PHISHING BASED SOCIAL ENGINEERING DETECTION IN AI

NURSYUHADAH BINTI AHMAD SUDERMAN

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

**BORANG PENGESAHAN STATUS LAPORAN**

JUDUL: [PHISHING BASED SOCIAL ENGINEERING DETECTION IN AI]

SESI PENGAJIAN: [2024 / 2025]

Saya: \_\_\_\_[NURSYUHADAH BINTI AHMAD SUDERMAN]\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

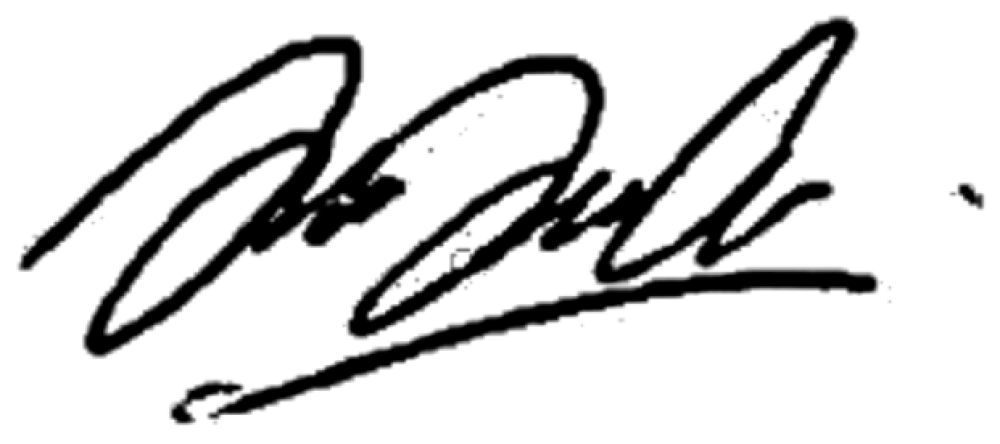
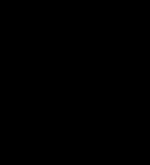
mengaku membenarkan tesis Projek Sarjana Muda ini disimpan di Perpustakaan Universiti Teknikal Malaysia Melaka dengan syarat-syarat kegunaan seperti berikut:

1. Tesis dan projekadalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan unituk tujuan pengajian sahaja.
3. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. \* Sila tandakan (✓)

|  |  |  |
| --- | --- | --- |
| \_\_\_\_\_\_\_\_\_\_ | SULIT | (Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972) |
| \_\_\_\_\_\_\_\_\_\_ | TERHAD | (Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi / badan di mana penyelidikan dijalankan) |
| \_\_\_\_✓\_\_\_\_ | TIDAK TERHAD |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |
| (TANDATANGAN PELAJAR) |  | (TANDATANGAN PENYELIA) |
| Alamat tetap: E-7-9 P/A DESA REJANG PERSIARAN REJANG, TAMAN SETAPAK JAYA 53300, WILAYAH PERSEKUTUAN KUALA LUMPUR |  | DR ZAHEERA ZAINAL ABIDIN |
|  |  |  |
| Tarikh: \_\_\_\_27-06-2025\_\_\_\_\_\_\_\_\_\_\_\_ |  | Tarikh: \_\_\_26-06-2025\_\_\_\_\_\_\_\_\_\_\_\_ |

CATATAN: \* Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa.



CATATAN: \* Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak

PHISHING BASED SOCIAL ENGINEERING DETECTION IN AI

NURSYUHADAH BINTI AHMAD SUDERMAN

This report is submitted in partial fulfillment of the requirements for the

Bachelor of [Computer Science (Computer Security)] with Honours.

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2025

# DECLARATION

I hereby declare that this project report entitled

[PHISHING BASED SOCIAL ENGINEERING DETECTION IN AI]

is written by me and is my own effort and that no part has been plagiarized

without citations.

STUDENT : NURSYUHADAH BINTI AHMAD SUDERMAN Date : 27-06-2025

I hereby declare that I have read this project report and found

this project report is sufficient in term of the scope and quality for the award of

Bachelor of [Computer Science (Computer Security)] with Honours.

SUPERVISOR : DR ZAHEERA BINTI ZAINAL ABIDIN Date : 26-06-2025

# DEDICATION

This work is dedicated to Allah Almighty, my creator, pillar, and source of wisdom, knowledge, and insight. Throughout this journey, He has been my source of strength, and I have only been able to fly on His wings.

To my beloved parents, Ahmad Suderman Bin Mohd Yaacob and Norayzuazlin Binti Mohd Ayob whom I wholeheartedly dedicate this study to. Thank you for being my source of inspiration and for your unwavering support and trust. To my brothers, whom I hold very dear to my heart.

To my supervisor, Dr Zaheera binti Zainal Abidin, for supporting and believing in me to accomplish this project. Finally, I want to thank my friends for being there for me during my bachelor's degree, sharing words of advice and support to help me accomplish this study.

# ACKNOWLEDGEMENTS

First and foremost, First and foremost, I would like to praise and thank Allah the Almighty, the Most Gracious, and the Most Merciful for His blessing given to me in completing this project. Peace and salutations to the Prophet Muhammad S.A.W, who has led all humanity from darkness to light.

I would like to express my heartfelt appreciation to Dr. Zaheera binti Zainal Abidin for her assistance in completing this assignment successfully. Her insightful thoughts have provided various inspirations and enlightenment for this project.

I'd also want to thank my dear parents and sisters for their encouragement and support during this effort. Their help and concern had given me confidence while I worked on this project. Finally, I am grateful to my friends who are ready to share their wisdom and have directly or indirectly guided me during this project.

# ABSTRACT

Social engineering attacks pose a significant and growing threat in the cybersecurity landscape, exploiting human psychology rather than technical vulnerabilities to manipulate individuals into compromising sensitive information. Among various social engineering techniques, phishing has emerged as the most prevalent and damaging, driven by its scalability, low execution cost, and evolving tactics such as spear phishing and smishing. As traditional rule-based detection methods struggle to keep up with these dynamic threats, Artificial Intelligence (AI) offers a robust and adaptive solution. This study explores AI-based approaches—specifically Machine Learning (ML), Deep Learning (DL), and Reinforcement Learning (RL)—for the detection of phishing attacks. It provides a comprehensive literature review, analysis of current research trends, and detailed examination of relevant datasets and algorithms including Support Vector Machines (SVM), Long Short-Term Memory (LSTM), and Deep Q-Networks (DQN). The research highlights the advantages of AI in automating threat detection, enhancing accuracy, and enabling systems to adapt to new attack vectors. By focusing on phishing as a primary social engineering threat, this work aims to contribute to the development of intelligent, scalable, and real-time defenses against increasingly sophisticated cyber threats.

# ABSTRAK

Serangan kejuruteraan sosial merupakan ancaman yang semakin ketara dalam landskap keselamatan siber, di mana ia mengeksploitasi psikologi manusia dan bukannya kelemahan teknikal untuk memanipulasi individu agar mendedahkan maklumat sensitif. Dalam pelbagai teknik kejuruteraan sosial, serangan *phishing* telah muncul sebagai bentuk yang paling meluas dan merosakkan, didorong oleh keupayaannya untuk berkembang secara besar-besaran, kos pelaksanaan yang rendah, serta taktik yang sentiasa berubah seperti *spear phishing* dan *smishing*. Oleh kerana kaedah pengesanan berasaskan peraturan tradisional semakin sukar menandingi ancaman yang dinamik ini, Kecerdasan Buatan (AI) menawarkan penyelesaian yang kukuh dan adaptif. Kajian ini meneroka pendekatan berasaskan AI—khususnya Pembelajaran Mesin (ML), Pembelajaran Mendalam (DL), dan Pembelajaran Penguatan (RL)—dalam pengesanan serangan *phishing*. Ia merangkumi ulasan literatur yang komprehensif, analisis *trend* penyelidikan semasa, serta pemeriksaan terperinci terhadap set data dan algoritma yang relevan termasuk *Support Vector Machines* (SVM), *Long Short-Term Memory* (LSTM), dan *Deep Q-Networks* (DQN). Kajian ini menekankan kelebihan AI dalam mengautomasikan pengesanan ancaman, meningkatkan ketepatan, dan membolehkan sistem menyesuaikan diri dengan vektor serangan baharu. Dengan memberi tumpuan kepada *phishing* sebagai ancaman utama kejuruteraan sosial, kajian ini bertujuan menyumbang kepada pembangunan sistem pertahanan yang pintar, berskala, dan masa nyata terhadap ancaman siber yang semakin canggih.

