National College of Ireland

A Resilient NLP-Based Detection System of Phishing Emails Leveraging Deep learning Technique

Industry Internship
MSc. Cybersecurity

Oluwafunsho John Alabi
Student ID: X22126899

School of Computing
National College of Ireland

Supervisor: Vikas Sahni

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Oluwafunsho John Alabi
Student ID:	X22126899
Programme:	M.Sc. Cybersecurity **Year:** 2023
Module:	Industry Internship
Supervisor:	Vikas Sahni
Submission Due Date:	5th January 2024
Project Title:	A Resilient NLP-Based Detection System of Phishing Emails leveraging Deep Learning Techniques
Word Count:8049	**Page Count 21**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:

Date:

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	☑
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☑
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☑

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Resilient NLP-Based Detection System of Phishing Emails leveraging Deep learning Technique

Oluwafunsho Alabi
X22126899

Abstract

Phishing attacks have become more sophisticated over time, causing a significant threat to both individuals and organizations globally. As attackers create new techniques, it's essential to stay ahead of the trend and invest in robust cybersecurity measures.

A promising approach is to leverage natural language processing (NLP) and deep learning techniques to create a cutting-edge detection system. This research aimed to do just that by analyzing current methods, identifying gaps, and introducing innovative solutions to improve phishing email detection accuracy. The NLP-based system has enormous potential in detecting deceiving content and neutralizing novel threats. To ensure its efficacy, it was subjected to rigorous testing and analysis, simulating real-world scenarios. This significantly contributed to the field of cybersecurity by strengthening defenses against phishing attacks and fostering a safer online environment for everyone. This is better than previous detection systems as it is more accurate, achieving an 84% accuracy.

Keywords: NLP Detection System, Deep learning technique, Resilient, Cybersecurity, Ethical considerations.

1 Introduction

The increase in phishing attacks has become an ever-challenging cybersecurity problem in the world due to its frequency and complexity. Cyber attackers use improved approach to deceive people and organization into financial loss and data exposure. (Naqvi et al., 2023).

By masquerading as trustworthy entities, phishing emails lure unsuspecting recipients to divulging critical information, such as login credentials and financial data. These breaches are not merely a technological failure but also a manipulation of human vulnerabilities, necessitating cutting-edge security responses. This research proposed an innovative system, fortified by deep learning algorithms, to discern and mitigate the pernicious impact of phishing schemes (Naqvi et al., 2023; Das et al., 2021).

In other to tackle this issue, the aim of this research was to establish a strong system that detects phishing emails by making use of NLP and deep learning techniques.

Cybercriminals are causing a lot of financial loss and exposing personal and corporate information due to human susceptibility, social engineering techniques, and technological vulnerabilities through phishing attacks. To combat these phishing attacks, a dynamic and flexible security solution must be implemented. This research tackles these issues mentioned using a resilient NLP-based detection system that harnesses deep learning techniques to recognize phishing emails and reduce its harmful consequences.

Through analysis, practical tests, and ethical considerations this research would aim to make a safer digital environment, strengthening defenses against dynamin attacks from phishing emails.

Phishing attacks which capitalize on human vulnerabilities and limitations in technology has become a common form of cybercrime. The previous technique of detecting these attacks faced various setbacks as it relied on traditional rules and machine learning based approaches to tackle the ever-changing tactics of cybercriminals. To properly combat phishing attacks, there is a significant need for an advanced, adaptable, and accurate detection system. The use

of natural language processing (NLP) and deep learning techniques to develop a novel detection system.

Research Question: How can an NLP-based detection system and deep learning techniques improve the resilience of the detection system against emerging phishing attacks?
The proposed solution involves developing an NLP-driven detection system utilizing deep learning algorithms. Through the analysis and categorization of email content, this system aims to precisely recognize possible phishing threats.

The Objectives of this study are:

1. To assess current machine learning pattern in phishing email detection.
2. To enhance feature extraction from email content through NLP.
3. To implement and compare deep learning algorithms for the classification of emails.
4. To assess the effectiveness of the developed system through rigorous validation methods.
5. To contribute to cybersecurity by improving phishing detection rates.

1.1 Rationale for the Study

Phishing attacks are a growing threat, with sophisticated techniques bypassing conventional cybersecurity measures. The necessity for advanced detection systems is imperative, as traditional rule-based and machine learning approaches fall short against the dynamic nature of phishing strategies (Das et al., 2021). This research will leverage NLP and deep learning to address these challenges, providing a novel approach to cybersecurity.

1.2 Significance of the Study

This study is to significantly reduce the success rate of phishing attacks by improving detection mechanisms. It aimed at contributing to safeguarding sensitive information and mitigating financial losses due to fraud. The outcomes could shape the development of cybersecurity protocols and user training programs, thus enhancing organizational resilience against social engineering threats (Benavides-Astudillo et al., 2023).

1.3 Summary of the Chapter

This chapter introduced the research, outlining its objectives, rationale, and significance. It has set the stage for a detailed exploration of the literature surrounding phishing detection and the role of NLP and deep learning in cybersecurity, which will be the focus of the next chapter. The subsequent literature review will delve into existing methods and identify gaps that this research aims to fill, thus laying the groundwork for the proposed detection system (Salloum et al., 2022).

The next section, the literature review, will scrutinize current methodologies, evaluate their effectiveness, and establish the theoretical foundation for the proposed system. It will critically analyze the role of NLP in feature extraction and the application of deep learning techniques in enhancing the detection of phishing emails (Egozi and Verma, 2018; Liang et al., 2017). Through this examination, the study will identify opportunities for innovation in phishing email detection and systematization, contributing to the advancement of the field.

2 Related Work

Research into phishing email detection has evolved from reliance on rule-based systems, which flag emails based on suspicious keywords and URLs, to more sophisticated machine learning models. Traditional methods, while initially effective, face challenges in keeping pace with the adaptive strategies of cybercriminals (Das *et al*., 2021). These rule-based systems often suffer from a lack of flexibility, resulting in misclassified emails and an inability to recognize new phishing tactics (Tang and Mahmoud, 2021).

2.1 Traditional Approach

The limitations of traditional detection methods have led researchers to explore advanced machine learning techniques, capable of learning and identifying complex patterns. Machine learning algorithms such as random forests and decision trees are now at the forefront of detecting phishing emails (Dinesh *et al*., 2023). These methods, powered by natural language processing (NLP), allow for more dynamic and contextual analysis of email content, enhancing detection accuracy (Haq *et al*., 2022).

The integration of NLP with machine learning facilitates a more nuanced understanding of language and email structure, thereby improving the identification of phishing attempts (Noah *et al*., 2022). PhisherCop, an automated tool for phishing detection, exemplifies the efficacy of NLP in cybersecurity (Yazhmozhi and Janet, 2019). Furthermore, the comparison of ensemble learning techniques, such as AdaBoost and Multiboosting, reveals their significant potential in phishing website detection (Subasi and Kremic, 2020). The literature review indicates a critical transition from rigid, rule-based approaches to adaptive, learning-oriented models, underlining the necessity of continual innovation in cybersecurity practices (Ukwen and Karabatak, 2021).

