PREDICTIVE MODELING AND RISK ASSESSMENT FOR COPD USING A MACHINE LEARNING APPROACH.

BY

S. M. Fardin Foyes Riad

ID: 202-15-3846

AND

Nahid Arman Fahim

ID: 202-15-3831

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

MS. NAZNIN SULTANA
Associate Professor
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised By

Dr. MD ZAHID HASAN
Associate Professor
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

July 2024

APPROVAL

This Project titled **"Predictive Modeling and Risk Assessment for COPD Using a Machine Learning Approach"**, submitted by **S. M. Fardin Foyes Riad and Nahid Arman Fahim** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **13-07-2024.**

BOARD OF EXAMINERS

Narayan Ranjan Chakraborty — **Board Chairman**
Associate Professor & Associate Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Md. Sadekur Rahman — **Internal Examiner 1**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Mr. Tanvirul Islam — **Internal Examiner 2**
Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Manowarul Islam — **External Examiner**
Associate Professor
Department of Computer Science and Engineering
Jagannath University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Ms. Naznin Sultana, Associate Professor, Department of Computer Science and Engineering, Daffodil International University**. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Ms. Naznin Sultana
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Dr. MD Zahid Hasan
Associate Professor
Department of CSE
Daffodil International University

Submitted by:

(S. M. Fardin Foyes Riad)
ID: 202-15-3846
Department of CSE
Daffodil International University

(Nahid Arman Fahim)
ID: 202-15-3831
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty for His divine blessing making it possible for us to complete the final year project successfully.

We are grateful and wish our profound indebtedness to **Ms**. **Naznin Sultana**, **Associate Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "Machine Learning" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to the **Dr. Sheak Rashed Haider Noori, Professor and Head of the Department of CSE**, for his kind help in finishing our project and also to other faculty members and the staff of the Department of CSE, Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

Chronic obstructive pulmonary disease (COPD) is a major worldwide health concern, impacting millions worldwide by its progressive and debilitating effects. This study aims to develop personalized machine learning models for accurate COPD risk assessment, leveraging a comprehensive dataset that includes clinical history, lifestyle factors, and environmental exposures. This project aims to improve the accuracy and efficiency of COPD diagnosis and risk assessment worldwide by utilizing several machine learning techniques, including Logistic Regression, Decision Tree, KNN, Naive Bayes, and AdaBoost. The study involves systematic data preprocessing, feature selection, and training of multiple algorithms to identify significant risk factors and their interactions across diverse populations. The primary goal is to create risk assessment tools that empower healthcare professionals and individuals to recognize and mitigate COPD risks early. Expected outcomes include improved clinical decision-making, early intervention strategies, and better patient outcomes in COPD management. Additionally, this research aims to make advanced diagnostic capabilities accessible to underserved communities, aligning with global health initiatives to reduce disparities and enhance healthcare equity. Through rigorous evaluation and comparative analysis, the study aims to develop reliable and precise prediction models that may be easily incorporated into real-life healthcare environments, thus revolutionizing COPD management and diagnosis. This research not only enhances the scientific comprehension of respiratory illness diagnoses but also has the potential to greatly influence public health outcomes, healthcare accessibility, and the general well-being of varied communities worldwide.

TABLE OF CONTENTS

Contents	Page No
Board of Examiners	ii
Declaration	iii
Acknowledgments	iv
Abstract	v
Table of Contents	vi-viii
List of Figures	ix-x
List of Tables	xi
CHAPTER 1: INTRODUCTION	**1-3**
1.1 Overview	1
1.2 Background and Present State	1-2
1.3 Problem Statement	2
1.4 Objectives	2
1.5 Scope and Limitations	2
1.6 Report Organization	3
1.7 Summary	3
CHAPTER 2: LITERATURE REVIEW	**4-8**
2.1 Overview	4
2.2 Related Works	4-6
2.3 Comparison between existing works	6-7
2.4 Open Issues	7-8
2.5 Summary	8

CHAPTER 3: METHODOLOGY/ REQUIREMENT ANALYSIS & DESIGN SPECIFICATION	**9-21**
3.1 Overview	9
3.2 Proposed Methodology/ System Design	9-12
3.2.1 Machine learning Models	12-15
3.2.2 Data Augmentation	15
3.3 Data Collection and Dataset Description	16-20
3.4 Hardware/ Software Requirement	20
3.5 Project Management and Financial Analysis	20-21
3.6 Summary	21
CHAPTER 4: IMPLEMENTATION	**22-29**
4.1 Overview	22
4.2 Train Model/ Prototype Design	22-27
4.3 System Testing/ Model Evaluation	27-29
4.4 Summary	29
CHAPTER 5: RESULT AND ANALYSIS	**30-43**
5.1 Overview	30
5.2 Experimental/ Simulation Result	30-38
5.3 Performance/ Comparative Analysis	39-43
5.4 Summary	43
CHAPTER 6: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	**44-46**
6.1 Impact on Life	44
6.2 Impact on Society & Environment	44-45
6.3 Ethical Aspects	45
6.4 Sustainability Plan	46
6.5 Summary	46

CHAPTER 7: CONCLUSION AND FUTURE WORK **47**

7.1 Conclusions 47

7.2 Further Suggested Works 47

7.3 Limitations/ Conflict of Interests 47

REFERENCES 48-50

APPENDIX A 51-56

LIST OF FIGURES

Figures	Page no
Figure 3.2.1: Proposed Methodology Flowchart	10
Figure 3.2.2.1: Dataset visualization	18
Figure 3.2.2.2: Dataset visualization after feature Engineering	19
Figure 4.2.1: Age Visualization	22
Figure 4.2.2: PackHistory Visualization	22
Figure 4.2.3: Age vs PackHistory	23
Figure 4.2.4: MWT1 Visualization	23
Figure 4.2.5: FEV1 Visualization	24
Figure 4.2.6: MWT1 vs FEV1	24
Figure 4.2.7: COPDSEVERITY DATA COUNT	25
Figure 4.2.8: Gender wise data Comparison	25
Figure 4.2.9: smoking category wise data	25
Figure 4.2.10: data Count AGEquartiles wise	26
Figure 4.2.11: Diabetics presence	26
Figure 4.3.1: Feature Importance	27
Figure 4.3.2: Correlation Matrix	28
Figure 5.2.1: Logistic Regression ROC Curve	30
Figure 5.2.2: Decision Tree ROC Curve	30
Figure 5.2.3: Random Forest ROC Curve	31
Figure 5.2.4: Ada Boost ROC Curve	31
Figure 5.2.5: KNN ROC Curve	32
Figure 5.2.6: Gradient Boosting ROC Curve	32
Figure 5.2.7: Naïve Bayes ROC Curve	33
Figure 5.2.8: LDA ROC Curve	33
Figure 5.2.9: LR Confusion Matrix	34
Figure 5.2.10: Decision Tree Confusion Matrix	34
Figure 5.2.11: Random Forest Confusion Matrix	35
Figure 5.2.12: Adaboost Confusion Matrix	35

Figure 5.2.13: KNN Confusion Matrix	36
Figure 5.2.14: LDA Confusion Matrix	36
Figure 5.2.15: Gradient Boosting Confusion Matrix	37
Figure 5.2.16: Naïve Bayes Confusion Matrix	37
Figure 5.3.1: LR Classification Report	39
Figure 5.3.2: Decision Tree Classification Report	39
Figure 5.3.3: Random Forest Classification Report	40
Figure 5.3.4: AdaBoost Classification Report	40
Figure 5.3.5: KNN Classification Report	40
Figure 5.3.6: Naïve Bayes Classification Report	41
Figure 5.3.7: LDA Classification Report	41
Figure 5.3.8: Gradient Boosting Classification Report	42

LIST OF TABLES

Tables	Page no
Table 2.3.1. Comparative analysis with previous work	7
Table 3.4.1: Estimated Cost for Machine Learning based Project	20
Table 5.3.1. Model Performance	38

LIST OF ACRONYMS

COPD	Chronic Obstructive Pulmonary Disease
ML	Machine Learning
KNN	K-Nearest Neighbour
LR	Logistic Regression
LDA	Linear Discriminant Analysis

CHAPTER 1

Introduction

1.1 Overview

The respiratory condition known as COPD or Chronic Obstructive Pulmonary Disease (COPD) is prevalent and significantly impacts both individuals affected by it and worldwide healthcare systems. [16]. Millions of people worldwide suffer from COPD, a serious lung disease that progresses over time. For patients to receive the best care and have a higher quality of life, early detection and risk assessment are crucial [17]. It refers to a variety of chronic respiratory disorders, including emphysema and chronic bronchitis, that are distinguished by persistent airflow limitation and respiratory symptoms such as coughing, wheezing, and dyspnea [18]. Globally, Chronic Obstructive Pulmonary Disease (COPD) is the primary cause of mortality and morbidity, causing estimated 251 million cases and 3.17 million deaths each year[19]. The number of people suffering from COPD is expected to rise further as a result of aging populations, continued vulnerability to risk elements such as tobacco smoking and indoor polluted air, and increased urbanization [20]. As a result, there is an urgent need for innovative approaches to improving early identification, risk assessment, and individualized COPD management methods.

This study aims to solve the various issues related to COPD using machine learning approaches. We intend to build prediction models that can detect individual susceptibility to COPD and assess associated risks using advanced algorithms and computational approaches such as K-nearest neighbors (KNN), logistic regression, Naive Bayes and decision trees. These models have the potential to improve our understanding of COPD progression, identify high-risk patients for specific treatment, and optimize resource allocation in healthcare systems [21].

1.2 Background and Present State

Chronic obstructive pulmonary disease or COPD is a significant concern for public health, especially in Bangladesh, where the prevalence rates are considerably higher than the world average [22]. Present techniques for diagnosing and treating COPD mostly depend on clinical evaluations and spirometry data, which, while successful, have restrictions in terms of early identification and individualized risk evaluation [23]. The progress of machine learning (ML) offers a chance to enhance the precision of forecasts and the capacity to diagnose Chronic Obstructive Pulmonary Disease or COPD in its first phase. This has the potential to decrease the rates of illness and death related to the condition [24].

1.3 Problem Statement

Although there have been notable improvements in medical research and treatment, Chronic Obstructive Pulmonary Disease or COPD continues to be One major contributor to mortality worldwide [16]. Conventional diagnostic techniques are frequently inadequate for detecting diseases at an early stage, particularly in areas with minimal resources such as rural Bangladesh. There is a pressing requirement for efficient predictive modeling systems that can utilize existing data to appropriately evaluate the risk of COPD. These techniques need to be customized to specifically target the distinct demographic and environmental characteristics that are widespread in Bangladesh [22].

1.4 Objectives

The main objectives of this research are:

- To build machine learning models capable of successfully forecasting the risk of COPD.
- To determine the primary risk variables and their interactions that are particular to the population of Bangladesh.
- To develop risk assessment tools that are easy to use for healthcare professionals and patients.
- To improve early identification and individualized care of COPD.

1.5 Scopes and Limitations

Scopes:

- Application of various machine learning approaches to predict COPD risk.
- Use of a complete dataset containing clinical history, lifestyle characteristics, and environmental exposures.
- Development of a risk assessment tool suited for COPD.

Limitations:

- Limited availability and quality of data from rural regions.
- Potential biases in data due to socio-economic and cultural variations.
- Constraints in computing resources for lengthy model training.

1.6 Report Organization

This report on COPD Prediction structured into seven chapters. Chapter 1 gives an introduction to the research, including the background, problem description, objectives, scope, and limitations. Chapter 2 analyzes relevant studies and gives a comparative overview of existing research. Chapter 3 explains the technique, covering system design and requirements. Chapter 4 outlines the implementation process, including model training and assessment. Chapter 5 gives the results and analysis. Chapter 6 examines the societal effect, ethical elements, and sustainability strategies. Finally, Chapter 7 ends the findings and proposes areas for further study.

1.7 Summary

In summary, this research aims to leverage machine learning models to improve the prediction and management of COPD. By developing robust prediction models and specific risk assessment tools, we desire to promote early identification and personalized treatment options, eventually bettering patient outcomes and lowering the impact of COPD on healthcare systems.

CHAPTER 2
Literature Review

2.1 Overview

This chapter analyzes existing work on predictive modeling and risk assessment for Chronic Obstructive Pulmonary Diseases using machine learning methods. It explores the strategies employed, data sources utilized, and the performance of different algorithms. The evaluation gives a thorough overview of the present state of research and indicates gaps and opportunity for additional exploration.

2.2 Related Works

Research on COPD prediction and risk assessment has deployed a range of machine learning techniques, each with various degrees of effectiveness. Studies have utilized Support Vector Machines (SVM), Decision Trees, Logistic Regression, and more. Data sources range from clinical records and spirometry data to wearable device data and expert opinions. Much work has been done on COPD before, and we have read through some citable research that is related to our work.

Jorge L.M. Amaral et al. [1] developed a system for clinical decision assistance using forced oscillation (FO) measurements. To look for the best possible classifier, they compared five different classification algorithms: LBNC, KNN, DT, ANN, and SVM. SVM is the best algorithm, with a precision of 96%. Their gap was the use of this methodology for detecting early smoking-induced respiratory alterations. Peng et al. [2] Using objective clinical data, they built a C5.0 decision tree classifier to predict the outcomes of AECOPD hospitalized patients. They used data from 410 hospitalized AECOPD patients with 28 features. The suggested decision tree classifier technique achieved an overall accuracy of 80.3%. Their future studies need validation in bigger, more diverse populations.

Alam et al. [3] The study reported a high prevalence of COPD in Bangladesh, highlighting a significant public health concern. Using a handheld spirometer, They evaluated the respiratory function of 3744 randomly chosen people aged 40 or older from both rural and urban areas. According to GOLD criteria, the prevalence of COPD was 13.5%, whereas LLN criteria showed 10.3%. Smoking, biomass fuel use, and older age were identified as major risk factors. It was more common in rural areas and among males than females. Further study is required to solve limitations and provide better preventative and management measures. Joshe et al. [4] They used symptom data to build a machine learning model that predicted the prevalence of Chronic Obstructive Pulmonary Disease in Bangladesh. The research comprised 101 COPD patients from the medical facility. The highest performing models were Decision Tree and Logistic Regression, with accuracy rates of 95.24%. Further research is needed for broader validation and clinical implementation strategies.