# TABLE OF CONTENTS

[DECLARATION ii](#_Toc204531376)

[DEDICATION iii](#_Toc204531377)

[ACKNOWLEDGEMENTS iv](#_Toc204531378)

[ABSTRACT v](#_Toc204531379)

[ABSTRAK vi](#_Toc204531380)

[TABLE OF CONTENTS vii](#_Toc204531381)

[list of tables x](#_Toc204531382)

[list of figures xi](#_Toc204531383)

[List of Abbreviations xiii](#_Toc204531384)

[List of ATTACHMENTS xiv](#_Toc204531385)

[Chapter 1: INTRODUCTION 1](#_Toc204531386)

[1.1 Introduction 1](#_Toc204531387)

[1.2 Problem Statement 2](#_Toc204531388)

[1.3 Project Question 3](#_Toc204531389)

[1.4 Objective 4](#_Toc204531390)

[1.5 Project Scope 5](#_Toc204531391)

[1.6 Project Contribution 6](#_Toc204531392)

[1.7 Thesis Organization 6](#_Toc204531393)

[1.8 Summary 8](#_Toc204531394)

[Chapter 2: literature review 9](#_Toc204531395)

[2.1 Introduction 9](#_Toc204531396)

[2.2 Related Work 11](#_Toc204531397)

[2.2.1 Phishing 12](#_Toc204531398)

[2.2.2 Machine Language Approaches 13](#_Toc204531399)

[2.2.3 Deep Learning Approaches 18](#_Toc204531400)

[2.2.4 Reinforcement Learning Approaches 22](#_Toc204531401)

[2.3 Analysis 26](#_Toc204531402)

[2.3.1 Analysis of Dataset 26](#_Toc204531403)

[2.3.2 Performance metrics 29](#_Toc204531404)

[2.4 Critical Review 31](#_Toc204531405)

[2.5 Research Gap Analysis 46](#_Toc204531406)

[2.6 Project Solution 48](#_Toc204531407)

[Chapter 3: METHODOLOGY 49](#_Toc204531408)

[3.1 Introduction 49](#_Toc204531409)

[3.2 Methodology 49](#_Toc204531410)

[3.2.1 Requirement Analysis 50](#_Toc204531411)

[3.2.2 System Design 51](#_Toc204531412)

[3.2.3 Implementation 52](#_Toc204531413)

[3.2.4 Testing 52](#_Toc204531414)

[3.2.5 Deployment 52](#_Toc204531415)

[3.2.6 Previously proposed methodology 54](#_Toc204531416)

[3.3 Project Schedule and Milestones 55](#_Toc204531417)

[3.3.1 Project Milestones 55](#_Toc204531418)

[3.3.2 Project Gantt Chart 56](#_Toc204531419)

[3.4 Summary 57](#_Toc204531420)

[Chapter 4: DESIGN 58](#_Toc204531421)

[4.1 Introduction 58](#_Toc204531422)

[4.2 Design Framework of Hybrid Approach for Phishing Detection in AI 59](#_Toc204531423)

[4.3 Summary 61](#_Toc204531424)

[Chapter 5: IMPLEMENTATION 62](#_Toc204531425)

[5.1 Introduction 62](#_Toc204531426)

[5.2 Summary 62](#_Toc204531427)

[Chapter 6: TESTING AND ANALYSIS 62](#_Toc204531428)

[6.1 Introduction 62](#_Toc204531429)

[6.2 Result and Analysis 62](#_Toc204531430)

[6.3 Summary 62](#_Toc204531431)

[Chapter 7: CONCLUSION 63](#_Toc204531432)

[7.1 Introduction 63](#_Toc204531433)

[7.2 Project Summarization 63](#_Toc204531434)

[7.3 Project Contribution 63](#_Toc204531435)

[7.4 Project Limitation 63](#_Toc204531436)

[7.5 Future Works 63](#_Toc204531437)

[7.6 Summary 63](#_Toc204531438)

[references 64](#_Toc204531439)

[Bibliography 65](#_Toc204531440)

[APPENDICES 66](#_Toc204531441)

# list of tables

|  |  |
| --- | --- |
|  | **PAGE** |

[Table 1.2‑1 Problem Statement 2](#_Toc112842306)

[Table 1.3‑1 Project Question 3](#_Toc112842307)

[Table 1.4‑1 Project Objective 4](#_Toc112842308)

[Table 1.6‑1 Project Contribution 5](#_Toc112842309)

[Table 2.4‑1: Summary of Research Papers Reviewed 22](#_Toc112842310)

[Table 3.2‑1: Hardware Requirement 28](#_Toc112842311)

[Table 3.2‑2: Software Requirement 29](#_Toc112842312)

[Table 3.2‑3: Example of Comparison Table generated after analysis of data is finished 31](#_Toc112842313)

[Table 4.4‑1: Possible Scenarios 40](#_Toc112842314)

[Table 5.4‑1 below shows the value of bandwidth usage, latency and throughput respective to the three protocols studied. 56](#_Toc112842315)

[Table 5.4‑1: Comparison Table from Collected Data 56](#_Toc112842316)

[Table 5.5‑1: Implementation Status 57](#_Toc112842317)

# list of figures

|  |  |
| --- | --- |
|  | **PAGE** |

Figure 1: Literature Review Structure Diagram 8

Figure 2: Action varieties in Social Engineering incidents 9

Figure 3: CoAP Protocol Layers (payatu.com) 10

Figure 4: Confirmable Message (payatu.com) 11

Figure 5: Non-confirmable message (payatu.com) 11

Figure 6: Publish/Subscribe model in MQTT (Martí et al., 2019) 13

Figure 7: QoS0 (Sueda et al., 2019) 14

Figure 8: QoS1 (Sueda et al., 2019) 15

Figure 9: QoS2 (Sueda et al., 2019) 15

Figure 10: AMQP Architecture (Anusha, et al., 2017) 16

Figure 11: Sensors used to collect temperature, humidity and light (Corak et al., 2018) 18

Figure 12: Diagram of testbed’s processing unit (Corak et al., 2018) 18

Figure 13: Topology for (a) Experiment I and (b) Experiment II (Moraes et al., 2019) 19

Figure 14: Diagram of hardware used for the testbed (Pohl et al., 2018) 20

Figure 15: Experimental setup (Mijovic et al., 2016) 21

Figure 16 Waterfall Model 28

Figure 17: Data Processing and Obtaining Result 30

Figure 18: Example of Graph that compares three protocols based on Bandwidth Utilisation 31

Figure 19: Network System Architecture 36

Figure 20: Logical Design 38

Figure 21: Physical Design 39

Figure 22: Scenario 1: CoAP 40

Figure 23: Scenario 2: AMQP 40

Figure 24: Scenario 3: MQTT 41

Figure 25: CoAP server running at Server 44

Figure 26: CoAP client running at PC1 44

Figure 27: CoAP Client sending GET message 45

Figure 28: Server received CoAP messages from PC1 45

Figure 29: Wireshark Packet Capture (CoAP) 46

Figure 30: RabbitMQ installed in Server browser 46

Figure 31: Python code for sending message 47

Figure 32: Python code for receiving message 48

Figure 33: RabbitMQ Management Interface 48

Figure 34: Wireshark Packet Capture (AMQP) 49

Figure 35: Moisquitto loaded and running 50

Figure 36: Connection details 50

Figure 37: PC1 MQTTlens interface 51

Figure 38: PC2 MQTTlens interface 51

Figure 39: PC2 received 100 messages from PC1 52

Figure 40: Wireshark Packet Capture (MQTT) 52

Figure 41: Protocol Hierarchy for CoAP 53

Figure 42: Protocol Hierarchy for AMQP 53

Figure 43: Protocol Hierarchy for MQTT 54

Figure 44: Packet Length for CoAP 54

Figure 45: Packet Length for AMQP 55

Figure 46: Packet Length for MQTT 55

Figure 47: Conversations Statistics for CoAP 56

Figure 48: Conversations Statistics for AMQP 56

Figure 49: Conversations Statistics for MQTT 56

Figure 50: Bandwidth Usage Graph 60

Figure 51: Latency Graph 61

Figure 52: Throughput Graph 61

Figure 53: Average Packet Size, B (CoAP) from Wireshark 62

Figure 54: Average Packet Size, B (MQTT) from Wireshark 62

# List of Abbreviations

|  |  |  |
| --- | --- | --- |
| **IoT** | **-** | **Internet of Things** |
| **CoAP** | **-** | **Constrained Application Protocol** |
| **AMQP** | **-** | **Advanced Message Queuing Protocol** |
| **MQTT** | **-** | **Message Queue Telemetry Transport** |
| **TCP** | **-** | **Transmission Control Protocol** |
| **UDP** | **-** | **User Datagram Protocol** |

# List of ATTACHMENTS

|  |  |  |  |
| --- | --- | --- | --- |
|  | | | **PAGE** |
|  | | |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# INTRODUCTION

## Introduction

Social engineering attacks are among the most persistent and evolving cybersecurity threats, leveraging psychological manipulation to deceive individuals into disclosing confidential information or compromising systems. Techniques such as phishing, baiting, and pretexting exploit trust, urgency, or curiosity, making them difficult to detect using traditional, rule-based security systems. Phishing stands out as a dominant threat due to its scalability, ease of execution, and the increasing sophistication of its tactics.

This study explores the role of Artificial Intelligence (AI) in detecting social engineering attacks, with a primary focus on phishing. The research involves a comprehensive analysis of existing AI approaches—specifically Machine Learning (ML), Deep Learning (DL), and Reinforcement Learning (RL)—to evaluate their effectiveness, challenges, and potential in improving detection accuracy. Rather than developing a prototype, this study critically reviews current methods, datasets, and models to identify strengths, limitations, and areas for improvement in AI-based social engineering detection systems.

## Problem Statement

The problem that has been identified is summarized in Table 1-1 below:

Table 1.2‑1 Problem Statement

|  |  |
| --- | --- |
| **PS** | **Problem Statement** |
| **PS1** | Existing AI models struggle to detect the constantly changing nature of social engineering attacks, resulting in lower detection accuracy. |

**PS1: Existing AI models struggle to detect the constantly changing nature of social engineering attacks, resulting in lower detection accuracy.**

Existing AI models face challenges in accurately detecting social engineering attacks due to the dynamic and evolving tactics used by attackers. Machine learning has gradually introduced a huge basket of “AI capabilities” that can be harnessed for social engineering and phishing attacks (Schmitt & Flechais, 2024). This constant change reduces the effectiveness of traditional detection approaches, leading to lower overall detection accuracy.

## Project Question

Three Project Question (PQ) is constructed based on the problem statement that needs to be answered in this project. The summary of project question is shown in Table 1-2.

