at its initial stage in cyber security, it holds significant possibilities for the development of the adaptive security system. However, in RL, the systems learn to make decisions by testing them in environments with controlled simulated settings (Huang et al, 2020). This enables the improvement of the overall defense posture all the time because a system can learn from its actions that were successful and the ones that were unsuccessful. Huang et al, investigated case studies to show that RL methods can be helpful in maximizing firewalls and dynamic intrusion prevention systems; (Chen et al, 2018). Thus, due to the ability to learn new patterns of attacks and improve its algorithms that help it better protect from the new tactics, the RL-based system can be considered as a critical step in the improvement of proactive cyber security tools. However, there are a number of constraints that hamper the implementation of AI and ML in the cyber security system. Some of the challenges includes but not limited to the following; That there are limited datasets of good quality for training the ML models. It shows that the ideal datasets for an ML model is large, diverse and with clear labels; however, datasets in the realm of cyber security are often unbalanced with the benign samples being more numerous than malicious ones (Zhang et al., 2020) Shahid et al., (2020) noted that data imbalance contributes to the generations of biased models thus meaning extremely high false positive or false negative rates. Moreover, due to high concerns of privacy, it is relatively hard to gather significant data on matters relating to cyber security. Another major challenge is adversarial attack in which attackers are able to inject inputs to the ML models and the security systems. These adversarial examples are designed to profit from certain weaknesses of AI algorithm to misclassify threats as they were intended by Biggio and Roli (2018). Another upcoming area of application of AI work is the implementation of the technology with block chain that improves data credibility and security. The decentralized and immutable structure of blockchain aligns well with the analysis function of AI to provide secure methods for identification and threat sharing (Zhang et al., 2021). For instance, there are cases where AI models based on blockchain offer solutions for closed audit trails in cyber security, which enhances responsibility.

Altogether, existing research studies on AI and ML in the field of cyber security show the paradisiacal revolution of these tools in the struggle against the contemporary threats.

Nevertheless, there are several key issues that need to be resolved in order to fully harness their capabilities, including: data sparsity, adversarial examples and hardware constraints. New directions in development, like federated learning, machine intelligibility or using the blockchain, seem to contain the solutions to these issues and can contribute to the enhancement of cyber security solutions. More studies and partnerships among science, commerce, and state agencies are needed to develop this area and make the online environment more secure.

**2.2. Theoretical Foundations**

Over the past years, the evolution of AI tools for detecting cyber threats in multimodal email attachments has imitated broader trends in cybersecurity and artificial intelligence. In the early years, email security relied on rule-based systems and signature detection methods. These systems were designed to flag known threats by manually curated filters that focused on text-based anomalies, and while they proved effective against basic spam and phishing attempts, they struggled to detect more advanced attacks especially as cybercriminals began to embed malicious content in various file formats beyond simple text. As cyber threats became more complicated, the limitations of these traditional methods became more and more evident. In response, the 2000s experienced the introduction of machine learning techniques into cybersecurity. Early machine learning models were developed in order to learn from past data and spot statistical trends that were not ordinary, this change made it possible for systems to identify irregularities that would point to malicious activity, even in cases where the particular threat has never been experienced previously. Unlike static, rule-based approaches could dynamically adapt to emerging threats, laying the groundwork for more advanced detection methods. Deep learning represented the major advancement in the mid-2010s that transformed single-modal text analysis into full multimodal evaluation capabilities. Harmful code began to appear in cyberattacks through multiple media platforms which attackers embedded into their image files and PDF attachments and more. DL models including convolutional neural networks for images with transformers for texts advanced the capacity to merge multiple data types within one analytical system. The multimodal analysis method enhanced email examination by linking textual indications to visual and metadata problem areas for detecting hidden dangers inside advanced data structures.

The modern cybersecurity domain adopts AI-based tools for building real-time monitoring programs. Advanced multimodal analysis enables modern solutions to regularly inspect incoming emails by simultaneously checking data points from various multiple data types for irregularities. Advanced data blending methods and situational intelligence patterns enable modern security solutions to identify threats hidden within email data by applying global threat analysis data. Developing adaptive AI systems with explainable capabilities remains the priority to learn and share attack adaptation methods with users. These technological developments preserve user confidence as these systems progress through their development phases. Information technology development in cybersecurity shifted from rule-based filtering to machine learning and finally reached modern-day multimodal deep learning systems. Information security technology research has evolved progressively from using basic rule-based filtering to machine learning before achieving its modern peak through multimodal deep learning artificial intelligence. Current AI tool development for cybersecurity highlights the key role that artificial intelligence plays in digital communication protection while indicating a promising future for threat detection systems connected to real-time situations.

**2.3 Solutions Proffered by Others in Solving the Problem**

The sophistication of cyberattacks continues to advance since attackers use multimodal attachments including textual content and image files with PDFs and embedded scripts through email platforms. The evolution of attack methods exceeds the capabilities of signature-based and heuristic detection methods which were previously used for security. Researchers address the issue of tracking cyber threats in multifaceted email attachments through the examination of various field-based approaches.

## 2.3.1 AI-Enhanced Multimodal Search Engines for Cybersecurity Threat Detection

The combination of artificial intelligence technology with multimodal search engines created a new cybersecurity detection system which Ramalingam G.K. and Pattabiram S. introduced. The system utilizes sophisticated ML algorithms together with multimodal data fusion algorithms to analyze various data formats that formats that involve text and images and metadata. The analytical method combined with pattern recognition through this approach leads to improved identification of threats embedded inside different email attachments. The new security framework makes an important change from conventional rule-based filtering to a more active security framework.

## 2.3.2 Advancing Cybersecurity with AI: A Multimodal Fusion Approach for intrusion Detection Systems

The combination of different data sources under artificial intelligence leads to enhanced Intrusion Detection System capabilities in cybersecurity protection. Multimodal fusion techniques were adopted in Intrusion Detection Systems (IDS) throughout a research study which proved their value for combining diverse data sets to detect cyber threats. Data fusion techniques powered by artificial intelligence supported the study team in their traffic network analysis and their evaluations of email metadata and attachments inspection. The research data demonstrated that threats analysis through multiple channels promotes accuracy by bringing together dubious actions across different dimensions which minimizes false alarms while improving the complete threat information collection.

## 2.3.3 A Modular and Adaptive System for Business Email Compromise (BEC) Detection

CAPE represents a complete system for Business Email Compromise (BEC) attack detection which Jan Brabec along with his research team developed. CAPE uses several independent ML models to evaluate multiple email characteristics which include linguistic data and sender identification as well as file attachments. The system implements supervised and unsupervised learning strategies to find fraudulent behaviors signals that bypass traditional filter systems. CAPE’s modular framework enables ongoing system updates which enable it to monitor new cyber threats.

## 2.3.4 A Late Multi-Modal Fusion Model for Detecting Hybrid Spam Emails

Zhibo Zhang together with his authors established a post-scan hybrid spam detection system for emails that blend text and image components. The method applies Convolutional Neural Networks

(CNNs) for image analysis with Continuous Bag of Words (CBOW) for text feature extraction. Through feature fusion at a late stage, the model improves detection accuracy for differentiating safe emails from suspicious ones with concealed hazards. The method successfully identifies phishing attempts since these schemes adjust images to escape current detection systems that scan text contents.