2.2 Machine Learning and Deep Learning Techniques

Machine learning (ML) models, trained on datasets like those provided by Razaulla *et al*. (2023) and Alkhalil *et al*. (2021), have been pivotal in classifying emails by learning from historical instances of phishing. These models can extrapolate general patterns from past examples to detect phishing attempts. The efficacy of such ML techniques is dependent on high standard and a varied nature of training datasets. Inadequate or biased datasets may lead to overfitting, causing the model to falter when encountering new phishing strategies (Das *et al*., 2021). Moreover, conventional ML features may miss certain fine points of the linguistic cues that characterize phishing emails.

To address these limitations, the research turns to natural language processing (NLP) and deep learning to extract nuanced features and provide context-rich analysis. Laying emphasis on the importance of NLP in identifying significant patterns within the email content that are indicative of phishing activities is necessary. Sentiment analysis and text classification, prove effective in discerning phishing emails, going beyond the capabilities of traditional ML approaches (Haq *et al*., 2022).

The implementation of NLP in phishing detection is crucial for uncovering context-specific indicators that traditional methods may overlook. This research aims to harness these advanced NLP techniques, integrated with deep learning, to develop a sophisticated detection system. Such a system is expected to enhance accuracy and efficiency, thereby improving the resilience of cybersecurity measures against the threats posed by phishing attacks (Noah *et al*., 2022).

By maximizing the potential of NLP, the detection system significantly reduces false positives and increase the reliability of phishing detection, fortifying defenses, and providing

robust protection in the cybersecurity arena (Subasi and Kremic, 2020; Ukwen and Karabatak, 2021; Yazhmozhi and Janet, 2019).

2.3 Advanced Detection Techniques

The current methodologies employed in detection are being significantly advanced through the integration of Natural Language Processing (NLP) and deep learning techniques. The emergent use of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) has shown to surpass the traditional NLP strategies (Das *et al*., 2021). These sophisticated models are adept at autonomously learning from the data's hierarchical representations, thereby improving the detection of phishing characteristics.

Deep learning techniques, when amalgamated with NLP, provide a robust framework for processing the intricate and unstructured nature of email content (Choudhary *et al*., 2022). This combination allows for the discernment of subtle linguistic nuances that traditional methodologies may miss. Moreover, these models adapt and learn from the language and context changes, thereby becoming more flexible in identifying new phishing strategies. The scalability of deep learning models is a considerable advantage, letting the system to manage vast volumes of data – a critical component in the dynamic landscape of cybersecurity threats (Taye, 2023).

2.4 NLP and Deep Learning Synergy

The proposed detection system in this research aims to amalgamate the feature extraction capabilities of NLP with the classification strength of deep learning models. This hybrid system is designed to address the limitations of traditional methods by recognizing and mitigating sophisticated phishing tactics intelligently (Vinayakumar *et al*., 2019). The system, through NLP, will extract more profound linguistic features from phishing emails, thereby strengthening the precision in detection. Concurrently, the integration of deep learning models ensures adaptability to the continually evolving phishing methodologies.

2.5 Cybersecurity Resilience

This research initiative aligns with the pressing need for an advanced and flexible phishing email detection system. It seeks to augment the cybersecurity domain by reflecting upon both the contributions and limitations of previous studies. By synthesizing NLP and deep learning techniques, the proposed system aims to establish an effective countermeasure against phishing threats.The proposed system's capacity to learn and adapt is critical for its success in real-world applications. It is designed to be not just reactive but predictive, using the insights gathered from NLP to foresee potential phishing attacks (Haq *et al*., 2022). The system will utilize these predictive capabilities to provide a more nuanced and comprehensive defense against phishing, distinguishing it from the conventional detection systems (Subasi and Kremic, 2020).

To summarise, the proposed detection system's design leverages the synergistic potential of NLP and deep learning. It is a response to the complex and evolving nature of phishing attacks, aiming to deliver a resilient solution capable of adapting to and mitigating these cybersecurity risks. The system's robustness is predicated on its ability to learn from and respond to the ever-changing tactics employed by cybercriminals, ensuring a state-of-the-art defense mechanism against phishing (Ukwen and Karabatak, 2021; Yazhmozhi and Janet, 2019).

The subsequent section will discuss the methods employed to design this system.

3 Research Methodology

For a comprehensive practical implementation of this research a semi-structured interview was employed in the qualitative research to gain a proper insight of real-world phenomena from the perspectives of practitioners.

3.1 Approach Overview

This section delves into the adopted strategy for phishing email detection, harnessing the capabilities of Natural Language Processing (NLP) synergized with deep learning mechanisms. Understanding the complexities in the modus operandi of cybercriminals, the study elucidates the foundational elements and methods that constitute the designed system. The primary phase in this approach involves the application of NLP to distil meaningful features from email content. This process involves refining the raw textual data of emails using methods such as tokenization and named entity recognition, among others (Liang *et al*., 2017). By converting this unstructured data into a structured format, the system is poised to discern significant patterns and linguistic markers indicative of phishing undertakings (Egozi & Verma, 2018). These extracted features are pivotal, equipping the subsequent deep learning models with the necessary insights to differentiate between phishing and legitimate emails effectively.

3.2 Process Flow

To address the challenge of precisely sorting out phishing emails from genuine ones, the study adopted a robust methodology leveraging both Natural Language Processing (NLP) and machine learning techniques. At the outset, the research relied on the Association of Computational Linguistics' phishing email dataset provided by Radev (2008). This dataset, comprising email content in English paired with binary labels denoting if the email was genuine, underwent intensive data preprocessing. As a preliminary step, the dataset was loaded into a Pandas DataFrame, a tool acknowledged for its data analysis capabilities (McKinney & Team, 2015). Ensuring data integrity was paramount, prompting the removal of any NaN values (Joel, Doorsamy, & Paul, 2022).

The email content underwent several preprocessing steps. These included converting the text to lowercase to standardize it, removing punctuation to simplify the text, and eliminating stop words, which often don't contribute meaningfully to classification tasks (Liang *et al*., 2017). Subsequently, the pre-processed text was converted into a numerical form using the Term Frequency-Inverse Document Frequency (TFIDF) vectorizer, a method that captures the significance of terms within individual emails concerning their frequency across the dataset (Liu *et al*., 2018).

Following preprocessing, the dataset was bifurcated into training and testing subsets, maintaining a 80-20 split. This division allows models to be trained on one subset and validated on another, an approach rooted in conventional machine learning practices (Muraina, 2022). However, stratification was employed on the target variable to ensure the consistent distribution of labels across both sets.

