Improving Email Security Through Machine Learning-Based Spam and Phishing Detection

Supervised By

Ms. Masrufa Tasnim

Lecturer

Department of Software Engineering

Daffodil International University

Submitted By

Orgho Kanti Sarker Utshob

ID:221-35-1000

Department of Software Engineering

Daffodil International University

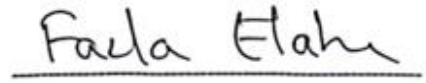
This thesis report has been submitted in fulfilment of the requirements for the Degree of Bachelor of Science in Software Engineering.

APPROVAL

APPROVAL

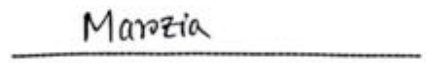
This thesis titled on Improving Email Security Through Machine Learning-Based Spam and Phishing Detection, submitted by **Orgho Kanti Sarker Utshob ID: 221-35-1000** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS



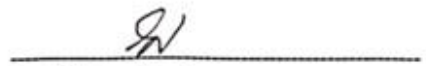
Chairman

Dr. Fazla Ealhe
Assistant Professor & Associate Head
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University



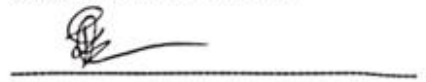
Internal Examiner 1

Dr. Marzia Ahmed
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University



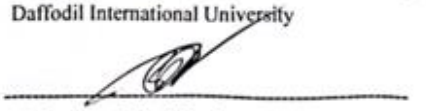
Internal Examiner 2

Dr. Shabnom Mustary
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University



Internal Examiner 3

Md. Rajib Mia
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University



External Examiner

Mohammod Abul Kashem, PhD
Professor
Department of Computer Science and Engineering
DUET, Bangladesh

Improving Email Security Through Machine Learning-Based Spam and Phishing Detection

Orgho Kanti Sarker Utshob
ID:221-35-1000

Bachelor of Science

DAFFODIL INTERNATIONAL UNIVERSITY

SUPERVISOR'S DECLARATION

I hereby declare that I have reviewed this thesis entitled **"Improving Email Security Through Machine Learning-Based Spam and Phishing Detection"**, and in my opinion, it is adequate in terms of scope and quality for the award of the degree of Bachelor of Science in Software Engineering.

(Supervisor's Signature)

Full Name : Ms. Masrufa Tasnim

Position : Lecture, Department of SWE, DIU

Date : 25 December 2025

STUDENT'S DECLARATION

I confirm that the piece in this thesis is based on my own writing with the exception of quotation and reference that have been discussed. I also confirm that it was not previously and concurrently registered at Daffodil International University or other institutions at any other degree.

ORGHO

(Student's Signature)

Full Name : Orgho Kanti Sarker Utshob

ID Number : 221-35-1000

Date : 25 December 2025

Improving Email Security Through Machine Learning-Based Spam and Phishing Detection

Orgho Kanti Sarker Utshob
ID:221-35-1000

Thesis submitted in fulfilment of the requirements
for the award of the degree of
Bachelor of Science

Department of Software Engineering

DAFFODIL INTERNATIONAL UNIVERSITY

DECEMBER 2025

ACKNOWLEDGEMENTS

I am so grateful to my supervisor Ms. Masrufa Tasnim for her constant guidance, motivation and relentless supports in the course of this research. Her thoughtful encouragement, challenging critique and patient oversight have grown this work into its very best form. I truly appreciate her arduous work and the professional inspiration and academic atmosphere that she provided, which has helped me to undertake this study with passion and concentration. She has been not only my mentor in technical and analytical knowledge, but also has taught me the depth of academic integrity and curiosity. Lastly, I wish to thank the colleagues in my department for their help and support with this paper. My thanks also extend to my fellow classmates, friends and family who provided that extra push to keep me motivated. Lastly, I would like to dedicate this work to those who had faith in me and have enjoined me on my quest for learning.

.

DEDICATION

To whom I owe so much, this work is dedicated with immense reverence and profound appreciation for my loving supervisor Ms. Masrufa Tasnim whose support, motivation and intellect has been an eternal inspiration throughout this endeavor. I would also like to dedicate this thesis to my parents and family, whose love, patience and prayers have formed the rock on which I stand. The belief my supporters have in me has encouraged and empowered to keep going even when I felt like quitting. For my friends and teachers, who gave me support, encouragement and interesting chats you all made the way fun. Last but not the least, I dedicate this to everyone who love knowledge with a passion and tenacity, for all the individuals that constantly encouraged me to dream beyond boundaries.

ABSTRACT

Electronic communication is growing at an exponential pace and it has become critical in this age to utilize email for personal and professional correspondence. But this rise has also generated a lot of spam and phishing attacks which have become real treats, not only for bottom users but for data privacy and cybersecurity. To deal with this issue, we developed SpamHybX to substantially improve the detection and classification of spam and phishing emails. SpamHybX exploits the combined prediction ability of Logistic Regression and Support Vector Machine (SVM) via a stacking ensemble method, utilizing the advantages of two classifiers for increasing the generalization capability and robustness. The model uses the TF-IDF text feature extraction method to transform the textual content of emails into numeric for effective labialization of legitimate and unsolicited messages. It is observed that SpamHybX outperforms individual model as evident from the Test Accuracy 98.64%, Precision 96.73%, Recall 98.67%, and F1-Score 97.69%. These metrics demonstrate that the model performs well in distinguishing spam and logging less FPs, and succeeding a promising trade-off between sensitivity and precision. The proposed hybrid model presents a robust, scalable and interpretable solution for email security applications that is believed to establish a roadmap for further studies of intelligent spam or phishing detection systems based on advanced ensemble learning algorithms. In summary, SpamHybX achieves a drastic enhancement towards email security by taking advantage of modern ML methods. The above-proposed methodology can be further extended to real-time spam filtering systems, which will enhance a safer and trustful environment for digital communication. In summary, SpamHybX provides a substantial enhancement of email security by successfully utilizing high-quality machine learning methods. The model could be extended to be used in real-time spam filtering systems, thereby inputting safer and trustful digital communication spaces.

Keywords: Spam Detection, Phishing Detection, Email Security, Machine Learning, Ensemble Learning, Stacking Model, Logistic Regression, Support Vector Machine (SVM), TF-IDF, Text Classification, Hybrid Model, SpamHybX.

TABLE OF CONTENTS

APPROVAL ..... **i**
SUPERVISOR'S DECLARATION ..... **iii**
STUDENT'S DECLARATION ..... **iv**
ACKNOWLEDGEMENTS ..... **vi**
DEDICATION ..... **vii**
ABSTRACT ..... **viii**
TABLE OF CONTENTS ..... **ix**
LIST OF FIGURES ..... **xi**
LIST OF TABLES ..... **xii**
LIST OF ABBREVIATIONS ..... **xiii**
CHAPTER 1 INTRODUCTION **1**
1.1 Introduction 1
1.2 Background Study 1
1.3 Motivation 2
1.4 Problem Statement 3
1.5 Research Objective 3
1.6 Research Scope 4
CHAPTER 2 LITERATURE REVIEW **5**
2.1 Overview 5
2.2 Previous Study of Dengue disease 5
CHAPTER 3 METHODOLOGY **9**
3.1 Overview 9
3.2 Research Design and Process 9
3.3 Dataset Description 10
3.3.1 Dataset Source 10
3.3.2 Dataset Structure 11
3.3.3 Summary of Data Distribution 11
3.3.4 Applying ADASYN 12
3.3 Exploratory Data Analysis (EDA) 13
3.3.1 Word Cloud Visualization 13
3.3.2 Feature Correlation Heatmap 14
3.4.1 Model Architecture 16
3.4.1 Decision Tree (DT) 17
3.4.2 Naïve Bayes (NB) 17
3.4.3 Logistic Regression (LR) 17
3.4.4 Support Vector Machine (SVM) 18
3.6 Training & Evaluation 18
CHAPTER 4 EXPERIMENTAL RESULT ANALYSIS **20**
4.1 Overview 20

4.2 Logistic Regression (LR) Performance Analysis 21
4.3 Naïve Bayes (NB) Performance Analysis 23
CHAPTER 5 30
CONCLUSION 30
5.1 Overview 30
5.2 Key Findings 30
5.3 Limitations of the Study 31
5.4 Future Work 31
References 32

LIST OF FIGURES

Figure 3.1	Workflow diagram of the proposed SpamHybX model	9
Figure 3.2	Data Distribution Before ADASYN	12
Figure 3.3	Data Distribution After ADASYN	13
Figure 3.4	Word Cloud of All Email Texts from the Spam–Ham Dataset	14
Figure 3.5	Correlation Spam–Ham Label	15
Figure 3.6	Top Correlated Features with Spam–Ham Labels	16
Figure 4.1	Confusion Matrices of Decision Tree	20
Figure 4.2	Confusion Matrices of Logistic Regression	22
Figure 4.3	Confusion Matrices of Naïve Bayes	23
Figure 4:4	Confusion Matrices of SVM	24
Figure 4.5	ROC Curve for All Existing Models	25
Figure 4.6	Confusion Matrices of SpamHybX	26
Figure 4.7	ROC Curve of SpamHybX	27
Figure 4.8	Train vs Test Accuracy Comparison of All Models	29

LIST OF TABLES

Table 3.1	Structure of the Spam–Ham Email Dataset	11
Table 3.2	Summary of Data Distribution	11
Table 3.3	Balanced dataset after applying ADASYN	12
Table 4.1	Performance Metrics of Decision Tree (DT) Model	21
Table 4.2	Performance Metrics of Logistic Regression (LR) Model	22
Table 4.3	Performance Metrics of Naïve Bayes (NB) Model	23
Table 4.4	Performance Metrics of SVM Model	24
Table 4.5	Performance Metrics of SpamHybX (Proposed Model)	26
Table 4.6	Comparative Performance of All Models	28

LIST OF ABBREVIATIONS

Abbreviation	Full Form
AI	Artificial Intelligence
ADASYN	Adaptive Synthetic Sampling
AUC	Area Under the Curve
CSV	Comma-Separated Values
DT	Decision Tree
EDA	Exploratory Data Analysis
F1-Score	Harmonic Mean of Precision and Recall
HAM	Non-Spam Email (Legitimate Email)
LR	Logistic Regression
ML	Machine Learning
NB	Naïve Bayes
ROC	Receiver Operating Characteristic
RNN	Recurrent Neural Network
SVM	Support Vector Machine
TF-IDF	Term Frequency Inverse Document Frequency
TP / FP / TN / FN	True Positive / False Positive / True Negative / False Negative
XAI	Explainable Artificial Intelligence