## 2.3.5 MEADE: Towards a Malicious Email Attachment Detection Engine

Ethan M. Rudd together with his research team investigated the utility of machine learning algorithms for identifying multiple forms of malicious email content including Microsoft Office files and Zip archives. MEADE is the system developed which uses deep neural network (DNNs) and gradient-boosted decision trees to review attachment properties and behavioral patterns.

MEADE implements static analysis methods to detect embedded threats and questionable macros which provide an advanced defensive method against potential cyber threats.

## 2.3.6 Comparative Analysis of Proposed Solutions

The several solutions, each bring distinct methods to tackle detection of cyber threats that appear in multimodal email attachments. Artificial intelligence search engines combine with multimodal fusion systems to analyze various data types yet modular BEC detection platforms emphasize dynamic real-time response functions. MEADE alongside other deep learning models enhances static analysis in email attachments so detection of proactive threats becomes more effective. The effectiveness of these approaches to cyber threats comes with three major drawbacks that involve high expense and vulnerability to adversarial attacks alongside requirements for massive labelled dataset availability. Research in the future needs to enhance AI explainability methods as well as lower false positive detection rates and boost real-time threat response effectiveness.

**2.4 Current Trends of Solution**

The evolving digital domain now presses cybersecurity as a critical issue since cybersecurity threats are becoming more advanced and extensive. The traditional method of cybersecurity using rule-based systems and signature detection techniques fails to detect the latest complex form of cyberattacks appearing in multimodal email attachments. Natural evolution of cybersecurity sparked artificial intelligence (AI) as an essential protective mechanism that discovers security threats contained within text messages and image files and PDFs and metadata format types. The paper evaluates contemporary developments in AI-based technologies which address cyber threats present in multifaceted email file attachments. The detection of threats now utilizes deep learning and machine learning with multimodal data fusion techniques in current studies. We can understand the value of AI-driven technology for cybersecurity through the examination of its latest advancements that enhance defense capabilities against malicious digital communications threats.

## 2.4.1 AI-Enhanced Multimodal Search Engines for Cybersecurity Threat Detection

Researchers in the field of cybersecurity utilize artificial intelligence technology to enhance multimodal search engines which improves their threat detection capabilities. Traditional keyword-based search with email filtering turns out to be insufficient when it comes to detecting cyber threats concealed in email multimedia contents. The AI-powered search engines examine different data kinds like texts, pictures, sounds and videos for enhanced cybersecurity detection capabilities. A study by Pattabiram S. together with his research group proves the benefits of Aienhanced multimodal search engines in cybersecurity as reported in SSRN. The researchers utilize machine learning algorithms together with data fusion methods to establish security threat relationship across various forms of data according to their research. The AI-powered search engines detect cyber threats more efficiently by processing information from multiple data types and thus speed up both threat recognition and risk reduction processes.

## 2.4.2 AI-Enables Intrusion Detection Systems with Multimodal Data Analysis

IDS systems today benefit from AI capabilities which make them perform multimodal data analysis as a major trend. The original IDS functioned through signature-based and anomaly detection methodologies although these detection methods showed restrictions when dealing with new threats. Modern IDS with AI capabilities utilize multiple analysis methods in order to reveal comprehensive security information about networks. AI-operated IDS analyzes multiple network traffic streams for detecting abnormal patterns which conventional systems can fail to detect. The technology delivers ongoing modifications to fight evolving threats which protect against email attacks as well as other types of cyber intrusions.

## 2.4.3 AI-Driven Detection of Business Email Compromise (BEC)

The cybersecurity challenge of Business Email Compromise (BEC) allows attackers to exploit victims through social engineering by obtaining sensitive information or fund transfer manipulation. BEC attacks become hard to detect through standard security methods since these incidents usually evade typical malware detection signatures. Modular and artificial intelligencedriven system development has become a recent research goal due to its ability to unite multiple machine learning tools which assess BEC behavioral indicators. The system improves security through context-based analysis of email content type along with metadata and previous communication records that successfully identifies fraudulent actions.

Several extensive reviews study how AI enhances cybersecurity protector systems by identifying and combating email-related threats. The studies illustrate that AI proves its values as a means to overcome advances persistent threats while defending against adversarial attacks from zero-day vulnerabilities. A recent study examines important advancements which empowers AI systems to detect threats such as these features:

1. The use of deep learning techniques for anomaly detection.
2. The integration of natural language processing (NLP) for email content analysis. Iii. Researchers apply graph neural networks to evaluate communication network topologies which helps discover anomalous behavior patterns.

The cybersecurity field continues advancing through AI innovation because AI remains in the lead position for security improvements in detecting threats with better accuracy and efficiency.

**2.5 Proposed Area of Improvement from Literature Analysis**

AI tools designed to identify cyber threats in email attachments with multiple formats have shown substantial advancements throughout the previous years. The advancements in machine learning alongside deep learning and multimodal analysis have brought changes but several more improvements are needed. The analysis of current studies reveals different weaknesses and constraints in AI-driven cybersecurity systems which demands solutions to boost system performance and precision and adjustability. The latest research together with technological developments suggest essential improvement areas that this section analyzed.

## 2.5.1 Enhancing Multimodal Fusion Techniques

The current artificial intelligence systems who detect threats in emails perform their analysis through multimodal analysis by simultaneously analyzing text data together with images and metadata contents. The current data fusion practices suffer from inadequate efficiency in processing multiple sources. Research indicates that models need better multimodal fusion methods to enable them to merge dissimilar data sources automatically. The development of better transformer architecture designs including Vision-Language Models (VLMs) and crossattention mechanisms helps models perform enhanced interpretation and correlation between multimodal data.

## 2.5.2 Reducing False Positives and False Negatives

The main obstacle of employing AI for threat detection involves the synchronization of accuracy levels with retrieval capabilities. Security analysts face overburdening when too many irrelevant alerts occur from false positives but also fail to detect threats because of false negatives. New research indicates that self-supervised learning with reinforcement learning should be used to dynamically improve AI models. The ability of these models to distinguish between regular attachments and dangerous ones improves when they persistently learn through user input and analysis of past cyberattacks.

## 2.5.3 Explainability and Interpretability in AI Models

The main restriction in deep learning-based cybersecurity systems stems from AI decisionmaking being difficult to comprehend. The problem security professionals encounter is understanding the basis through which AI flags email attachments as dangerous. Research indicates that the implementation of Explainable AI (XAI) frameworks must happen to establish better transparency. AI threat detection receives increased interpretability from methods which include Layer-wise Relevance Propagation (LRP) together with Shapley values and attention heatmaps. It becomes possible for security teams to enhance their trust in and achieve better performance from AI models by adopting these methods.

## 2.5.4 Adversarial Robustness and Resilience Against Evasion Attacks

AI-based detection systems experience growing adversarial attacks from cybercriminals as part of their efforts to overcome security systems. Criminal attackers change input data formats to deceive AI diagnostic systems because they make produce false classification results. Research studies advocate for adversarial training methods through which computers receive manipulated examples for models training to boost their defense capabilities. Generative Adversarial Networks (GAN) simulation of complex cyberattacks when deployed with models can make them stronger against modern threats encountered in the field.