The modelling phase witnessed the exploration of various algorithms, from Decision Trees and AdaBoost to Neural Networks, k-Nearest Neighbours, and Support Vector Machines. Each of these models was carefully trained, optimized, and evaluated based on several metrics, with the Area Under the ROC Curve emerging as a primary evaluation metric (Pencina *et al*., 2008). For instance, Decision Trees, known for their interpretability, were trained at different depths, and their performance was gauged using AUC (Raschka, 2018). Similarly, ensemble methods like AdaBoost combined multiple weak learners to forge a more potent classifier, enhancing the model's ability to discern phishing emails (Subasi & Kremic, 2020). Neural networks, especially the Multi-Layer Perceptron (MLP), were trained using

varying learning rates, and their performance was visualized employing learning curves (Goyal, Pandey, & Jain, 2018). Furthermore, the k-NN algorithm's proficiency was evaluated by altering the number of neighbors (Hnini *et al*., 2020), while SVM classifiers were trained using different kernel functions to identify the most efficacious one (Kumar, Chatterjee, & Díaz, 2020).

Visualization tools played a crucial role throughout the research. They provided insights into model behaviours, showcasing relevant metrics like AUC against hyperparameters, aiding in understanding the underlying trade-offs (Park *et al*., 2020). A comparative analysis was provided, pitting all models against each other to discern the most effective one for the task at hand.

In essence, this research methodology, steeped in empirical practices and supported by an array of advanced techniques, seeks to create a resilient phishing email detection system. Through rigorous experimentation and comprehensive evaluation, it aims to fortify the cybersecurity domain, offering enhanced protection against phishing attacks.

3.3 Tools and Test Data

https://datasetsearch.research.google.com/search?src=0&query=%20Radev%20(2008)&docid=L2cvMTF0bXFyZHYzcQ%3D%3D.

3.4 Evaluation Approach

The evaluation of models in this research is anchored on a multifaceted approach to ensure robustness and accuracy. Primarily, the Area Under the Curve metric was utilised, gauging the models' proficiency in differentiating between positive and negative classes in binary classification tasks, such as phishing email detection (Pencina *et al*., 2008). To complement AUC, other key metrics including accuracy, precision, recall, F1 score, and the ROC curve were also employed, providing a comprehensive performance assessment (Raschka, 2018).

Computational efficiency was also closely examined. Not only was the model's performance pivotal, but its computational demands, encapsulated in training and prediction durations, were deemed essential for understanding its suitability in real-time scenarios (Reif, Shafait, & Dengel, 2011).

Learning curves, which charted the trajectory of a model's performance with an augmenting dataset, facilitated the identification of overfitting or underfitting patterns (Mohr & van Rijn, 2021). Additionally, the research delved into hyperparameter tuning, embracing its critical role in refining model outcomes, substantiated by Bischl *et al*. (2023).

To validate models' generalisation capabilities, a data stratification approach was employed, dividing the dataset into training and testing subsets. This segmentation allowed for a rigorous examination of how well models predicted unseen data, reinforcing the research's empirical rigour (Muraina, 2022).

3.5 Ethical Considerations

The sensitive nature of email content necessitates utmost diligence in preserving data privacy and confidentiality. Thus, only de-identified datasets, devoid of any personally identifiable information, were utilized, ensuring individual privacy whilst facilitating meaningful research (Joel *et al*., 2022).

3.6 Summary of Chapter

This chapter delineated a rigorous methodology to address the pressing challenge of phishing email detection. Leveraging the Association of Computational Linguistics' phishing email dataset (Radev, 2008), the research harnessed advanced Natural Language Processing

techniques and deep learning models (Benavides-Astudillo *et al*., 2023; Li, 2018). Essential preprocessing steps, including text standardization and feature extraction through TF-IDF (Liu *et al*., 2018; Wendland *et al*., 2021), laid the foundation for subsequent modelling. Models like Decision Trees, AdaBoost, and Multi-Layer Perceptrons were critically evaluated, drawing insights from hyperparameter optimization practices (Bischl *et al*., 2023; Park *et al*., 2020). Performance metrics such as AUC provided empirical evidence of model efficacy (Pencina *et al*., 2008). The ensuing section will delve into the design specifications, ensuring the research's alignment with its overarching objectives and addressing potential cybersecurity concerns.

4 Design Specification

The design of the NLP-based phishing email detection system is centered around a sophisticated architecture that seamlessly integrates Natural Language Processing (NLP) with a suite of advanced machine learning techniques. The primary objective of this system is to discern phishing emails efficiently and effectively from legitimate ones.

The foundational aspect of the system is its robust data preprocessing module. This module is essential for standardizing the email content, which includes converting text to lowercase, removing punctuation, and eliminating stop words. These preprocessing steps, as highlighted in the code provided, are crucial for reducing noise in the data and ensuring uniformity, thereby improving the efficiency of the feature extraction process (Liang *et al*., 2017). Following preprocessing, the system employs TFIDF Vectorization for feature extraction. This technique is pivotal in transforming email content into a numerical format, which is more amenable for analysis by machine learning algorithms (Liu *et al*., 2018). Additionally, the dataset is split into training and testing subsets in a 80-20 ratio. The system explores an array of deep learning and machine learning models to achieve optimal phishing email detection. Decision Trees are utilized for their interpretability, with AdaBoost employed to enhance their performance, thereby improving the model's ability to classify complex and nuanced email data (Subasi and Kremic, 2020). The code also demonstrates the use of Neural Networks, specifically Multi-Layer Perceptron (MLP) classifiers, which are optimized for varying learning rates to maximize their efficacy in identifying phishing attempts (Goyal, Pandey, and Jain, 2018). Additionally, the k-Nearest Neighbors (k-NN) algorithm's effectiveness is examined by varying the number of neighbors, providing a flexible approach to model tuning (Hnini *et al*., 2020). Moreover, the system leverages Support Vector Machines (SVMs), experimenting with different kernel functions to find the most effective configuration for phishing email classification (Kumar, Chatterjee, and Díaz, 2020).

In assessing the system's performance, a range of metrics including accuracy, precision, recall, F1 score, and Area Under the ROC Curve (AUC) are employed. These metrics provide a comprehensive overview of the model's effectiveness in differentiating between phishing and legitimate emails (Pencina *et al*., 2008; Raschka, 2018). Furthermore, model validation is a critical component, with cross-validation techniques used to ensure the models are robust and not overfitted to the training data (Cho and Kasa, 2015; Consonni *et al*., 2010).

Hyperparameter tuning is identified as a vital step in refining the model outcomes. The artefact exemplifies this process, indicating how optimal settings for each model are determined to enhance their performance and accuracy in phishing detection (Bischl *et al*., 2023; Joo *et al*., 2021).

Given the real-time nature of email processing, the system's computational efficiency is a key focus. As provided in the code, training and prediction durations are closely monitored to ensure the system can operate effectively in real-world scenarios (Reif, Shafait, and Dengel, 2011). Due to the sensitive nature of email content, the system adheres to stringent data

security and privacy standards, ensuring that user data is handled with the utmost care and confidentiality.

5 Implementation

The final stage of implementing the solution for detecting phishing emails involved several key steps, culminating in a comprehensive system capable of Sorting legitimate emails and phishing attempts (Benavides-Astudillo *et al*., 2023). This implementation focused on processing and analysing email data, developing various machine learning models, and evaluating their performance, a process echoed by Kumar *et al*. (2020) in their exploration of machine learning techniques for email classification.