Table 1.3‑1 Project Question

|  |  |  |
| --- | --- | --- |
| **PS** | **PQ** | **Project Question** |
| PS1 | PQ1 | |  | | --- | |  |  |  | | --- | | What are the existing phishing techniques for detecting social engineering attacks in an AI environment? | |
| PQ2 | How accurate are existing phishing techniques in detecting social engineering attacks? |
| PQ3 | How does the accuracy performance of the proposed approach compare with existing phishing techniques in detecting social engineering attacks? |

**PQ1: What are the existing phishing techniques for detecting social engineering attacks in an AI environment?**

This question explores the range of phishing detection methods, including machine learning, deep learning, and reinforcement learning approaches, which are commonly used to identify social engineering attacks in digital communications.

**PQ2: How does the accuracy performance of the proposed approach compare with existing phishing techniques in detecting social engineering attacks?**

This question focuses on measuring the performance of existing phishing detection approaches, considering metrics such as accuracy, precision, recall, and overall detection rates.

**PQ3: How does the accuracy performance of the proposed approach compare with existing phishing techniques in detecting social engineering attacks?**

[nanti buat balik]

## Objective

Based on the project questions formulated in the previous section, appropriate project objectives (PO) are developed. The Project Objective (PO) is summarized in Table 1-3.

Table 1.4‑1 Project Objective

|  |  |  |  |
| --- | --- | --- | --- |
| **PS** | **PQ** | **PO** | **Project Objective** |
| PS1 | PQ1 | PO1 | To investigate existing phishing techniques for detecting social engineering attacks in AI environment. |
| PQ2 | PO2 | To analyze accuracy performance of phishing in detecting social engineering attacks. |
| PQ3 | PO3 | To evaluate the performance of proposed approach in detecting social engineering attacks. |

**PO1: To investigate existing phishing techniques for detecting social engineering attacks in AI environment.**

Understanding current phishing detection techniques is essential for identifying the strengths and limitations of existing approaches. This investigation provides a foundation for developing more effective AI-driven methods to counter social engineering attacks.

**PO2: To analyze accuracy performance of phishing in detecting social engineering attacks.**

Evaluating the accuracy of phishing detection techniques is crucial for assessing their reliability and effectiveness in real-world scenarios. This analysis will highlight the strengths and weaknesses of existing methods.

**PO3: To evaluate the performance of proposed approach in detecting social engineering attacks.**

[nanti buat balik]

## Project Scope

The project’s scopes are focused on the following aspects:

1. The AI tools considered for phishing attack detection are Machine Learning (ML), Deep Learning (DL), and Reinforcement Learning (RL).
2. The performance of the AI model will be evaluated based on accuracy, precision, recall, and F1-score.
3. The platform used for training and testing the AI models is WEKA and KNIME

## Project Contribution

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **PS** | **PQ** | **PO** | **PC** | **Project Contribution** |
| PS1 | PQ1 | PO1 | PC1 | Comprehensive review of AI techniques (ML, DL, RL) for phishing detection. |
| PQ2 | PO2 | PC2 | A summary of performance metrics (accuracy precision, recall, F1-score) in current AI models. |
| PQ3 | PO3 | PC3 | Proposed hybrid AI framework to improve detection accuracy beyond baseline models. |

## Thesis Organization

This report is divided into six chapters, namely Chapter 1: Introduction, Chapter 2: Literature Review, Chapter 3: Methodology, Chapter 4: Design, Chapter 5: Implementation, Chapter 6: Discussion and Chapter 7: Conclusion.

**Chapter 1: Introduction**

This chapter discusses the project's introduction, background, research problem, research questions, objectives, scope, significance of the project, and report organization. It sets the stage for the problem addressed and outlines the goals of the project in the context of AI-based social engineering attack detection.

**Chapter 2: Literature Review**

This chapter provides a review of relevant literature, discussing previous research on phishing attacks, existing detection methods, and AI techniques such as Machine Learning (ML), Deep Learning (DL), and Reinforcement Learning (RL). It also identifies gaps in the current literature and frames the research questions and objectives.

**Chapter 3: Methodology**

This chapter outlines the research methodology adopted for the project, including the datasets used, AI techniques implemented (ML, DL, RL), and the performance evaluation metrics. It provides a step-by-step explanation of the model development, data preprocessing, feature extraction, and model training processes.

**Chapter 4: Analysis and Design**

This chapter explains the project's implementation in detail, including how the project was executed and how the results were generated.

**Chapter 5: Implementation**

This chapter will cover the installation and configuration of the project's environment, detailing the software used and how it was set up for the current stage of implementation.

**Chapter 6: Testing**

This section focuses on testing and validating the project using the dataset to obtain results. It also presents a comprehensive discussion of the findings and outcomes.

**Chapter 7: Conclusion**

This chapter includes the introduction, a summary of the project, its contributions, limitations, suggestions for future work, and the conclusion.

## Summary

The problem description, aim, scope of the project, relevance, and projected output of the project are all clearly outlined in this chapter. The description of the related works to this project is detailed later in Chapter 2.

# literature review

## Introduction

Social engineering is a psychological manipulation technique where cybercriminals exploit human trust, fear, or urgency to deceive individuals into performing actions that compromise security, such as disclosing confidential information or enabling system access. Instead of relying on traditional hacking tools to exploit software or hardware vulnerabilities, social engineering exploits human behavior, making it one of the most dangerous and successful forms of attack in the modern threat landscape. Techniques include phishing, baiting, pretexting, vishing, and quid pro quo.

Among these techniques, phishing remains the most widespread and damaging form of social engineering attack, and its prevalence has been supported by numerous global studies and industry analyses. Research shows that phishing is consistently ranked as the most common social engineering threat due to its effectiveness, ease of execution, and adaptability across communication channels (Alzahrani et al., 2021; Basit et al., 2022). Attackers exploit digital platforms, particularly email, to impersonate legitimate institutions and deceive users into providing sensitive information.

What sets phishing apart from other social engineering methods is its ability to scale rapidly and target both individuals and organizations with minimal resources. Unlike baiting or pretexting, phishing campaigns can be automated and distributed widely using email bots and phishing kits available on underground forums. These features make phishing a recurring top threat in annual cybersecurity threat reports by multiple academic and professional organizations.

The continued evolution of phishing tactics, from generic spam to personalized spear phishing and smishing (SMS phishing), has contributed to its resilience and growth. This explains why, compared to other forms of social engineering attacks, phishing is the focus of this research. It provides not only a well-documented attack vector but also a diverse set of datasets and practical applications for testing AI-based detection models.

As the complexity and frequency of phishing attacks grow, the need for intelligent, adaptive defense systems becomes critical. Traditional security measures often fall short, especially against zero-day phishing attacks or personalized spear phishing. This has led to a strong shift toward Artificial Intelligence (AI), specifically Machine Learning (ML), Deep Learning (DL), and Reinforcement Learning (RL), as a robust solution for phishing detection. AI-based models can process large volumes of data, identify hidden patterns, and adapt over time, enabling more accurate and timely threat detection.

Given phishing's prominence and evolving nature, this study focuses exclusively on AI-based phishing detection. Compared to other forms of social engineering, phishing offers both a greater volume of research data and clearer use cases for AI implementation, making it an ideal focal point for academic and practical analysis. The subsequent sections provide a deep dive into AI techniques, available software tools, datasets, and recent research that contribute to building effective, intelligent phishing detection systems.



Figure 1: Literature Review Structure Diagram (BETULKAN)

## Related Work

Recent advancements between 2020 and 2024 have demonstrated the growing impact of AI in combating phishing threats. Researchers have developed increasingly sophisticated models leveraging ML, DL, and RL to classify, detect, and respond to phishing activities in real time.

Rahman et al. (2022) introduced a hybrid model combining Decision Trees and anomaly detection, achieving improved phishing email detection rates in enterprise environments. Their work highlighted the effectiveness of ensemble methods in capturing varied attack patterns. Aljohani and Hossain (2021) employed the BERT transformer model to detect phishing content by analyzing contextual semantics within emails. Their results showed significant accuracy improvements over traditional ML models.

Zhao et al. (2020) used Long Short-Term Memory (LSTM) networks to detect phishing attempts in mobile text messages. LSTM’s ability to process sequential information proved critical in identifying suspicious messages that mimic legitimate communications. Nasir et al. (2023) proposed a CNN-based phishing detection system that integrated user behavior logs with content analysis. The combination of behavioral and textual features enhanced detection performance, particularly for spear-phishing cases.

Asker and Essa (2024) emphasized the potential of unsupervised models, such as Autoencoders, to identify phishing emails without labelled data. Their findings suggest these models are well-suited for environments where phishing strategies evolve rapidly and labelled examples are scarce. Zhang et al. (2023) explored the use of Reinforcement Learning to create adaptive phishing detection systems. Their RL-based framework optimized decision-making policies in real-time, improving resilience against evolving attack vectors while minimizing false positives.

These studies demonstrate a transition from static rule-based systems to adaptive, intelligent models that integrate language understanding, behavior analysis, and real-time learning. They underscore the importance of hybrid approaches and model interpretability in building effective phishing detection solutions.

### Phishing

Phishing is a cyber-attack technique in which malicious actors impersonate trustworthy entities to deceive victims into revealing sensitive information, such as usernames, passwords, credit card details, or login credentials. These attacks often take the form of emails, websites, or messages that closely resemble legitimate communication from known organizations. The purpose of phishing is typically to commit fraud, gain unauthorized access to systems, or install malware.