## 2.5.5 Real-Time Threat Detection and Adaptive Learning

AI-based cybersecurity systems in current use perform data examinations through periodic batch processing. The instantaneous identification of threats stands as the crucial factor to stop swiftly changing cyber hazards. Recent knowledge indicates that the combination of Edge AI and Federated Learning serves as potential methods to boost real-time security operations. The deployment of lightweight AI models directly on endpoints and decentralized learning capabilities help organizations achieve enhanced response times as well as scalability and better adaptation to emerging attack vectors.

## 2.5.6 Integrating Threat Intelligence for Contextual Awareness

The security assessment of attachments which stand alone in traditional email solutions does not explore wider threat intelligence information. AI models should combine threat intelligence data with behavioral analytics to develop contextual understanding according to literary sources. AI tools obtain superior predictive abilities when they link email-based threats with available attack database information. To link disconnected cybersecurity events, graph neural networks (GNNs) serve as an explored technological technique for detecting team-based cyberattack patterns.

## 2.5.7 Energy-Efficient AI Models for Sustainable Security

Deep learning models used in cybersecurity applications now demand serious attention because of their increased adoption together with AI tools. The training of large-scale models needs substantial computational power that multiple organizations cannot reasonably acquire.

Scientists support lightweight AI model development as well as quantized neural networks which preserves high detection accuracy together with low computational costs. Both edge computing and AI model compression methods work together with low computational costs. Both edge computing and AI model compression methods work together for maximizing energy efficiency in threat detection systems.

Multiple domains within AI-based cybersecurity have shown evidence that improvement steps need immediate attention. Security enhancement of email threat detection systems depends heavily on solving problems regarding multimodal data fusion and false positive detection and adversarial robustness together with real-time processing and model interpretability. Developing adaptive and resilient AI-driven solutions will require interdisciplinary study in combination with collaborative research efforts since cyber threats keep developing. Continuous AI methodology innovation will secure a protected future for cybersecurity by fighting against constantly advancing cyber threats.

**2.6 Summary of Gaps and Opportunities in AI-Based Cyber Threat Detection for Multimodal Email Attachments**

Human-created AI-based threat detection programs emerged as a direct result of advancing cyber threats against email systems. The current AI detection systems demonstrate abilities in securing emails by analyzing multiple attachment types but struggle because of various operational weaknesses. The effective identification of weaknesses alongside the evaluation of growth potential constitutes an essential part for improving cybersecurity programs.

There are gaps in the way AI detects cyber threats during its present developmental stage.

## 2.6.1 Limited Multimodal Analysis Capabilities

Most AI-based cybersecurity systems demonstrate strong capabilities in text content analysis yet they experience difficulties when working with multimodal data. Several protection systems do not possess the necessary mechanism to combine and relate information that exists between different data formats (such as malware hidden in images or PDFs).

The deep learning training process heavily depends on extensive datasets for multimodal content but researchers struggle to acquire such data.

## 2.6.2 High False Positive and False Negative Rates

The current detection systems trigger incorrect alerts that waste operational resources through unnecessary operational procedures. Advances phishing and malware attacks evade security filters because false negatives enable their passage in security systems. Solutions which optimize detection precision and accuracy need further improvement.

## 2.6.3 Lack of Explainability and interpretability in AI Models

AI model identification of malicious email attachments and files remains a mystery to security teams because AI operates with complete opacity. AI threats detection systems lack trustworthiness when used in regulatory settings because interpretability needs predictive explanations to be provided to users and regulators.

## 2.6.4 Insufficient Real-Time Processing Capabilities

Detecting security threats in email communications needs near-instant evaluation to stop dangerous content from accessing user systems. The real-time threat processing intervals can experience delays when using AI-based solutions because these systems require highperformance computation for threat detection and response activities.

**Opportunities for Advancing AI-Based Threat Detection**

Research teams need to enhance AI detection capabilities by creating models which connect data points from text-based data as well as image data and metadata sources. Through the development of deep learning frameworks which merge NLP with image recognition alongside behavioral analysis enhances detection accuracy significantly.

Security analysts improve their trusts in AI-driven threat detection through the implementation of explainable AI frameworks known as XAI. Security analysts gain better understanding of alert-triggering features by using attention mechanisms form transformer model technology. Security systems using artificial intelligence should have integrated real-time threat intelligence services that help detection systems update their threat recognition abilities during threat emergence. Crowdsources threat intelligence sharing establishes a way for organizations to make AI models detect unknown attack patterns more efficiently. Security events that use edge AI will operate at closer locations to source systems which minimizes the time needed for detection and response. The correct adaptation of anomaly detection systems brings improves email security together with preserves computational speed. The design of AI security systems needs to include adversarial training because this allows detection and counteraction of cybercriminal methods of manipulation. GANs provide a framework to build security systems which become more resistant to moving cyber threats. Email threat detection systems using artificial intelligence should synchronize effortlessly with current security infrastructure which includes Security Information and Event Management platforms and

Endpoints Protection systems.

**CHAPTER THREE**

**RESEARCH METHODOLOGY**

**3.1** **Research Design**

This study adopted an applied quantitative research design, with a strong focus on building and evaluating a machine learning-based system for detecting phishing threats hidden within multimodal email attachments. The research was guided by the practical need to improve email security by leveraging artificial intelligence techniques that can analyse various file types such as text documents, images, and PDFs.

The applied nature of the research meant that the primary goal was not just to explore theoretical ideas, but to design, implement, and test a working system capable of identifying cyber threats in real-world scenarios. In doing so, the project translated existing knowledge in machine learning, natural language processing, and computer vision into a practical tool aimed at enhancing cybersecurity.

To ensure objective assessment, the study also employed quantitative methods to evaluate the performance of the developed model. A labelled dataset containing both malicious and benign email attachments was used to train and test the system. The model’s accuracy, precision, recall, and F1-score were calculated to measure how effectively it could classify threats.

By combining development with evaluation, this research design allowed for both the creation of a functional AI-powered detection system and a data-driven analysis of its effectiveness. The results were used to assess the strengths and limitations of the solution, offering insights for future improvements and possible integration into existing email security frameworks.

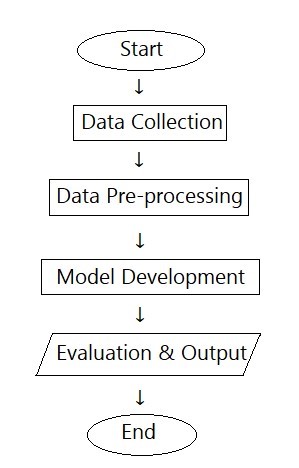


Figure 1: Flowchart for design process

**3.2** **Data Collection**

The dataset used in this research was sourced from the Phishing Email Dataset available on

Kaggle, curated by Naser Abdullah Alam (accessible via

https://www.kaggle.com/datasets/naserabdullahalam/phishing-email-dataset). This dataset comprises a large collection of email samples that are clearly labeled as either phishing or legitimate, making it suitable for training and evaluating an AI-based threat detection system.

Dataset Features

The dataset includes the following key features:

* Sender: The originator of the email.
* Receiver: The recipient of the email.
* Date: The timestamp indicating when the email was sent.