Biswas et al. [5] The study identified several risk factors for COPD in rural Bangladeshi women, including biomass fuel use, smoking, and certain cooking habits. Their study was conducted among 250 women in different villages of Chittagong, Bangladesh. In rural women over 40, the overall COPD prevalence was 20.4%. A systematic review was conducted by Ipsita Sutradhar et al. [6] To determine the prevalence and risk factors of COPD. In low socioeconomic groups, rural areas, males, and biomass fuel users, the prevalence rate was higher. It has been determined that the use of tobacco-related substances, contact with biomass fuel, aging, and a history of asthma are the main risk factors for COPD. A supervised prediction system was trained by Swaminathan et al. [7] using the opinions of physicians in a clinically and statistically thorough set of patient cases. A panel of doctors who each examine similar patients in a representative Accuracy of the model is evaluated using patient validation set.. Israa Mohamed et al. [8] aimed to predict patient readmission for Chronic Obstructive Pulmonary Disease using machine learning techniques.. The main indicators for assessing the prediction power of the models in each time frame were assumed to be AUC and ACC. They compare the performance of various machine learning algorithms for predicting 30-days readmission in COPD patients. Using databases on patient complaints and lung function measurement.

Badnjevic et al. [9] developed a sophisticated diagnostic tool able to differentiate people with COPD, asthma, or normal lung capacity. Over a period of two years they collected 3657 patients data and from these they used to verify 1650 patients data. Regarding other machine learning techniques, gradient boosting proved to have the greatest accuracy. Sanchez-Morillo et al. [10] presented a systematic review with the goal of assessing the current algorithms and concentrating on how well they work to identify exacerbations and direct course of treatment decisions for COPD sufferers utilizing home telemonitoring treatments. Alamgir et al. [11] aimed to assess the staging, spirometric, and clinical characteristics of asthma and COPD in smokers and nonsmokers. The study included 232 smokers and 100 nonsmokers.

Spathis et al. [12] The writers evaluated the potential of machine learning techniques for diagnosing asthma and COPD, showing promising results with specific algorithms and data features. The dataset consists of 132 patient records with 22 different attributes. Several machine learning algorithms were tested, Naive Bayes, logistic regression, neural networks, SVMs, K-Nearest Neighbor, decision trees, and random forest are some of the algorithms used. Among these algorithms, the Random Forest classifier achieved the highest precision of 97.7%. There is a limitation of working with a small sample size and an age limitation, for which future work is needed. Goto et al. [13] The study intends to enhance the prediction of ED disposition for patients who are suffering outbreaks of asthma or COPD exacerbations using machine learning techniques. The data for the years 2007-2015 was utilized from the National Hospital and Ambulatory Medical Care Survey (NHAMCS). A comparison was conducted amongst four machine learning algorithms: random forest, lasso regression, gradient boosting and Deep Neural Network. Boosting and Random Forest algorithm performed best with a C-statistic of 0.80 and 0.83. For wider validation, clinical use, and resolving potential constraints, more study is required.

The goal of Chia-Tung Wu et al. [14] aimed to design and analyze a machine learning system that would utilize data from wearable devices to anticipate patients likelihood of suffering an acute exacerbation of chronic obstructive pulmonary disease (AECOPD) within seven days. They tested the prediction accuracy of deep neural network models and machine learning models such as random forest, decision trees, k-nearest neighbor, linear discriminant analysis, adaptive boosting, and others using a set of input characteristics. Among these models, the DNN model performed the best, with six metrics exceeding 93%. Zeng et al. [15] Author developed a model to predict severe exacerbations of COPD. They looked at 278 potential characteristics and 43,576 data points from University of Washington Medicine facilities between 2011 and 2019 in order to create the model. Gradient boosting, random forest, neural networks, and logistic regression were the four algorithms they tested to forecast severe COPD exacerbations. The best method has gradient boosting with AUC of 0.866.

2.3 Comparison between existing works

Different studies employ various prediction algorithms (SVM, C5.0, Decision Tree, Logistic Regression, etc.) with varying accuracies (77.70% - 97.70%). Data sources include hospital records, spirometry data, wearable device data, and physician views. Some studies concentrate on specific demographics (such as rural Bangladesh) or applications (such as readmission prediction).

Table 2.3.1 Comparative analysis with previous work

SL No	Author Name	Used Algorithm	Best Accuracy with Algorithm
1.	Jorge L.M. Amaral et al. [1]	Linear Bayes Normal Classifier, KNN, Decision Trees, ANN and SVM	SVM = 96%
2.	Peng et al. [2]	Decision Tree Classifier, Classification and Regression Tree (CART) and the Iterative Dichotomiser 3 (ID3) model	Decision Tree Classifier = 80.3%
3.	Joshe et al. [4]	Logistic Regression, Decision Tree, KNN, Naive Bayes, Adaboost	Decision Tree and Logistic Regression=95.24%
4.	Ipsita Sutradhar et al. [6]	Binary Logistic Regression, Multivariate Logistic Regression, Conditional Logistic Regression	______
5.	Swaminathan et al. [7]	Gradient Boosting, Logistic Regression	97%
6.	Israa Mohamed et al. [8]	Logistic Regression, Random Forest, SVM, Gradient Boosting, Neural Network	_________
7.	Badnjevic et al. [9]	Random Forests, Gradient Boosting, Logistic Regression, ANN	Gradient Boosting=98.33%
8.	Sanchez-Morillo et al. [10]	Decision Trees, Logistic Regression	_________
9.	Spathis et al. [12]	Naive Bayes, Logistic Regression, Neural Networks, SVMs, K-Nearest Neighbour, Decision Trees, and Random Forest	Random Forest = 97.7% precision
10.	Goto et al. [13]	Random Forest, Lasso Regression, Boosting (Gradient Boosted Decision Tree), and Deep Neural Network	Random Forest= 0.83
11.	Chia-Tung Wu et al. [14]	Random Forest, Decision Trees, KNN, Linear Discriminant Analysis, Adaptive Boosting, DNN	DNN=93%
12.	Zeng et al. [15]	Logistic regression, random forest, neural network, and gradient boosting	gradient boosting =86%

2.4 Open issues

Despite significant progress, several challenges remain in COPD predictive modeling:

- **Data Quality and Availability:** High-quality, detailed datasets are crucial for training robust models. However, data from rural and resource-limited settings are often incomplete or inconsistent.

- **Model Generalization:** Ensuring that models generalize successfully to diverse populations and environments is important. Models trained on certain datasets may not perform as well when applied to new data.

- **Integration of Diverse Data Sources:** Combining data from clinical records, spirometry, wearable devices, and expert opinions can enhance model accuracy but also introduces complexity in data integration and preprocessing.

- **Interpretability and Transparency:** Clinical acceptability and confidence depend on machine learning models being interpretable and transparent as they get more complicated.

2.5 Summary

The literature review highlights the diverse methodologies employed in COPD prediction and the varying degrees of success achieved by different machine learning algorithms. While advanced techniques like SVMs and CNNs show promise in achieving high accuracy, simpler models like logistic regression provide valuable insights into individual risk factors. The review identifies several open issues, including data quality, model generalization, and the need for interpretable models. These insights lay the foundation for the proposed methodology in the subsequent chapters, aimed at addressing these challenges and improving COPD prediction and risk assessment.

CHAPTER 3

Methodology/ Requirement Analysis & Design Specification

3.1 Overview

The methodology chapter outlines the systematic process used in this research to develop predictive models for COPD risk assessment. The processes followed from data collecting and preprocessing to model training, evaluation, and deployment are discussed in this chapter. We applied a number of machine learning methods to determine the top prediction models. Each phase was carefully developed to ensure the resilience and accuracy of the models. The chapter is divided into several subsections including system design, hardware and software requirements, project management, and financial analysis. The approaches applied try to solve the particular difficulties presented by the data and the particular background of COPD in Bangladesh.

3.2 Proposed Methodology/ System Design

Starting with data collecting and preparation, the suggested approach consists in many main phases. The dataset was obtained from Kaggle and includes various features relevant to COPD, such as clinical history, lifestyle factors, and environmental exposures. Data preprocessing steps included handling missing values, normalization, and feature selection. We then tested many machine learning models like AdaBoost, Random Forest, K-Nearest Neighbors (KNN), Logistic Regression, Decision Trees, Naive Bayes, Linear Discriminant Analysis, and Gradient Boosting. Each algorithm was trained on the preprocessed dataset, and hyperparameter tuning was performed to optimize performance.

The system design includes an architectural framework where the data pipeline, model training, and evaluation modules are integrated. Leveraging Pandas, Scikit-learn, and Matplotlib, we coded and visualized using Python and Google Collab notebooks. Accuracy, precision, recall, and F1-score comprised the models' assessment measures. The best-performing models were selected for deployment in a user-friendly risk assessment tool designed for healthcare professionals.

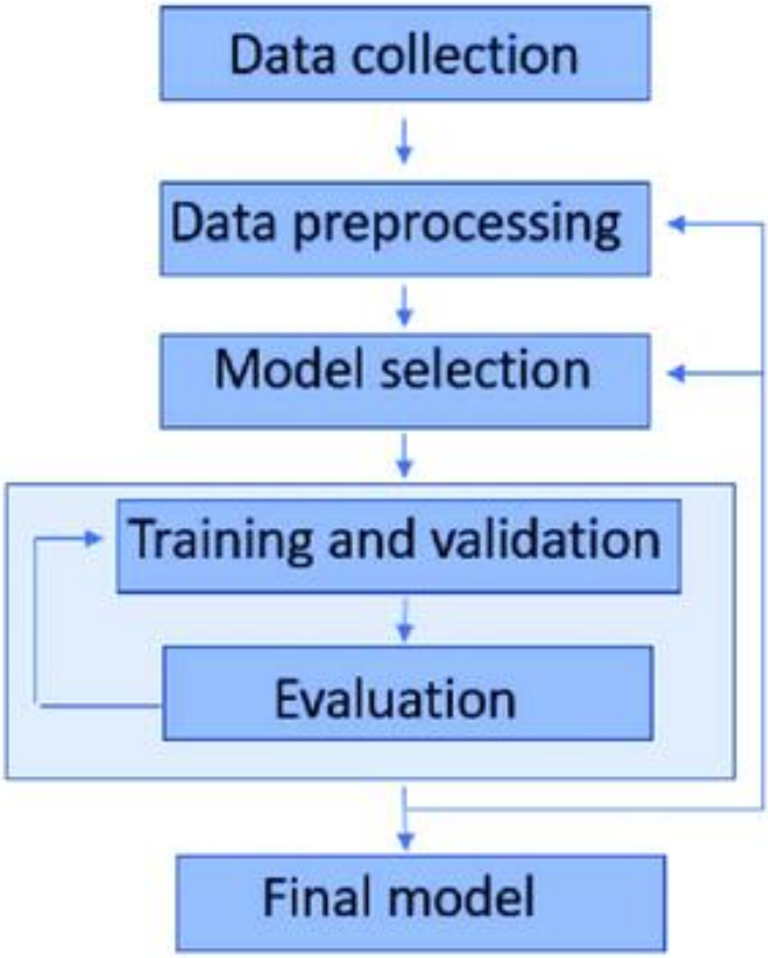

Figure 3.2.1: Proposed Methodology Flowchart

Data Collection:

Data collecting, from many sources, starts the process with unprocessed raw data. Databases, API, web scraping, sensor, survey, or other data sources can all provide this information. Since they directly affect the capacity of the model to learn and provide accurate forecasts, the quality and volume of the gathered data are very important. Sometimes this stage could call for data integration from several sources to provide an all-encompassing dataset.

Data Preprocessing:

Data must be preprocessed to guarantee it is fit for analysis after it has been gathered. This phase entails various chores:

- **Cleaning**: Removing null values, correcting errors, and detection of outliers.
- **Transformation**: Organizing information into a format machine learning systems may find simple interpretation. This might call for standardizing or

normalizing numerical features, encoding categorical variables, and feature engineering creates of new features.

- **Splitting**: Creating training, validation, and test sets out of the data. The model is trained using the training set; the validation set tunes hyperparameters and evaluates the model throughout development; the test set evaluates the performance of the finished model.

Model Selection:

The choice of the suitable machine learning algorithms comes next using the preprocessed data. The kind of problem—classification, regression, clustering—as well as the data's features determine the method of choosing among algorithms. Among the common techniques are linear regression, logistic regression, decision trees, random forests, support vector machines, k-nearest neighbors, and several forms of neural networks. In practice, multiple algorithms may be tested to determine which performs best.

Training and Validation:

The selected models are trained on the training dataset. During training, the model learns patterns and relationships within the data by altering its parameters to minimize a loss function. Validation entails testing the model's performance on the validation set to verify it generalizes effectively to unknown data and does not overfit. Techniques like as cross-validation can be used to acquire a more trustworthy assessment of the model's performance. This stage generally involves hyperparameter tuning, where the hyperparameters of the model are tuned to increase performance.

Evaluation:

After training and validation, the model's performance is tested using the test dataset or other validation procedures. Evaluation metrics differ based on the type of problem. Metrics including accuracy, precision, recall, F1 score, and ROC-AUC are employed. The evaluation step is crucial for determining whether the model meets the performance requirements and is ready for deployment.

Final Model:

Once a model has been trained, verified, and assessed effectively, it becomes the final model. This model is then deployed to generate predictions on new, unknown data. Deployment might entail integrating the model into a production environment, setting up APIs, or establishing user interfaces for end-users to interact with the model. Continuous monitoring and maintenance are necessary to guarantee the model stays accurate and effective over time. This may entail retraining the model with fresh data, upgrading the model architecture, or fine-tuning hyperparameters as appropriate.

3.2.1 Machine Learning Models

A machine learning model is a computational system specifically created to identify patterns and generate predictions or judgments by analyzing data. It operates by training on a dataset, learning from its features and outputs to develop a mathematical representation of the underlying relationships. This model can then process new, unseen data to generate accurate predictions or classifications. Machine learning models exhibit a range of complexity, spanning from basic linear regressions to sophisticated neural networks. Each model is well-suited for certain tasks, including regression, classification, clustering, and reinforcement learning. Essential components of a machine learning model are data preparation, model selection, training, validation, and testing. Properly implemented, these models can significantly enhance decision-making processes across diverse fields like healthcare, finance, and technology by providing data-driven insights and automating complex tasks.

Logistic regression: Logistic regression is a machine learning technique employed to solve binary classification issues, predicting one of two potential outcomes by modeling the likelihood that an input belongs to a certain class using the logistic (sigmoid) function. It combines input features linearly and transforms them into probabilities, which are thresholded to determine class labels. Its simplicity and interpretability, efficiency with large datasets, and reduced overfitting (especially with L1 or L2 regularization) make it popular. Contrary to its title, it is employed for categorization, not regression, and is applied in fields like medicine, finance, and marketing for tasks such as predicting disease presence, credit scoring, and customer segmentation.

Decision Tree: A decision tree is a widely used machine learning technique for solving classification and regression problems. It divides a dataset into smaller groups based on the values of input features, creating a tree-like model that represents a series of decisions. Every internal node evaluates a certain feature, each branch shows the test conclusion, Each leaf node exhibits a class label or a continuous value. Decision trees are capable of handling both numerical and categorical data, and they are straightforward to analyze and represent. Additionally, decision trees are non-parametric. However, they can overfit, particularly with deep trees, although this may be handled by pruning, creating a maximum depth, or applying ensemble methods like Random Forests and Gradient Boosting. They are employed in different applications, such as medical diagnosis and credit risk assessment, due to their capacity to represent complicated data linkages and rapid, efficient model creation.