Phishing continues to be the most prevalent type of social engineering attack globally. Its popularity among attackers stems from its high success rate, scalability, and ability to exploit human behavior rather than technical vulnerabilities. According to the 2023 Verizon Data Breach Investigations Report, phishing accounted for more than one-third of all social engineering-related incidents. The Anti-Phishing Working Group (APWG) also reported over 1.2 million phishing attacks in a single quarter, indicating a steady and alarming upward trend. These attacks have evolved from generic bulk emails to more targeted approaches such as spear phishing, business email compromise (BEC), and smishing (SMS phishing), which are increasingly difficult to detect. The figure below shows that Pretexting is now more prevalent than Phishing in Social Engineering incidents. However, when we look at confirmed breaches, Phishing is still on top (Verizon, 2023).



**Figure 2: Action varieties in Social Engineering incidents (Verizon, 2023)**

Due to the sophistication and frequency of phishing, AI-based techniques have become central to its detection. Machine Learning approaches such as Support Vector Machine (SVM) and Random Forest are commonly used due to their strong classification capabilities on structured datasets. These models analyze features such as email headers, URL characteristics, and content patterns., models like Long Short-Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT) have proven effective in understanding the semantic structure of emails, enabling the identification of sophisticated phishing attempts that evade traditional filters.

Recently, Reinforcement Learning (RL) has emerged as a powerful approach for dynamic and adaptive phishing detection. Unlike ML and DL, which learn from static datasets, RL enables an agent to learn optimal policies through interactions with an environment, receiving feedback in the form of rewards or penalties. Techniques such as Q-learning, Deep Q-Networks (DQN), and Policy Gradient Methods are used to develop adaptive systems that can fine-tune detection thresholds based on evolving user behavior and attack strategies. These RL-based methods are especially useful in real-time systems, where phishing tactics change frequently and require continuous learning to maintain detection accuracy.

### Machine Language Approaches

**Machine Learning (ML)** is a core subset of Artificial Intelligence (AI) that focuses on creating algorithms and models that enable computers to automatically learn and improve from experience without being explicitly programmed for specific tasks. At its essence, ML empowers systems to identify patterns, make predictions, and adapt to new data through iterative learning processes. This adaptability is achieved by training models on vast amounts of labelled or unlabelled data and optimizing performance based on measurable outcomes (GeeksforGeeks, 2023).

Modern ML approaches are categorized into three main types: supervised learning, where models learn from labelled data; unsupervised learning, where models uncover hidden structures in unlabelled data; and reinforcement learning, where agents learn optimal behaviors by interacting with environments and receiving feedback in the form of rewards or penalties (Kelleher, 2020). ML is widely applied across domains such as cybersecurity, finance, healthcare, and e-commerce due to its ability to process and learn from high-dimensional, complex datasets.

In the context of cybersecurity, ML plays a pivotal role in the detection and mitigation of threats like phishing attacks, malware, and network intrusions. Techniques such as Support Vector Machines (SVM), Random Forest, and Naïve Bayes have demonstrated high effectiveness in classifying malicious content by learning from historical attack patterns and extracting relevant features from structured or textual data (Alzahrani et al., 2021; Kaur & Arora, 2023).

#### Support Vector Machines (SVM)

Support Vector Machine (SVM) model as a hyper plane in which several classes are represented. So that we may reduce the amount of mistakes we make, SVM will create the hyper plane in an iterative fashion. SVM's purpose is to classify datasets such that a maximum marginal hyper plane may be discovered (Jain & Gupta, 2023).

The high accuracy of SVMs in phishing detection has been demonstrated in various studies. For instance, research by Gupta et al. (2023) highlights that SVMs, when properly tuned, can achieve remarkable detection rates and low false positive rates, making them a reliable choice for cybersecurity applications. However, the effectiveness of SVMs comes with certain challenges. The considerable parameter adjustment necessary to maximize the model's performance is one of its primary drawbacks. It can take some time to carefully choose parameters like the kernel parameters and the penalty parameter (C) to balance the choices between variance and bias.

The computational expense associated with Support Vector Machines (SVMs), particularly when using non-linear kernels, can be significant. Training an SVM model involves solving a complex optimization problem, which can be computationally intensive, especially when dealing with large datasets typically encountered in phishing detection systems. This computational burden can limit the practicality of SVMs in real-time phishing detection, where fast processing is crucial to ensure timely responses and effective protection. Despite these challenges, the high precision and robustness of SVMs in identifying complex patterns in data make them a valuable asset in phishing detection frameworks.

To mitigate these limitations, hybrid approaches that combine SVMs with other machine learning techniques have been explored. These methods aim to capitalize on the strengths of SVMs while addressing their computational drawbacks. For example, integrating SVMs with feature selection techniques can reduce data dimensionality, lessening the computational load and improving real-time applicability. As cyber threats continue to evolve, the role of SVMs in phishing detection remains critical, providing a balance of high accuracy and adaptability necessary for effective cybersecurity defences.



Figure 3: SVM algorithm (Jain & Gupta, 2023)

#### Random Forest

Random Forest is a well-known machine learning method that performs excellently when dealing with regression and classification problems. It creates a lot of decision trees during training, outputs the mode of classes for classification tasks or the mean prediction for regression tasks and increases accuracy and generalization ability by voting or averaging over all trees. Because of the ensemble technique, the model resists overfitting. Because Random Forest can handle big datasets with high dimensionality and is flexible and effective, it is widely employed. It is also resistant to noise and outliers and provides insights into feature importance. Visualization techniques can aid in model evaluation and decision-making (Gunjan & Prasad, 2024).

**Random Forest Technique**

Random Forest (RF) is a powerful ensemble learning algorithm that has shown strong performance in phishing detection due to its ability to handle complex patterns and noisy data. It operates by building multiple decision trees using different random subsets of the training dataset—a process known as bootstrapping. Each tree is trained independently, which introduces diversity and reduces the risk of overfitting. This is especially beneficial in phishing detection, where malicious patterns can be subtle and varied across different instances. Once the forest is constructed, each decision tree contributes to the prediction process. In classification tasks such as distinguishing between phishing and legitimate activities, each tree casts a "vote" for a class label. The final decision is based on majority voting, where the class with the most votes becomes the output. This collective decision-making mechanism improves both accuracy and stability, making Random Forest well-suited for phishing detection systems that require consistent performance under diverse threat scenarios. Refer to the figure below, which illustrates how multiple decision trees independently classify an instance and then combine their outputs through majority voting to determine the final class.



**Figure 4: Random Forest Algorithm**

### Deep Learning Approaches

**Deep Learning (DL)** is a specialized subfield of Machine Learning (ML) that utilizes artificial neural networks with multiple layers—referred to as **deep neural networks**—to automatically learn complex patterns from large volumes of data. Inspired by the structure and function of the human brain, DL systems can model intricate relationships and hierarchical representations in data, enabling them to outperform traditional ML algorithms in tasks involving high-dimensional, unstructured inputs such as images, audio, and text.

In the context of cybersecurity, and particularly phishing detection, deep learning techniques offer significant advantages. Unlike conventional ML models that rely heavily on manually engineered features, DL models can perform **automatic feature extraction**, allowing them to learn hidden patterns in raw email content, URLs, or user interaction logs. Common DL models applied in phishing detection include:

* **Long Short-Term Memory (LSTM)**: A type of Recurrent Neural Network (RNN) that is effective in capturing sequential dependencies in phishing emails or SMS messages.
* **Transformer models (e.g., BERT)**: These models understand the contextual meaning of words in sentences, making them highly effective in detecting sophisticated and linguistically deceptive phishing emails.

The adaptability and high accuracy of DL models make them increasingly vital for building robust, intelligent, and real-time phishing detection systems.

#### Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture designed to effectively capture long-term dependencies in sequential data. In the context of phishing website detection, LSTM plays a crucial role in modeling temporal patterns and contextual features that may not be apparent in static data. As described by Elberri et al. (2024), LSTM was integrated with a Convolutional Neural Network (CNN) in a hybrid deep learning architecture aimed at improving phishing detection accuracy. While CNN excels at extracting spatial features from grayscale images generated from URL and web content data, LSTM contributes by analyzing the sequential behavior and dependencies within the data. This combination allows the model to better understand complex relationships, such as those found in URL structures and embedded script behavior typical of phishing attacks. The study demonstrated that the CNN-LSTM hybrid model outperformed standalone CNN and LSTM models in terms of accuracy, sensitivity, and precision, highlighting LSTM’s effectiveness in enhancing the classification of phishing and legitimate websites.

**S****tructure of an LSTM neural network cell**

LSTM is a variant of the RNN deep learning architecture designed specifically for tasks like time series analysis and classification. LSTM effectively uses a gating mechanism to deal with vanishing gradient problems in the training process.The LSTM memory cell has four gates named forgetting f, input gate i, control gate c, and output gate o. The fundamental configuration of the LSTM cell is presented in Figure 4 and it consists of the output of the previous memory cell Ct-1. (Elberri et al., 2024)

This neural network uses components such as the input signal at each time step Xt, the current memory cell Ct output, the previously hidden unit Ht − 1, and the currently hidden unit Ht. The forget gate determines the way in which the contribution from the previous time step is incorporated, resulting in a value ranging from zero to one for each datapoint in Ct-1. The input gate regulates the amount of input that is stored in the memory cell from the current time step. Meanwhile, the control gate updates the memory cell contents from Ct-1 to Ct. The output gate dictates the extent to which the internal state influences the external state at the current time step. The symbol ⊗ represents the element-wise multiplication of vector elements, while ⊕ signifies the summation of vector along with the application of the σ (sigma) function. To formulate the LSTM artificial neural network. (Elberri et al., 2024)

Diagram of a diagram of a cell state

AI-generated content may be incorrect.