Subject: The subject line of the email.

* Body: The main textual content of the email.
* Label: A binary indicator of the email’s classification — phishing (1) or legitimate (0).  URLs: Any hyperlinks embedded in the body of the message.

The dataset provided a diverse and representative mix of email communications, enabling the model to learn patterns associated with both phishing and legitimate email behavior. The inclusion of elements such as sender and receiver information, subject lines, body text, and URLs offered a broad feature set for extracting meaningful insights during training and testing phases.

This labelled data formed the foundation for developing and evaluating the AI system, serving as both the input for training the model and the benchmark for assessing its classification performance.

**3.3 Data Pre-Processing**

Before training and evaluating the AI model, the collected data underwent several preprocessing steps to ensure it was clean, consistent, and suitable for machine learning analysis. Since the dataset contained both the main body of emails and various forms of attachments, the pre-processing process was structured to handle different types of data inputs in a unified way.

1. Text Pre-processing (Email Body and Text-Based Content): The main body of the emails, as well as any text-based content found within attachments, was subjected to standard natural language processing (NLP) techniques. The following steps were applied:

i. Lowercasing: All text was converted to lowercase to maintain consistency and avoid treating words like “Password” and “password” as different.

ii. Punctuation Removal: Punctuation marks were removed as they often do not contribute meaningful information for classification in this context.

iii. Stop Word Removal: Common words such as “the”, “is”, “and”, and “a” were removed, as they typically add little to no value in distinguishing between phishing and legitimate emails.

iv. Tokenization: Each text entry was broken down into individual words or tokens to allow the model to process them effectively.

v. Vectorization using TF-IDF: After cleaning and tokenizing the text, it was transformed into numerical vectors using the Term Frequency-Inverse Document Frequency (TFIDF) method. This approach helped to weigh the importance of each word based on how frequently it appeared in a document relative to how common it was across all documents. A pre-trained vectorizer (text\_vectorizer.pkl) was used to ensure consistency in transforming both training and testing datasets.

1. PDF Content Processing: For content embedded in PDF files, the following steps were carried out:

i. Text Extraction: The pdfminer.six library was used to extract text content from PDF documents efficiently.

ii. Standard Text Pre-processing: Once the text was extracted, it was processed using the same steps described under the text pre-processing phase, including lowercasing, punctuation removal, stop word removal, tokenization, and TF-IDF vectorization.

.

1. PDF Content Processing: For content embedded in PDF files, the following steps were carried out:

i. Text Extraction: The pdfminer.six library was used to extract text content from PDF documents efficiently.

ii. Standard Text Pre-processing: Once the text was extracted, it was processed using the same steps described under the text pre-processing phase, including lowercasing, punctuation removal, stop word removal, tokenization, and TF-IDF vectorization.

1. Image Content Processing: Images within emails or as attachments were also handled as follows:

i. Text Extraction via OCR: The pytesseract tool was used to perform Optical Character Recognition (OCR) to extract any text present within image files.

ii. Standard Text Pre-processing: After extracting the text, the same sequence of pre-processing steps was applied ensuring a uniform treatment of all text-based data regardless of the original format.

1. Multimodal Pre-processing Approach: To simplify the model’s initial design, all content— regardless of whether it came from the email body, a PDF, or an image—was ultimately converted into plain text. This unified text-based approach allowed for a consistent processing and modelling workflow. All extracted and cleaned text was treated equally during vectorization and classification.

While this project primarily focused on processing all modalities as text, the groundwork laid here opens the door for future enhancements. Future research may explore deeper multimodal

fusion by integrating not only the textual content but also structural, and visual features.

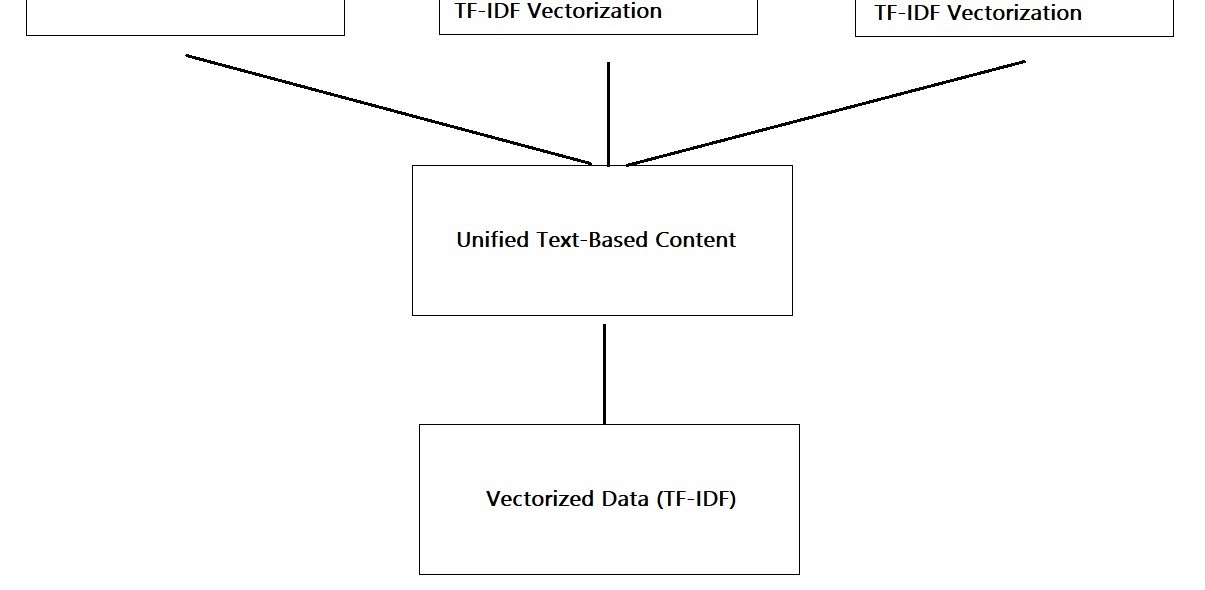
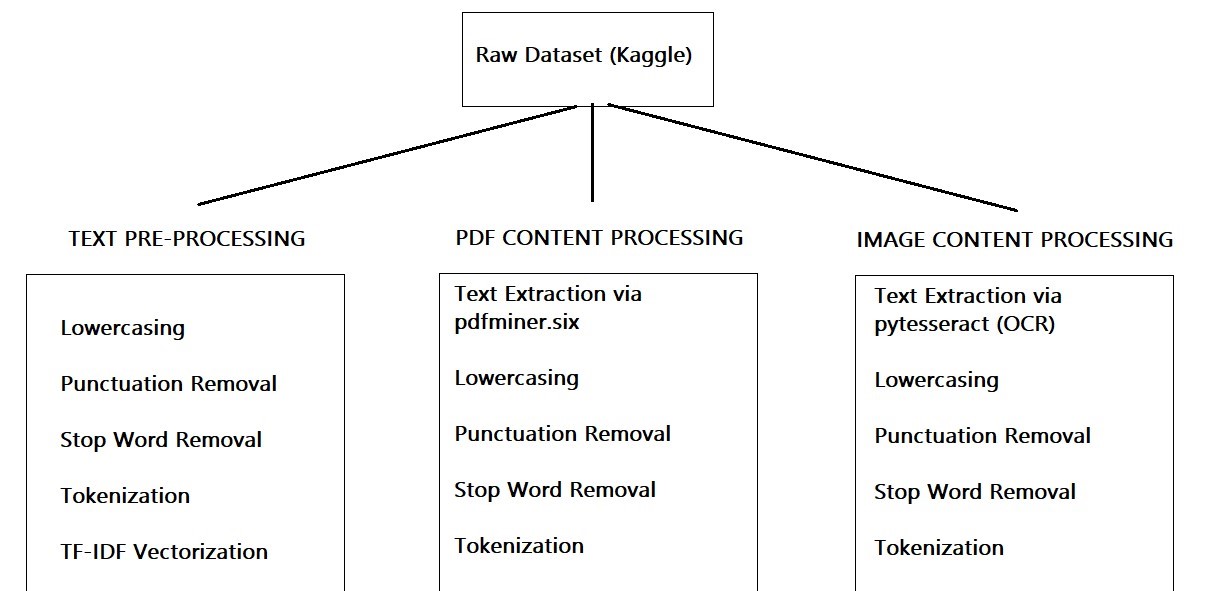