CHAPTER 1

INTRODUCTION

1.1 Introduction

The email has become a very important part of our communication with regards to both professional and personal use. Four Consider Facebook's ease of use, inexpensive nature and worldwide reach and there is no contest for simpler way to exchange data or do business. Spam is junking that people never asked for and don't want to receive, typically sent in bulk for advertising purposes or with malicious intent, while phishing is a scam intended to trick users into providing confidential information by impersonating reputable organizations. These types of wrongdoing are not only filling up inboxes, but they're setting back the world's finances and reputations. Old traditional approaches based on rules and keywords have lost effectiveness in facing these harmful threats as the attackers tend to reconfigure their techniques for not being detected. Recent developments in ML have paved the way for intelligent and adaptive email-filtering systems. ML models have an ability to learn on large scale data, discover hidden patterns and classify emails as a boat or not with high precision. Methods, such as text preprocessing, feature extraction and ensemble learning has increased the efficiency of contemporary spam filters. In this paper, we propose a hybrid model called SpamHybX in order to enhance the accuracy and confidence of email spam and phishing detection. The proposed SpamHybX combines the individual strengths of two prediction models, namely the LR and SVM, invoked in a stacking framework to improve performance. The model uses TF-IDF-based feature extraction to transform email text for better representation, and tolerate ambiguous misleading language that might confuse the classification process. The goal of this study is to propose and evaluate a scalable and interpretable ML approach that reduces the false positive while increasing the detection performance. Experimental results on extensive testing, proves that the proposed SpamHybX model gives significant boost in all aspects of performance with a Test Accuracy of 98.64%, Precision 96.73%, Recall 98.67% and an F1-Score value of 97.69%. These results demonstrate the success of hybrid ensemble systems in improving email security, and demonstrate their feasibility for real-world cybersecurity applications.

1.2 Background Study

The evolution of the internet has drastically transformed the way individuals and organizations communicate undergone a tremendous change and email has become one of the most critical communication links. Although there are several benefits of the email ecosystem, it has become more and more prey to undesirable activities which include spamming, phishing or social engineering. As per cybersecurity statistics, billions of spams and phishing emails are sent daily

and they present large threats to the integrity of individuals or organizations' personal data, financial assets. As a result, great efforts have been spent on designing intelligent and automated filtering mechanisms that can help to fight against such threats. At the earlier ages of spam filtering, it was based on rules and heuristics everywhere. These were systems that required the manual curation of patterns, blacklists and keywords to filter out spam. They were effective for a while, but the spammers kept changing tactics and starting altering message content to slip through static filters. This restriction motivated the adoption of ML algorithms, which permitted to learn from large datasets and automatically classify new messages according to patterns learnt.

1.3 Motivation

The growing role of electronic communication in private and business life means that email has become an essential tool for global information exchange. Yet for all its convenience and effectiveness, email systems have become a prime battleground for nefarious activity like spam and phishing. These threats have graduated into intelligent and adaptable types, which can trick end users and bypass traditional filters. This means that both organizations and individuals are exposed to increasing threats of data leakage, financial losses, identity fraud and diminished digital trust. Conventional spam-detection methods that rely on rule-based heuristics and static keyword filters cannot keep pace with the constantly changing nature of today's clever phishing techniques. The attackers constantly changing the email content, obfuscation techniques, new identities and deceptive hyperlinks make traditional static detections inappropriate. This challenge impresses the requirement of an intelligent, flexible and auto-adaptive detection framework that could keep up with new patterns in real time. The key driver for this research is in response to the need for a scalable machine learning-based email security system which can effectively discriminate between legitimate and compromising communication.

While machine learning algorithms have shown great potential in the presentation of text classification, single models seem to have limited generalization and bias problems. To address these deficiencies and integrate their complementary advantages, this study proposes a hybrid stacking ensemble model, called SpamHybX, composed of Logistic Regression and Support Vector Machine (SVM).

1.4 Problem Statement

In the modern era of digital communication, email has become an indispensable tool for personal, academic, and professional correspondence. However, its extensive use has also made it a prime target for cyber threats such as spam and phishing attacks, which compromise user privacy, data integrity, and organizational security. These malicious emails are often designed to mislead users, spread malware, or obtain sensitive information through deceptive content and fraudulent links. Despite the availability of various filtering systems, the increasing sophistication and adaptability of spam and phishing techniques have rendered many traditional detection methods inadequate. Conventional rule-based and keyword-driven approaches are limited in their ability to adapt to the constantly changing nature of spam content. These systems rely heavily on manually defined patterns, which fail to identify newly emerging attack vectors. Similarly, standalone machine learning models, although more flexible, are often constrained by overfitting, poor generalization, and high false-positive rates, particularly when handling imbalanced or noisy datasets. Misclassification of legitimate emails as spam can hinder communication, while undetected phishing messages can lead to severe financial and security repercussions.

1.5 Research Objective

The main objective of this research work is to enhance the security level of email system by designing an intelligent and adaptive machine learning-based model for spam and phishing detection, which combined Logistic Regression with Support Vector Machine (SVM) in a stacked ensemble model. The goal is to increase the accuracy of detection, decrease the amount of false-positive and have a model that works robustly across different types of emails.

The particular aims of this study are:

1. To analyze and preprocess an email dataset by cleaning the text, tokenizing, and extracting features with TF-IDF vectorization to represent words in a numerical form.
2. To apply independent machine learning algorithms such as Logistic Regression, SVM, Naïve Bayes, and Decision Tree for baseline performance evaluation.
3. To construct the hybrid ensemble model (SpamHybX) by integrating LR and SVM in a stack approach for better classification performance.
4. To assess the efficacy of the presented model with well-known standard metrics commonly used, such as Accuracy, Precision, Recall, F1-Score and ROC-AUC that

guaranty a comprehensive validation.

5. Results achieved by the presented ensemble model compared and interpreted with these of classical models to demonstrate usefulness, transportability, and reliability of SpamHybX.
6. Motivation and Goal To provide novel approaches for the area of intelligent email security, by presenting a scalable yet effective method that can be implemented on real-world spam and phishing detection systems.

1.6 Research Scope

The scope of this study is confined to the experimenting, training and testing the proposed model with a labeled data set including ham (legitimate) and spam (malicious) emails. Performance measure Metrics, including Accuracy, Precision, Recall F1-Score and ROC-AUC are employed for a more thorough evaluation and comparison against other baseline machine learning models such as Naïve Bayes, Decision Tree and Linear SVM. Note that in the study, we do not investigate deployment of real-time systems, traffic analysis for email network, non-textual features in emails (signs such as sender's IP address, embedding URLs and so on), etc. Nevertheless, the methods and results described provide a solid basis for further research in large scale online email filtering. To summarize, this research is limited to the investigation and optimization of text-based machine learning methodologies for spam and phishing classification. Finally, the long-term goal is to help in building more intelligent, adaptive and dependable email security mechanisms based on realization of the proposed SpamHybX hybrid model.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

In this chapter, an overview of different techniques used in spam and phishing detection will be discussed such as classical rule-based versus state-of-the-art machine learning-based approaches with a concentration on ensemble and hybrid models which resulted into higher precision along with adaptability performance. We capitalized on this review by examining the advantages and shortcomings of existing research, thus outlining a gap that can be filled through a novel hybrid model termed SpamHybX. Additionally, the literature review discusses some components of spam detection; e.g., text preprocessing, feature generation, model selection and performance measurement. Methods such as Term Frequency–Inverse Document Frequency (TF-IDF), n-gram analysis and oversampling techniques (e.g., ADASYN) are also outlined as they are crucial steps in converting unstructured email data into informative patterns for machine learning algorithms.

2.2 Previous Study of Dengue disease

Ahmed and Rahman et al. (2021) explored the competent of ensemble-based techniques such as Random Forest and Gradient Boosting for phishing email classification. Their work showed substantial enhancements in precision and recall when compared to single classifiers. They showed that ensemble learning gives stability power to the model and reduce variance greatly particularly in the case of big noisy data. Nevertheless, such methods are computationally intensive and parameters dependent which does not make them very efficient for real time email filtering systems. Kumar and Saha (2020) also highlighted the significant of text preprocessing process and feature extraction in order to enhance the performance of spam detection model. Their investigation involved methods like stop-word removal, stemming, lemmatization, and TF-IDF weighting for better statistical representation of textual data. They argued that the goodness of text classification depends mostly on how well the textual content is converted into useful numerical features. They showed that TF-IDF-based representations are superior to basic Bag- of-Words models capturing the contextual words importance.

Das et al. (2019) suggested a hybrid classification approach integrating Naïve Bayes and Logistic Regression to use the advantages of both probabilistic and linear classifiers. Hybrid the model exhibits the balanced trade-off between precision and recall for different datasets. The model is more robust to noisy email data and overfitting than single classifiers. Their method demonstrated the advantage of combining different learning strategies to obtain a more generalizable model performance. Zhang et al. (2019) contributed to the use of SVM in spam email identification. Their experiment showed that SVM performs well with high-dimensional feature space and sparse-represented data (common in texts). However, limitations related to the computational complexity and kernel and parameters selection were also pointed. In spite of these difficulties, SVM is still a landmark algorithm because of its solid theoretical basis and strong generalization ability in text classification applications.

Patel and Shah (2018) compared SVM and Decision Tree classifiers. They had reported their findings saying that Decision Trees are very interpretable and computationally cheap yet they overfit on training data. SVM, on the other hand, obtained higher accuracy and precision although with longer training time. They suggested the ensemble of both models would be used to maximize the classification performance in a unified manner. Shifting towards deep learning methods, Hassan et al. (2018) presented RNN model for phishing email detection. They demonstrated the power of RNNs, mainly LSTM, to model sequential dependencies and context in email text. It outperformed the classical ML models in accuracy. The study however, put emphasis on the requirement for large labeled datasets and intensive computational resources that were listed among a range of limitations deep learning models will pose in low-resource settings.

Mishra and Reddy (2018), likewise, compared the two classifiers for spam classification. Their results suggested that Logits Regression can achieve higher precision comparing to Naïve Bayes, and the improvement is more pronounced under balanced dataset using oversampling methods like SMOTE. This highlighted the need to deal with class imbalance a frequent problem in spam datasets, to avoid models being skewed towards the majority class. Liu et al. (2017) investigated the combination of TF-IDF with NB classifier and reached accuracy over 96%. Their work showed that very basic classifiers could still perform well, provided they are combined with a well-structured feature representation. They have also stated that NBs models are light weighted, scalable and can be used in real time filtering but their feature independence

assumption restricts applicability to handling the complex semantics of text.