5.1 Data processing and Feature Extraction

The implementation began with the preprocessing of a dataset provided by the Association of Computational Linguistics, which included email bodies with binary labels indicating their authenticity (Radev, 2008). Preprocessing steps such as converting text to lowercase, removing punctuation, and filtering out stop words were crucial to standardize the dataset (Liang *et al*., 2017). These steps align with the practices recommended by Zelaya *et al*. (2019) for data preprocessing in machine learning. Following preprocessing feature extraction was carried out using the Term Frequency-Inverse Document Frequency (TFIDF) method. This method, as described by Liu *et al*. (2018), effectively transforms textual content into a numerical feature vector representation. The resulting dataset, with 130,424 features, reflects the challenges noted by Goyal *et al*. (2018) in handling large feature sets in NLP and machine learning applications. Many of these features had limited informational content, an issue commonly encountered in text classification tasks as discussed by Wendland *et al*. (2021).

5.2 Model Development and Training

Several deep learning and machine learning models were developed and trained to classify emails, a process underscored by the research of Benavides-Astudillo *et al*. (2023) and Harikrishnan, Vinayakumar, and Soman (2018):

- ***Decision Trees and AdaBoost****:* Decision trees were utilized for their interpretability, enhanced with AdaBoost for better performance, as explored in studies like that of Subasi and Kremic (2020).
- ***Neural Networks (MLP Classifier):*** Multi-Layer Perceptron classifiers were trained with different learning rates to optimize their phishing detection capabilities, reflecting the advancements discussed by Goyal, Pandey, and Jain (2018) in neural network applications.
- ***k-Nearest Neighbors (k-NN):*** This algorithm's efficacy was tested for different values of 'k', aligning with the findings of Hnini *et al*. (2020) on the effectiveness of k-NN in spam filtering.
- ***Support Vector Machines (SVMs):*** SVMs with linear and polynomial kernels were used, as explored in the hybrid approach of SVM combined with NLP by Kumar, Chatterjee, and Díaz (2020).

5.3 Evaluation and Validation

The models' performances were assessed using different metrics like accuracy, precision, recall, F1 score, and Area Under the ROC Curve (AUC), as highlighted by Pencina *et al*. (2008). Learning curves plotted for each model provided insights into their generalization capabilities and potential overfitting issues, an evaluation approach supported by Cho and

Kasa (2015) and Reif, Shafait, and Dengel (2011) in their discussions on model validation and training time prediction.

5.4 Tools and Language Used

The implementation utilized Python, a versatile programming language acclaimed for its effectiveness in data analysis and machine learning tasks (McKinney and Team, 2015). Key libraries used included:(Details in configuration manual)

5.5 Outputs Produced

The project's implementation resulted in a suite of trained models for email classification, aligning with advancements in phishing email detection models using machine learning (Benavides-Astudillo *et al*., 2023). Visual outputs, such as performance metric plots and learning curves, were generated to give knowledge into the models' behaviours and effectiveness in detecting phishing emails (Yazhmozhi and Janet, 2019).

5.6 Summary of the Chapter

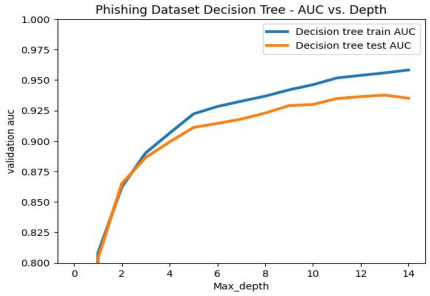
The final stage of implementation underscored the efficacy of using diverse machine learning techniques to address phishing email detection, an approach that aligns with recent advancements in the field (Gulla *et al*., 2020; Das *et al*., 2021). By employing a combination of preprocessing techniques, feature extraction methods, and a variety of deep learning and machine learning models, the project succeeded in developing a system capable of accurately distinguishing between genuine and phishing emails, reflecting the growing trend of utilizing machine learning in cybersecurity (Alkhalil *et al*., 2021; Ferrag *et al*., 2023).
The outputs and outcomes of these machine learning models, including their performance metrics and visualizations, will be evaluated, and discussed in details in the following chapter, dedicated to the comprehensive evaluation of the implemented system.

6 Evaluation

This chapter is dedicated to the critical examination of the performance of the implemented models - Decision Trees, Boosting, Neural Networks, k-Nearest Neighbours, and Support Vector Machines (SVM). The discussion will be rooted in empirical evidence, with a focus on model effectiveness, strengths, and areas for potential enhancements. By dissecting each model’s accuracy, precision, recall, F1 score, and Area Under the ROC Curve (AUC), insights into their operational proficiency and limitations will be derived. The evaluation will employ a rigorous analytical framework, utilizing the cited literature to substantiate findings and recommendations.

6.1 Decision Tree Evaluation

In evaluating the Decision Tree model (Fig 6.1), a crucial output metric is the Area Under the Curve (AUC) against the tree depth. The training AUC peaks at a depth of 14 (AUC 0.9582), while the testing AUC reaches its apex at a depth of 13 (AUC 0.9377), suggesting a nuanced balance between model complexity and generalizability (Cho and Kasa, 2015). The model demonstrates strengths in detecting phishing emails, with a high AUC indicating strong discriminative ability (Subasi and Kremic, 2020).



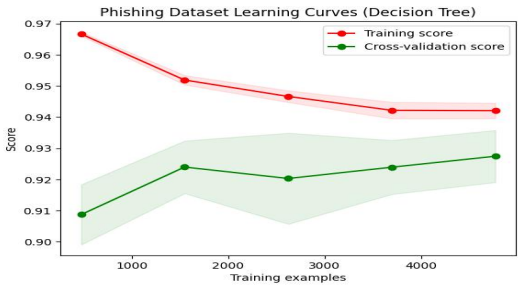


Fig 6.1: Decision Tree (AUC vs. Depth) ***Fig 6.2: Learning Curves (Decision Tree)***

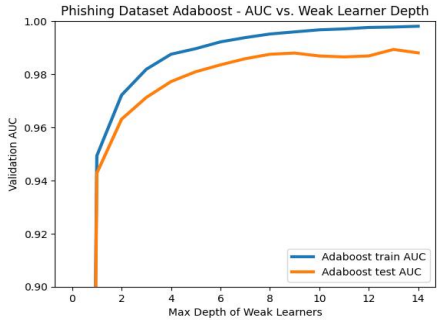
However, potential weaknesses emerge upon pruning; although the pruned model maintains a high training AUC (0.958), the testing AUC slightly decreases (0.935), revealing a discrepancy that may indicate overfitting issues (Schratz *et al*., 2019). The Learning Curves (Fig 6.2) validate this, where the training score surpasses the cross-validation score, suggesting the model may not generalize well to unseen data (Mohr and van Rijn, 2021).