Random Forest: Random Forest is a collective learning strategy that boosts prediction accuracy and resilience for classification and regression problems by merging numerous decision trees. It constructs numerous decision trees from random subsets of the training data using bootstrap aggregating (bagging) and produces the mode (classification) or mean (regression) of the various trees, avoiding overfitting. Random Forest also introduces randomness in feature selection, ensuring diverse trees that collectively improve accuracy and stability. It handles large datasets, high dimensionality, and missing data well. Its high accuracy and ability to capture complex patterns make it popular in finance, healthcare, and marketing.

AdaBoost: AdaBoost, or Adaptive Boosting, is an ensemble learning algorithm that enhances the performance of weak classifiers by combining them into a strong classifier. It sequentially trains multiple weak learners, typically decision stumps, and emphasizes instances misclassified by previous learners by adjusting their weights. This adaptive weighting focuses on harder-to-classify examples, improving accuracy. The final model aggregates predictions of all weak learners, weighted by their performance, to make the final decision. AdaBoost reduces both bias and variance, making it efficient for binary classification and adaptable to multi-class issues. It achieves high accuracy and is robust to overfitting, particularly with weak learners that underfit. AdaBoost is utilized in different disciplines such as text classification, face detection, and bioinformatics.

K-Nearest Neighbors (KNN): The k-Nearest Neighbors (KNN) approach is a quick and effective solution for both regression and classification applications. It performs on the concept that comparable cases are near one other. For classification, While for regression KNN forecasts the value by averaging the values of its k nearest neighbors, it gives a class label to a new data point depending on the majority class among its k nearest neighbors. KNN is non-parametric and lazy, meaning it doesn't assume any data distribution and only does calculation when queried. This simplicity makes KNN straightforward to construct and comprehend, but it can be computationally demanding for big datasets. Despite this, KNN is useful in pattern recognition, recommendation systems, and anomaly detection because to its versatility and ability to handle varied data sources.

Linear Discriminant Analysis (LDA): A statistical method for categorization and dimensionality reduction is called linear discriminant analysis (LDA). It identifies the linear combination of characteristics that best distinguishes various classes by increasing the separation between classes while decreasing within-class variation. This is done by computing the mean and variance of each class and selecting the axes that optimize the ratio of between-class variation to within-class variance. LDA is most effective when class distributions are well-separated and normally distributed. It reduces computational cost, eliminates noise and redundancy, and improves classification performance. Common applications include face recognition, customer segmentation in marketing, and medical diagnosis, valued for its interpretability and efficiency in enhancing model accuracy.

Gradient Boosting: Gradient Boosting is an collective learning approach used for classification and regression applications that enhances prediction accuracy by repeatedly merging many weak learners, often Decision Trees. Each subsequent tree corrects the faults of the prior ones by focusing on residuals, and the final model is a weighted sum of all trees. This approach employs gradient descent to minimize the loss function, lowering mistakes progressively. Known for its great accuracy and capacity to handle many data sources, Gradient Boosting is frequently used in search engine ranking, financial modeling, and disease prediction.

Accuracy:

Accuracy is a single metric used to measure the performance of a classification model. Accuracy refers to the proportion of accurate predictions produced by the model.

Accuracy equation:

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

Precision:

Precision relates to the probability of accurately detecting a positive label and the number of correctly detected positive instances. Precision quantifies the fraction of accurately anticipated positive cases. Accuracy is helpful when false positives are more significant than false negatives.

Precision equation:

$$Precision = \frac{TP}{TP + FP}$$

Recall:

Recall or Sensitivity assesses how often positive expected occurrences were labeled accurately. Recall is an essential metric when FN prevails over FP. F1-score is an additional statistic for classification accuracy that takes into consideration both recall and precision. Since accuracy and recall are harmonic means, the F1 score offers a comprehensive knowledge of these two metrics. It achieves its optimum when Precision and Recall are equal.

Recall equation:

$$Recall = \frac{TP}{TP + FN}$$

F-1 Score:

F1 score is a recall-precision metric of classification accuracy. As F1-score is the harmonic mean of Precision and Recall, it gives a whole point of view.

F-1 Score equation:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

3.2.2 Data Augmentation

Data augmentation is a method used in machine learning to boost the diversity and quantity of data used for learning without actually obtaining additional data. This is performed by applying a number of adjustments to the current dataset, such as rotations, translations, flips, scaling, and color changes in images, or synonym replacement, insertion, and deletion in text data. These transformations create new, modified versions of the original data, helping to increase the model's capacity to generalize and reduce overfitting.

By artificially increasing the dataset, data augmentation helps models to learn more robust characteristics and become more resilient to fluctuations and distortions in real-world data. It is particularly beneficial in fields like computer vision and natural language processing, where collecting and labeling large amounts of data can be expensive and time-consuming. Data augmentation not only enhances model performance but also makes the training process more efficient by leveraging existing data more effectively, leading to improved accuracy and dependability in many applications such as image recognition, audio processing, and text categorization.

3.3 Data Collection and Dataset Description

3.3.1 Data Collection

We have obtained the dataset from Kaggle. The data in this dataset were collected from diverse sources to guarantee a complete and robust dataset for predicting and assessing COPD risk. We have increased the data (Figure 3.2.2.2) with data augmentation techniques. The sources include:

1. **Hospital Records:**
   - **Source:** Local hospitals and healthcare facilities.
   - **Data:** Patient demographics, medical history, clinical test results, and COPD diagnosis.
   - **Collection Method:** Extracted from electronic health records (EHRs) with appropriate consent and anonymization.

2. **Spirometry Data:**
   - **Source:** Spirometry tests conducted at hospitals and clinics.
   - **Data:** Results of the pulmonary function test comprising measurements of forced vital capacity (FVC) and forced expiratory volume in one second (FEV1).
   - **Collection Method:** Directly obtained from spirometry devices, ensuring standardization and accuracy.

3. **Wearable Device Data:**
   - **Source:** Wearable health monitoring devices.
   - **Data:** Continuous tracking of respiratory rate, heart rate, physical activity levels, and environmental exposure.
   - **Collection Method:** Collected through device APIs and synchronized with patient records.

4. **Physician Opinions:**
   - **Source:** Surveys and interviews with pulmonologists and primary care physicians.
   - **Data:** Expert assessments of patient conditions, risk factors, and treatment plans.
   - **Collection Method:** Structured questionnaires and recorded interviews.

3.3.2 Dataset Description

The compiled dataset consists of multiple features that are critical for the prediction and risk assessment of COPD. Below is a detailed description of the dataset:

1. **Patient Demographics:**
   - **Age:** Numeric value representing the age of the patient.
   - **Gender:** Categorical variable indicating the patient's gender.
   - **Smoking History:** Categorical variable representing smoking status

2. **Clinical Test Results:**
   - **FVC (Forced Vital Capacity):** Numeric figure representing the amount of air a person may forcibly expel after taking the deepest inhale possible.
   - **FEV1 (Forced Expiratory Volume in one second):** Numeric number showing the volume of air breathed in the first second of a forceful expiration.
   - **FEV1/FVC Ratio:** Numeric value representing the ratio of FEV1 to FVC, an important diagnostic measure for COPD.

3. **Wearable Device Data:**
   - **Respiratory Rate:** Numeric value indicating the number of breaths per minute.
   - **Heart Rate:** Numeric value indicating the number of heartbeats per minute.
   - **Physical Activity Levels:** Numeric values representing activity metrics such as steps taken, calories burned, and active minutes.
   - **Environmental Exposure:** Numeric values indicating exposure to pollutants and allergens.

4. **Expert Assessments:**
   - **COPD Severity:** Categorical variable indicating the severity of COPD (e.g., mild, moderate, severe, very sevre) based on physician assessment.

Outcome Variable:

- **COPD Diagnosis:** numerical variable indicating the MILD (1), MODERATE (2), SEVRE (3), VERY SEVRE (4) of COPD.

Here is an explanation of all features of the dataset:

1. **ID**: Patient identifier.
2. **AGE**: Age of the patient.
3. **Pack History**: History of cigarette smoking quantified in pack-years.

4. **COPDSEVERITY**: Severity of COPD categorized (e.g., SEVERE, MODERATE, VERY SEVERE).
5. **MWT1**: Distance walked during the first 6-minute walk test in meters.
6. **MWT2**: Distance walked during the second 6-minute walk test in meters.
7. **MWT1Best**: Best distance walked between the first and second walk tests in meters.
8. **FEV1**: Forced Expiratory Volume in one second, measured in liters.
9. **FEV1PRED**: Percentage of the predicted FEV1 value based on age, gender, height, and ethnicity.
10. **FVC**: Forced Vital Capacity, the total volume of air exhaled throughout the FEV test.
11. **FVCPRED**: Percentage of the predicted FVC value.
12. **CAT**: COPD Assessment Test score, indicating the impact of COPD on a patient's health status.
13. **HAD**: Hospital Anxiety and Depression Scale score.
14. **SGRQ**: St. George's Respiratory Questionnaire score, showing the influence of respiratory illness on overall health, everyday living, and subjective well-being.

15. **AGEquartiles** in the dataset categorizes patients into one of these four age groups. For example:
    - represent the youngest 25% of the patients.
    - represent the second youngest 25% of the patients.
    - represent the second oldest 25% of the patients.
    - represent the oldest 25% of the patients.

16. **copd**: A numerical value representing the presence and severity of COPD.
17. **gender**: Gender of the patient (typically 0 for female, 1 for male).
18. **smoking**: Smoking status (e.g., 0 for non-smoker, 1 for former smoker, 2 for current smoker).
19. **Diabetes**: Presence of diabetes (0 for no, 1 for yes).
20. **muscular**: Presence of muscular diseases (0 for no, 1 for yes).
21. **hypertension**: Presence of hypertension (0 for no, 1 for yes).
22. **AtrialFib**: Presence of atrial fibrillation (0 for no, 1 for yes).
23. **IHD**: Presence of ischemic heart disease (0 for no, 1 for yes).

index	ID	AGE	PackHistory	COPDSEVERITY	MWT1	MWT2	MWT1Best	FEV1	FEV1PRED	FVC	FVCPRED	CAT	HAD	SGRQ	AGEquartiles	gender	smoking	Diabetes	muscular	hypertension
0	58	77	60	SEVERE	120	120	120	1.21	36	2.4	98	25	8	69.55	4	1	2	1	0	0
1	57	79	50	MODERATE	165	176	176	1.09	56	1.64	65	12	21	44.24	4	0	2	1	0	0
2	62	80	11	MODERATE	201	180	201	1.52	68	2.3	86	22	18	44.09	4	0	2	1	0	0
3	145	56	60	VERY SEVERE	210	210	210	0.47	14	1.14	27	28	26	62.04	1	1	2	0	0	1
4	136	65	68	SEVERE	204	210	210	1.07	42	2.91	98	32	18	75.56	1	1	2	0	1	1
5	84	67	26	MODERATE	216	180	216	1.09	50	1.99	60	29	21	73.82	2	0	1	1	0	0
6	93	67	50	SEVERE	214	237	237	0.69	35	1.31	48	29	30	77.44	2	0	1	1	0	0
7	27	83	90	SEVERE	214	237	237	0.68	32	2.23	77	22	2	45.41	4	1	2	1	0	0
8	114	72	50	MODERATE	231	237	237	2.13	63	4.38	80	25	6	69.61	3	1	1	1	0	0
9	152	75	6	SEVERE	226	240	240	1.06	46	2.06	75	31	20	55.56	3	0	2	0	1	0
10	166	76	6	SEVERE	226	240	240	1.1	46	2.06	75	31	20	55.56	3	0	2	1	1	0
11	139	59	28	VERY SEVERE	240	230	240	0.68	29	2.02	73	23	18	55.23	1	0	2	0	0	0
12	49	64	30	VERY SEVERE	246	234	246	0.45	17	1.56	45	15	22	50.53	1	1	2	1	0	0
13	80	74	75	MILD	213	270	270	1.79	83	2.62	102	24	19	45.0	3	0	1	0	0	0
14	162	70	103	MODERATE	264	271	271	1.2	52	2.09	121	12	1	39.66	2	0	2	0	0	0
15	162	71	105	MODERATE	264	271	271	0.72	52	2.09	121	12	1	39.66	2	0	2	0	0	0
16	41	69	78	SEVERE	243	273	273	1.46	46	3.33	81	15	10	28.86	2	1	2	1	0	0
17	1	55	109	MODERATE	281	273	281	1.54	72	2.15	85	28	29	76.5	1	0	2	0	0	0
18	153	72	15	VERY SEVERE	289	267	289	0.6	24	1.81	55	30	6	38.74	3	1	2	0	0	0
19	167	72	15	VERY SEVERE	289	267	289	0.89	24	1.81	55	30	6	38.74	3	1	2	0	0	0
20	138	74	24	VERY SEVERE	267	295	295	0.51	27	2.06	89	26	30	71.21	3	0	2	0	0	0
21	48	75	40	VERY SEVERE	297	291	297	0.79	30	1.81	51	22	9	35.79	3	1	1	0	0	0
22	110	69	15	SEVERE	300	300	300	0.91	39	2.9	105	29	9	58.78	2	0	2	0	0	0
23	15	73	75	SEVERE	300	243	300	1.46	3	2.37	53	18	14	34.71	3	1	2	1	0	0
24	70	75	45	MILD	282	300	300	2.35	85	4.12	112	23	19	58.25	3	1	2	0	0	0
25	63	80	67	MODERATE	300	305	305	1.77	58	2.77	67	32	26	67.66	4	1	2	0	0	0
26	81	76	38	SEVERE	281	312	312	1.06	32	3.11	70	25	11	56.8	4	1	1	0	0	0
27	11	73	31	MODERATE	336	336	336	1.88	70	2.71	70	26	13	66.51	3	1	2	0	1	0
28	31	77	75	MODERATE	322	338	338	1.92	68	2.66	69	16	2	36.39	4	1	2	0	1	0
29	68	88	1	SEVERE	350	350	350	1.3	37	2.0	70	11	1	47.2	4	1	2	1	0	1
30	97	44	30	SEVERE	360	368	368	1.66	42	3.08	63	30	17	72.24	1	1	2	0	0	0
31	127	82	45	SEVERE	345	369	369	1.18	46	2.57	73	10	6	37.04	4	1	2	0	1	0

Figure 3.3.2.1: Dataset visualization

Figure 3.2.2.1 illustrates the initial distribution of the dataset variables before feature engineering, providing a comprehensive overview of the data's characteristics.