**Figure 5: Structure of an LSTM neural network cell (Elberri et al., 2024)**

A math equations with black text

AI-generated content may be incorrect.

**Figure 6: Equations of LSTM (Elberri et al., 2024)**

#### Bidirectional Encoder Representations from Transformers (BERT)

Bidirectional Encoder Representations from Transformers (BERT) is a powerful deep learning model that has shown great effectiveness in detecting phishing attacks, particularly those involving social engineering through text-based communication. Phishing often relies on carefully crafted language to manipulate users into revealing sensitive information or clicking malicious links. BERT’s unique bidirectional architecture allows it to analyze both the left and right context of each word in a sentence, enabling it to detect subtle linguistic patterns and deceptive cues used in phishing attempts. In recent studies, BERT has been applied to phishing detection by transforming message content into 768-dimensional contextual embeddings using the [CLS] token, which captures the overall semantics of the text. These embeddings, when combined with other extracted features such as URLs, email addresses, or phone numbers, are used as inputs for neural network classifiers to distinguish phishing messages from legitimate ones. Compared to traditional text classification techniques, BERT has demonstrated superior performance in identifying context-based phishing strategies, making it a valuable AI tool for building intelligent systems that can proactively detect and prevent social engineering-based cyber threats.



**Figure 7: Transformer Model Architecture (Jain et al., 2025)**

### Reinforcement Learning Approaches

Reinforcement learning (RL) is one of the sub-domains of machine learning. The goal is to let the agent learn how to act based on the environmental state to maximize the expected long-term rewards, where the learning problem can usually be modeled as Markov decision problems (MDPs). Figure 8 shows the interactive feedback loop between the agent and the environment (Wang et al.,2020). The potential of reinforcement learning (RL) in phishing detection remains under explored. RL can enable continuous learning and adaptation, improving detection systems and complementing existing approaches. Research efforts are necessary to develop and validate RL-based methods, driving the creation of next-generation phishing detection systems that are more adaptive, accurate, and efficient (Prasanna B T Nandeesha H D, 2025).



**Figure 8: Interaction between the agent and the environment: at each time step, after the agent observes the environment, it chooses an action according to its policy. After the action is executed, the environment gives a reward signal to the agent and transit to a new state.**

Reinforcement learning (RL) agents are generally trained in episodes, each consisting of a certain number of steps. Given an episode, the sequence of states, actions, and rewards builds the trajectory or rollout of π. Let *k* be the index assigned to an episode; the *cumulative discounted reward* is defined as = . Then, the objective function to be optimized can be indicated as and the maximization problem, which the agent tries to solve, aims at finding for all and .

#### Q- Learning

Q-learning is a value-based reinforcement learning algorithm. The goal of Q-learning is to learn the optimal action-selection policy for an agent interacting with an environment. The agent learns this by updating a table of values called the Q-table where each entry represents the value of taking a particular action in each state (Kovalchuk, 2024).

The Q-learning algorithm uses the following formula to update the Q-value for a state-action pair based on figure below:

A math equation with black text

AI-generated content may be incorrect.

**Figure 9: Q – Learning equation (Kovalchuk, 2024)**

Where:  
- *Q*(*s, a*) is the Q-value for state s and action a.  
- alpha is the learning rate, controlling how much new information overrides old information.  
- *R* is the immediate reward for taking action an in state *s*.  
- gamma is the discount factor, representing the importance of future rewards.  
- (*s’, a’*) is the maximum Q-value for the next state *s’*, representing the best possible reward achievable from that state.

#### Deep Q-Learning

Deep Q-Learning or Deep Q Network (DQN) is an extension of the basic Q-Learning algorithm, which uses deep neural networks to approximate the Q-values. Traditional Q-Learning works well for environments with a small and finite number of states, but it struggles with large or continuous state spaces due to the size of the Q-table. Deep Q-Learning overcomes this limit by replacing the Q-table with a neural network that can approximate the Q-values for every state-action pair (Amin, 2024).

A diagram of a network

AI-generated content may be incorrect.

Figure 10: Structure of DQN

**Concepts of Deep Q-Learning**

1. **Q-Function Approximation**: Instead of using a table to store Q-values for each state-action pair, DQN uses a neural network to approximate the Q-values. The input to the network is the state, and the output is a set of Q-values for all possible actions.
2. **Experience Replay**: To stabilize the training, DQN uses a memory buffer (replay buffer) to store experiences (state, action, reward, next state). The network is trained on random mini batches of experiences from this buffer, breaking the correlation between consecutive experiences and improving sample efficiency.
3. **Target Network**: DQN introduces a second neural network, called the target network, which is used to calculate the target Q-values. This target network is updated less frequently than the main network to prevent rapid oscillations in learning.
4. **Bellman Equation in DQN**: The update rule for DQN is based on the Bellman equation, like Q-Learning:



Figure 11: Bellman equation *i*

Where:

* *θ* are the weights of the main Q-network,
* *θ−* are the weights of the target Q-network,
* *s* is the current state,
* *a* is the action taken,
* *r* is the reward received,
* *s′* is the next state,
* (*s′,a′*) is the maximum Q-value for the next state.

## Analysis

### Analysis of Dataset

|  |  |
| --- | --- |
| Feature | Description |
| URLLENGTH | Length of the full URL (longer URLs can be suspicious). |
| DOMAINLENGTH | Length of the domain only. |
| ISDOMAINIP | If domain is an IP address instead of a name (1 = Yes, 0 = No). |
| TLD | Top-Level Domain (e.g., .com, .org). |
| TLDLENGTH | Length of the TLD. |
| TLDLEGITIMATEPROB | Probability that the TLD is used in legitimate websites. |

**Table 1: Basic URL features**

|  |  |
| --- | --- |
| Feature | Description |
| CharContinuationRate | Measures character consistency; lower values may indicate randomness. |
| URLSimilarityIndex | How similar the URL is to a known legitimate one (likely from reference data). |
| URLCharProb, LetterRatioInURL, DegitRatioInURL, SpacialCharRatioInURL | Various measures of character distribution. |

**Table 2: Statistical & Lexical features**

|  |  |
| --- | --- |
| Feature | Description |
| HasObfuscation | Whether the URL contains confusing or misleading elements. |
| NoOfObfuscatedChar, ObfuscationRatio | Measures on how heavily obfuscated the URL is. |

**Table 3: Obfuscation and Manipulation Indicators**

|  |  |
| --- | --- |
| Feature | Description |
| NoOfEqualsInURL, NoOfQMarkInURL, NoOfAmpersandInURL, NoOfOtherSpecialCharsInURL | These characters often appear in phishing query strings. |

**Table 4: Special characters in URL**

|  |  |
| --- | --- |
| Feature | Description |
| IsHTTPS | Whether it uses HTTPS (1 = Yes, 0 = No). |
| LineOfCode, LargestLineLength | Measures of the HTML code complexity. |
| HasTitle, Title, DomainTitleMatchScore, URLTitleMatchScore | Checks alignment between page title and URL/domain. |

**Table 5: HTTPS and Web Page Content**

|  |  |
| --- | --- |
| Feature | Description |
| HasFavicon, HasSubmitButton, HasPasswordField, HasHiddenFields | Phishing pages often fake visuals or hide elements. |
| NoOfImage, NoOfCSS, NoOfJS | Counts of resources included—can signal legitimacy. |

**Table 6: Visual & Form elements**

|  |  |
| --- | --- |
| Feature | Description |
| NoOfURLRedirect, NoOfSelfRedirect | Phishing pages often redirect to obscure their origin. |
| Robots, IsResponsive | Site behavior and SEO configuration. |

**Table 7: Redirection & Network Behaviour**

|  |  |
| --- | --- |
| Feature | Description |
| Bank, Pay, Crypto | Binary flags if the site seems related to banks, payment, or cryptocurrency. |

**Table 8: Suspicious Intent Flags**

### Performance metrics

These metrics evaluate how well the machine learning models identify phishing websites or emails:

1. Accuracy

A black background with white text

AI-generated content may be incorrect.

**Figure 12: accuracy formula**

* TP: True Positives (phishing sites correctly detected)
* TN: True Negatives (legitimate sites correctly detected)
* FP: False Positives (legitimate sites misclassified as phishing)
* FN: False Negatives (phishing sites missed)

1. Precision

A black background with white text

AI-generated content may be incorrect.

**Figure 13: Precision formula**

1. Recall (Sensitivity)

A math equation with white text

AI-generated content may be incorrect.

**Figure 14: Recall formula**

1. F1 score - Harmonic mean of Precision and Recall

A black and white sign with white text

AI-generated content may be incorrect.