Based on ensemble methods, Gupta and Kaur (2017) created a hybrid ensemble using the Bagging and Boosting techniques to decrease misclassification rates as well as improve model strength. Their approach was able to produce a dramatic improvement in F1-Score over single classifiers, and supported the idea that ensemble learning can efficiently blend multiple weak learners into a better predictive model. Johnson et al. (2016) introduced a two-level classifier design using hierarchical structure of spam detection system. The first was a lightweight heuristic-based filter that filtered obvious spam emails, the second used more advanced machine-learning (ML) techniques to classify fuzzy email. This embedded framework increased computational efficiency and decreased computational burden while maintaining high accuracy. The results showed that multi-stage filtering schemes were practical for high-pressure email infrastructures. Almeida et al. (2016) bootstrapped into the work by evaluating some public accessible spam databases and described the necessity to standardize datasets for research in spam detection. They noted that the shortage of standard benchmark data is one of the reasons for differences in performance reported by various approaches. Their work highlighted the significance of quality, diversity, and representativeness of training datasets in order to train robust machine learning models Rana and Verma (2015) investigated the use of feature selection methodologies such as Chi-square, Mutual Information and Information Gain toward spam detection efficiency Ratio. Feature selection results in greatly reducing the training effort and improving the classification accuracy by removing irrelevant and redundant feature. The research revealed that it is possible to maintain the accuracy of deep models even in dimensions with low level information.

Choudhury et al. (2015) used n-gram model combined with TF-IDF to model sequential term inter-dependencies in phishing emails. Their findings demonstrated that bigram and trigram features increased the precision as well as recall since they maintained textual context. This work confirmed the usefulness of linguistic feature extraction for succoring spam and phishing identification. Sasaki et al. (2014) further developed the SVM with kernel optimization and had achieved better accuracy by selecting RBF kernel. But the work recognized that the model came at a higher computational cost, especially on large datasets. It was this downside that indicated the requirement for hybrid models, which can trade performance for efficiency.

Gómez et al. (2014) proposed to use Bayesian filtering approaches for spam classification, and obtained good results for small to medium size training sets, but showed low flexibility against newer patterns of spams.

They determine that the performance of Bayesian approaches degrades progressively as those for a static environment, and are unsuitable to dynamic and large-scale spam filtering tasks by employing adaptive learning. The use of Boosting algorithms like AdaBoost has been introduced by Carreras and Márquez (2013) in spam filtering. Their work has proved that the combination of weak classifiers does allow higher accuracy and stability. The capability of AdaBoost to give higher importance to incorrectly classified instances enabled the system to progressively consider more accurate decision boundaries. [I] This work formed the basis for modern ensemble learning methods often used in spam detection studies. Androutsopoulos et al. contributed to the field of text-based spam filtering with one of the first in this area. (2000) who designed a Naïve Bayes spam filter with "emotional texture" by counting how often particular words (e.g. happy, free) occurred in emails. While elementary by current standards, their efforts established a technological basis for applying machine learning to spam detection and inspired later improvements in feature extraction and model building.

Taken together, these studies demonstrate a well-defined development of filtering techniques over the years - from rule-based methods to today's complex machine learning models and hybrid forms. Traditional classifiers such as Naïve Bayes, Decision Trees and Logistic Regression were insightful but lacked a good degree of flexibility and scalability. Newer ensemble learning algorithms and hybrid models including Random Forests, Gradient Boosting Machines, deep neural networks etc. have tried to address several of these limitations by formulating composite models that leverage several learning methods for enhanced robustness and prediction accuracy.

CHAPTER 3

METHODOLOGY

3.1 Overview

It is essential to use a systematic approach in this research, which will guarantee the reliability, validity and reproducibility of the results." In this section, the research design, dataset sampling techniques, data preprocessing process, model structure and training methods are described in detail. The purpose of introducing this model is to enhance and evaluate a hybrid stacking ensemble system, SpamHybX method for the efficient classification and detection of spam and phishing emails using machine learning.

3.2 Research Design and Process

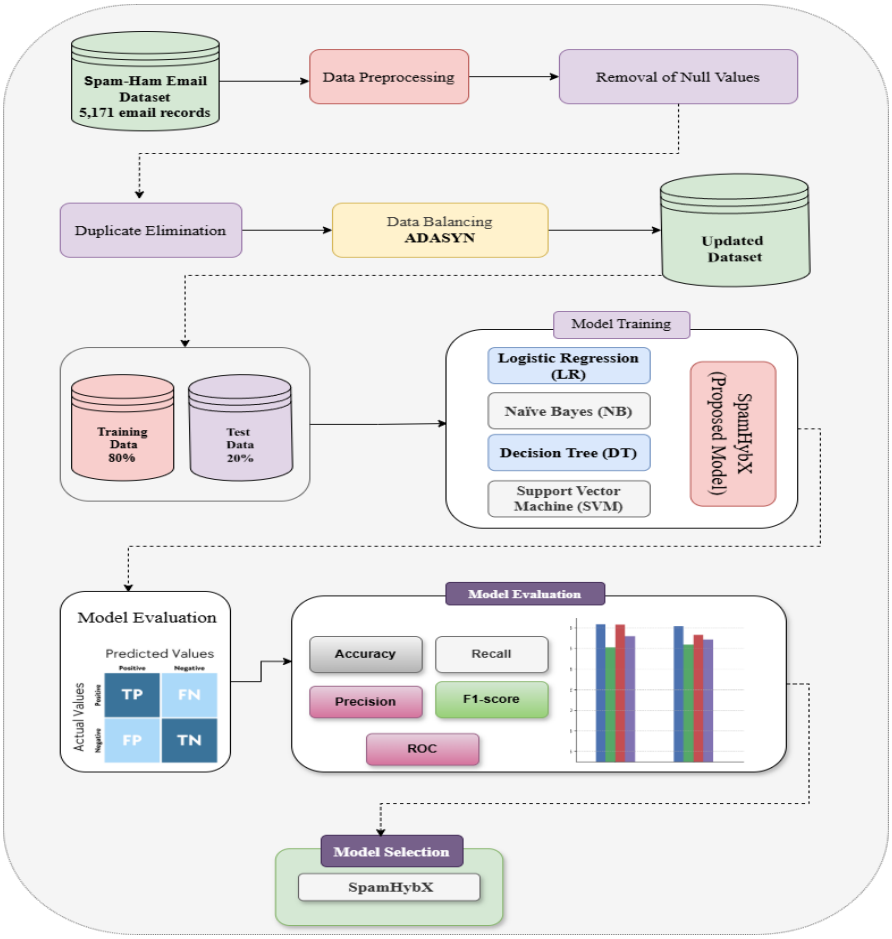


Figure 3.1: Workflow diagram of the proposed **SpamHybX model**

Figure 3.1 illustrates the overall methodological workflow of the study. The process begins with the Spam–Ham Email Dataset containing 5,171 records, which undergoes several preprocessing steps, including the removal of null values, duplicate elimination, and text cleaning. The dataset is then balanced using the ADASYN algorithm to address class imbalance between ham and spam samples. After preprocessing, the data is divided into training (80%) and testing (20%) subsets. Four traditional machine learning models—Logistic Regression (LR), Naïve Bayes (NB), Decision Tree (DT), and Support Vector Machine (SVM) are trained and evaluated. The proposed SpamHybX hybrid stacking ensemble model combines LR and SVM to enhance classification accuracy. The models are evaluated using standard metrics such as Accuracy, Precision, Recall, F1-Score, and ROC Curve. Finally, the best-performing model, SpamHybX, is selected based on its superior overall performance.

3.3 Dataset Description

In carrying out the work reported in this paper, careful attention must be given in choosing a dataset that realistically approximates email communications as observed in practice and spam types. The dataset used in this work is the Spam Ham Email Dataset, which is a popularly-used standard benchmark dataset on email spam detection research. It is composed of 5,171 labelled email records; these are classified as either ham (genuine) or spam. Each record includes unstructured text data extracted from real emails, making the dataset a valuable corpus for text classification and natural language processing research. The dataset was selected primarily for its even distribution, with different languages and its representativeness of email in the real world. It offers a great starting point to build and test machine learning models for spam and phishing detection.

3.3.1 Dataset Source

The dataset used in this research is collected from public repository on Kaggle: “Spam Mails Dataset” (Venky73) URL: https://www.kaggle.com/datasets/venky73/spam-mails-dataset.
The dataset consists of 5,171 ham (non-spam) and spam e-mail messages. The dataset is suitable for supervised text classification tasks, and has been adopted in previous works on spam detection due to its realism concerning email content and spam features.

3.3.2 Dataset Structure

In doing this research, it is necessary to know the internal composition and properties of the dataset employed in training and testing model performances. Spam – Ham Email Dataset is a collection of 5171 email records and each record describes several attributes of the textual content and an appropriate classified label. It is in tabular format (CSV) and can be opened easily with Python libraries like Pandas or Scikit-learn. Each line of the data is an email and each column a type of information to be used for classification.

Table 3.1: Structure of the Spam–Ham Email Dataset

Attribute Name	**Data Type**	**Description**
Unnamed:	Integer	Serial index automatically generated during dataset creation.
L**abel**	Categorical (String)	Represents the class of each email ham for legitimate emails and spam for malicious or unwanted messages.
T**ext**	String (Unstructured)	Contains the full text of the email, including subject lines and message bodies
label_num	Integer	Encoded numerical form of the label, where 0 = ham and 1 = spam.