index	AGE	PackHistory	MWT1Best	FEV1	FEV1PRED	FVC	FVCPRED	CAT	HAD	SGRQ	AGEquartiles	gender	smoking	Diabetes	muscular	hypertension	AtrialFib	IHD	copd
0	77	60	120	1.21	36	2.4	98	25	8	69.55	4	1	2	1	0	0	1	0	3
1	79	50	176	1.09	56	1.64	65	12	21	44.24	4	0	2	1	0	0	1	1	2
2	80	11	201	1.52	68	2.3	86	22	18	44.09	4	0	2	1	0	0	1	0	2
3	56	60	210	0.47	14	1.14	27	28	26	62.04	1	1	2	0	0	1	1	0	4
4	65	68	210	1.07	42	2.91	98	32	18	75.56	1	1	2	0	1	1	0	0	3
5	67	26	216	1.09	50	1.99	60	29	21	73.82	2	0	1	1	0	0	1	0	2
6	67	50	237	0.69	35	1.31	48	29	30	77.44	2	0	1	1	0	0	1	0	3
7	83	90	237	0.68	32	2.23	77	22	2	45.41	4	1	2	1	0	0	1	0	3
8	72	50	237	2.13	63	4.38	80	25	6	69.61	3	1	1	1	0	0	1	0	2
9	75	6	240	1.06	46	2.06	75	31	20	55.56	3	0	2	0	1	0	0	0	3
10	76	6	240	1.1	46	2.06	75	31	20	55.56	3	0	2	1	1	0	1	0	3
11	59	28	240	0.68	29	2.02	73	23	18	55.23	1	0	2	0	0	0	1	0	4
12	64	30	246	0.45	17	1.56	45	15	22	50.53	1	1	2	1	0	0	1	0	4
13	74	75	270	1.79	83	2.62	102	24	19	45.0	3	0	1	0	0	0	1	0	1
14	70	103	271	1.2	52	2.09	121	12	1	39.66	2	0	2	0	0	0	0	0	2
15	71	105	271	0.72	52	2.09	121	12	1	39.66	2	0	2	0	0	0	0	0	2
16	69	78	273	1.46	46	3.33	81	15	10	28.86	2	1	2	1	0	0	1	0	3
17	55	109	281	1.54	72	2.15	85	28	29	76.5	1	0	2	0	0	0	0	0	2
18	72	15	289	0.6	24	1.81	55	30	6	38.74	3	1	2	0	0	0	0	0	4
19	72	15	289	0.89	24	1.81	55	30	6	38.74	3	1	2	0	0	0	0	0	4
20	74	24	295	0.51	27	2.06	89	26	30	71.21	3	0	2	0	0	0	0	0	4
21	75	40	297	0.79	30	1.81	51	22	9	35.79	3	1	1	0	0	0	0	0	4
22	69	15	300	0.91	39	2.9	105	29	9	58.78	2	0	2	0	0	0	0	0	3
23	73	75	300	1.46	3	2.37	53	18	14	34.71	3	1	2	1	0	0	0	0	3
24	75	45	300	2.35	85	4.12	112	23	19	58.25	3	1	2	0	0	0	0	0	1
25	80	67	305	1.77	58	2.77	67	32	26	67.66	4	1	2	0	0	0	0	0	2
26	76	38	312	1.06	32	3.11	70	25	11	56.8	4	1	1	0	0	0	0	0	3
27	73	31	336	1.88	70	2.71	70	26	13	66.51	3	1	2	0	1	0	0	0	2
28	77	75	338	1.92	68	2.66	69	16	2	36.39	4	1	2	0	1	0	1	0	2
29	88	1	350	1.3	37	2.0	70	11	1	47.2	4	1	2	1	0	1	0	1	3
30	44	30	368	1.66	42	3.08	63	30	17	72.24	1	1	2	0	0	0	0	0	3
31	82	45	369	1.18	46	2.57	73	10	6	37.04	4	1	2	0	1	0	0	0	3

Figure 3.3.2.2: Dataset visualization after feature Engineering

Figure 3.2.2.2 shows the dataset after feature engineering, highlighting the transformed and selected features used for model training.

Data Summary:

- **Total Records:** 540 entries.
- **Feature Count:** 23 features, including demographics, clinical test results, wearable device data, and expert assessments.

By integrating data from diverse sources, this comprehensive dataset provides an excellent foundation for constructing accurate and effective machine learning algorithms for COPD risk prediction and evaluation.

3.4 Hardware/ Software Requirement

The development and execution of this project required specific hardware and software configurations. On the hardware side, a computer with at least 8GB of RAM and a multi-core processor was essential to handle data processing and model training efficiently. For larger datasets or more complex models, access to a high-performance computing environment or cloud-based resources such as Google Colab or AWS is beneficial.

The software prerequisites included Python 3.7 or above, along with important libraries such as Pandas for data processing, Scikit-learn for machine learning techniques, Matplotlib and Seaborn for data visualization, and google colab Notebook for an interactive scripting environment. Additionally, version control was controlled using Git, and collaborative work was enabled through sites like GitHub. The integration of these technologies and resources provided a smooth workflow from data pretreatment to model deployment.

3.5 Project Management and Financial Analysis

Effective project management is vital for the effective completion of this research. We adopted Agile methodologies, with the project divided into sprints that lasted two weeks each. Each sprint focused on specific tasks such as data preprocessing, model training, or documentation. Regular meetings were held to review progress and address any challenges.

Financial analysis involved budgeting for hardware upgrades, software licenses, and cloud computing resources. The primary costs were associated with acquiring high-performance computing resources for model training and cloud storage for data. Additionally, we allocated funds for miscellaneous expenses such as publication fees and conference participation. The financial plan ensured that all project activities were adequately funded without exceeding the allocated budget.

Table 3.4.1: Estimated Cost for Machine Learning based Project

SN	Components	Estimated Cost (BDT)
01.	Hardware/Infrastructure	35,000-40,000
02.	Software and Tools	5000-7000
04.	Documentation and Report Writing	1000-2000
05.	Miscellaneous (e.g., licenses, tools)	500-1000
06.	Contingency (10% of total)	500-1000
Total Estimated Cost		**42,000 -51,000**

3.6 Summary

In summary, this chapter highlighted the systematic process employed for constructing predictive models for COPD risk assessment. From data collection and preprocessing to model training and assessment, each step was meticulously planned and executed. The hardware and software requirements were specified to ensure efficient data processing and model development. Project management techniques and financial analysis were defined to assure the project's successful and timely completion. The methodologies discussed in this chapter lay the foundation for the subsequent implementation and results analysis chapters.

CHAPTER 4

IMPLEMENTATION

4.1 Overview

The implementation chapter details the practical steps taken to realize the proposed methodology. This includes the training of machine learning models, prototype design, and system testing. Each section within this chapter discusses the procedures, tools, and techniques used to develop, train, and assess the prediction models. The implementation phase is crucial as it translates theoretical designs into functional models that can be tested and validated.

4.2 Train Model

The first step in the implementation process was to train the machine learning models using a dataset obtained from Kaggle, which contained comprehensive data on COPD-related features. The data underwent extensive preprocessing to ensure its suitability for training the models. This preprocessing involved several key steps: handling missing values, normalizing data, and selecting relevant features. Handling missing values was crucial to maintain the integrity of the dataset, as incomplete data could lead to biased or inaccurate model predictions. Normalizing the data ensured that features were on a similar scale, which is important for many machine learning algorithms. Feature selection helped in identifying the most significant variables that contribute to the prediction of COPD, thereby enhancing the models' functionality and efficiency in the process.

To identify the most effective model for COPD prediction, we implemented a variety of machine learning approaches. These techniques included Logistic Regression, Decision Trees, Random Forest, AdaBoost, K-Nearest Neighbors (KNN), Naive Bayes, Linear Discriminant Analysis, and Gradient Boosting. Each algorithm has its strengths and unique characteristics, which made it important to evaluate multiple approaches. To evaluate the models' performance on unseen data, the dataset was divided into training and testing sets, ensuring that the evaluation metrics were not biased by overfitting. Hyperparameter tuning was an essential part of this process. To tune the hyperparameters for every model, we employed cross-validation and grid search approaches. Grid search involves systematically working through multiple combinations of hyperparameter values, cross-validation ensures that the model is assessed using many subsets of the training set, to provide a reliable indicator of the model's performance.

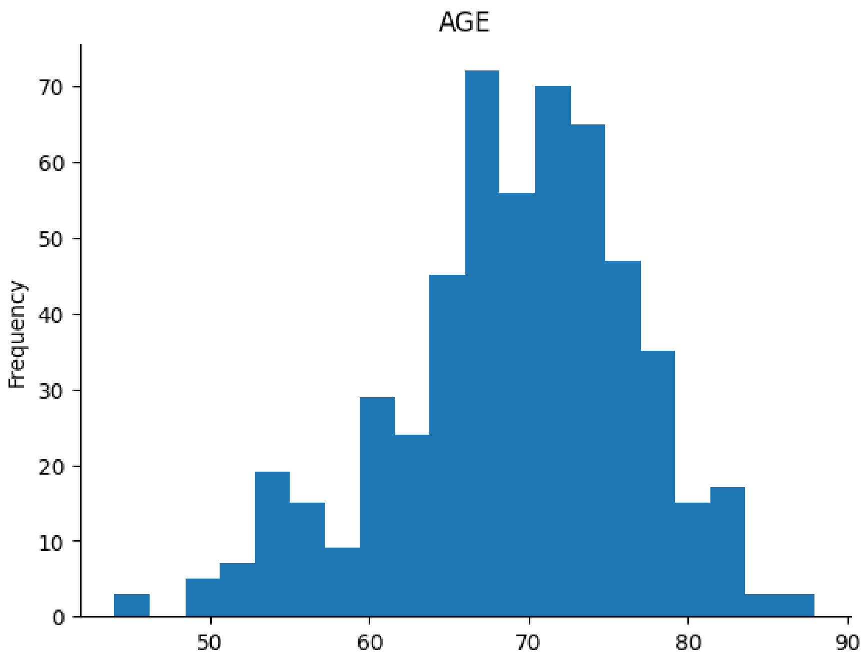

Figure 4.2.1: Age Visualization

Figure 4.2.1 depicts the age range of the study participants, illustrating the demographic spread within the dataset.

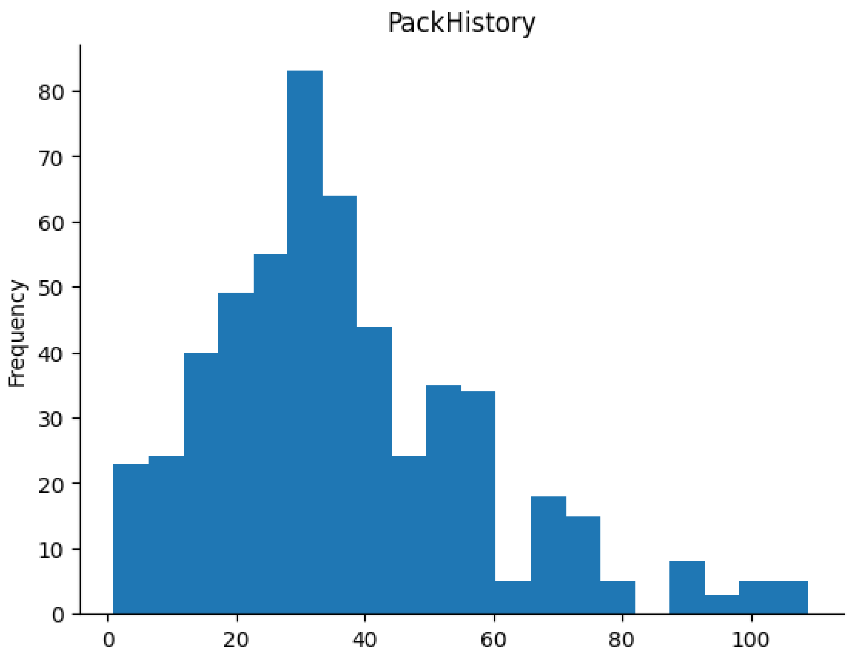

Figure 4.2.2: PackHistory Visualization

Figure 4.2.2 visualizes the pack history of smoking among participants, an important risk factor for COPD.

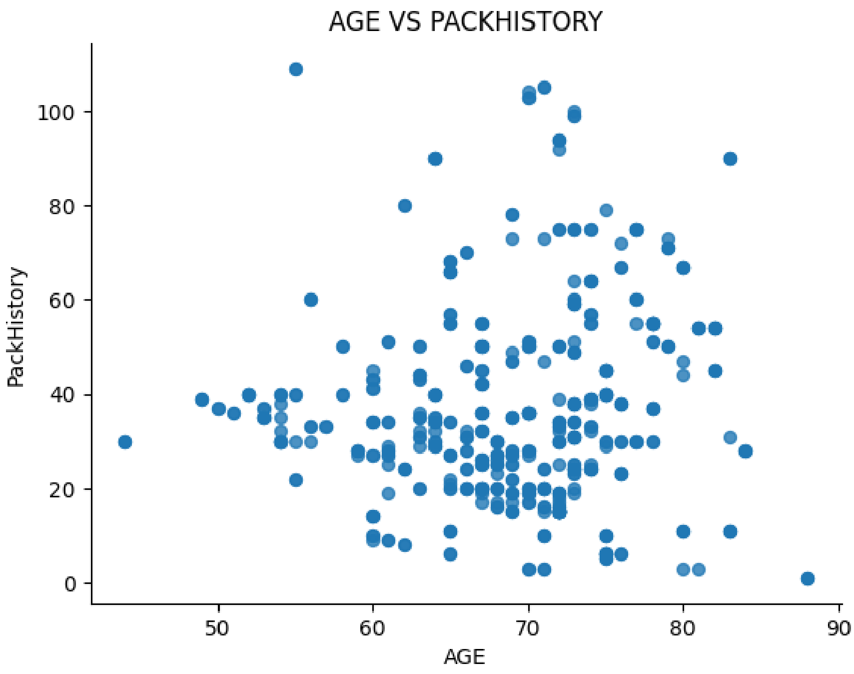

Figure 4.2.3: Age vs PackHistory

Figure 4.2.3 demonstrates scatter plot and the relationship between age and pack history, helping to identify correlations between these two variables.