**Figure 15: F1 score formula**

## Critical Review

Primary studies on Machine Learning techniques.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Ref | Applied Approach | Used Algorithm | Used Data set | Main Findings | Limitation/Challenges |
| Rashid et al., 2020 | Machine Learning | Support Vector Machine | Alexa, Common Crawl archive (5000 URL) | The suggested method categorizes phishing and legal websites with 95.66% of Accuracy. | The study is very shallow and has used only one classifier, i.e., SVM, and five features for detecting phishing websites. A small data set was collected using GNU and Python scripts. Moreover, only one performance metric, i.e., Accuracy, was used for model evaluation. |
| Basit et al., 2020 | Machine Learning | Random Forest  K-Nearest Neighbor Decision tree  Artificial Neural Network | UCI machine learning repository 11,055 instances 30 features | The combination of K-Nearest Neighbors and Random Forest classifier detects phishing attacks with 97.33% accuracy. | The study has not used multiple data sets to evaluate their ensemble model. Further, the UCI dataset is open source and has normalized features. It does not include the Original URL. The study has also not included any feature selection procedure. The study has picked the open-source data set and existing ML algorithms for their study. It still needs to include the calibration values of each selected ML approach. |
| Saha et al., 2020 | Machine Learning | Random Forest  Decision tree | Kaggle 11,504 URL 32 attributes | The highest Accuracy of 97.00% was achieved through the Random Forest classifier. | The study has used only two Machine Learning approaches and only a single dataset. They used the PCA feature selection technique for analyzing data set characteristics. The study intended to use CNN for anticipating phishing attacks. They did not compare the results with existing equivalent techniques. It is a very shallow study. |
| Kasim, 2021 | Machine Learning & Heuristic | Support Vector Machine LightGBM  Multilayer Perceptron Convolution Neural Network | ISCXURL-2016 2978 instances and 77 different features | The current technique uses the Light Gradient Boosted Machine model to classify the features encoded with SAE-PCA at a rate of 99.60% accuracy. | The study has done the experiment on a limited dataset of 2978 instances, and PCA has reduced the feature selection from 77 to 20; these are also very limited. |
| Geyik et al., 2021 | Machine Learning | Decision tree  Logistic Regression Naive Bayes  Random Forest | PhishTank  Alexa  Common-crawl | The highest Accuracy produced by the Random Forest classifier is 83.0%. | Performance achieved with the dataset is very low compared to other studies with similar classifiers and datasets. |

**Table 9: Primary studies on Machine Learning techniques**

Primary studies on Deep Learning techniques

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Ref | Applied Approach | Used Algorithm | Results | Main Findings | Limitation/Challenges |
| Bagui et al., 2019 | Deep Learning | LSTM, CNN, and  WordEmbedding DL | Word embedding DL Accuracy = 98.89%, F1 = NA  DT, NB, SVM, CNN, LSTM Accuracy= < 97.50%, F1 = NA | Showed the context of the email is important in detecting phishing email. | Cannot detect malicious link or attachments inside email body. |
| Giri et al., 2022 | Deep Learning | GloVe + CNN, and BERT + FCN | GloVe + CNN Accuracy = 98%, F1 = 0.9749  BERT+FCN Accuracy = 96%, F1 = 0.9576 | Compare the combination of word embedding techniques and DL architectures. | Cannot detect malicious link or attachment inside email body. BERT model has limitation of maximum 512 tokens (words length). |
| Ramprasath et al., 2023 | Deep Learning | RNN with LSTM cells | RNN Accuracy =99.1%, F1 = 0.958  SVM Accuracy = 98.2%, F1 = 0.932  CkNN Accuracy =98.1%, F1 = 0.928 | N/A | Cannot detect malicious link or attachment inside email body. No dropout layer to prevent overfitting. |
| Valecha et al., 2021 | Deep Learning | Bi-LSTM | Accuracy = 95.97%, F1 = 0.9569 | Phishing detection based on gain and loss persuasion cues of text context. | Cannot detect malicious link or attachment inside email body. Manual coding of persuasion cues labels and manual hyperparameter tuning. |
| Paradkar, 2023 | Deep Learning | LSTM, Bi-SLTM, CNN | CNN Accuracy =98.05%, F1 = 0.9826  LSTM Accuracy = 97.32%, F1 = 0.9786  Bi-LSTM Accuracy = 98.04%, F1= 0.9825  NB, LR, SVM, DT Accuracy = <73.23%, F1 = NA | N/A | Cannot detect malicious link or attachment inside email body. |
| Divakarla et al., 2023 | Deep Learning | LSTM, Bi-LSTM, CNN+RNN | LSTM Accuracy = 98.8%, F1 = 0.987  Bi-LSTM Accuracy = 95.4%, F1= 0.95  CNN+RNN Accuracy = 97.9%, F1 =0.956 | N/A | Cannot detect malicious link or at tachment inside email body. No dropout layer to prevent overfitting. |
| Gholampour et al., 2023 | Deep Learning | ALBERT, ROBERTA, BERT, DE BERTA, DEBERT, SQ, and YOSO | BERT and its variants Accuracy = 98% ~ 99%, F1 = 0.92 ~ 0.97 | Developed new adversarial ham/phish dataset. Proposed ensemble method with KNN as shield model to assign correct label before feeding to DL models. | Cannot detect malicious link or attachment inside email body. Require high computational resources. Maximum tokens of BERT an AL BERT is 512. Dataset is small and model may overfitted. |
| Bountakas et al., 2021 | Deep Learning | BERT | Balance dataset: Word2Vec+RF Accuracy = 98.95%, F1 = 0.9897  Other combination Accuracy = <97%, F1 = 0.9744  Imbalanced dataset: Word2Vec+LR Accuracy = 98.62%, F1= 0.9241  Other combination Accuracy = < 98.42%, F1 = 0.8996 | Compare the combination of NLP techniques and ML models. | Cannot detect malicious link or at attachment inside email body. BERT model has a limitation of maximum 512 tokens. |
| Qachfar et al., 2022 | Deep Learning | BERT | BERT F1=0.991 to 0.998  RF, DT, SVM, SGD, KNN, GNB, LR, LSTM, CNN F1=0.72 to 0.99 | Propose method to reduce the impact of imbalanced data by adding synthetic training data. | Cannot detect malicious link or attachment inside email body. BERT model has a limitation of maximum 512 tokens. |
| Heet al., 2024 | Deep Learning | LSTM, Bi-SLTM | LSTM-XGB Accuracy = 98.35%, F1 = 0.9824  L-SVM, L-GNB, L-DTC Accuracy = <97%, F1 = <0.96% | Double-layer detection mechanism for both phishing and insider threats. | Cannot detect image links embedded in phishing emails. Dataset is small and model may be overfitted. |
| AbdulNabi et al., 2021 | Deep Learning | BERT | BERT Accuracy = 97.30%, F1 = 0.9696  BiLSTM Accuracy = 96.43%, F1 = 0.96  KNN and NB Accuracy <94%, F1 = <0.94% | N/A | Cannot detect malicious link or attachment inside email body. Maximum input sequence length is 300. |
| Muralidharan et al., 2023 | Deep Learning | BERT, CNN | Accuracy = 99.2%, F1 = 0.941 | Ensemble learning to analyze all email segments including attachment. | Inference time to process and may need high computing resources. |
| Lee et al., 2021 | Deep Learning | CNN-LSTM, BERT | RF-BERT+RF-CNN+LSTM AUPRC=0.9997, F1 =NA  RF-Word2Vec+LSTM-CNN+LSTM AUPRC = 0.9851, F1 = NA | Propose modular architecture to analyze all components of email except attachment. | Cannot detect malicious attachment inside email body. BERT model has a limitation of maximum 512 tokens (words length). Require high computational resources. |

**Table 10: Primary studies on Deep Learning techniques**

Primary studies on Reinforcement Learning

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Ref | Applied Approach | Used Algorithm | Result | Main Findings | Limitation/Challenges |
| H. Kamal et al., 2024 | Reinforcement Learning | Q-Learning | Achieved high detection accuracy with faster convergence | Proposed Q-Learning model for phishing detection. Demonstrated that RL outperforms traditional ML models. | Limited to URL-based features. Dataset and real-world application scalability were not fully addressed. |
| Chatterjee and Namin, 2024 | Reinforcement Learning | Q-Learning | High detection rates with optimized policy | Successfully applied Q-Learning for detecting malicious websites, demonstrating high reward optimization. | Focused on specific phishing datasets, potentially limiting generalization. |
| Nguyen and Reddi, 2023 | Reinforcement Learning | Deep Q-Network (DQN) | High adaptability and faster threat response. | Showcased DRL’s potential in cybersecurity with real-time learning capabilities | High computational cost and complexity in real-world deployment. |
| Chatterjee et al., 2023 | Reinforcement Learning | Deep Q-Learning | High accuracy and faster response time | Demonstrated improved phishing detection accuracy with DRL models. | Potential overfitting and lack of interpretability in complex scenarios. |

**Table 11: Primary studies Reinforcement techniques**

**Table 12: Comparison of Methods Used in Previous Studies**





**Figure 16: Bar chart of methods used in previous studies**

A colorful pie chart with numbers

AI-generated content may be incorrect.

**Figure 17: Pie chart of methods used in previous studies**

## Research Gap Analysis

Despite significant advancements in AI-driven phishing detection systems, considerable limitations persist in current methodologies. Much of the existing research primarily concentrates on either machine learning (ML) or deep learning (DL) methods in isolation, disregarding the potential benefits of hybrid models that combine both techniques. Additionally, many models are trained and evaluated on outdated or limited datasets, which do not adequately reflect the complexity and evolving nature of modern phishing attacks. While reinforcement learning (RL) offers a promising pathway for enabling systems to adapt to emerging threats, its application in phishing detection has mainly remained underexplored. Furthermore, deep learning models like BERT face practical challenges, including input length constraints and substantial computational costs, which impede their deployment in real-time scenarios.

Additionally, many current detection systems struggle to identify advanced phishing attempts that involve obfuscated links or malicious attachments. The absence of standardized benchmarks and comparative studies further complicates the fair evaluation and refinement of existing solutions. These challenges underscore the need for more flexible, scalable, and adaptive detection systems capable of effectively addressing real-world phishing threats.

Despite these advancements, existing AI-based phishing detection methods still face fundamental challenges due to their reliance on either machine learning (ML) or deep learning (DL) techniques alone. Machine learning models, such as Support Vector Machines (SVM), are highly proficient in structured data analysis but often struggle to process unstructured content, such as text from phishing emails or fraudulent websites. On the other hand, deep learning models like Long Short-Term Memory (LSTM) networks excel at analyzing sequential and unstructured data; however, they are computationally intensive and require substantial amounts of training data, making them less feasible for real-time applications.