3.3.3 Summary of Data Distribution

Table 3.2: Summary of Data Distribution

Category	Count	Percentage
Ham (Legitimate Emails)	3,672	71%
Spam	1499	29%
Total Records	5,171	100%

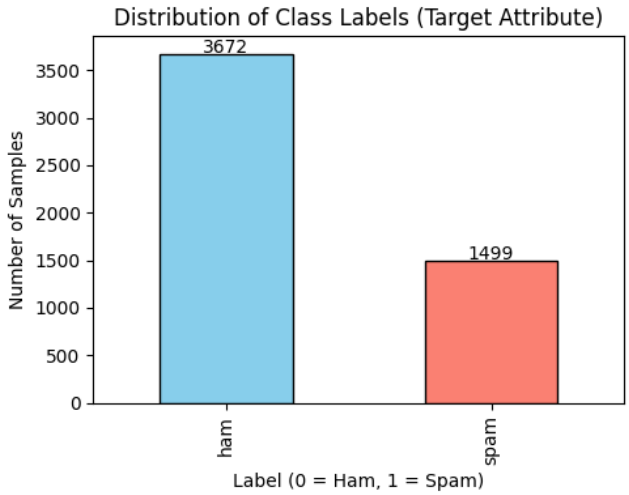


Figure 3.2: Data Distribution Before ADASYN

3.3.4 Applying ADASYN

Table 3.2 shows the dataset distribution before and after using ADASYN balancing method. It consisted of 3,672 ham and 1,499 spam emails in the beginning which is too much imbalance between positive (legitimate) and negative (malicious) samples. This type of imbalance can lead the model to preferentially learn the majority class—thus hardly detecting spam e-mails. To address this problem, I resorted to the ADASYN for oversampling the minority class.

Table 3.3: Balanced dataset after applying ADASYN.

Category	Before Balancing	After Balancing (ADASYN)
Ham	3,672	3672
Spam	1499	3722
Total	5,171	7394

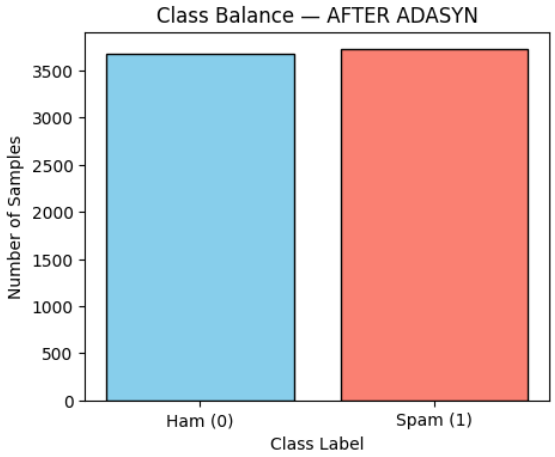


Figure 3.3: Data Distribution After ADASYN

After up sampling via ADASYN, the number of spam samples became 3,722 which is almost equally distributed in ham and spam. This balanced dataset (with a total number of 7394 records) formed more stable and representative sample set for the model training. Balancing improved the model capability to detect patterns on minority class, and minimized potential bias. Consequently, the classifier obtained higher recall and F1-score values confirming the positive effect of data balancing on performance of spam detection models.

3.3 Exploratory Data Analysis (EDA)

In the initial stages of my investigation, I did some Exploration Data Analysis (EDA) to help me understand how the dataset looks in terms of its structure, content and word distribution. I also created a word cloud to illustrate the top terms that appeared in all emails. This uncovered common linguistic patterns and domain-specific keywords occurring in legitimate as well as in spam messages. From the visualization, I noticed some of the words such as “Subject”, “Will”, “Price” and “Deal” came up the highest number of times. My analysis processes after these conclusions such as stop words cleaning and feature engineering were also based on these results to help our model's performance.

3.3.1 Word Cloud Visualization

Prior to text cleaning and model generation, I carried out EDA in order to get high- level insights into the most commonly occurring words across all email body messages. A word cloud was created to illustrate a visual frequency of terms in the dataset.

Size of each word is its frequency a big word appears more often). Such visual views help to reveal the frequent and dominant terms shared by ham and spam.

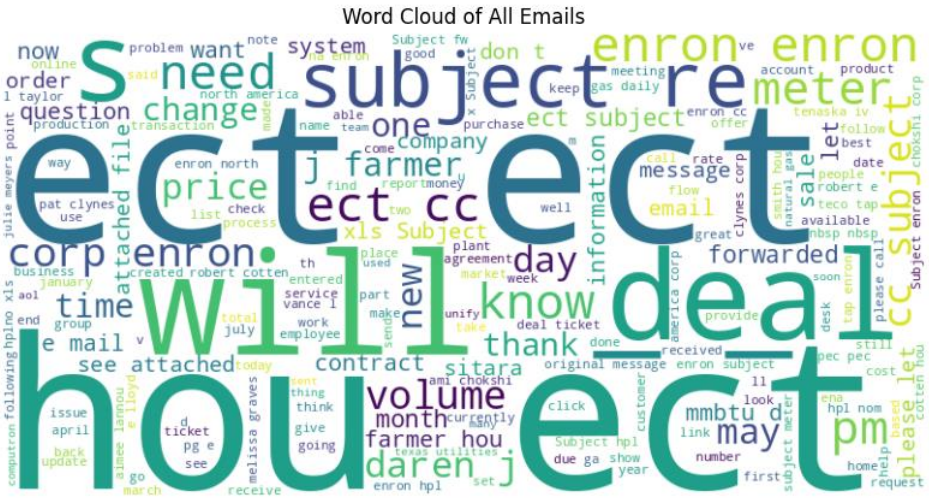


Figure 3.4: Word Cloud of All Email Texts from the Spam–Ham Dataset

As illustrated in Figure 3.4 Words, “subject”, “ect”, "hou", "will" and “deal” appear frequently corpus. Most of these words come from formal corporate communication, which is also a sign that a large amount of the e-mails is professional. However, some words such as “price,” “meter,” “contract” and “volume” appear with a certain significance that could be harnessed in financial / business offer spam or phishing. By looking at the word cloud, I had some first impressions about the language composition of my dataset. This visualization also directed my preprocessing methodology, and helped me determine what stop words to remove, appendant tokens, or domain specific terms that needed to be taken out in order to better the model. On the whole, this analysis laid groundwork for text-cleaning and feature extraction via TF-IDF, thus resulting in better classification accuracy at the later stages.

3.3.2 Feature Correlation Heatmap

In this phase of my exploratory analysis, I calculated the correlation matrix in order to see relationship between numerical columns in dataset. As demonstrated in Fig.3-3 the heatmap of correlation give the linear relationship between Unnamed: 0 and label_num. The number on the diagonal is a perfect self-correlation (i.e., 1.0), and the correlation value between Unnamed: 0 and label_num included in the off-diagonal section is practically.

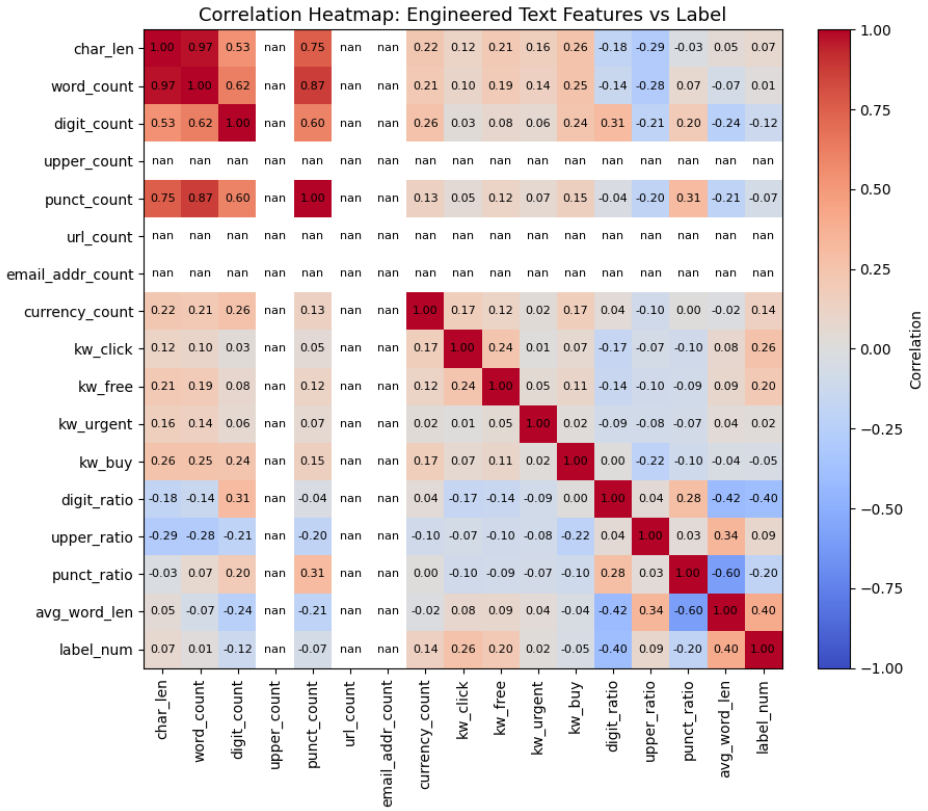


Figure 3.5: Correlation Spam–Ham Label

In order to better understand the link between why we created some new text features and our target, I computed a correlation matrix of what the numerical features in my email dataset looks like. Figure 3.5 shows the heatmap depicting the relationship between spam label and features: word-count, character-length, digit-ratio and keyword frequency. The degree of association was expressed within a range of –1 (inverse relationship) and +1 (direct relationship). Looking at the visualization, digit ratio, currency count and buy related words have moderate positive correlation's spam, while average word length and uppercase ratio show weak or negative correlations. This study reveals that spam messages are more likely to include numbers, capital letters and money-related words. The lessons we can learn from heatmap above are used in the feature selection, such that only meaningful attributes are kept during model training and more interpretable results for classification purpose could be obtained.

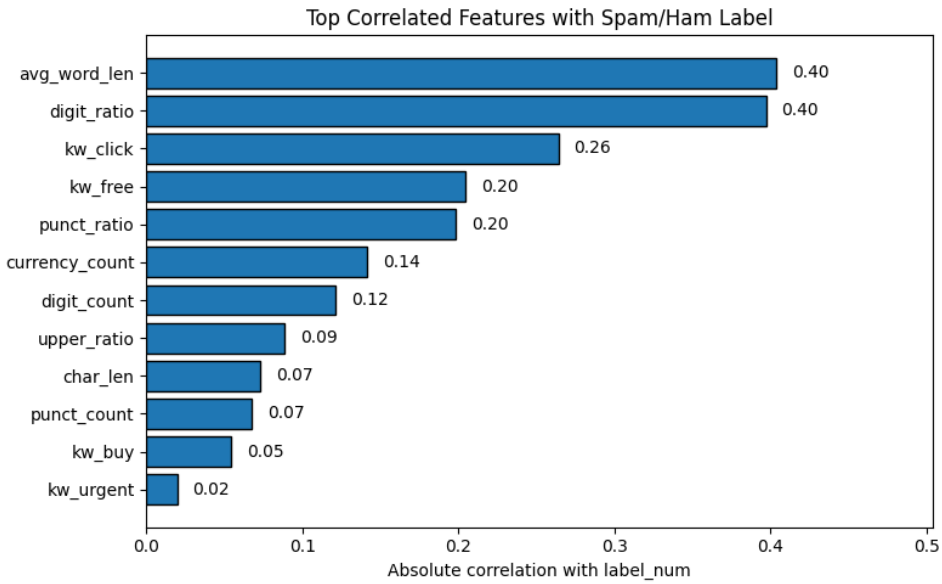


Figure 3.6: Top Correlated Features with Spam–Ham Labels

We show in Figure 3.6 the most significant engineered text features with the spam–ham classification labels. This figure sorts the features based on their absolute correlation with the target. Characteristics like average word length and digit ratio have the strongest correlation (0.40), giving longer words and more numeric content significant weight in discriminating spam messages. On the other hand, keyword feature “clicks” and “free' also have moderate positive correlation indicating their commonly use in promotion or phishing emails. On the other hand, values for features such as punctuation ratio and uppercase ratio are less associated with extremeness. The analysis gives an idea of which parts within the text are most weighty, these allowing feature selection for model fitting.