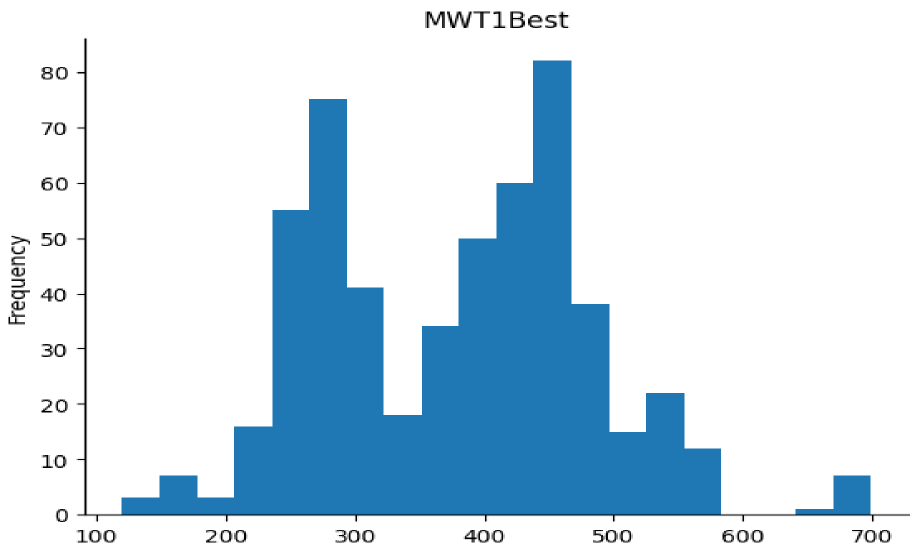

Figure 4.2.4: MWT1 Visualization

Figure 4.2.4 illustrates the distribution of the six-minute walk test (MWT1) results, a measure of functional exercise ability.

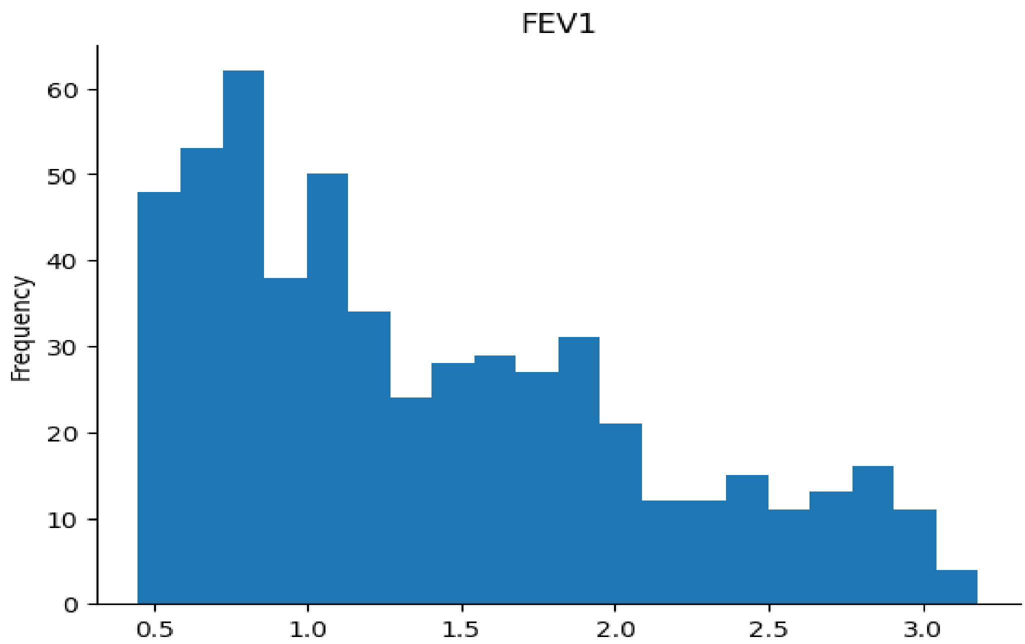

Figure 4.2.5: FEV1 Visualization

Figure 4.2.5 presents the distribution of forced expiratory volume (FEV1) in one second, a critical measure in the diagnosis of COPD.

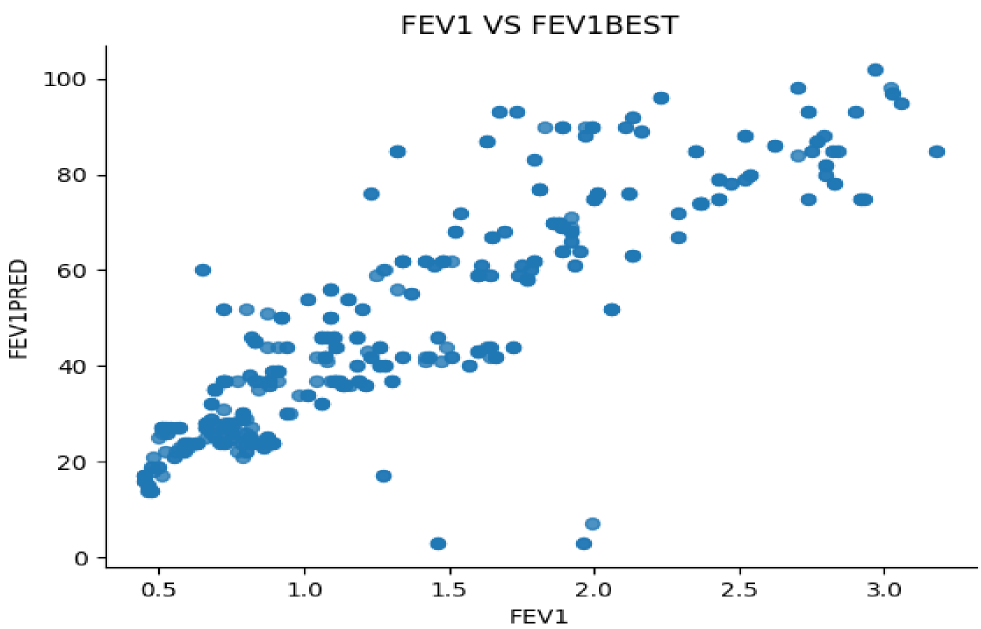

Figure 4.2.6: MWT1 vs FEV1

Figure 4.2.6 shows the relationship between MWT1 and FEV1, highlighting potential correlations between exercise capacity and lung function.

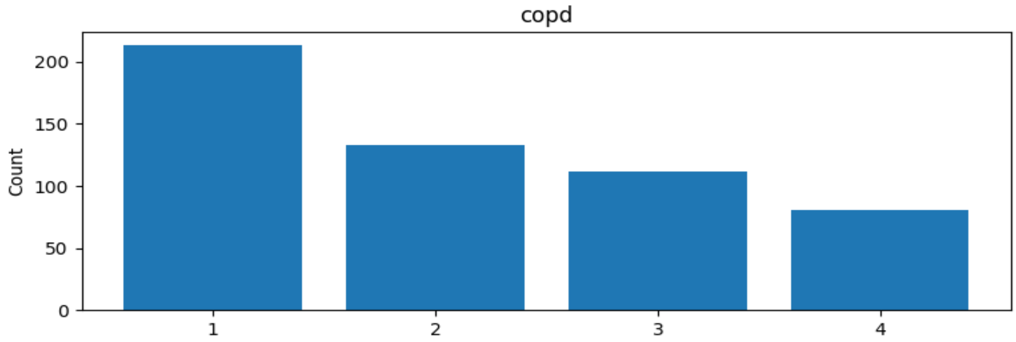

Figure 4.2.7: COPDSEVERITY DATA COUNT

Figure 4.2.7 displays the frequency distribution of COPD severity levels among the participants.

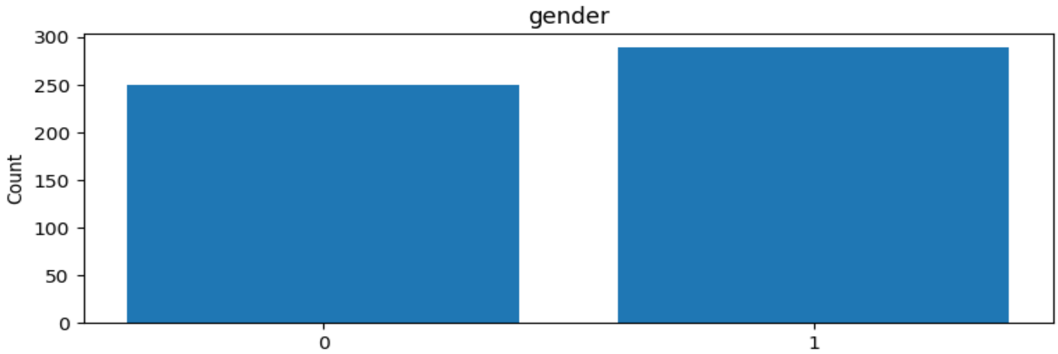

Figure 4.2.8: Gender wise data Comparison

Figure 4.2.8 compares the distribution of male and female participants in the dataset.

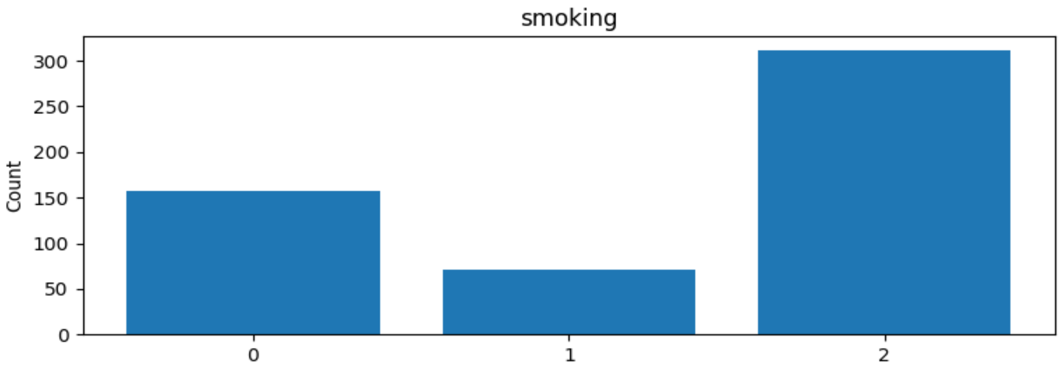

Figure 4.2.9: smoking category wise data

Figure 4.2.9 categorizes participants based on their smoking history, crucial for assessing COPD risk.

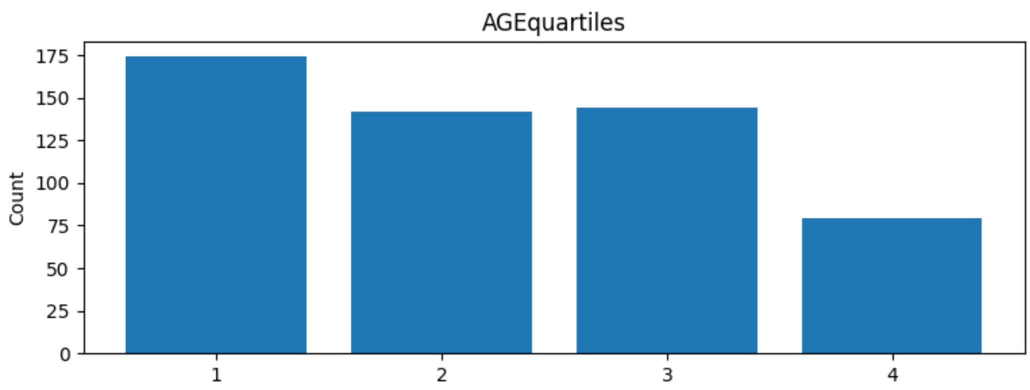

Figure 4.2.10: data Count AGEquartiles wise

Figure 4.2.10 shows the distribution of participants divided into quartiles based on age.

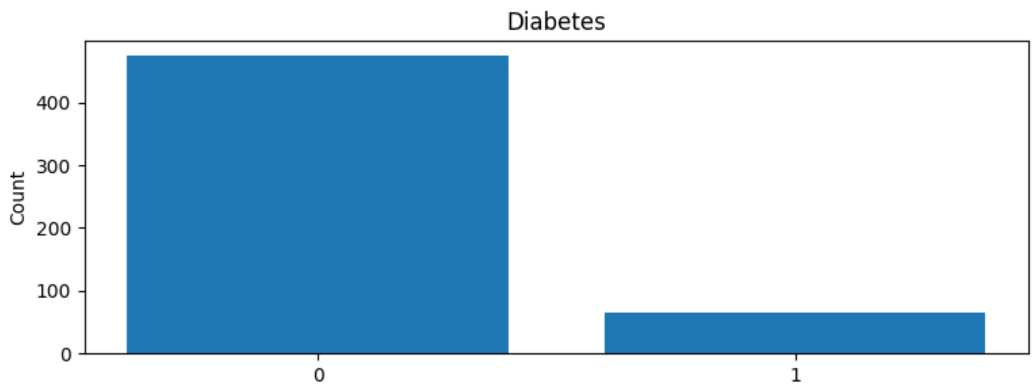

Figure 4.2.11: Diabetics presence

Figure 4.2.11 illustrates the prevalence of diabetes among the study participants, a potential comorbidity affecting COPD.

4.3 System Testing/ Model Evaluation

System testing and model assessment are key processes to assure the dependability and correctness of the produced models. Several metrics, including accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC), were used to evaluate the models. These measurements gave a detailed assessment of each model's performance.

We also conducted error analysis to determine the common types of errors that the models made. This analysis helped in refining the models and improving their performance. The final models were then integrated into the prototype, and end-to-end testing was conducted to ensure that the entire system functioned as expected.

```
copd             1.000000
smoking          0.625886
SGRQ             0.235626
AtrialFib        0.204480
CAT              0.156550
HAD              0.156081
gender           0.127117
AGEquartiles     0.100416
Diabetes         0.060777
AGE              0.053183
hypertension     0.005575
muscular         0.000662
PackHistory     -0.004733
MWT1Best        -0.348182
FVC             -0.379919
FVCPRED         -0.397771
FEV1            -0.473703
FEV1PRED        -0.505903
IHD             -0.587730
Name: copd, dtype: float64
```

Figure 4.3.1: Feature Importance

Figure 4.3.1 ranks the significance of several attributes in forecasting COPD risk as established by the model.