This research aims to bridge these limitations by employing a hybrid solution that combines both machine learning (ML) and deep learning (DL) approaches, effectively leveraging the strengths of each. Machine learning models, such as Support Vector Machines (SVM), are well-suited for handling structured data, like URLs, 47 while deep learning models, like Long Short-Term Memory (LSTM), can efficiently process complex unstructured data, like the content of phishing emails or web pages. The hybrid approach not only improves adaptability to dynamic phishing tactics but also enhances detection accuracy and minimizes false positives by utilizing the complementary strengths of both techniques. By integrating the advantages of both machine learning (ML) and deep learning (DL), this hybrid model offers greater robustness and scalability compared to single-model approaches.

In alignment with this objective, the study by Rao et al. (2019) introduced a multi-model ensemble combining LSTM and SVM for phishing detection on mobile devices. The proposed hybrid model demonstrated superior performance compared to traditional methods, achieving an accuracy of 97.30% on its dataset and 98.50% on a benchmark dataset. This study highlighted that integrating deep learning for feature extraction with machine learning for classification significantly improved detection speed, accuracy, and robustness, particularly in addressing zero-day phishing attacks.

Building on these findings, the hybrid model proposed in this research combines the speed and interpretability of machine learning with the contextual learning capabilities of deep understanding, making it more adaptable to the evolving nature of phishing threats.

In conclusion, while both machine learning and deep learning models have inherent limitations when used independently, the hybrid AI solution offers a more comprehensive and practical approach to phishing detection. This makes it a more viable option for addressing the sophisticated and rapidly evolving landscape of social engineering attacks.

## Project Solution

To address the limitations identified in prior research, particularly the low adaptability of single-model approaches and the inability to detect complex phishing patterns, this project proposes a hybrid Machine Learning (ML) and Deep Learning (DL) solution using the KNIME Analytics Platform. The hybrid approach combines the structured pattern recognition strengths of traditional ML algorithms such as Support Vector Machine (SVM) with the advanced feature extraction capabilities of Deep Learning models like Long Short-Term Memory (LSTM). KNIME is selected as the development environment due to its flexibility, visual workflow design, and native support for both ML and DL through integration with Keras. In this solution, structured features from URLs and website metadata are processed using ML classifiers, while unstructured textual content, such as email bodies or page content, is passed to an LSTM model trained to capture sequential dependencies typical of phishing language. The outputs from both paths are then combined using an ensemble mechanism or decision logic to improve overall detection accuracy. This dual-model strategy leverages the speed and interpretability of ML with the contextual learning strength of DL, enabling a more robust and scalable detection system that adapts to the evolving nature of phishing attacks. By deploying this hybrid model within KNIME, the project ensures a practical, modular framework suitable for academic experimentation and real-world application.

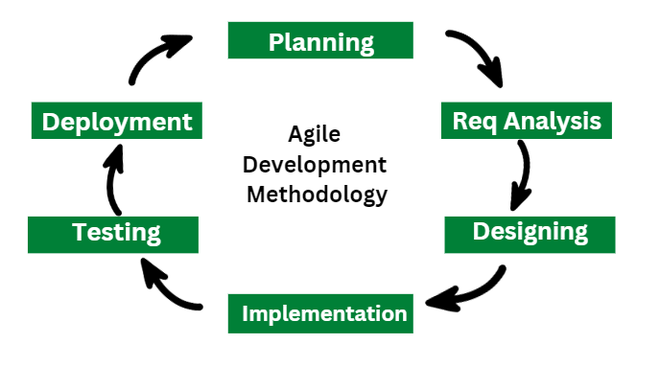
# METHODOLOGY

## Introduction

This chapter outlines the methodology used to develop and evaluate a hybrid Machine Learning (ML) and Deep Learning (DL) phishing detection framework using the KNIME Analytics Platform. The process involves data preprocessing, model training, evaluation, and system integration. The hybrid approach leverages structured and unstructured data to enhance detection accuracy. Additionally, the Agile methodology is adopted to enable iterative development, continuous feedback, and adaptive refinement of the system based on performance evaluations.

## Methodology

The project adopts a hybrid methodological strategy combining Agile software development for planning and iterative refinement with experimental workflow development in KNIME, implementing a modular pipeline where structured URL features are classified using ML models like SVM, while unstructured email content is analyzed using LSTM, with all outputs ensembled for the final phishing detection decision.



**Figure 19: Agile Methodology**

### Requirement Analysis

#### Hardware Requirement

A laptop is used as a workstation for all tasks, from researching to documenting. The laptop specifications are shown in the table below.

Table 13: Hardware Requirement

|  |  |
| --- | --- |
| **Specification** | **Description** |
| Processor Type | AMD Ryzen 7 4700U with Radeon Graphics 2.00 GHz |
| Operating System | Windows 11 Home Single Language Version: 24H2 |
| Operating System Type | 64-bit operating system, x64-based processor |
| RAM | 8.00 GB (7.42 GB usable) |
| Storage | 477 GB |
| Display Resolution | 1920 x 1080, 144 Hz |

#### Software Requirement

This project's development includes the usage of some software. The software used in this project is listed in the table below.

Table 14: Software Requirement

|  |  |
| --- | --- |
| **Software** | **Description** |
| Windows 11 | An environment of operating system used for project execution. |
| KNIME Analytics Platform | An open-source platform for data analytics, machine learning, and deep learning, used in this research to implement a Phishing detection using hybrid AI solution. |
| Microsoft Word 365 | Software is used to complete project reporting and documentation. |
| Microsoft Excel 365 | For datasets collections. |
| Microsoft Visio 2021 | Software used to create structure diagram |

### System Design

The proposed system is designed to accept both structured and unstructured inputs, which are processed through parallel pipelines. The input layer handles structured data such as URL features including length, number of special characters, and domain information, and unstructured data such as email text or website content. In the processing layer, the ML pipeline in KNIME applies models like SVM to classify structured features, while the DL pipeline implemented in KNIME with Keras uses LSTM to analyze the contextual semantics of unstructured text. The ensemble layer then combines the outputs from both pipelines using decision rules such as majority voting or weighted scoring. Finally, the output layer delivers the final classification result, indicating whether the input is phishing or legitimate.

### Implementation

This project was implemented utilising the KNIME Analytics Platform, which offers an easy-to-use framework for creating processes that enable deep learning and machine learning. Using conventional machine learning models like Support Vector Machine (SVM), which are accessible in KNIME, the structured data such as URL features was analysed. KNIME's Keras plugin was used to incorporate a Long ShortTerm Memory (LSTM) model for the analysis of unstructured data, including email text. A decision process was used to integrate the predictions made by the ML and DL models after training to provide the final classification outcome. Iterative testing and improvement were possible since the whole process from data input and cleaning to model training and evaluation was controlled inside KNIME's modular workflow interface.

### Testing

The system is tested using test data that was not used during training. Accuracy shows how correct the model is. Precision shows how many predicted phishing cases are phishing. Recall shows how many real phishing cases are correctly found. F1-score combines precision and recall. The results are compared between ML models, DL model, and the combined hybrid model.

### Deployment

Once all data from the structured and unstructured sources have been processed through their respective pipelines, the results are presented and analysed about other related research studies. A critical evaluation is conducted based on the findings. Graphs are plotted for key performance metrics, including accuracy, precision, recall, and F1-score, to compare the performance of individual ML and DL models, as well as the ensemble model. A comparison table is also provided to summarise the results across all models and metrics. Examples of the graphical and tabular results are illustrated in the figure below.

A table with numbers and text

AI-generated content may be incorrect.

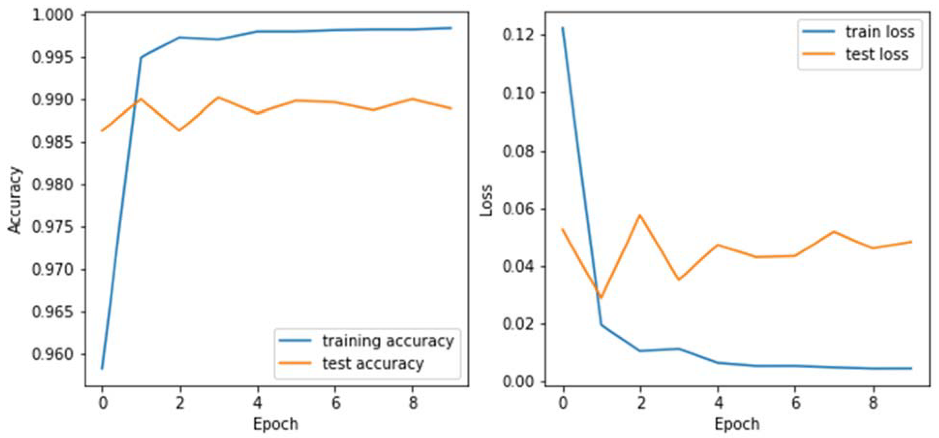
**Figure 19: Performances of selected machine learning models**

**(Sameen et al., 2020)**

A screenshot of a graph

AI-generated content may be incorrect.