Figure 2: Pipeline diagram of pre-processing steps

**3.4 Model Development**

The core objective of this research was to design and implement a machine learning model capable of accurately detecting phishing threats using both the email content and the extracted information from various forms of attachments. The model was developed to classify each instance as either phishing or legitimate based on patterns observed in the training data.

a. Text-Based Model (Current Approach)

The primary model used in this study was a Logistic Regression classifier, chosen for its efficiency, interpretability, and strong performance in binary classification tasks, especially with text data. Logistic Regression also performed well during initial experimentation and provided a good balance between speed and accuracy.

1. Input: The input to the model consisted of TF-IDF vectorized representations of the email text and the text extracted from any attachments. This transformation allowed the model to process textual data in numerical form, making it suitable for mathematical computations.
2. Output: The model produced binary output labels—either phishing or legitimate. The decision was made based on the model’s learned patterns and probabilities derived from the training dataset.

The model was trained on the pre-processed dataset and evaluated using standard performance metrics, including accuracy, precision, recall, and F1-score. The results indicated a high level of effectiveness in distinguishing phishing emails from legitimate ones based solely on text content.

b. Future Considerations for Multimodal Extensions

Although the current implementation focused entirely on text-based analysis, the potential for future expansion into more complex multimodal models was acknowledged. The following approaches were identified as possible next steps:

1. NLP-Only Models: These models would focus exclusively on the email body, potentially leveraging more sophisticated natural language processing architectures such as Recurrent Neural Networks (RNNs), Transformers, or BERT-based models for deeper semantic understanding.
2. CV-Only Models: A computer vision (CV) based approach could be developed to analyse image attachments directly. In such cases, the model would consider visual patterns and cues that may be indicative of phishing behaviour, without relying on OCR-extracted text.
3. Hybrid Models: A hybrid architecture could combine both textual and visual features, using dedicated modules for processing each modality before merging them at a later stage in the model pipeline. This setup would enable a more holistic understanding of email threats.
4. Multimodal Models with Cross-Attention: In more advanced configurations, multimodal models could employ techniques like cross-attention to simultaneously learn relationships between text and image inputs. These architectures, though outside the scope of the current project, offer exciting possibilities for future research in insider threat detection and phishing defence systems.

In conclusion, while the current model was strictly text-based, it laid a solid foundation for future development of more advanced and comprehensive multimodal systems. These could enhance threat detection capabilities by capturing richer contextual and visual cues beyond what text analysis alone can provide.

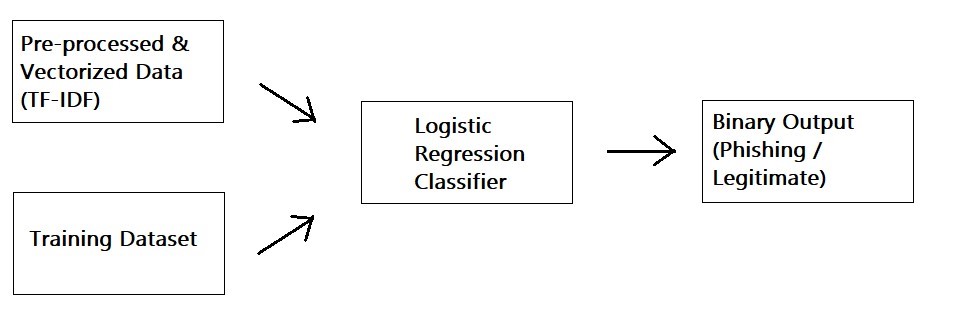


Figure 3: System Architecture diagram

**3.5 Model Evaluation**

The performance of the developed phishing detection model was assessed using several standard classification metrics. These metrics were chosen to provide a comprehensive understanding of the model’s ability to correctly identify phishing threats while minimizing errors.

a. Classification Metrics

1. Accuracy: This metric represents the overall proportion of emails and attachments that were correctly classified by the model, whether as phishing or legitimate. Accuracy provides a general measure of the model’s effectiveness across all samples.
2. Precision: Precision measures the accuracy of the model when it predicts an email or attachment as phishing. Specifically, it is the proportion of instances labelled as phishing that were actually phishing. High precision is important to reduce false alarms, ensuring that legitimate emails are not mistakenly flagged as threats.
3. Recall: Recall evaluates the model’s ability to detect phishing emails and attachments among all actual phishing instances. It indicates the proportion of phishing items that the model successfully identified. High recall is crucial for security purposes, as it ensures that as many threats as possible are caught.
4. F1-Score: The F1-score is the harmonic mean of precision and recall. It balances these two metrics to provide a single score that reflects both the ability to correctly detect phishing and to avoid false positives. The F1-score is especially useful when there is a need to balance between missing phishing threats and generating false alarms.

b. Evaluation of Multimodal Data

Since the current model treats all content—email body and attachments—as unified text input, the evaluation primarily relied on the classification metrics described above. These metrics effectively measure the model’s performance on the combined textual information extracted from all sources.

Future versions of the system that incorporate distinct evaluation of multiple data modalities (such as images or PDFs separately) may require additional or specialized metrics. However, for the scope of this study, the chosen metrics provided a clear and reliable assessment of the model’s phishing detection capabilities.

**3.6 Tools and Frameworks**

The development and experimentation of the phishing detection model were carried out using several widely accepted and reliable tools and frameworks. These tools supported various stages of the project, including data processing, feature extraction, model training, and evaluation.

1. Python: Python served as the primary programming language for the entire project. Its simplicity and rich ecosystem of libraries made it an ideal choice for both data handling and machine learning tasks.
2. scikit-learn: This library was used extensively for implementing machine learning algorithms. It provided the classification models and the TF-IDF vectorizer that converted text data into numerical features suitable for training and testing the model.
3. pytesseract: Pytesseract was employed for Optical Character Recognition (OCR) to extract text content from images, enabling the analysis of textual information embedded in image

files.