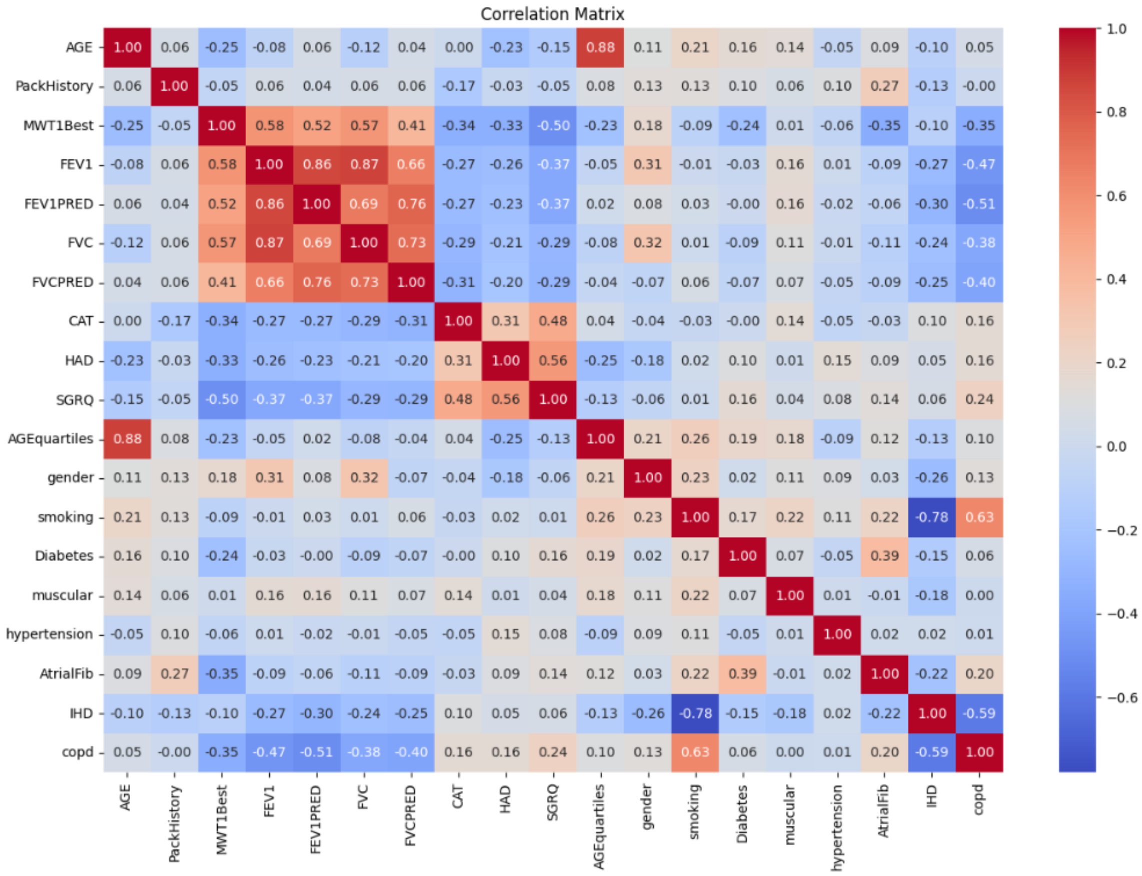

Figure 4.3.2: Correlation Matrix

Figure 4.3.2 shows the correlation between different variables in the dataset, identifying potential multicollinearity issues.

4.4 Summary

In summary, the implementation chapter covered the practical steps taken to develop and train the machine learning models for COPD risk assessment. This included data preprocessing, model training, hyperparameter tuning, and prototype design. System testing and model evaluation were conducted using various metrics to ensure the reliability and accuracy of the models. The implementation phase successfully translated the proposed methodology into functional models and a user-friendly prototype, setting the stage for the results and analysis chapter.

CHAPTER 5

RESULT AND ANALYSIS

5.1 Overview

This chapter provides the research results acquired from the implemented models and analyzes their performance. It includes experimental results, comparative analysis, and performance metrics for each machine learning model. The objective is to evaluate the models' accuracy for COPD risk to identify which model performs best. The research also covers possible areas for improvement and the practical consequences of the findings.

5.2 Experimental/ Simulation Result

The experimental results were obtained by training and testing the machine learning models on the preprocessed dataset. The effectiveness of each model was assessed using a hold-out test set and cross-validation techniques.

Logistic Regression 94%, Decision Trees 96%, Random Forest 98%, AdaBoost 83%, K-Nearest Neighbors (KNN) 87%, Naive Bayes 82%, Linear Discriminant Analysis 86%, and Gradient Boosting 96%.

The model for Logistic Regression attained an accuracy of 94%, with the values of precision and recall are 95% and 94% respectively. The Decision Tree model showed an accuracy of 96%, with a precision of 97% and a recall of 96%. The Random Forest model achieved an accuracy of 98%, with precision and recall values of 98% and 98% respectively. The AdaBoost model had an accuracy of 83%, with the values of precision and recall are of 86% and 83%. The K-Nearest Neighbors (KNN) model achieved an accuracy of 87%, with precision and recall values of 90% and 88% respectively. The Naive Bayes model had an accuracy of 82%, with precision and recall values of 86% and 82%. The Linear Discriminant Analysis model showed an accuracy of 86%, with a precision of 87% and a recall of 86%. Finally, the Gradient Boosting model achieved an accuracy of 96%, with precision and recall values of 97% and 96% respectively.

The experimental results indicated that the Random Forest model was the most effective in predicting COPD risk, followed by the Logistic Regression and Decision tree models. These results were consistent across different evaluation metrics and cross-validation folds.

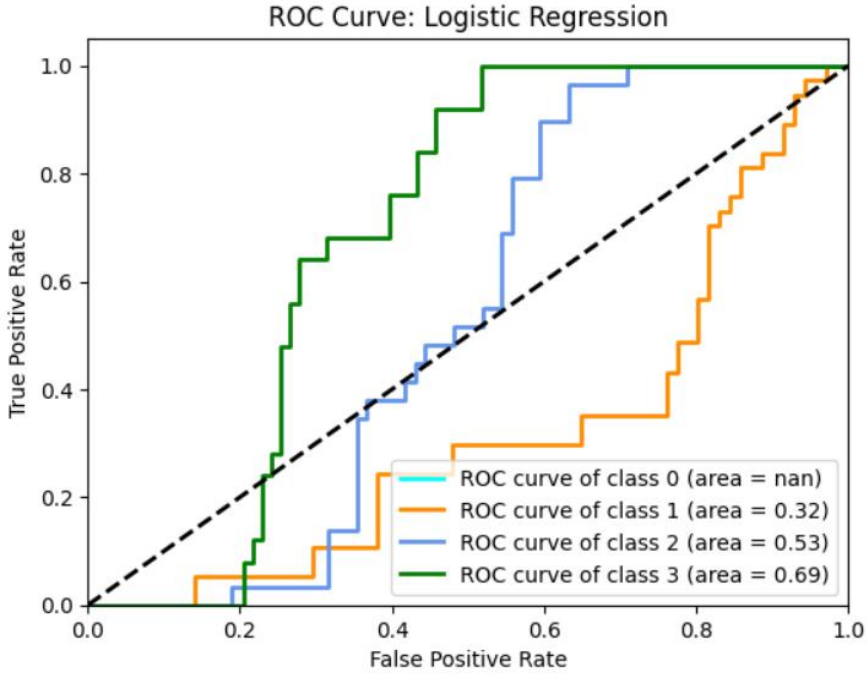

Figure 5.2.1: Logistic Regression ROC Curve

Figure 5.2.1 presents the ROC curve of Logistic Regression model, indicating its performance in distinguishing between COPD and non-COPD cases.

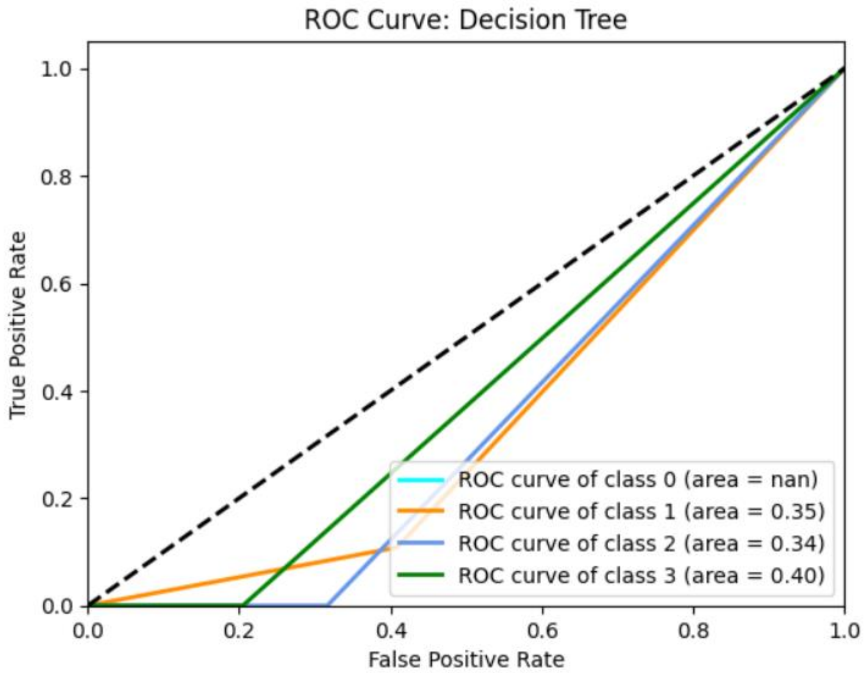

Figure 5.2.2: Decision Tree ROC Curve

Figure 5.2.2 shows the ROC curve of Decision Tree model, illustrating its classification performance.

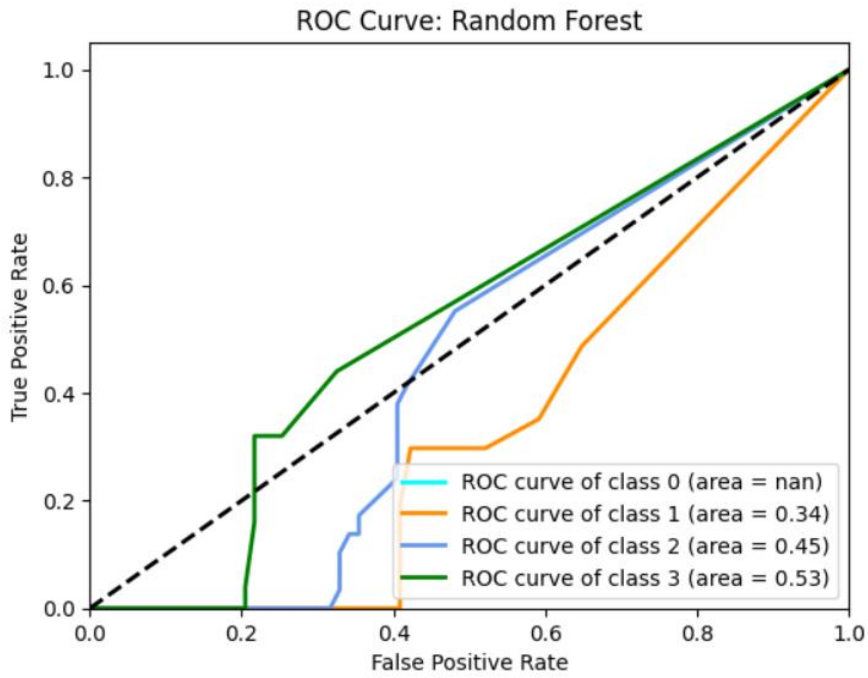

Figure 5.2.3: Random Forest ROC Curve

Figure 5.2.3 displays the ROC curve of Random Forest model, highlighting its accuracy and predictive power.

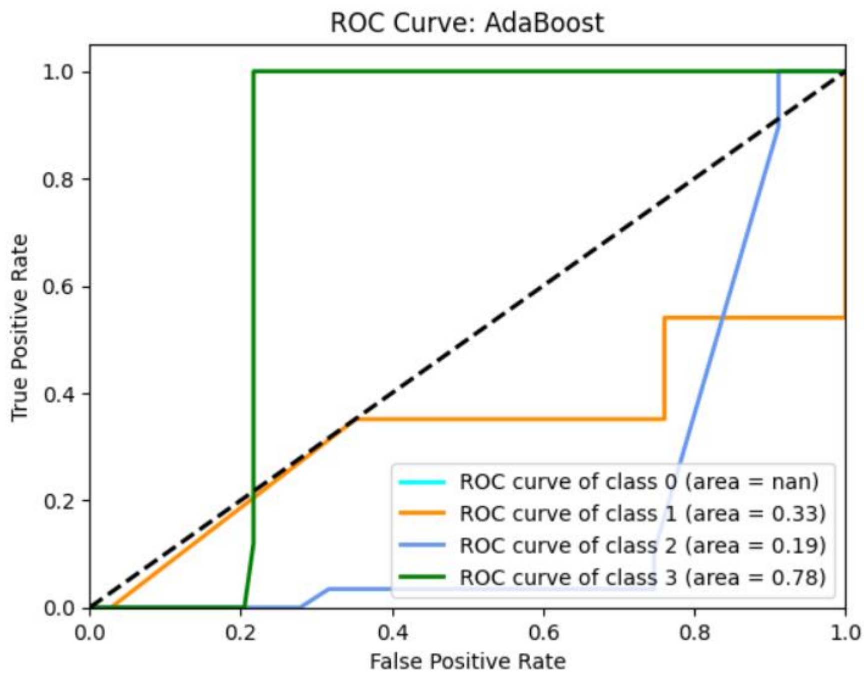

Figure 5.2.4: Ada Boost ROC Curve

Figure 5.2.4 presents the ROC curve of AdaBoost model, demonstrating its effectiveness in classification.

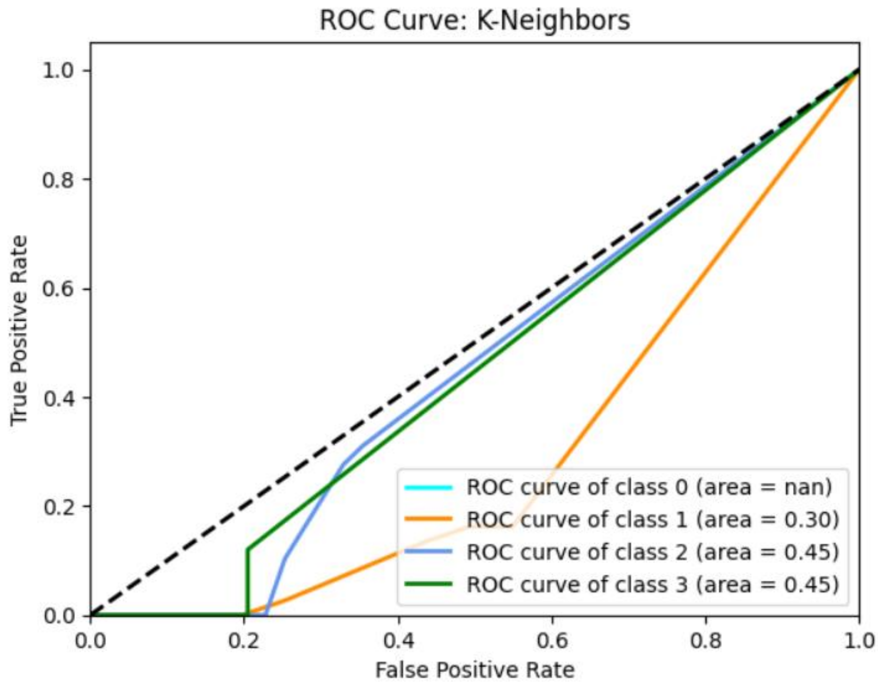

Figure 5.2.5: KNN ROC Curve

Figure 5.2.5 shows the ROC curve of K-Nearest Neighbors model, indicating its performance metrics.