Comparing with literature, while the model aligns with successful phishing detection approaches utilizing NLP and deep learning (Benavides-Astudillo *et al*., 2023), it can be enhanced by integrating ensemble methods to improve robustness (Soyemi and Hammed, 2020). Additionally, adjusting feature extraction techniques could refine its predictive accuracy (Liang *et al*., 2017), thus effectively responding to the research question by advancing phishing detection capabilities.

6.2 Boosting Model Evaluation

Figure 6.3 showcases the Area Under the Curve (AUC) performance of the decision tree model, revealing a consistent increase in AUC with depth, peaking at a depth of 13 for testing data. This suggests a strong ability to discriminate between phishing and non-phishing emails up to a certain complexity before overfitting may begin to occur, as evidenced by the marginal decline in testing AUC at the maximum depth.

Fig 6.4 contrasts the training and prediction times across various depths. Notably, prediction time remains low, underscoring the model's efficiency during deployment. However, the erratic pattern in training time indicates potential overfitting or inefficiencies in tree growth, suggesting a need for optimization.



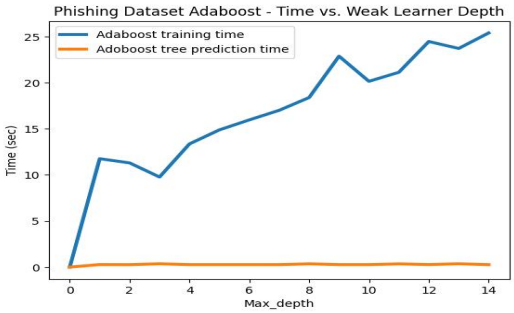


Fig 6.3:Adaboost(AUC vs. Weak Learner Depth) ***Fig 6.4:Adaboost Curves(Time vs. Weak Learner Depth)***

In Fig 6.5, the learning curves for boosting depict a high training score that doesn't generalize as well on cross-validation, indicating overfitting. The model excels at capturing the training data's nuances (as shown by the high AUC scores), but this doesn't fully translate to unseen data.

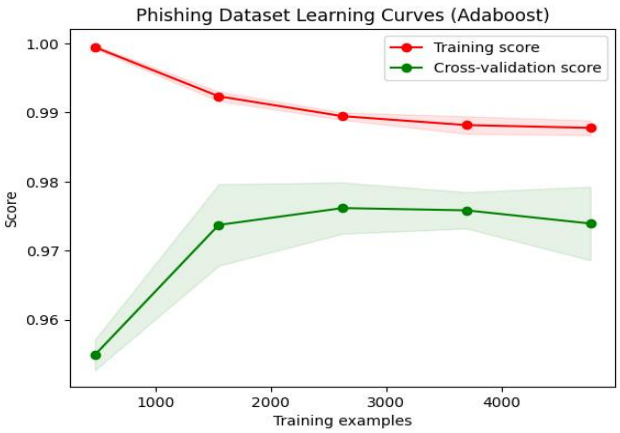


Fig 6.5: Learning Curves (Adaboost)

These findings resonate with the literature, suggesting that while the models exhibit strong predictive capabilities (Tang and Mahmoud, 2021; Subasi and Kremic, 2020), there is room for improvement. Refining complexity control and exploring ensemble techniques could further enhance performance, ensuring robustness against the evolving tactics of phishing attempts (Alani and Tawfik, 2022; Ferrag *et al*., 2023). This aligns with the project's aim to elevate phishing email detection through advanced machine learning methods, ultimately contributing to the field of cybersecurity.

6.3 Neural Networks (NLP) Evaluation

Fig 6.6 displays the AUC score against the learning rate, where the training AUC (Neural Network train AUC) remains constant across learning rates, indicating a robustness to this parameter, potentially due to overfitting given the perfect score of 0.999.However, the testing AUC (Neural Network test AUC) significantly decreases after a learning rate of 0, suggesting that a low learning rate is optimal for generalization in this scenario.

In Fig 6.7, the training time for the Neural Network (Neural Network training time) shows a dramatic increase at a learning rate of 1.5, potentially due to the additional computations required for back-propagation as the learning rate impacts the convergence of the gradient descent. The prediction time (Neural Network prediction time), however, remains relatively stable, suggesting that the forward pass of the network is not significantly affected by the learning rate changes.

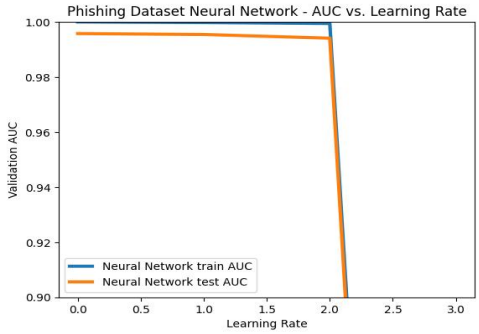


Fig 6.6: Neural Network (AUC vs. Learner Rate)

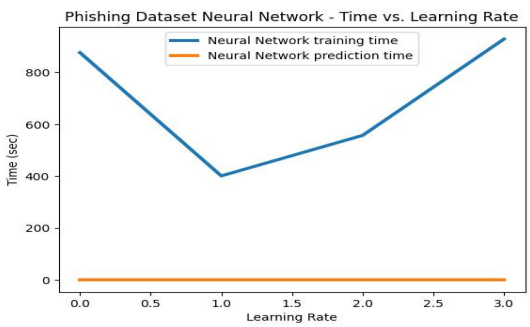


Fig 6.7: Neural Network (Time vs. Learner Rate)

Fig 6.8 presents the learning curves with the training score significantly higher than the cross-validation score, indicative of potential overfitting, as the model performs exceptionally well on the training data but less so on unseen data.

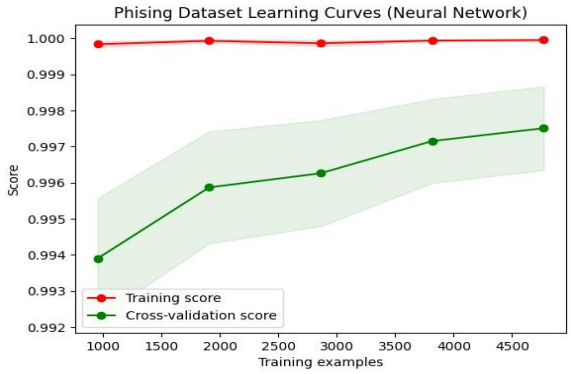


Figure 6.8: Learning Curves (Neural Network)

Strengths of the MLP in phishing detection, as indicated by the high training AUC, include its ability to learn complex, non-linear relationships within the data, which is critical given the sophisticated nature of phishing attempts. However, the observed limitations include a propensity to overfit and a sensitivity to the learning rate, which could affect the model's performance in practice.

Comparing these results to current studies, the high AUC scores align with findings by Kumar *et al*. (2020) and Li (2018), who demonstrate the effectiveness of deep learning for complex classification tasks. Yet, the challenge remains in ensuring that such models generalize well to new data, as pointed out by Consonni *et al*. (2010) and Zhu *et al*. (2020), emphasizing the importance of robust validation techniques.