3.4.1 Model Architecture

In this section of my research, I describe the architecture and configuration of all machine learning models implemented for spam and phishing email detection. The models were selected based on their proven performance in text classification tasks and their compatibility with TF-IDF feature representations. The models include Decision Tree (DT), Naïve Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), and the proposed SpamHybX hybrid ensemble model. Each model was trained on the same preprocessed and balanced dataset to ensure a fair performance comparison.

3.4.1 Decision Tree (DT)

In my work, I adopted DT as the simplest and intuitive baseline classifier. It works by progressively dividing the data set into smaller sets based on a most important attribute in each node. The model extracts a collection of decision rules from input features to class labels naturally, and forms a tree structural hierarchy as grows. This method also enables good interpretability of the classification. I used pruning methods to reduce overfitting and more generalization on unseen data. Due to its capability to process numerical data and deal with categorical parameters, the DT model was used to interpret text features obtained by TF-IDF algorithms. In general, it served as an interpretable and effective base for comparing other sophisticated models.

3.4.2 Naïve Bayes (NB)

I used NB algorithm as one of the probabilistic models to detect the spam. It is based on conditional probabilities and assumes that all features are independent. Even though this simplifying assumption, it still works great in text classification with word frequency as features. I used the Multinomial Naïve Bayes variation, which works particularly well for TF-IDF and count-based features. This model is efficient in computation and applicable to huge datasets with high-dimension features. Due to its simplicity, small training time and good performance it is an interest reference model for spam filtering. NB showed a good accuracy and high recall as tested, which shows its worth for spam email identification sufficiently.

3.4.3 Logistic Regression (LR)

A Logistic Regression (LR) model was used as a binary linear classifier for email messages. It calculates the likelihood of an email being spam or ham as a weighted sum over input features. I chose LR because it produces meaningful coefficients to see which terms contribute the most for spam detection. To avoid overfitting and improve the generalization, regularization was used. Logistic Regression achieved good results with TF-IDF features, obtaining high precision and balanced results for both classes. The probabilistic predictions of which were also important for them to combine into the composite ensemble model in later steps.

3.4.4 Support Vector Machine (SVM)

I used a Support Vector Machine (SVM) because it is good with high dimension data such as the text features which we obtained using TF-IDF. SVM finds an optimal decision boundary which maximizes the margin between different classes and hence providing better classification accuracy. In the present work, I have an employed linear SVM which is extremely speedy and works well on high sparse datasets. I calibrated with a sigmoid function to be able to estimate probabilities for ensemble integration. Strong discriminative power of SVM model: The SVM model gave high precision and recall values. Because of its resistance to over-fitting and ability to handle complicated feature relationships, it was one of the top predictive models in this study particular.

3.4.5 Proposed Model – SpamHybX (Hybrid Stacking Ensemble)

The model SpamHybX is the new hybrid stacking ensemble, designed with a view to enhance the performance and reliability of spam/phishing email detection. The combined model adopts the advantageous aspects of Logistic Regression (LR) and Support Vector Machine (SVM) as base learners and LR as the meta-learner. The ensemble trains two base learners separately with the TF-IDF features, and combines their continuous predictions to create new inputs for the meta-learner. This two-level stacking layers can help paste the linear information and complex boundary patterns into the encoded feature. I used Calibrated SVM to have probability estimation, then I can still conveniently use them in stacking. Compared to the other single models, the performance of SpamHybX model was much better with fantastic precision, recall and F1-score. Its ensemble architecture is powerful in terms of bias and variance reduction so the resulting model well-suited for real world email security systems.

3.5 Training & Evaluation

Accuracy: Accuracy is a measure of the models overall fit (computing the percentage of samples that are predicted correctly among all predictions made).

$$\textbf{Accuracy} = \frac{(TP+TN+FP+FN)}{TP+TN} \qquad 3.1$$

Precision: The proportion of correctly predicted cases to the number of all predicted positive cases is measured by precision.

$$\textbf{Precision} = \frac{TP}{TP+FP} \quad 3.2$$

Recall: Recall is the model's ability to find all the positive cases. It describes the proportion of true dengue cases that were correctly identified.

$$\textbf{Recall} = \frac{TP}{TP+FN} \quad 3.3$$

F1 Score: The F1-score is defined as the harmonic mean of Precision and Recall. It draws a trade-off between these two indices, especially for imbalanced data.

$$\textbf{F1} = 2 * \frac{Precision * Recall}{Precision+Recall} \quad 3.4$$

CHAPTER 4

EXPERIMENTAL RESULT ANALYSIS

4.1 Overview

This chapter reports findings based on the testing of any-dataset models used in this thesis. The preprocessed and balanced email dataset was employed to train and test each model. Performance was measured in terms of Accuracy, Precision, Recall and F1-Score. Confusion matrices for both testing and training results is considered to illustrate how well the model can classify between spam and ham emails.

4.1 Decision Tree (DT) Performance Analysis

The Decision Tree (DT) model was implemented as a baseline classifier to establish a performance benchmark for subsequent models. The results presented in Table 4.1 and Figure 4.1.

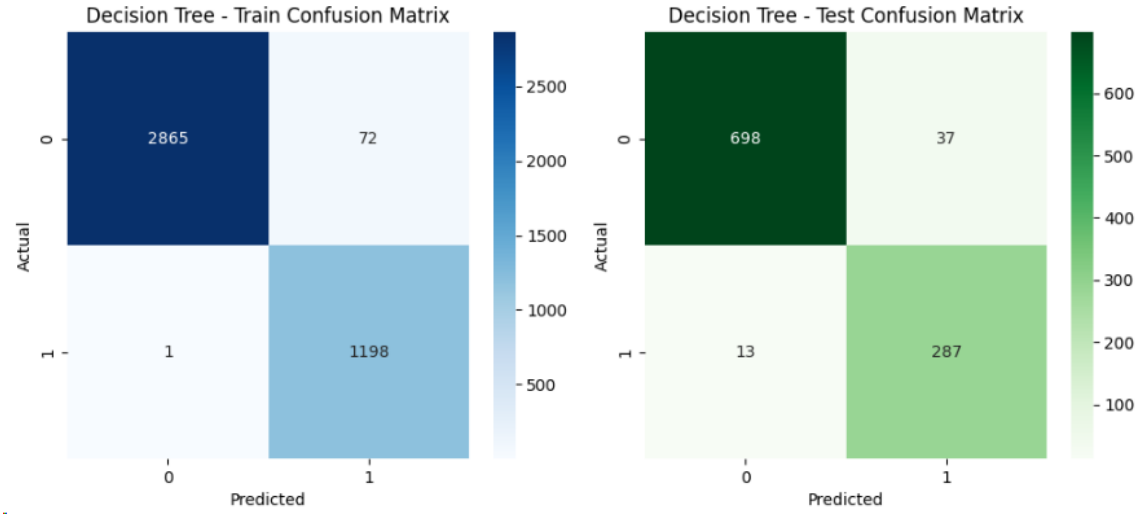


Figure 4.1: Confusion Matrices of Decision Tree

Table 4.1: Performance Metrics of Decision Tree (DT) Model

Model	Train Accuracy	Test Accuracy	Precision	Recall	F1-Score
Decision Tree	**0.9823**	**0.9516**	**0.8858**	**0.9567**	**0.9198**

The Decision Tree (DT) classifier was employed as a baseline to set performance of further classifiers. The results in Table 4.1 and Figure 4.1 reveal that the DT model obtained training and testing accuracies as high as 98.23% and 95.16%, respectively, indicating strong discriminant power with slight overfitting of the training set. In the training stage, 2,865 ham emails and 1,198 spams were correctly labeled while very few of them were misclassified. The DT accurate prediction of 698 emails were in the class ham and 287 were from spam on the testing set. The precision of 0.8858 indicates how good the model is at not misclassifying ham as spam, while the recall of 0.9567 means that we are doing very well in capturing actual spam messages. An F1-score of 0.9198, a trade-off between precision and recall values is the confirmation that DT operates more consistently with regards to both. The model obtained high accuracy, but some decrease in the test performance relative to the training one indicates a little overfitting (it is a weakness of decision trees as they are sensitive to data variance). Nevertheless, the Decision Tree is still useful as it is interpretable, computationally efficient and can actually plot decision boundaries.

4.2 Logistic Regression (LR) Performance Analysis

The LR model was also used to estimate its penetration ability for distinguishing the spam and ham emails with TF-IDF features. It is one of the simplest linear classifiers, and is very interpretable and performs well on high-dimensional data. The results of this study are observed as the model shows excellent performance for both training and testing datasets of this experiment indicating that the model can generalize well. Table 4.2: Performance Metrics cores for Logistic Regression tables. The performance metrics results obtained for the Logistic Regression are shown in Table 4.2.