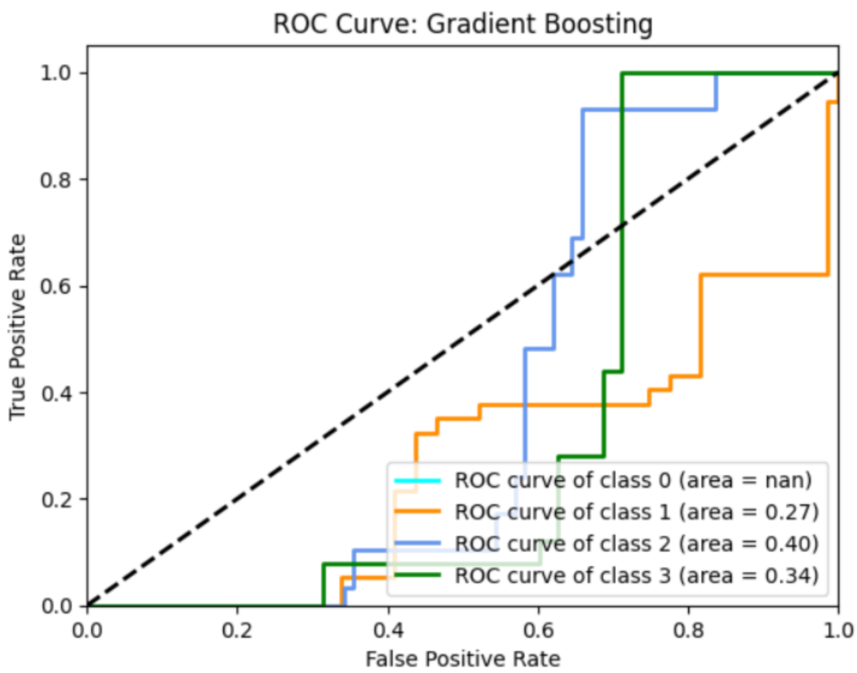

Figure 5.2.6: Gradient Boosting ROC Curve

Figure 5.2.6 illustrates the ROC curve of Gradient Boosting model, showcasing its prediction capabilities.

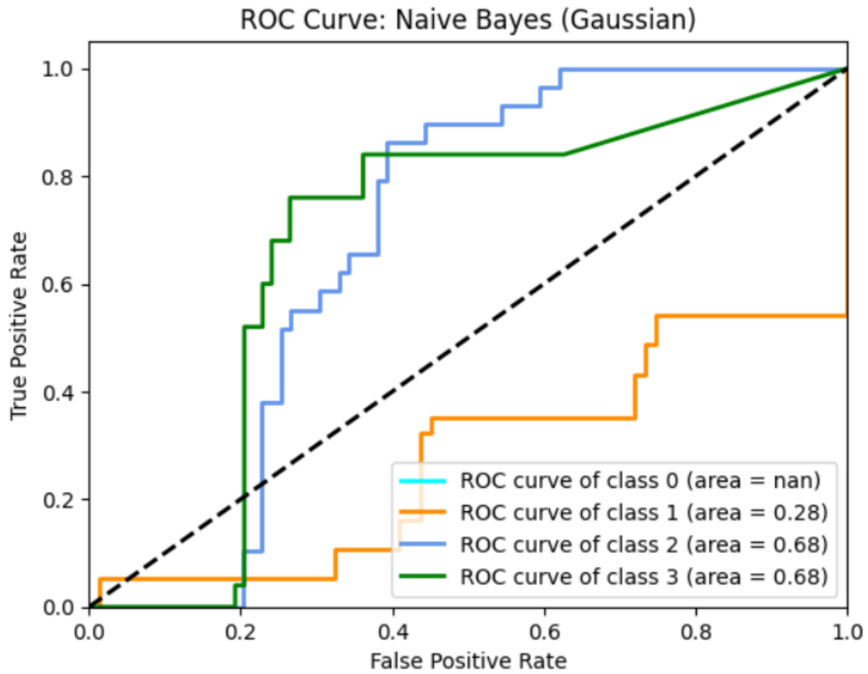

Figure 5.2.7: Naïve Bayes ROC Curve

Figure 5.2.7 presents the ROC curve of Naive Bayes model, evaluating its classification accuracy.

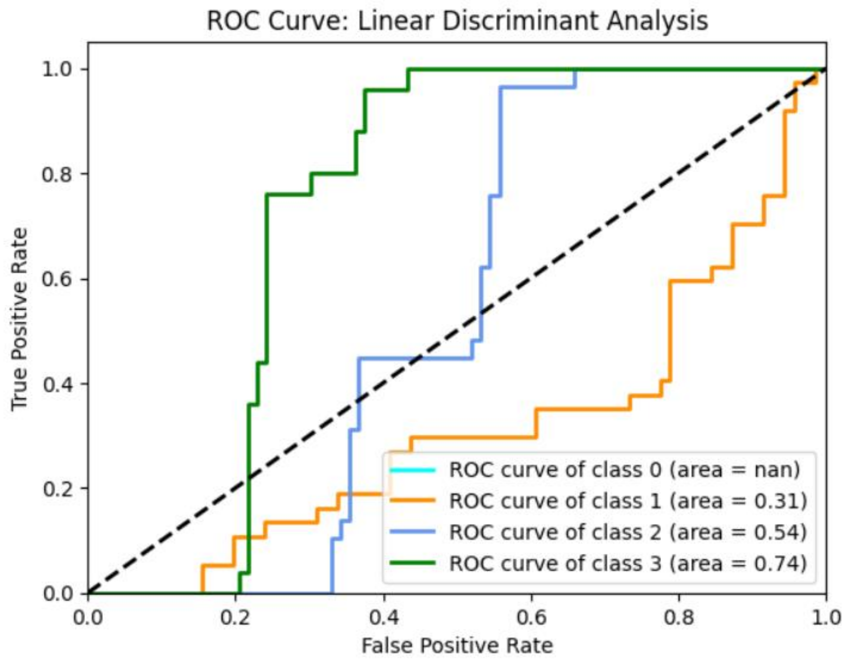

Figure 5.2.8: LDA ROC Curve

Figure 5.2.8 shows the ROC curve of Linear Discriminant Analysis model, indicating its performance in classification tasks.

Confusion Matrix:

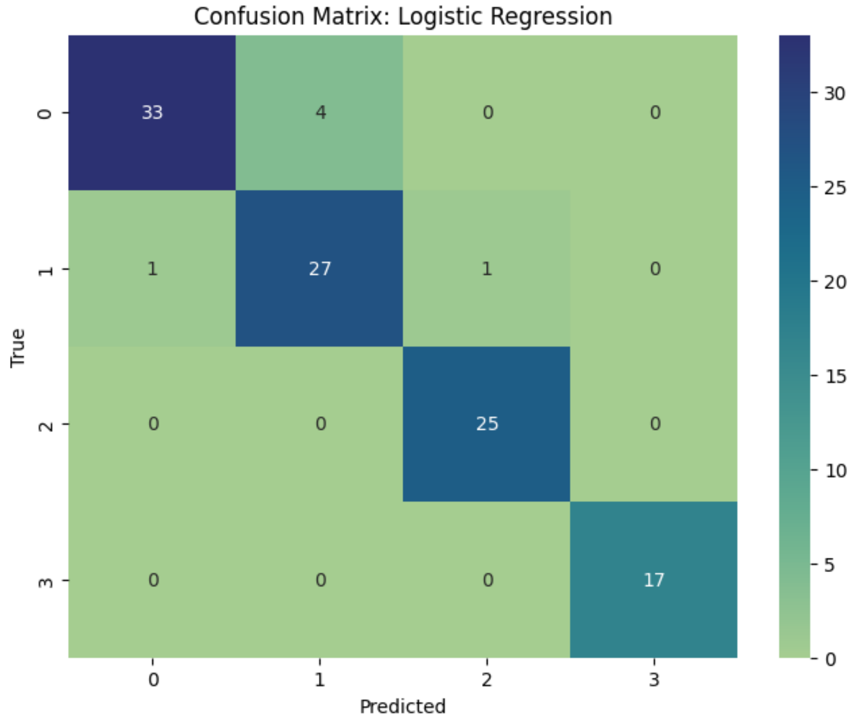

Figure 5.2.9: LR Confusion Matrix

Figure 5.2.9 provides the confusion matrix for the Logistic Regression model, detailing true positives, false positives, true negatives, and false negatives.

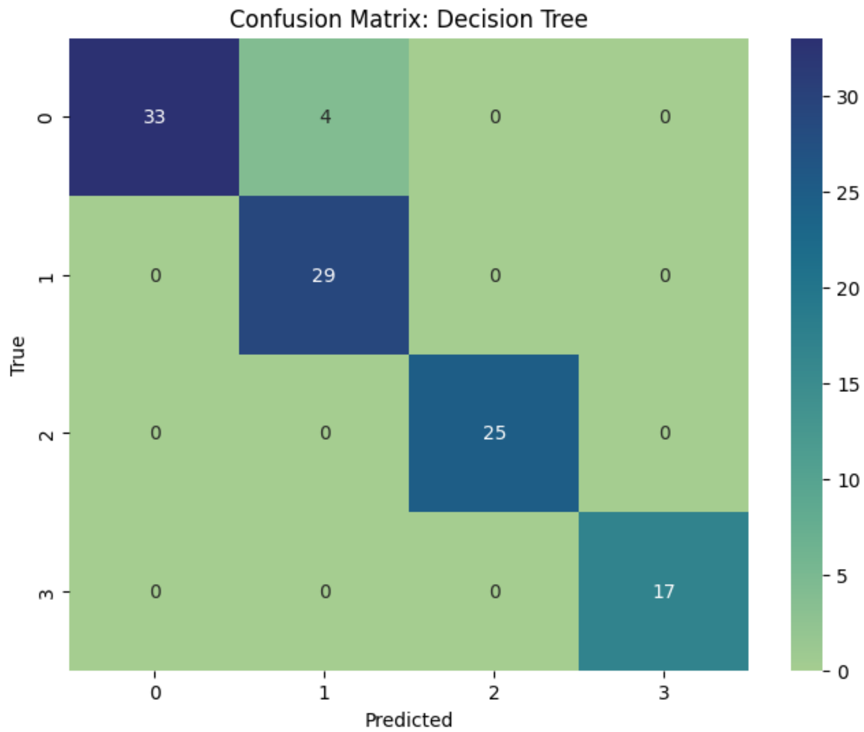

Figure 5.2.10: Decision Tree Confusion Matrix

Figure 5.2.10 provides the confusion matrix for the Decision Tree model, summarizing its classification outcomes.

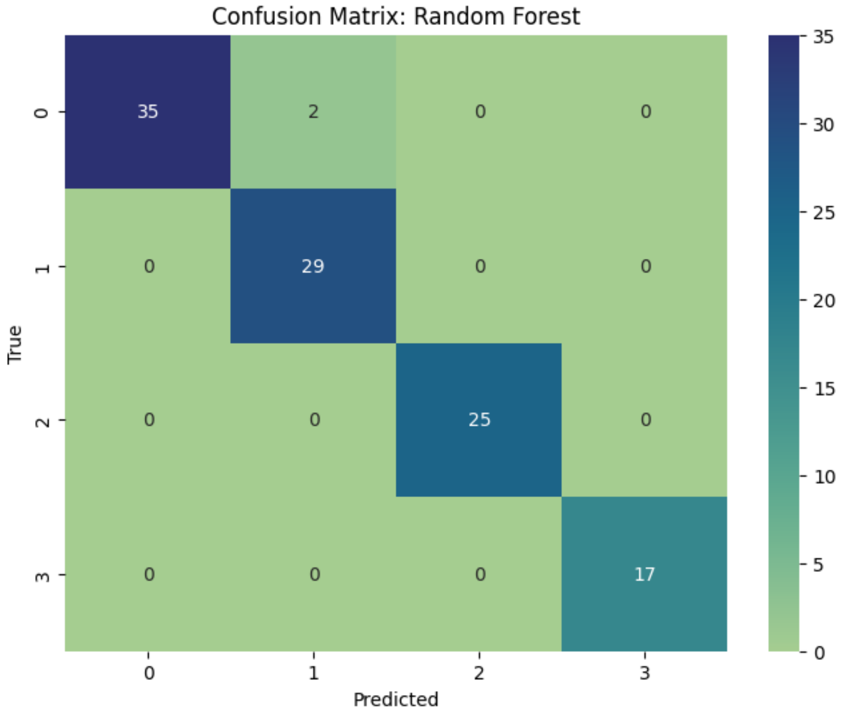

Figure 5.2.11: Random Forest Confusion Matrix

Figure 5.2.11 displays the confusion matrix for the Random Forest model, indicating its predictive accuracy.

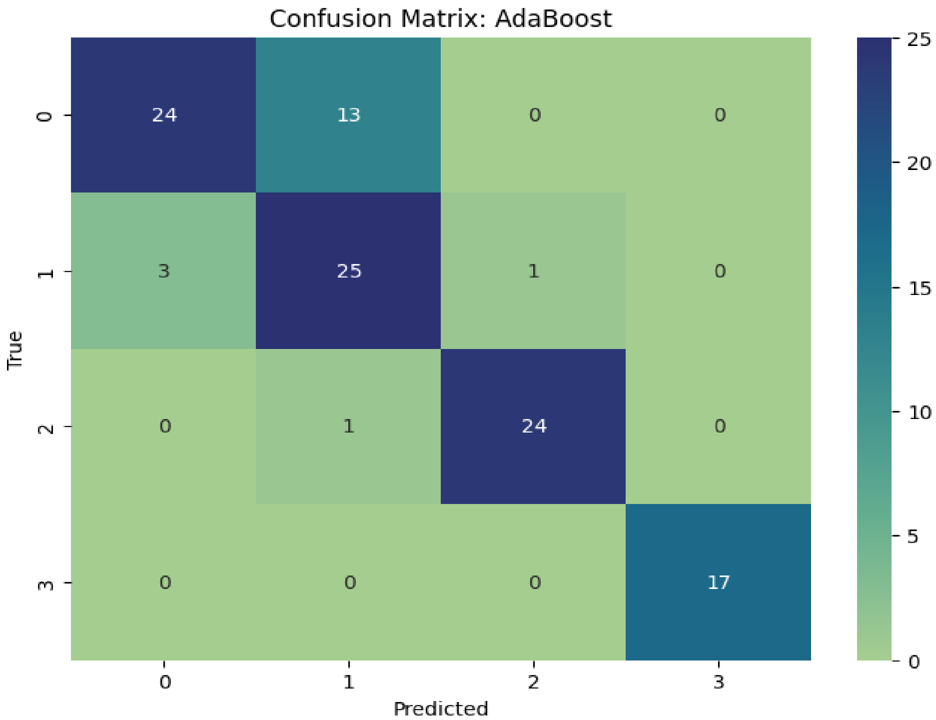

Figure 5.2.12: Adaboost Confusion Matrix

Figure 5.2.12 presents the confusion matrix for the AdaBoost model, detailing its performance metrics.