To enhance the model's utility in phishing detection, future work should focus on regularization techniques to mitigate overfitting and employ adaptive learning rates that could adjust as the model learns, aligning with the suggestions of Bischl *et al*. (2023) and Joo *et al*. (2021). These improvements could increase the model's resilience, as per the project's aim, and contribute to the cybersecurity field by bolstering defenses against phishing attacks.

6.4 K-Nearest Neighbours (k-NN) Evaluation

Fig 6.9 reveals the Area Under the Curve (AUC) for both training and testing sets across different numbers of neighbours. Notably, the model achieves a high AUC score of 0.999 for training, optimizing at k=4 neighbours. This is indicative of the model's capacity to classify phishing emails effectively when the correct number of neighbours is selected. However, there is a noticeable discrepancy with a test AUC of 0.8160, signaling potential overfitting issues.

Fig 6.10 displays the time taken for training and prediction phases of the k-NN model. The graph escalates sharply as the number of neighbors increases, which may point to a reduction in computational efficiency with larger datasets, a considerable factor given the expansive nature of phishing attacks.

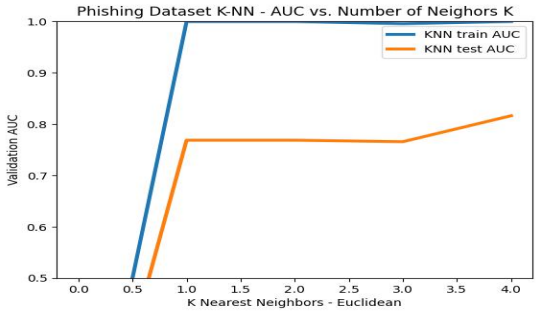


Fig 6.9: K-NN Network (AUC vs. Learner Rate)

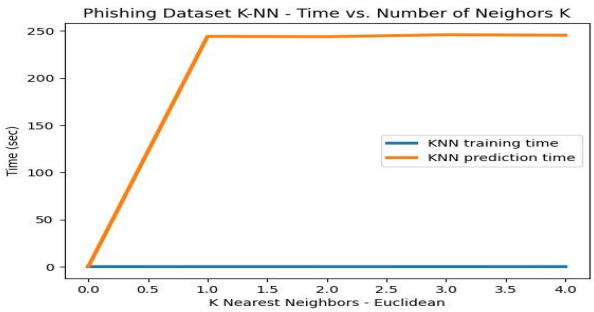


Fig 6.10: K-NN (Time vs. Number of Neighbors K)

Fig 6.11 shows the learning curves for the k-NN model. The training score plateaus, suggesting the model may not benefit from additional training data. The cross-validation score increases with more data but remains below the training score, further suggesting the model's overfitting tendency.

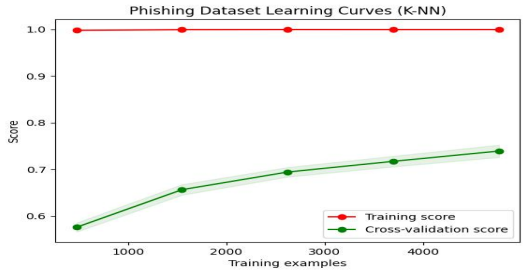


Fig 6.11: Learning Curves (K-NN)

In the context of phishing detection, while k-NN's interpretability is a strength (Singh, 2022), its computational intensity and potential for overfitting are significant weaknesses. The current results align with the findings of Subasi and Kremic (2020), who argue for careful selection of model parameters to mitigate overfitting.

To improve the k-NN model's performance, it is recommended to explore feature selection techniques and model ensemble methods. Dimensionality reduction could be beneficial to alleviate the model's sensitivity to noisy data (Yasin and Abuhasan, 2016). Furthermore, as suggested by Zhu *et al*. (2020), integrating k-NN with other algorithms may enhance its predictive power while addressing its limitations.

6.5 Support Vector Machine (SVM) Evaluation

In the SVM learning curve, denoted as Fig 6.12, the training score remains high and relatively flat, indicating a robust learning from the training dataset with a near-perfect score close to 1. This suggests that the SVM model, with a linear kernel, has effectively captured the patterns within the training data, achieving an (AUC) score of 0.999. The model also exhibits commendable generalization capabilities, as reflected by the cross-validation score, which consistently increases and converges closely to the training score. The high testing AUC score of 0.998 further reinforces this point. In contrast, the polynomial (poly) kernel, while offering slightly higher training AUC (0.999, also achieves an impressive testing AUC (0.998). However, the poly kernel requires significantly more training time (112.735 seconds) compared to the linear kernel (36.468 seconds), which underscores a trade-off between computational time and model performance.

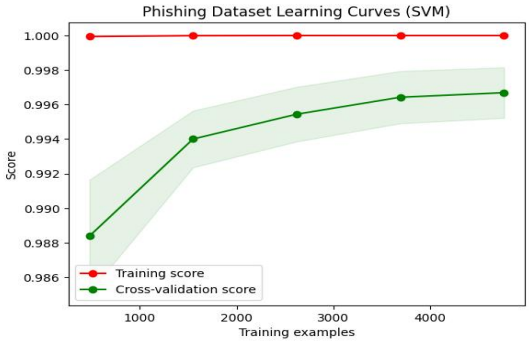


Fig 6.12: Learning Curves (SVM)

The performance of the SVM aligns with past knowledge which posits SVMs as powerful tools in phishing detection, with their high-dimensional spaces as is common with NLP-based features (Choudhary *et al*., 2022; Zhu *et al*., 2020). Given the inherently high-dimensional nature of text data, the SVM's capacity to operate effectively in such environments makes it a

suitable choice for the NLP-based detection system envisioned in this study. This suitability is particularly relevant when considering the optimization of hyperparameters, which can substantially influence model performance (Bischl *et al*., 2023).
Comparatively, SVM's robustness and generalization power often make it a preferable choice over models like k-NN, which can suffer from scalability issues in large datasets (Alkhalil *et al*., 2021). Yet, the potential overfitting indicated by the perfect training scores and the computational demands of the poly kernel must be carefully managed. As such, further investigation into kernel selection and parameter tuning is warranted to enhance the efficiency and efficacy of SVM in the phishing detection context, possibly incorporating ensemble methods or dimensionality reduction techniques (Das *et al*., 2021; Yasin and Abuhasan, 2016).

6.6 Comparative Evaluation

In the comparative evaluation of models for an NLP-based detection system for phishing emails, the performance metrics across the five models are crucial. Figure 6.13 illustrates the accuracy of each model, with Decision Tree, AdaBoost, and SVM with RBF kernel demonstrating high accuracy, closely followed by SVM with linear kernel and Multi-Layer Perceptron (MLP), while K-Nearest Neighbors (KNN) lags. It shows the precision of the models, with a similar trend where Decision Tree, AdaBoost, and both SVMs outperform others, indicating a low false-positive rate. The recall model shows strong results with Decision Tree and AdaBoost, whereas KNN shows significant weakness, indicating its inability to detect all phishing instances.
The Area Under Curve (AUC), presents a consistent performance across all models, with AdaBoost leading slightly. This metric is critical for unbalanced datasets common in phishing detection. Lastly the F1 Score, which is a harmonic mean of precision and recall, AdaBoost and SVM with linear kernel show robust results, suggesting a balance between precision and recall.