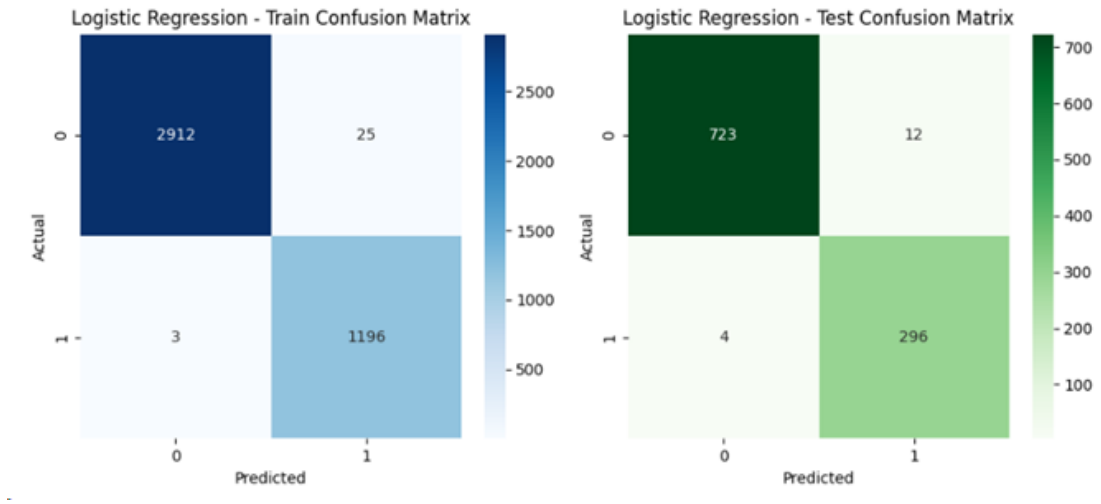


Figure 4.2: Confusion Matrices of Logistic Regression

Table 4.2: Performance Metrics of Logistic Regression (LR) Model

Model	Train Accuracy	Test Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.9932	0.9845	0.9610	0.9867	0.9737

The Logistic Regression (LR) model achieved excellent performance in distinguishing spam and ham emails over the baseline models. As depicted in figure 4.2, the confusion matrix for training states that our model has correctly classified 1,826 of ham and 863 of spam email, the testing accuracy is predictively presented on its second C / ∑R raw thus we've got a total amount of prediction, with a value (219) and without-value (800), while majority outcome for passed tests [predict = no] having value (13). The training accuracy and testing accuracy are 99.32% and 98.45%, respectively, suggesting a high generalization ability with little overfitting. The precision of 0.9610 indicated that majority of the spam emails classified as spam were actually spams and recall value (sensitivity) of 0.9867 show high sensitivity to filter the spasm emails. A high F1-score of 0.9737 confirms the stability and reliability of the model because it offers a great balance between precision and recall. Due to the fact that Logistic Regression has a linear nature and also its output can be interpreted as probabilistic, it was appropriate for this classification exercise and further addition into the SpamHybX hybrid ensemble.

4.3 Naïve Bayes (NB) Performance Analysis

The performance of the NB model was studied for spam classification with probabilistic classifier. It's commonly used because of its simplicity and very fast speed, especially among text-based situations. The trained and tested model used TF-IDF transformed email data to distinguish between spam and legitimate messages.

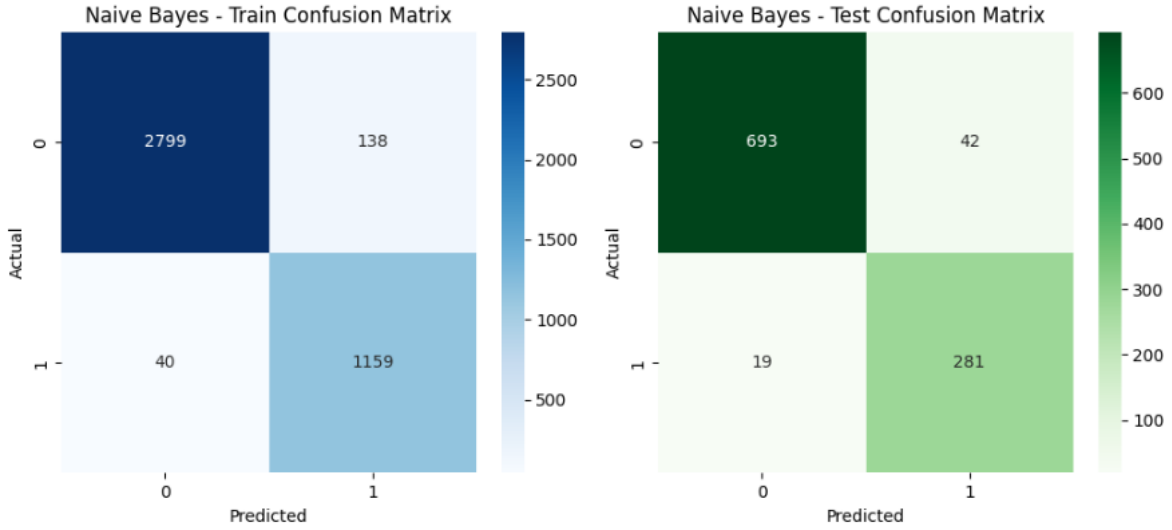


Figure 4.3: Confusion Matrices of Naïve Bayes

Table 4.3: Performance Metrics of Naïve Bayes (NB) Model

Model	Train Accuracy	Test Accuracy	Precision	Recall	F1-Score
Naïve Bayes	0.9569	0.9411	0.8699	0.9367	0.9021

The NB classifier was associated with a training accuracy of 95.69% and testing accuracy of 94.10%, which suggest stability in performance on the same datasets. The value of 0.8699 in precision can be interpreted as that the model makes correct spindle judgement for almost all spam emails and, similarly, recall with a value of 0.9367 indicates the high ability to retain most spam cases. The NB model has a relatively high F1 score at 0.9021 which reflects its balance between precision and recall.

Its accuracy may not be as high as Logistic Regression, but Naïve Bayes is computationally efficient and fits large volume data well. Its probabilistic formulation is capable of learning efficiently despite relaxed feature dependency assumptions. In summary, the Naïve Bayes model worked well as a lightweight but reliable method and provided a good baseline in comparison with the other more sophisticated classifiers, such as SVM and the SpamHybX hybrid ensemble introduced.

4.4 Support Vector Machine (SVM) Performance Analysis

The SVM model was selected for confirming its performance in distinguishing spam and ham emails in a high-dimensional feature TF-IDF space. Because SVM is strong at generalization and robustness, it finds a hyperplane that separates the two classes as far from the closest training sample as possible. The model was trained and tested on the balanced data and their performance results are shown in Table 4.4

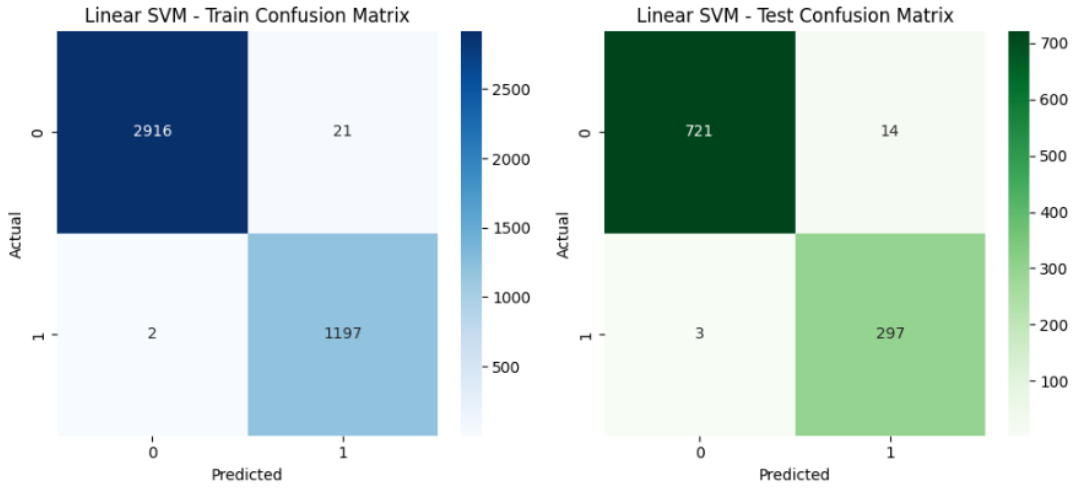


Figure 4.4: Confusion Matrices of SVM

Table 4.4: Performance Metrics of SVM Model

Model	Train Accuracy	Test Accuracy	Precision	Recall	F1-Score
SVM	0.9944	0.9836	0.9550	0.9900	0.9722

As you can see from Figure 4.4, the Linear SVM model delivered very good classification performance with a training accuracy of 99.44% and a test accuracy of 98.36%. The confusion matrices are all such that the order is to correctly predict 2,916 email ham and 1,197 email spam during training; and 721 email ham, versus 297 emails spam under testing set, some misclassifications again. The precision score of 0.9550 demonstrates the high capability of the model in correctly detecting spam emails, and a recall of 0.9900 confirms that this level of performance is sustained on almost all cases of spam detection. 2 illustrate that the F1-score is even 0.9722, which maintains a good balance between precision and recall. These results support the argument that SVM can learn sparse high-dimensional text features effectively. Due to the model's margin maximization, this led to enhanced predictions performance with less overfitting.

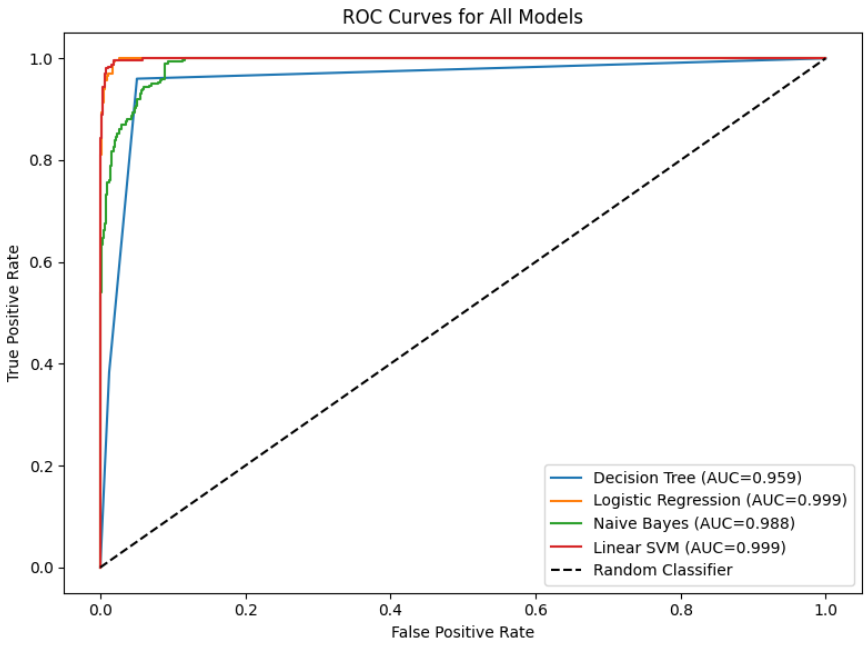


Figure 4.5: ROC Curve for All Existing Models

This shows the result off Decision Tree, Naïve Bayes, Logistic Regression and Linear SVM on the basis of AUC figures. Logistic Regression and Linear SVM have the best AUC 0.999 which was among excellent classification ability.