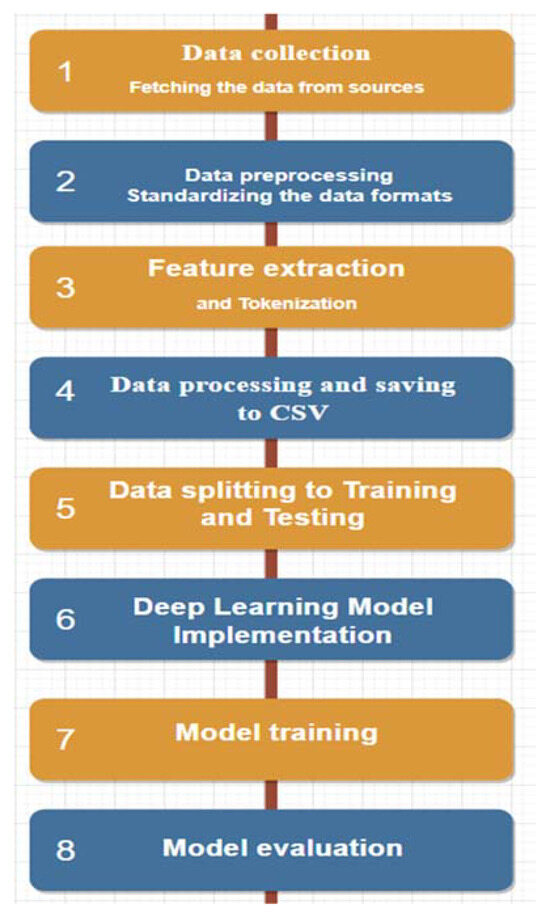
**Figure 20: LSTM classification report (Atawneh & Aljehani, 2023)**



**Figure 21: LSTM model training and testing accuracy**

**(Atawneh & Aljehani, 2023)**

### Previously proposed methodology



**Figure 22: Methodology for phishing detection**

**(Atawneh & Aljehani, 2023)**

## Project Schedule and Milestones

### Project Milestones

A milestone is a project marker that denotes a shift or stage of progress. Thus, project milestones is essential in keeping track of upcoming events or goals across the timeline.

**Table 15: Research Milestone of PSM 1**

|  |  |  |
| --- | --- | --- |
| **Week** | | **Activity** |
| 1 | | * Proposal discussion * Proposal assessment & verification |
| 2 | | * Chapter 2(Analysis Begins) |
| 3 | | * Chapter 2 * Chapter 1 |
| 4 | | * Chapter 2 |
| 5 | | * Chapter 2 |
| 6 | | * Chapter 2 * Chapter 1 |
| 7 | | * Chapter 2 (last discussion before mid-semester break) |
| ***Mid- Semester Break*** | | |
| 8 | * Chapter 2 * Chapter 3 | |
| 9 | * Chapter 3 | |
| 10 | * Chapter 4 | |
| 11 | * Chapter 4/ Project Demo | |
| 12 | * Chapter 4/ Project Demo | |
| 13 | * Project Demo * PSM 1 Report | |
| 14 | * Project Demo * Final report submission and slide presentation for evaluation by Supervisor and Evaluator | |
| 15 | * Revision week | |
| 16 | * Final Exam | |
| 17 | * Final Exam | |
| 18 | * Intallation of the tools might be needed for this project (Knime, Anaconda Navigator, python3 package) * Do corrections for the report that has been commented by the Evaluator | |

### Project Gantt Chart

Gantt Charts provide a thorough overview of the project from start to finish, as well as all the activities required to complete the project. It aids in demonstrating how far the tasks have progressed.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **PHASE** | **WEEK** | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 |
| Requirement Analysis |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| System Design |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Implementation |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Testing |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Deployment |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

## Summary

This chapter discusses the methodology employed in this research, focusing on the application of hybrid solutions, which combine machine learning and deep learning techniques, to detect phishing attacks based on URL and email content. To ensure a structured and timely analysis, this chapter also includes a detailed project milestone and Gantt Chart. This structured figure provides a clear timeline and outlines key activities, helping to manage the analysis’s progress effectively. The systematic approach outlined in this methodology ensures that the research is carried out efficiently and thoroughly. The upcoming chapter will delve into the research implementation, providing detailed insight into how the methodologies discussed are practically applied to achieve the analysis’s objectives.

# DESIGN

## Introduction

This chapter presents the system design of the proposed hybrid Machine Learning and Deep Learning phishing detection framework developed using the KNIME Analytics Platform. The design phase is crucial as it translates the project’s conceptual framework into a structured and implementable architecture. It involves both logical and physical designs that define how data flows through the system, how different components interact, and how the system integrates machine learning classifiers with deep learning models to produce accurate phishing detection results. The objective of this chapter is to provide a clear and detailed overview of the system’s internal structure, including the data processing pipeline, model training paths, and ensemble decision logic. The design considerations ensure the system is modular, scalable, and adaptable for real-world deployment while remaining aligned with the project objectives and addressing the gaps identified in the literature review.

## Design Framework of Hybrid Approach for Phishing Detection in AI

A black background with a black square

AI-generated content may be incorrect.

**Figure 23: Design methodology framework for hybrid approach for phishing detection**

The process begins at the Data Collection stage, where phishing and legitimate data are gathered from multiple sources. Structured data includes features extracted from URLs such as length, presence of IP addresses, number of memorable characters, and other manually engineered indicators. Unstructured data, on the other hand, consists of email content or website text, typically presented in raw text format.

Following data collection, the pipeline branches into two processing flows. The first flow is the ML Preprocessing step, which handles the structured URL-based data. This step involves feature selection, normalization, and encoding, which prepares the data for input into traditional machine learning classifiers. Models such as Support Vector Machine (SVM) are trained and evaluated within this segment to predict whether the structured features point to a phishing attempt.

Simultaneously, the second flow processes unstructured text data through the DL Preprocessing phase. This involves cleaning the text, removing stopwords, tokenisation, and transforming it into sequences suitable for neural networks. The cleaned and encoded text is then passed to a Long Short-Term Memory (LSTM) network, a type of recurrent neural network effective in understanding sequential and contextual information in text.

Once both the ML and DL models produce their predictions, the outputs are forwarded to the Ensemble Layer. This layer combines the predictions from both pipelines using a decision fusion strategy, such as majority voting or weighted averaging, to improve overall classification accuracy and robustness.

The final step is the Output Layer, which presents the final decision—whether the analysed input is classified as phishing or legitimate. This decision can then be used for reporting, alerting users, or feeding into further security systems for automated responses.

Each stage of the pipeline plays a critical role in ensuring the system can effectively handle different data types and enhance phishing detection performance through complementary strengths of ML and DL models.

## Summary

To summarize, this chapter is essential as it outlines a straightforward approach to developing a phishing detection system that utilizes both machine learning and deep learning techniques. It presents the overall system architecture, detailing how structured and unstructured data are processed through parallel pipelines, and describes the modular design, data preprocessing methods, model selection, and ensemble strategy. The evaluation plan and deployment process are also addressed, including how performance metrics will be measured and visualized. The next phase, implementation, will be guided by the analysis and design established in this chapter.

# IMPLEMENTATION

## Introduction

This chapter showcases the real-world execution of the suggested hybrid phishing detection system utilizing the KNIME Analytics Platform. The initiative combines Machine Learning (ML) and Deep Learning (DL) methods to identify phishing attacks from structured URL information and unstructured email/text data, respectively. The ML part utilizes Support Vector Machine (SVM) to analyze structured features like URL length, special character presence, and domain indicators, whereas the DL part uses Long Short-Term Memory (LSTM) networks through Keras integration in KNIME to examine email and text patterns characteristic of phishing attempts.

The execution was done in a modular way, utilizing KNIME's graphical interface for constructing workflows based on nodes. Distinct pipelines were created for structured and unstructured data, with each including preprocessing, model training, and prediction phases. The outputs of both models are merged using an ensemble logic layer to deliver a more precise and dependable phishing detection conclusion.

This chapter details the procedures followed to set up and adjust the working environment, construct and link the KNIME workflows, and run the hybrid detection system. It additionally outlines the process of importing, cleaning, and splitting the dataset, along with the training and evaluation of the ML and DL models. This implementation aims to showcase the practicality and efficiency of a hybrid AI-driven approach for addressing phishing-related social engineering threats.

## Software Development Environment Setup

The development and implementation of the hybrid phishing detection model were carried out using the **KNIME Analytics Platform**, a powerful open-source software designed for data analytics, machine learning, and deep learning workflows. The selection of KNIME was motivated by its modular, node-based interface, ease of integration with Python and Keras, and strong support for both structured and unstructured data processing.

The environment setup involved several essential installations and configurations to ensure a smooth development process for both Machine Learning (ML) and Deep Learning (DL) models:

### KNIME Analytics Platform

* VersionUsed: KNIME Analytics Platform 5.4.4
* KNIME was installed as the core platform for visual workflow design. It allows for seamless creation, execution, and evaluation of ML and DL models without the need for extensive coding.
* The KNIME Deep Learning **-** Keras Integration extension was installed via the KNIME Extension Installer to enable the use of LSTM models.
* The KNIME Python Integration was also installed to support scripting and Python-based nodes where necessary.

### Python and Keras Integration

* **Python Version**: Python 3.9.15 (via Anaconda Distribution)
* **Keras Version**: 2.13+ (with TensorFlow as backend)
* Python was installed and configured through **Anaconda Navigator** to manage packages and environments. A dedicated KNIME Python environment was created and linked to KNIME for seamless execution of Python and Keras nodes.

Required libraries installed include:

* tensorflow
* keras
* numpy
* pandas
* scikit-learn

## Software Configure Management

# TESTING AND ANALYSIS

## Introduction

## Result and Analysis

## Summary

# CONCLUSION

## Introduction

## Project Summarization

## Project Contribution

## Project Limitation

## Future Works

## Summary

# references

# Bibliography

# APPENDICES