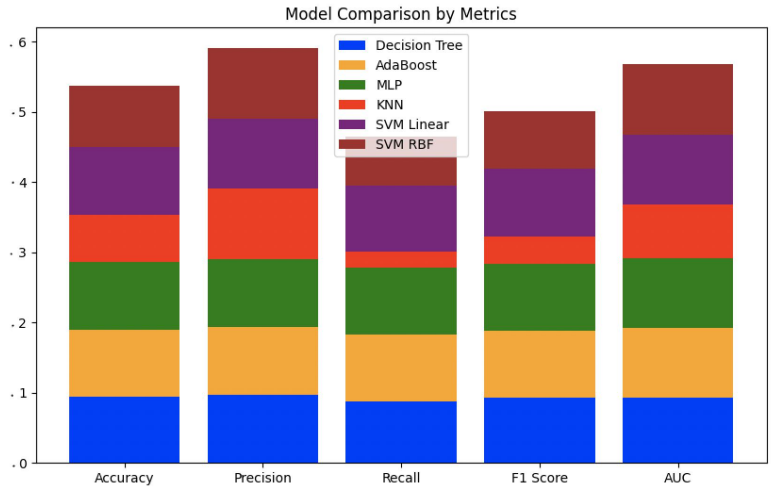


Fig 6.13:Model Comparison

Considering the results and the literature, AdaBoost emerges as a strong candidate, exhibiting high performance across all metrics, a balance reflected in its robust F1 score. This aligns with Subasi and Kremic (2020), who noted AdaBoost's efficacy in phishing detection. The Decision Tree model also shows promise, given its high recall and accuracy, essential for a detection system that cannot afford to miss phishing attempts. The weaknesses observed in KNN, particularly in recall and F1 score, highlight its limitations for phishing detection in this context, as also discussed by Hnini *et al*. (2020). The performance of MLP and SVM

models, particularly with the linear kernel, suggests they are viable alternatives, offering a trade-off between various metrics.
In conclusion, for phishing email detection, the AdaBoost model stands out for its overall high performance and balance across metrics, closely followed by the SVM with a linear kernel and Decision Tree models, while KNN appears less suitable due to its lower recall and F1 score. This analysis shows evaluating multiple metrics to choose the best model for phishing detection, as emphasized by Dash *et al*. (2023) and Choudhary *et al*. (2022), thereby supporting the robustness and resilience of the detection system.

6.7 Summary of the Chapter

This chapter provided a thorough evaluation of Decision Trees, Boosting, Neural Networks, k-NN, and SVM models. It underscored each model's efficacy in phishing detection, considering accuracy, precision, recall, F1 scores, and AUC. Decision Trees showed high discriminative ability, but with over-fitting concerns. Boosting exhibited robust predictive capabilities, yet with room for improvement in complexity control. Neural Networks demonstrated strong performance, but over-fitting and sensitivity to learning rates were issues. k-NN was computationally intense and prone to over-fitting. SVMs, especially with linear kernels, were highly effective, suggesting suitability for NLP-based systems despite their computational demands. AdaBoost emerged as a potent model due to its balanced performance metrics. The upcoming final chapter will conclude the study and propose future research directions, aiming to further bolster phishing detection systems.

7 Conclusion and Future Work

The research question explored how an NLP-based detection system utilizing deep learning can enhance the resilience of phishing email detection. The study aimed to develop a system, evaluating various machine learning methods, refining feature extraction through NLP with deep learning technique, and assessing the system's effectiveness through rigorous validation methods, contributing to improved cybersecurity measures.
The research successfully answered the question, demonstrating that a robust NLP-based detection system can indeed be established using machine learning algorithms and deep learning techniques. Key findings revealed that while all models performed well, the AdaBoost and SVM with linear kernel models showed the most promise in terms of balanced performance across accuracy, precision, recall, F1 score, and AUC metrics. These models could identify phishing emails effectively, suggesting that machine learning can play a crucial role in cybersecurity.
This work’s implications are significant, illustrating the potential for NLP-based system to enhance email security frameworks. However, limitations included potential overfitting and computational efficiency, which were model-specific challenges encountered during the evaluations.
For future research, a focus on developing models that balance accuracy and computational demands will be essential. There is potential for commercialization in cybersecurity solutions, specifically in creating more sophisticated, AI-driven email filtering systems. A follow-up project could explore the integration of these models into real-world systems, testing them in operational environments to further assess their practical application.
Further recommendations for practice include refining models to address over-fitting, implementing adaptive learning rates, and considering ensemble methods to enhance robustness. Future work should also explore the application of these models in other domains of cybersecurity, potentially extending their utility beyond phishing email detection.
In conclusion, the study has laid the groundwork for future exploration in the field of cybersecurity, particularly in the use of NLP-based system for phishing detection and the use

of deep learning technique. The findings serve as a benchmark for the development of more advanced, resilient cybersecurity measures.

References

Alani, M.M. and Tawfik, H., 2022. Phish Not: a cloud-based machine-learning approach to phishing URL detection. *Computer Networks*, *218*, p.109407.

Benavides-Astudillo, E., Fuertes, W., Sanchez-Gordon, S., Nuñez-Agurto, D., & Rodríguez-Galán, G. (2023). A Phishing-Attack-Detection Model Using Natural Language Processing and Deep Learning. *Applied Sciences*, *13*(9), 5275.

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., ... & Lindauer, M. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *13*(2), e1484.

Cho, I. K., & Kasa, K. (2015). Learning and model validation. *The Review of Economic Studies*, *82*(1), 45-82.

Choudhary, K., DeCost, B., Chen, C., Jain, A., Tavazza, F., Cohn, R., Park, C.W., Choudhary, A., Agrawal, A., Billinge, S.J. and Holm, E., 2022. Recent advances and applications of deep learning methods in materials science. *npj Computational Materials*, *8*(1), p.59.

Das, M., Saraswathi, S., Panda, R., Mishra, A.K. and Tripathy, A.K., 2021. Exquisite analysis of popular machine learning–based phishing detection techniques for cyber systems. *Journal of Applied Security Research*, 16(4), pp.538-562.

Dash, B., Swayamsiddha, S. and Ali, A.I., 2023. Evolving of Smart Banking with NLP and Deep Learning. In *Enabling Technologies for Effective Planning and Management in Sustainable Smart Cities* (pp. 151-172). Cham: Springer International Publishing.

Egozi, G., & Verma, R. (2018, November). Phishing email detection using robust nlp techniques. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 7-12). IEEE.