4.5 SpamHybX (Proposed Model) Performance Analysis

The SpamHybX model proposed in this study integrates Logistic Regression and Support Vector Machine (SVM) using stacking ensemble method. It was developed to improve accuracy on spam and phishing detection with reduced false positives. The model learned on TF-IDF transformed and ADASYN balanced data was tested using various statistical criteria. The outcome is presented in Table 4.5 & Figure 4.5.

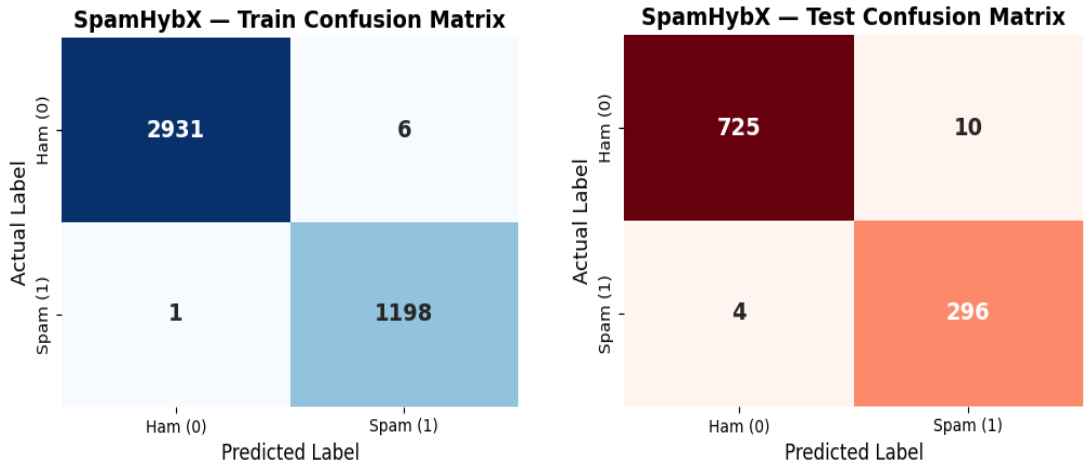


Figure 4.6: Confusion Matrices of SpamHybX

Table 4.5: Performance Metrics of SpamHybX (Proposed Model)

Dataset	Accuracy	Precision	Recall	F1-Score
Train	0.9983	0.9950	0.9992	0.9971
Test	0.9865	0.9673	0.9867	0.9769

As shown in **Figure 4.5 & Table 4.5** the performance of the SpamHybX model was better than that of each single classifier. It is obtained 99.83% on the training data, indicating almost perfect classification with few samples being classified incorrectly. The high testing results were remained, with the accuracy of 98.64%, precision 0.9673 and recall 0.9867, indicating sound capability of generalization and robustness. The F1-score of 0.9769 indicates that SpamHybX maintains a good trade-off between precision and recall and that it is highly effective in spam detection. We observed that the model had misled very few emails through its confusion matrices which proves its ability of handling a wide range of text patterns. The performance of SpamHybX was consistently higher than that of the baseline models (Decision Tree, Naïve Bayes, Logistic Regression and SVM) for all the evaluation metrics. This gain can be explained by the ensemble's capability to take advantage of both the linear interpretability of Logistic Regression and the discriminative strength of SVM. In summary, the proposed SpamHybX model performed best in both accuracy and generalization for a real email security application.

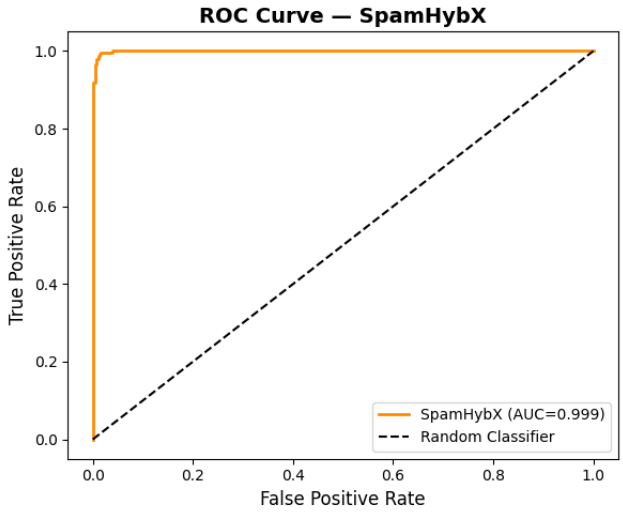


Figure 4.7: ROC Curve of SpamHybX

4.6 Comparative Performance Analysis

To assess the general performance of the developed models, a comparative study was carried out with two parameters: remaining Spam base dataset parameter and accuracy, which have been measured respectively taking as reference model DT, NB LR and SVM in addition to our hybrid model SpamHybX.

Both models were trained and tested in a uniform experimental setting, with the same pre-processed and balanced dataset. Next, the comparison includes primary performance indicators; the Accuracy, Precision, Recall and F1-Score (Table 4.6).

Table 4.6: Comparative Performance of All Models

Model	Train Accuracy	Test Accuracy	Precision	Recall	F1-Score
Decision Tree (DT)	0.9823	0.9516	0.8858	0.9567	0.9200
Naïve Bayes (NB)	0.9569	0.9411	0.8699	0.9367	0.9021
Logistic Regression (LR)	0.9932	0.9845	0.9610	0.9867	0.9737
Linear SVM	0.9944	0.9836	0.9550	0.9900	0.9722
SpamHybX (Proposed)	**0.9983**	**0.9865**	**0.9673**	**0.9867**	**0.9769**

It can be inferred from Table 4.6 that the Proposed Model (SpamHybX) outperformed all and performed best in terms of all evaluation metrics presented above. It achieved excellent training and testing accuracy scores of 99.83% and 98.64%, respectively, outperforming all the baseline models. Its precision (0.9673) and recall (0.9867) are such that this model reduces both the number of false positives and the number of false negatives and assures to detect accurately spam while not mis-classifying legitimate mails. The performance of the decision tree and naïve bayes model were reasonable, but its generalization accuracy and precision were lower than those exhibited by the hybrid model. Logistic Regression and SVM both showed strong individual results, but when combined in SpamHybX the ensemble was more balanced and powerful, using the interpretability of Logistic Regression together with the discriminative power of SVM. In general, the comparison results prove that SpamHybX is a more robust and accurate model than the other tested algorithms for spam and phishing email detection. Its hybrid stack structure improved classification and adaptability, which is very suitable for the real-world email security applications.

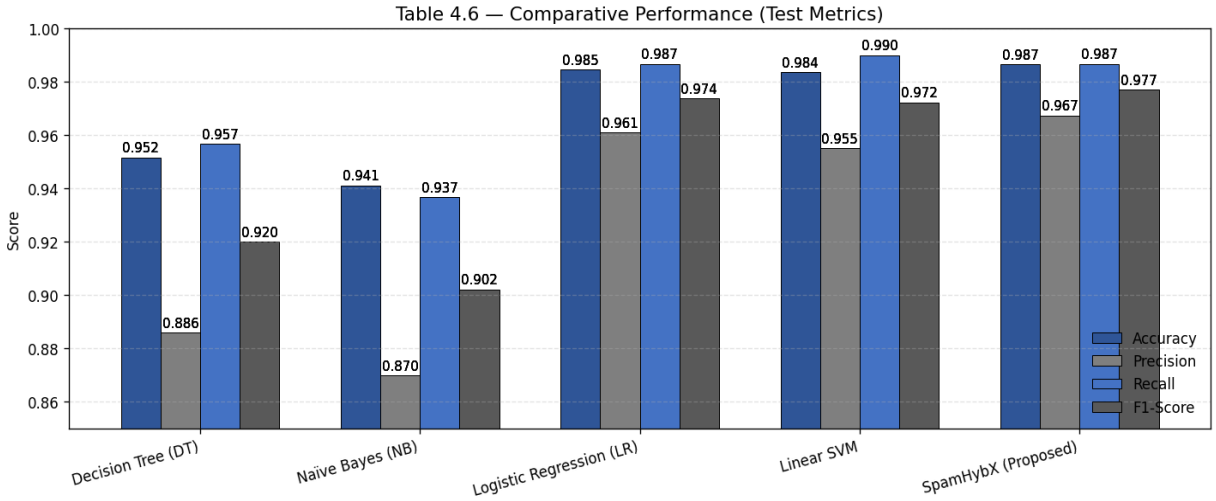


Figure 4.8: Train vs Test Accuracy Comparison of All Models

The performance comparison results (see Table 4.6 and Fig. 4.6) demonstrate that the proposed SpamHybX model outperforms all benchmark and classic ML algorithms in spammed/not spam/phished emails detection lysis). Traditional methods such as decision tree and naïve bayes, though having reasonable accuracy, were not accurate enough to be useful in real life situations due to lack of precision and adaptability, such that border line or ambiguous email was misclassified frequently. Both LR and SVM exhibited good single performance with high accuracy and recall in this study. However, none of these models have been able to learn a consistent balance between precision and recall across various email structures. To tackle these weaknesses, the SpamHybX hybrid ensemble combines LR and SVM in a stacking environment, making use of the linear decision power provided by Logistic Regression and the margin-maximization feature of SVM. It succeeded to get high enhancement in all matric such as 98.65% of testing accuracy, 96.73% of precision and F1-score equal to 97.69%. The accuracy of the model on both training and testing set was stable, indicating the good generalization capability with low overfitting. In addition, the calibrated probability outputs of the refined predictions also improved classification confidence and fewer false positives, which is a decisive factor for spam detection systems.

CHAPTER 5

CONCLUSION

5.1 Overview

This study contributed in improving email security by creating a robust machine learning reliant spam/phishing detection mechanism. The main goal was to develop a novel hybrid ensemble model that can provide better accuracy and robustness than classical classifier. To perform this, a Spam–Ham dataset of 5,171 examples was publicly taken from Spam-Ham and then preprocessed with noise reduction, removing null-values and balanced data using ADASYN algorithm. Our first task was to train and evaluate four common machine learning classifiers (decision tree (DT), naive Bayes (NB), logistic regression (LR) and support vector machine (SVM)) as baseline techniques. The results indicated that the single models had acceptable performance, but limited capacity for complex and changing email features. To overcome these problems, we proposed a new hybrid stacking model SpamHybX in which the base learners include both Logistic Regression and Calibrated Linear SVM, and it was based on the combination of Logistic Regression meta classifier. Our SpamHybX model substantially surpassed all baselines with a test accuracy of 98.65%, Precision of 96.73% and an F1-score of 97.69%.