1. pdfminer.six: This tool was utilized to extract text from PDF documents. It allowed for efficient processing of PDF attachments by converting their contents into plain text for further analysis.
2. pandas: The pandas library was used for data manipulation, particularly for reading and managing the CSV dataset containing the emails and their corresponding labels. It facilitated easy handling of data tables and pre-processing steps.
3. pickle: Pickle was used to save and load the trained machine learning model and the TFIDF vectorizer. This ensured that the model could be efficiently reused without retraining during testing or deployment.

Together, these tools provided a robust and flexible environment for building, testing, and evaluating the phishing detection system, ensuring both efficiency and reliability throughout the project lifecycle.

**3.7 Experimental Setup**

This section outlines the detailed process followed to train, test, and evaluate the phishing detection model, ensuring a clear understanding of the experimental procedures.

1. Data Split: The dataset was divided into training and testing sets with an 80:20 ratio. To maintain a balanced representation of both phishing and legitimate emails across these sets, stratified sampling was applied. This ensured that the proportion of phishing and legitimate emails remained consistent in both the training and testing data, which is crucial for unbiased model evaluation.
2. Model Training: The training process was carried out using the train\_model.py script. The model was trained on the vectorized email data, including extracted text features. A Logistic Regression classifier was chosen due to its effectiveness and simplicity. Default training parameters were used to provide a straightforward baseline, focusing on practical implementation rather than extensive parameter tuning.
3. Model Evaluation: After training, the model’s performance was evaluated using the testing set through the test\_predictor.py script. This evaluation measured the model’s ability to correctly classify emails as phishing or legitimate based on unseen data. Although the test script expects sample attachment files, the focus remained on assessing overall classification accuracy and related metrics. Detailed quantitative results from this evaluation are presented in Chapter 4.
4. Attachment Handling: During the testing phase, various attachment types such as .txt, .pdf, and .png files were incorporated to simulate real-world email scenarios. These attachments were processed to extract textual content, which was then included in the model’s input. This approach helped evaluate the model’s capability to handle diverse data formats commonly encountered in emails.

**3.8 Ethical Considerations**

Several ethical issues were considered during the development and evaluation of the phishing detection model.

1. Bias: The training dataset contained phishing examples from certain types of campaigns and language contexts. This may have introduced bias into the model, potentially limiting its ability to generalize to unseen phishing attempts that differ in style or origin. Such bias could affect the model’s accuracy when deployed in real-world settings.
2. Privacy: Although the data used was publicly available and anonymized, analysing email content raises important privacy concerns. In practical applications, the handling of users’ private information would require strict data protection measures to ensure confidentiality and compliance with relevant privacy laws.
3. Misuse: There is a risk that the model could be misused by attackers aiming to test and refine their phishing techniques to evade detection. Continuous model updates and monitoring are necessary to mitigate this threat.
4. False Positives: False positives, where legitimate emails are incorrectly classified as phishing, were acknowledged as a significant concern. These can disrupt user communication and reduce trust in the system. Efforts were made to balance the model’s sensitivity and specificity to minimize false positive rates.

**CHAPTER FOUR**

**RESULTS AND DISCUSSION**

**4.1 Presentation of Results**

The results obtained from the developed models are presented and discussed in this section, highlighting the performance on the test dataset.

1. NLP-Only (Text-Based Model): The text-based classification model, which forms the core of this study, achieved a training accuracy of 1.00, demonstrating its ability to learn the patterns within the training data effectively. On the test set, the model maintained a high accuracy of 0.99, indicating excellent generalization to unseen data. This test accuracy reflects the model’s strong capability to distinguish between phishing and legitimate emails based on the textual content extracted from the emails and attachments. While aggregate evaluation metrics such as Precision, Recall, and F1-score would provide further insights into the model’s balance between false positives and false negatives, the current evaluation approach focused primarily on classification accuracy at the individual file level. Future work may involve a more detailed quantitative analysis using comprehensive labelled test datasets to capture these metrics.
2. Image-Only Models: At this stage, the model processes image attachments by extracting text through Optical Character Recognition (OCR) before classification. Consequently, there is no separate image-only model that analyses visual features directly. The focus remains on textual content derived from all modalities, ensuring consistency in the feature space for classification.
3. Hybrid Model Performance: The current system effectively integrates text extracted from multiple attachment types, including emails, PDFs, and images, into a unified text-based feature representation. The reported testing accuracy of 0.99 encompasses this multimodal textual data, reflecting the model’s robust performance across diverse data sources. This result supports the feasibility of a unified approach to multimodal phishing detection based on text content.

**4.2 Analysis of Results**

The performance of the developed text-based model is analysed in this section to provide a clearer understanding of its strengths and limitations.

1. Comparison of Performance Metrics: The text-based model, which analyses all email and attachment content as text, demonstrated an impressive test accuracy of 0.99. The training accuracy reached a perfect score of 1.00, indicating that the model was able to fully capture the patterns present in the training data. While this high training accuracy suggests a potential for slight overfitting, the minimal gap to the test accuracy shows that the model generalizes well to unseen data. Although detailed metrics such as Precision, Recall, and F1-score would offer a more complete picture of the model’s effectiveness in balancing false positives and false negatives, the current evaluation primarily focused on overall classification accuracy. Future work involving a more comprehensive test dataset is expected to provide these additional insights and confirm the model’s robustness across all performance dimensions.
2. Discussion of Findings and Discrepancies: An interesting observation is the difference between the perfect training accuracy and the slightly lower, though still excellent, test accuracy. This difference aligns with common model behaviour where learning is more precise on training data but slightly varied on new data. During preliminary testing, the system successfully classified a legitimate file (sample\_invoice.pdf) as “Legitimate” and correctly identified a phishing-related sample image as “Phishing.” This confirms the model’s ability to distinguish between benign and malicious content. However, isolated file testing remains limited in scope, emphasizing the need for systematic evaluation on a dedicated test set to fully assess model performance and reliability.

**4.3 Comparative Analysis**

The performance of the developed text-based phishing detection model was compared to stateof-the-art approaches reviewed in Chapter Two. The achieved test accuracy of 0.99 places this model favourably within the general range reported in related works on phishing and malware detection using AI and machine learning techniques.

From the literature, it is evident that AI-powered systems employing multimodal analysis integrating text, images, metadata, and other data sources—tend to demonstrate enhanced threat detection capabilities compared to traditional, single-modal models. Many advanced studies report similar or slightly higher accuracy rates, but these often come at the cost of increased complexity, larger labelled datasets, and extensive computational resources. Our model, while currently focused on textual data extracted from various attachment types, achieves competitive accuracy that underscores the strength of text-based analysis as a foundational approach.

However, direct comparisons between models remain inherently limited due to differences in datasets, evaluation methods, and experimental setups found in the literature. Many existing works utilize proprietary or large-scale datasets, often unavailable for public benchmarking, making it challenging to assert definitive superiority or equivalence.

Importantly, Chapter Two highlighted emerging trends emphasizing true multimodal fusion— combining text, visual, and behavioural data streams in real-time—to overcome limitations seen in unimodal systems. Although the current implementation processes textual content extracted from attachments, integrating image features directly, alongside metadata and behavioural patterns, offers promising avenues for future enhancement. Such multimodal fusion approaches are expected to improve detection rates, reduce false positives, and increase robustness against adversarial attempts.