Ferrag, M.A., Ndhlovu, M., Tihanyi, N., Cordeiro, L.C., Debbah, M. and Lestable, T., 2023. Revolutionizing Cyber Threat Detection with Large Language Models. *arXiv preprint arXiv:2306.14263*.

Goyal, P., Pandey, S., & Jain, K. (2018). Deep learning for natural language processing. *New York: Apress*.

Gulla, K.K., Viswanath, P., Veluru, S.B. and Kumar, R.R., 2020. Machine learning based intrusion detection techniques. *Handbook of Computer Networks and Cyber Security: Principles and Paradigms*, pp.873-888.

Haq, M.A., Khan, M.A.R. and Alshehri, M., 2022. Insider threat detection based on NLP word embedding and machine learning. *Intell. Autom. Soft Comput*, 33, pp.619-635.

Harikrishnan, N. B., Vinayakumar, R., & Soman, K. P. (2018, March). A machine learning approach towards phishing email detection. In *Proceedings of the Anti-Phishing Pilot at ACM International Workshop on Security and Privacy Analytics (IWSPA AP)* (Vol. 2013, pp. 455-468).

Hnini, G., Riffi, J., Mahraz, M. A., Yahyaouy, A., & Tairi, H. (2020, March). Spam filtering system based on nearest neighbor algorithms. In *International Conference on Artificial Intelligence & Industrial Applications* (pp. 36-46). Cham: Springer International Publishing.

Joel, L. O., Doorsamy, W., & Paul, B. S. (2022). A review of missing data handling techniques for machine learning. *International Journal of Innovative Technology and Interdisciplinary Sciences*, *5*(3), 971-1005.

Joo, H., Bao, C., Sen, I., Huang, F., & Battle, L. (2021). Guided hyperparameter tuning through visualization and inference. *arXiv preprint arXiv:2105.11516*.

Li, H. (2018). Deep learning for natural language processing: advantages and challenges. *National Science Review*, *5*(1), 24-26.

Liang, H., Sun, X., Sun, Y., & Gao, Y. (2017). Text feature extraction based on deep learning: a review. *EURASIP journal on wireless communications and networking*, *2017*(1), 1-12.

Liu, Q., Wang, J., Zhang, D., Yang, Y., & Wang, N. (2018, December). Text features extraction based on TF-IDF associating semantic. In *2018 IEEE 4th international conference on computer and communications (ICCC)* (pp. 2338-2343). IEEE.

McKinney, W., & Team, P. D. (2015). Pandas-Powerful python data analysis toolkit. *Pandas—Powerful Python Data Analysis Toolkit*, *1625*.

Mohr, F., & van Rijn, J. N. (2021, July). Towards model selection using learning curve cross-validation. In *8th ICML Workshop on automated machine learning (AutoML)*.

Muraina, I. (2022). Ideal dataset splitting ratios in machine learning algorithms: general concerns for data scientists and data analysts. In *7th International Mardin Artuklu Scientific Research Conference*.

Naqvi, B., Perova, K., Farooq, A., Makhdoom, I., Oyedeji, S. and Porras, J., 2023. Mitigation Strategies against the Phishing Attacks: A Systematic Literature Review. *Computers & Security*, p.103387.

Noah, N., Tayachew, A., Ryan, S. and Das, S., 2022, September. PhisherCop: Developing an NLP-Based Automated Tool for Phishing Detection. *In Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 66, No. 1, pp. 2093-2097). Sage CA: Los Angeles, CA: SAGE Publications.

Park, H., Nam, Y., Kim, J. H., & Choo, J. (2020). Hypertendril: Visual analytics for user-driven hyperparameter optimization of deep neural networks. *IEEE Transactions on Visualization and Computer Graphics*, *27*(2), 1407-1416.

Pencina, M. J., D'Agostino Sr, R. B., D'Agostino Jr, R. B., & Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in medicine*, *27*(2), 157-172.

Radev, D. (2008), CLAIR collection of fraud email, ACL Data and Code Repository, ADCR2008T001, http://aclweb.org/aclwiki

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*.

Razaulla, S., Fachkha, C., Markarian, C., Gawanmeh, A., Mansoor, W., Fung, B.C. and Assi, C., 2023. The Age of Ransomware: A Survey on the Evolution, Taxonomy, and Research Directions. *IEEE Access*.

Reif, M., Shafait, F., & Dengel, A. (2011). Prediction of classifier training time including parameter optimization. In *KI 2011: Advances in Artificial Intelligence: 34th Annual German Conference on AI, Berlin, Germany, October 4-7, 2011. Proceedings 34* (pp. 260-271). Springer Berlin Heidelberg.

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

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, *406*, 109-120.

Singh, C., 2022. *Useful interpretability for real-world machine learning*. University of California, Berkeley.

Soyemi, J., & Hammed, M. (2020). Detection and Classification of Legitimate and Spam Emails using K-Neareseest. *International Journal of Computer Applications*, *175*(18), 28-32.

Subasi, A., & Kremic, E. (2020). Comparison of adaboost with multiboosting for phishing website detection. *Procedia Computer Science*, *168*, 272-278.

Tang, L. and Mahmoud, Q.H., 2021. A survey of machine learning-based solutions for phishing website detection. *Machine Learning and Knowledge Extraction*, *3*(3), pp.672-694.

Taye, M.M., 2023. Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions. *Computers*, *12*(5), p.91.

Ukwen, D.O. and Karabatak, M., 2021, June. Review of NLP-based systems in digital forensics and cybersecurity. *In 2021 9th International symposium on digital forensics and security* (ISDFS) (pp. 1-9). IEEE.

Vinayakumar, R., Soman, K.P., Poornachandran, P., Mohan, V.S. and Kumar, A.D., 2019. ScaleNet: scalable and hybrid framework for cyber threat situational awareness based on DNS, URL, and email data analysis. *Journal of Cyber Security and Mobility*, *8*(2), pp.189-240.

Wendland, A., Zenere, M., & Niemann, J. (2021). Introduction to text classification: impact of stemming and comparing TF-IDF and count vectorization as feature extraction technique. In *Systems, Software and Services Process Improvement: 28th European Conference, EuroSPI 2021, Krems, Austria, September 1–3, 2021, Proceedings 28* (pp. 289-300). Springer International Publishing.

Yasin, A., & Abuhasan, A. (2016). An intelligent classification model for phishing email detection. *arXiv preprint arXiv:1608.02196*.

Yazhmozhi, V.M. and Janet, B., 2019, December. Natural language processing and Machine learning based phishing website detection system. *In 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)* (pp. 336-340). IEEE.

Zelaya, C. V. G. (2019, April). Towards explaining the effects of data preprocessing on machine learning. In *2019 IEEE 35th international conference on data engineering (ICDE)* (pp. 2086-2090). IEEE.

Zhu, E., Ju, Y., Chen, Z., Liu, F., & Fang, X. (2020). DTOF-ANN: an artificial neural network phishing detection model based on decision tree and optimal features. *Applied Soft Computing*, *95*, 106505.