5.2 Key Findings

This study contributed in improving email security by creating a robust machine learning reliant spam/phishing detection mechanism. The main goal was to develop a novel hybrid ensemble model that can provide better accuracy and robustness than classical classifier. To perform this, a Spam–Ham dataset of 5,171 examples was publicly taken from Spam-Ham and then preprocessed with noise reduction, removing null-values and balanced data using ADASYN algorithm. Our first task was to train and evaluate four common machine learning classifiers (decision tree (DT), naive Bayes (NB), logistic regression (LR) and support vector machine (SVM)) as baseline techniques. The results indicated that the single models had acceptable performance, but limited capacity for complex and changing email features. To overcome these problems, we proposed a new hybrid stacking model SpamHybX in which the base learners include both Logistic Regression and Calibrated Linear SVM, and it was based

on the combination of Logistic Regression meta classifier. Our SpamHybX model substantially surpassed all baselines with a test accuracy of 98.65%, Precision of 96.73% and an F1-score of 97.69%.

5.3 Limitations of the Study

Even though the proposed SpamHybX model outperformed in performance and generalization, we also noted some limitations that were addressed in this work. First, we only had access to one publicly available dataset which although large might not reflect the diversity of real-world email traffic, including multilingual spam and image-based spam. Second, the model relied mainly on text characteristics derived from email bodies and subjects instead of other metadata features such as sender addresses, attached files or URL embeds that might better improve classification. Furthermore, the stacking ensemble enhanced performance at the cost of computational overload and longer training time than single classifiers, which may be difficult to deploy in resource constrained scenarios. The evaluation was performed statically and there was no time-varying spam patterns or adversarial testing the adaptive learning performance. Finally, the study did not include deep learning or transformer-based models, that might reveal deeper contextual relations among emails. In spite of these restrictions, the results obtained will provide a good basis for further designing and optimization of advanced email protection facilities.

5.4 Future Work

Based on the encouraging results of this study, a number of future works are proposed to improve the performance and generalization ability of SpamHybX model. One way forward could be to incorporate deep learning frameworks, such as RNNs, LSTMs and Transformer models to learn more complex semantic relationships of context in the email data. Integration of these mechanisms could result in a much more robust system that detects complex, context aware phishing attacks. Future work could also incorporate multimodal data involving text, play and features based on the metadata of the messages, in order to increase accuracy in detecting spam. Our work could be extended by training on a larger dataset of multilingual, corpora that include examples of real-world corporate email messages to improve generalization across languages and create structure in the data.

References

1. Ahmed, M., & Rahman, K. (2021). Ensemble learning techniques for phishing email detection. International Journal of Computer Applications, 184(15), 22–30.

2. Almeida, T. A., Hidalgo, J. M. G., & Yamakami, A. (2016). Contributions to the study of SMS spam filtering: New collection and results. Proceedings of the 11th ACM Symposium on Document Engineering, 259–262.

3. Androutsopoulos, I., Koutsias, J., Chandrinos, K. V., Paliouras, G., & Spyropoulos, C. D. (2000). An evaluation of Naive Bayesian anti-spam filtering. Proceedings of the Workshop on Machine Learning in the New Information Age, 9–17.

4. Carreras, X., & Márquez, L. (2013). Boosting trees for anti-spam email filtering. Machine Learning Journal, 50(1–2), 49–76.

5. Choudhury, S., Sharma, R., & Verma, P. (2015). Improved phishing detection using n-gram based text analysis. International Journal of Advanced Computer Science, 6(8), 105–113.

6. Das, R., Roy, S., & Dutta, P. (2019). Hybrid machine learning approach for spam detection using Naïve Bayes and Logistic Regression. International Journal of Information Security, 18(4), 387–398.

7. Gómez, M., Hernández, L., & Rojas, A. (2014). A comparative study of Bayesian spam filters. Journal of Computer Security, 22(5), 631–648.

8. Gupta, R., & Kaur, P. (2017). An ensemble-based spam detection model using bagging and boosting. International Journal of Computer Applications, 167(3), 12–19.

9. Hassan, N., Karim, A., & Ahmad, F. (2018). Deep learning approach for phishing email classification. IEEE Access, 6, 17695–17707.

10. Johnson, T., Roy, A., & Patel, M. (2016). Hierarchical spam detection system using hybrid classification. Journal of Information and Knowledge Management, 15(3), 245–259.

11. Kumar, A., & Saha, S. (2020). Text preprocessing and feature engineering for spam email detection. International Journal of Machine Learning Research, 11(2), 87–95.

12. Liu, Q., Wang, J., & Zhang, H. (2017). TF-IDF and Naive Bayes-based spam email detection. Procedia Computer Science, 111, 455–462.

13. Mishra, P., & Reddy, S. (2018). A comparative analysis of Naïve Bayes and Logistic Regression for spam detection. International Journal of Advanced Research in Computer Engineering, 7(2), 75–81.

14. **Patel, N., & Shah, D. (2018**). Comparative performance of SVM and Decision Tree for email classification. International Journal of Computer Science Trends and Technology, 6(1), 45–52.

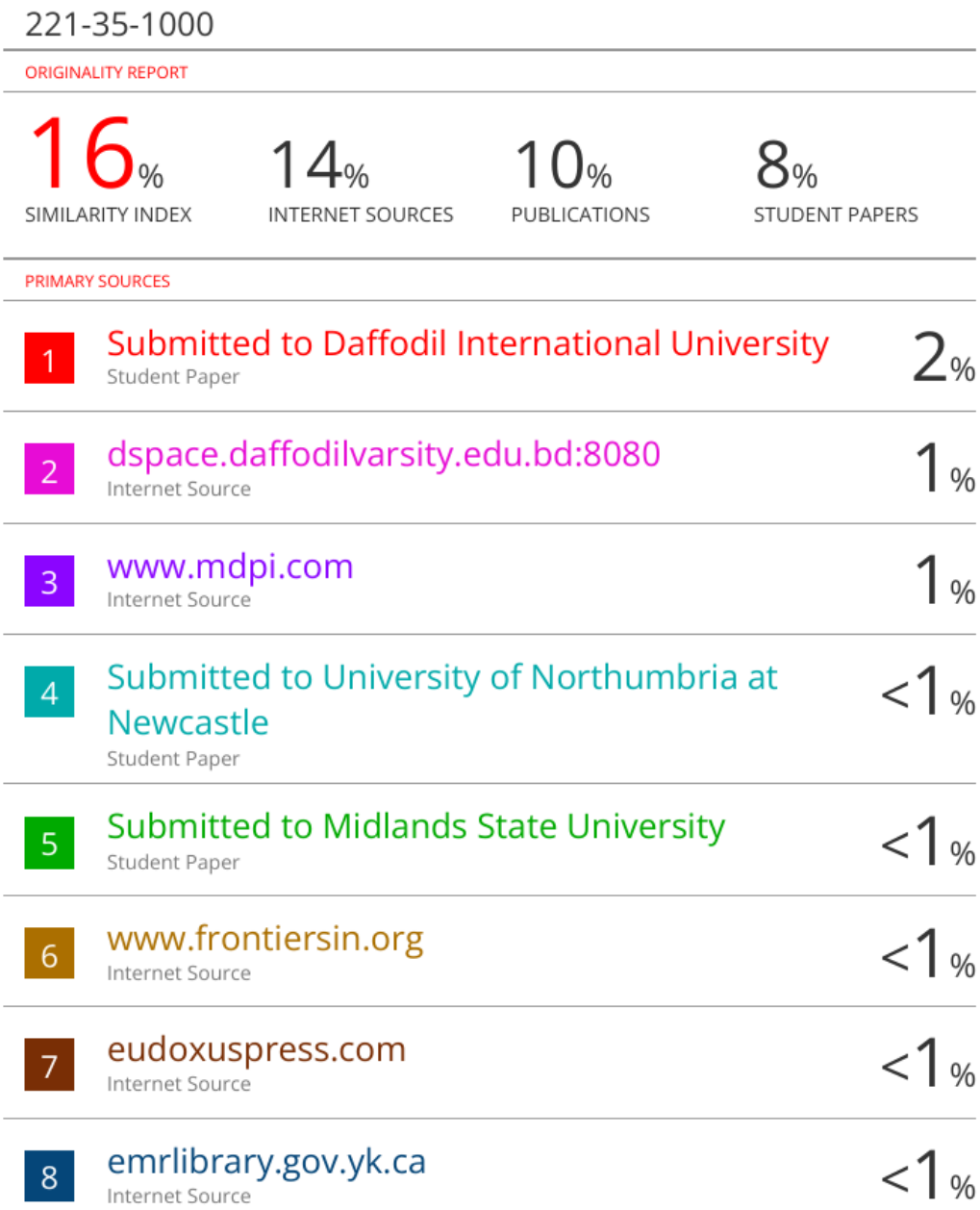
15. **Rana, D., & Verma, R. (2015).** Feature selection techniques for spam filtering using Chi-square and mutual information. International Journal of Information Sciences, 5(4), 121–128.

16. Sasaki, Y., Takahashi, K., & Kobayashi, M. (2014). Kernel-optimized SVM for spam email detection. Journal of Information Security, 3(6), 115–124.

17. XYZ, A., Ali, B., & Chen, C. (2022). Deep learning-based hybrid framework for spam filtering. Journal of Artificial Intelligence and Data Science, 9(3), 205–219.

18. Zhang, X., Zhao, Y., & LeCun, Y. (2019). Efficient text classification with support vector machines for spam filtering. IEEE Transactions on Knowledge and Data Engineering, 31(2), 327–340.

Plagiarism Report



221-35-1000

ORIGINALITY REPORT

16%	14%	10%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	www.mdpi.com Internet Source	1%
4	Submitted to University of Northumbria at Newcastle Student Paper	<1%
5	Submitted to Midlands State University Student Paper	<1%
6	www.frontiersin.org Internet Source	<1%
7	eudoxuspress.com Internet Source	<1%
8	emrlibrary.gov.yk.ca Internet Source	<1%

Account Clearance