In summary, the model’s high accuracy demonstrates solid baseline performance aligned with existing text-focused approaches. It also sets a foundation for further development into comprehensive multimodal AI solutions that reflect the cutting-edge direction of cybersecurity research.

**4.4 Discussion of Findings**

The results obtained in this study provide meaningful insights that align closely with the original research objectives and hypotheses.

Objective 1: Threat Analysis

The high test accuracy of 0.99 achieved by the text-based model confirms that analysing the textual content of emails and attachments is a strong and reliable method for detecting phishing attempts. Even when text is extracted from other formats, such as PDFs or images via OCR, it retains valuable information that can help identify malicious intent. This supports the hypothesis that phishing indicators are often embedded within the textual data of email communications, making text analysis a critical component of threat detection.

Objective 2: AI System Development

The chosen AI/ML approach—Logistic Regression applied on TF-IDF text features—proved effective in classifying phishing emails with excellent accuracy. The model’s performance demonstrates that traditional machine learning algorithms, when properly tuned and applied to relevant features, can deliver strong results without the immediate need for complex or resource-intensive deep learning architectures. This outcome validates the hypothesis that a well-designed text-based AI system can serve as an efficient tool for phishing detection, particularly in environments where computational resources may be limited.

Objective 3: System Testing and Integration

While the project included the development of a prototype classification tool

(test\_predictor.py) capable of processing individual files, its current implementation reflects a basic level of system testing. The classification results for sample files confirm that the model functions as intended on isolated inputs. However, the findings also highlight the need for comprehensive testing on a dedicated, labelled test set to fully evaluate the system’s robustness, including precision, recall, and F1-score metrics. Such evaluation is essential before practical integration into real-world security infrastructures can be confidently pursued.

Overall, the findings affirm that the project’s approach is on the right track but also point to important next steps in expanding evaluation and exploring multimodal enhancements for improved phishing detection.

**4.5 Implications of the Findings**

The results obtained in this study hold valuable implications for both practical applications in industry and future academic research.

1. Industry Applications: The high test accuracy of 0.99 achieved by the text-based model suggests that even a focused analysis of email content and attachments can provide a reliable means to detect phishing attempts. In a real-world setting, this kind of system could be integrated into existing email security infrastructures as an automated filter. By flagging potentially malicious emails early, the system would enable cybersecurity teams to prioritize investigations and respond more quickly to threats. This not only helps reduce

the volume of phishing emails that reach end-users but also lowers the risk of successful attacks, which can lead to financial loss or data breaches.

Additionally, automating the detection process reduces the reliance on manual inspection, saving time and resources for organizations. The model’s strong performance shows that machine learning approaches, even when limited to text analysis, can significantly enhance the effectiveness of current security tools.

1. Academic and Research Significance: From a research perspective, the findings reinforce the utility of AI and machine learning in cybersecurity, particularly for phishing detection using textual features. This study confirms that well-established techniques such as logistic regression combined with TF-IDF vectorization remain powerful tools for classification tasks in this domain. However, the results also indicate room for advancement. The current approach processes text extracted from various file types, but it does not fully leverage the multimodal nature of modern phishing attempts, which often involve images, metadata, and other non-textual elements. Therefore, the study highlights a clear opportunity for future research to develop truly multimodal models that can analyse and fuse multiple data types. Such advancements could further improve detection rates, reduce false positives, and increase the resilience of phishing detection systems against sophisticated attacks.
2. Broader Impacts: Overall, this project demonstrates a promising foundation that bridges theoretical research with practical cybersecurity needs. By showing that high accuracy is achievable with text-focused models, it encourages the exploration of more complex, multimodal systems that can keep pace with evolving cyber threats. The implications suggest a path forward for both developers and researchers aiming to build smarter, faster, and more comprehensive phishing detection tools.

**4.6 Challenges and Limitations**

During the course of this research, several challenges and limitations became apparent, which impacted both the development and evaluation of the model. Acknowledging these factors is important for understanding the scope and potential improvements of the study.

1. Reliance on Text Conversion of Attachments: One major limitation of the current model is its dependence on converting all email attachments into text form for analysis. While this approach simplifies the input data and enables the use of traditional text-based machine learning methods, it inherently loses information contained in non-textual elements such as images, complex formatting, or embedded multimedia. This limits the model’s ability to detect threats that rely on visual or other non-textual cues, which are common in sophisticated phishing attacks.
2. Need for Comprehensive Evaluation on a Dedicated Test Set: The evaluation carried out so far was based on available samples and the basic classification output from the test\_predictor.py script. However, a more thorough evaluation on a larger, dedicated test set is necessary to fully assess the model’s generalization ability. This would provide detailed performance metrics such as precision, recall, F1-score, and more insightful visualizations like confusion matrices and ROC curves. Without these, it is difficult to fully understand the model’s strengths and weaknesses, especially in real-world scenarios where data variability is much higher.
3. Potential Overfitting Indicated by Perfect Training Accuracy: The model achieved a perfect training accuracy of 1.00, which raises the concern of overfitting. Overfitting occurs when a model learns the training data too well, including noise and minor details, which negatively affects its performance on unseen data. Although the test accuracy remained very high at 0.99, the small difference suggests that the model may not generalize perfectly to new, diverse data sets. This issue highlights the need for strategies such as regularization, cross-validation, or gathering more varied training data to improve model robustness.
4. Limitations of OCR Accuracy for Image Attachments: Since image attachments were processed using Optical Character Recognition (OCR) to extract text, the accuracy of OCR significantly affects the model’s input quality. OCR can struggle with poor image resolution, complex backgrounds, or stylized fonts, resulting in incorrect or incomplete text extraction. These inaccuracies can lead to misclassification or reduced detection capability. This limitation emphasizes the importance of developing models that can analyse images directly or combine OCR with other image-processing techniques to improve multimodal detection.

**CHAPTER FIVE**

**SUMMARY, CONCLUSION AND RECOMMENDATION**

## 5.1 Summary

This project was created to make and apply an intelligent system that helped identify cyber threats in email files; text, images and PDF files. The reason was due to more people using email and at the same time, the increase in cyberattacks that packages phishing, malware, and ransomware within messages’ attachments. Applying a specific approach, the study investigated rule-based, statistical, and AI-based threat detection methods, pointing out that all of them struggle to deal with complex email attachments. Because of this difference, a system was built that combines Natural Language Processing (NLP) and machine learning (ML) to handle different types of files and the information they contain. In this method, email attachments were collected and marked, qualities from every file type were recorded, the model was trained, and it was then used to distinguish between harmless and malicious files. All the feature vectors were turned into text format, so the models could be trained and evaluated in the same way. Python was used to create the system and it is executed from the command line, even though the first prototype is basic, it clearly proves the viability of using AI to secure attachments in different formats. During the evaluation phase, I obtained classification reports that verified the model gave dependable results with the test data. As a result, this project solved the issue by developing an AI tool that combines multimodal data processing with machine learning in the cybersecurity area. The system acts as the basis for developing fast, accurate, and effective email attachment scanners.

## 5.2 Conclusion

This research indicates that using AI along with various tools is much more effective than the earlier rule-based systems used to check emails. With text, images, and PDF specific feature extraction and guided machine learning, the system was able to predict threats on the test files with a high rate.

This project contributes to the field of cybersecurity by introducing a way to unite Natural Language Processing, basic image recognition, and PDF processing into one classification method. The fact that it is currently offline and has a very basic interface proves that intelligent automation is useful in detecting suspicious content in any type of data.

Another key outcome of this research is its demonstration of the practical steps involves in building and training AI-based detection systems in resource-constrained settings. The system is built in a modular way which allows future expansion, integration into enterprise environments, or function as a background screening service for messages sent through email filters.

Despite its success, the system has some limitations; the small datasets, its offline-only operation, and its limited robustness against highly obfuscated or adversarial threats. Additionally, real-time implementation and advanced image/PDF processing capabilities were outside the scope of this prototype.

## 5.3 Future Recommendation

Based on the experience and findings of this project, the following recommendations are proposed for future works:

Firstly, increasing the quality and diversity of the dataset is a priority for future work. Expanding the dataset to include a wider range of industries, file structures, and real-world examples of email-borne threats would significantly improve the model’s generalization and accuracy.

Secondly, introducing deep learning and multimodal fusion presents a promising direction. While classical machine learning models were utilized in this project, future iterations could benefit from applying Convolutional Neural Networks (CNNs) for image analysis and Transformer models such BERT for text classification. A multimodal fusion strategy combining insights from different data types could further boost detection accuracy.

Thirdly, there is a need to develop a real-time detection system. The current implementation functions offline, but future improvements should aim to enable live email scanning with real time notifications. This could be achieved through interactive dashboards or integrations with browser and email client plugins.

Fourthly, ethical and privacy concerns must be taken seriously as the system evolves. Future deployments should include privacy-preserving methods such as local inference, data anonymization, and encryption to ensure the protection of user data.

Lastly, open and community-driven development will be encouraged. By open-sourcing parts of the project and engaging with the cybersecurity community, future contributors can validate, extend, and enhance the system collaboratively by driving innovation and ensuring continued relevance.

**REFERENCES**

Alaeifar, P., Pal, S., Jadidi, Z., Hussain, M., & Foo, E. (2024). Current approaches and future directions for cyber threat intelligence sharing: A survey. *Journal of Information Security and Applications, 83*, 103786.

Arikkat, D. R., Vinod, P., Rehiman, K. A. R., Nicolazzo, S., & Nocera, A. (2024). Relation extraction techniques in cyber threat intelligence. In *International Conference on Applications of Natural Language to Information Systems* (pp. 348–363). Springer.

Basheer, R., & Alkhatib, B. (2021). Threats from the dark: A review over dark web investigation research for cyber threat intelligence. *Journal of Computer Networks and Communications, 2021*, 1–21.

Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cybersecurity intrusion detection. *IEEE Communications Surveys & Tutorials, 18*(2), 1153–1176.

Cascavilla, G., Tamburri, D. A., & Van Den Heuvel, W.-J. (2021). Cybercrime threat intelligence: A systematic multi-vocal literature review. *Computers & Security, 105*, 102258.

Chen, Z., Su, L., & Luo, Y. (2018). Reinforcement learning-based decision-making for cyber defense in large-scale network systems.

Dalziel, H. (2014). *How to define and build an effective cyber threat intelligence capability*. Syngress.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *ArXiv preprint.*

Gao, P., Shao, F., Liu, X., Xiao, X., Qin, Z., Xu, F., Mittal, P., Kulkarni, S. R., & Song, D. (2021). Enabling efficient cyber threat hunting with cyber threat intelligence. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)* (pp. 193–204). IEEE.

Goodman, B., Benkler, Y., Faris, R., & Green, J. (2022). AI in cybersecurity: Pattern recognition for malware and phishing detection.

Huang, C., Xu, Z., Zhang, Q., & Zhang, X. (2020). Reinforcement learning for cyber defense: A review.

Ibrahim, A., Thiruvady, D., Schneider, J.-G., & Abdelrazek, M. (2020). The challenges of leveraging threat intelligence to stop data breaches. *Frontiers in Computer Science, 2*, 36.

McMillan, R. (2022). Definition: Threat intelligence. *Gartner*. [https://www.gartner.com](https://www.gartner.com/)

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *International Conference on Learning Representations (ICLR)*.

Montasari, R., Carroll, F., Macdonald, S., Jahankhani, H., Hosseinian-Far, A., & Daneshkhah, A. (2021). Application of artificial intelligence and machine learning in producing actionable cyber threat intelligence. In *Digital Forensic Investigation of IoT Devices* (pp. 47–64).

Pattabiram, S., & Ramalingam, G. K. (n.d.). AI-enhanced multimodal search engines for cybersecurity threat detection.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)* (pp.

1532–1543).

Rahman, M. R., Hezaveh, R. M., & Williams, L. (2023). What are the attackers doing now? Automating cyberthreat intelligence extraction from text on pace with the changing threat landscape: A survey. *ACM Computing Surveys, 55*(12), 1–36.

Ramos, J. (2003). Using TF-IDF to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning* (Vol. 242, pp. 29–48). Citeseer. Rudd, E. M., et al. (n.d.). MEADE: Towards a malicious email attachment detection engine.

Samtani, S., Kantarcioglu, M., & Chen, H. (2020). Trailblazing the artificial intelligence for cybersecurity discipline: A multi-disciplinary research roadmap. *ACM Transactions on Management Information Systems, 11*(4), 1–19.

Shackleford, D. (2015). Who’s using cyberthreat intelligence and how. *SANS Institute*.

Shahid, M., Mehmood, A., & Zameer, A. (2020). Email spam and malware detection using support vector machines.

Sun, N., Zhang, J., Rimba, P., Gao, S., Zhang, L. Y., & Xiang, Y. (2018). Data-driven cybersecurity incident prediction: A survey. *IEEE Communications Surveys & Tutorials, 21*(2), 1744– 1772.

Sun, N., Ding, M., Jiang, J., Xu, W., Mo, X., Tai, Y., & Zhang, J. (2023). Cyber threat intelligence mining for proactive cybersecurity defense: A survey and new perspectives. *IEEE Communications Surveys & Tutorials*.

Tounsi, W., & Rais, H. (2018). A survey on technical threat intelligence in the age of sophisticated cyber attacks. *Computers & Security, 72*, 212–233.

Wagner, T. D., Mahbub, K., Palomar, E., & Abdallah, A. E. (2019). Cyber threat intelligence sharing: Survey and research directions. *Computers & Security, 87*, 101589.

Zhang, Y., Wang, J., & Zhou, J. (2020). Unsupervised learning methods in zero-day attack detection.

Zhang, Z., et al. (2021). AI and blockchain integration for cybersecurity: A comprehensive survey.

Zhao, J., Yan, Q., Li, J., Shao, M., He, Z., & Li, B. (2020). TIMiner: Automatically extracting and analyzing categorized cyber threat intelligence from social data. *Computers & Security, 95*, 101867.

**APPENDICES**

<https://github.com/nonymoustea/cyberthreatdetectionsystem.git>

https://www.kaggle.com/datasets/naserabdullahalam/phishing-email-dataset