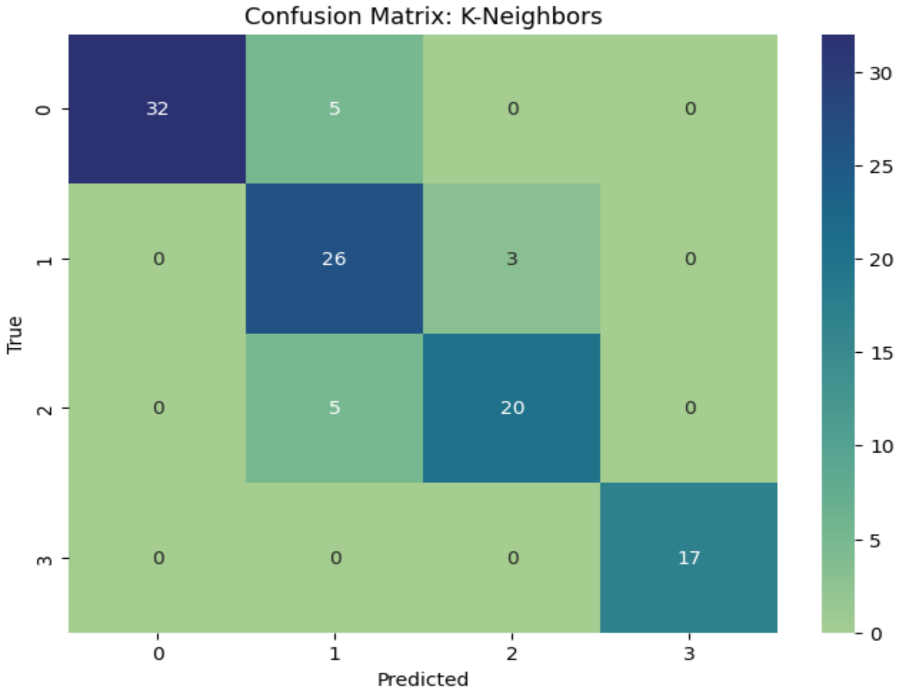

Figure 5.2.13: KNN Confusion Matrix

Figure 5.2.13 shows the confusion matrix for the KNN model, highlighting its classification accuracy.

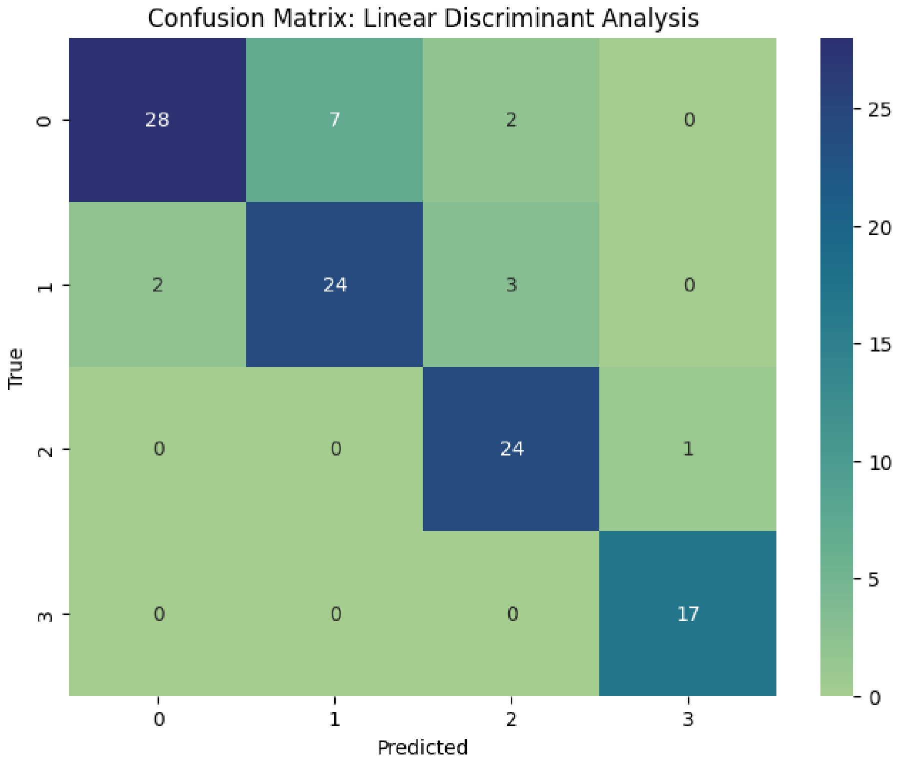

Figure 5.2.14: LDA Confusion Matrix

Figure 5.2.14 provides the confusion matrix for the LDA model, summarizing its predictive outcomes.

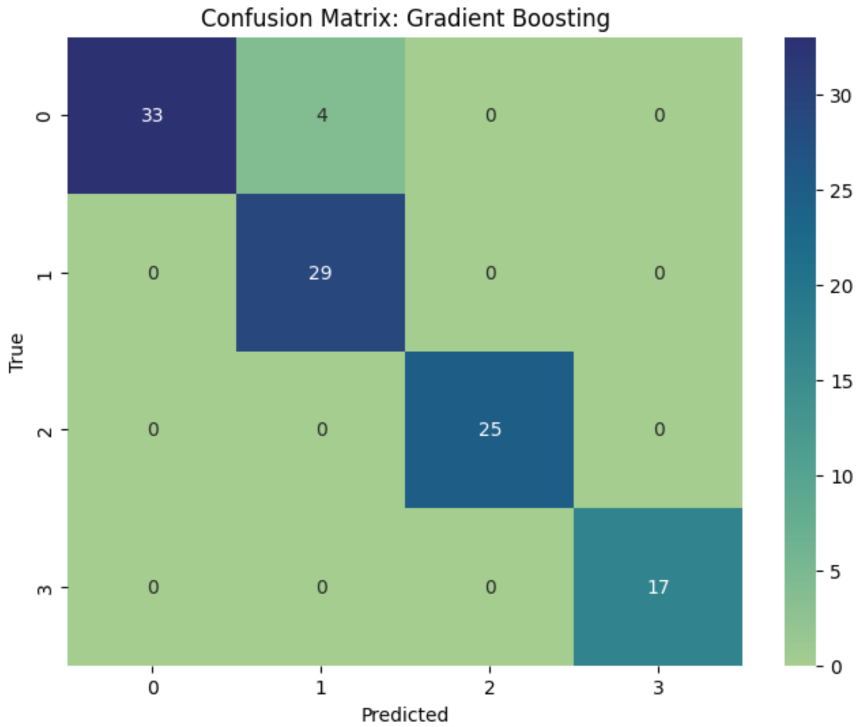

Figure 5.2.15: Gradient Boosting Confusion Matrix

Figure 5.2.15 presents the confusion matrix for the Gradient Boosting model, detailing its performance metrics.

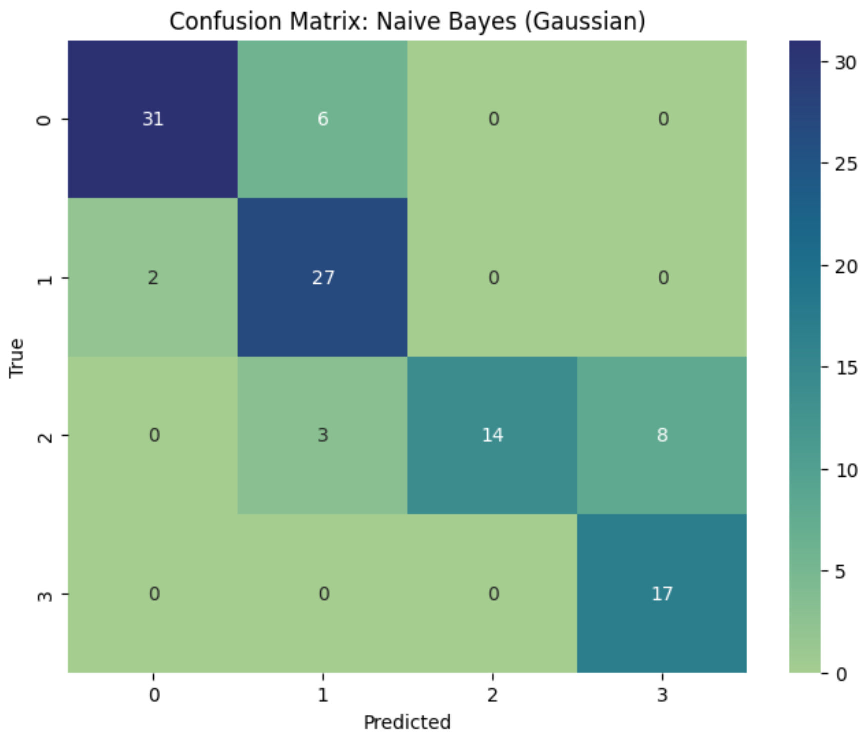

Figure 5.2.16: Naïve Bayes Confusion Matrix

Figure 5.2.16 shows the confusion matrix for the Naïve Bayes model, indicating its classification performance.

5.3 Performance/ Comparative Analysis

A comparative analysis of the models is performed to identify the strengths and weaknesses of each algorithm. According to the research, the Random Forest model performed better than the other models in terms of accuracy, precision and recall. Its ability to combine multiple weak classifiers into a strong one contributed to its superior performance.

The Logistic Regression model, while slightly less accurate than AdaBoost, provided valuable insights into the significance of individual risk factors. The KNN model performed well, particularly in identifying high-risk patients, but its performance decreased with an increasing number of neighbors. The Decision Tree model, though easy to interpret, was prone to overfitting, which affected its accuracy. The Naive Bayes model, despite its simplicity, showed decent performance but struggled with the assumption of feature independence.

Overall, the comparison research demonstrated how crucial it is to select the appropriate model for the particular needs and limitations of the application. The Random Forest model was recommended for deployment due to its superior performance and robustness.

Table 5.3.1. Model Performance

Architecture	Training Accuracy	Test Accuracy	Test Loss
Logistic Regression	0.9698	0.9444	0.0555
Decision Tree Classifier	1.0	0.9629	0.0370
Random Forest	1.0	98.148	0.018
Ada Boost	0.879	0.8333	0.1666
KNN	0.9443	0.879	0.120
Naïve Bayes	0.835	0.824	0.175
Linear Discriminant Analysis	0.886	0.861	0.138
Gradient Boosting	1.0	0.962	0.037

Classification Report:

Logistic Regression Classification Report:

	precision	recall	f1-score	support
1	0.97	0.89	0.93	37
2	0.87	0.93	0.90	29
3	0.96	1.00	0.98	25
4	1.00	1.00	1.00	17
accuracy			0.94	108
macro avg	0.95	0.96	0.95	108
weighted avg	0.95	0.94	0.94	108

Figure 5.3.1: LR Classification Report

Figure 5.3.1 presents the Logistic Regression model's classification report, which includes the precision, recall and F1-score.

Decision Tree Classification Report:

	precision	recall	f1-score	support
1	1.00	0.89	0.94	37
2	0.88	1.00	0.94	29
3	1.00	1.00	1.00	25
4	1.00	1.00	1.00	17
accuracy			0.96	108
macro avg	0.97	0.97	0.97	108
weighted avg	0.97	0.96	0.96	108

Figure 5.3.2: Decision Tree Classification Report

Figure 5.3.2 provides the Decision Tree model's classification report, which summarizes its performance indicators.

Random Forest Classification Report:

	precision	recall	f1-score	support
1	1.00	0.95	0.97	37
2	0.94	1.00	0.97	29
3	1.00	1.00	1.00	25
4	1.00	1.00	1.00	17
accuracy			0.98	108
macro avg	0.98	0.99	0.98	108
weighted avg	0.98	0.98	0.98	108

Figure 5.3.3: Random Forest Classification Report

Figure 5.3.3 presents the Random Forest model's classification report, which includes the precision, recall and F1-score.

AdaBoost Classification Report:

	precision	recall	f1-score	support
1	0.89	0.65	0.75	37
2	0.64	0.86	0.74	29
3	0.96	0.96	0.96	25
4	1.00	1.00	1.00	17
accuracy			0.83	108
macro avg	0.87	0.87	0.86	108
weighted avg	0.86	0.83	0.83	108

Figure 5.3.4: AdaBoost Classification Report

Figure 5.3.4 provides the classification report for the AdaBoost model, summarizing its performance metrics.

K-Nearest Neighbors (KNN) Classification Report:

	precision	recall	f1-score	support
1	1.00	0.86	0.93	37
2	0.72	0.90	0.80	29
3	0.87	0.80	0.83	25
4	1.00	1.00	1.00	17
accuracy			0.88	108
macro avg	0.90	0.89	0.89	108
weighted avg	0.90	0.88	0.88	108

Figure 5.3.5: KNN Classification Report

Figure 5.3.5 presents the classification report for the KNN model, detailing its precision, recall, and F1-score.

```
Naive Bayes Classification Report:
              precision    recall  f1-score   support

           1       0.94      0.84      0.89        37
           2       0.75      0.93      0.83        29
           3       1.00      0.56      0.72        25
           4       0.68      1.00      0.81        17

    accuracy                           0.82       108
   macro avg       0.84      0.83      0.81       108
weighted avg       0.86      0.82      0.82       108
```

Figure 5.3.6: Naïve Bayes Classification Report

Figure 5.3.6 provides the classification report for the Naïve Bayes model, summarizing its performance metrics.

```
Linear Discriminant Analysis Classification Report:
              precision    recall  f1-score   support

           1       0.93      0.76      0.84        37
           2       0.77      0.83      0.80        29
           3       0.86      0.96      0.91        25
           4       0.89      1.00      0.94        17

    accuracy                           0.86       108
   macro avg       0.86      0.89      0.87       108
weighted avg       0.87      0.86      0.86       108
```

Figure 5.3.7: LDA Classification Report

Figure 5.3.7 presents the classification report for the LDA model, detailing its precision, recall, and F1-score.

Gradient Boosting Classification Report:

	precision	recall	f1-score	support
1	1.00	0.89	0.94	37
2	0.88	1.00	0.94	29
3	1.00	1.00	1.00	25
4	1.00	1.00	1.00	17
accuracy			0.96	108
macro avg	0.97	0.97	0.97	108
weighted avg	0.97	0.96	0.96	108

Figure 5.3.8: Gradient Boosting Classification Report

Figure 5.3.8 provides the classification report for the Gradient Boosting model, summarizing its performance metrics.

5.4 Summary

In summary, this chapter presented the experimental results and performance analysis of the machine learning models developed for COPD risk assessment. The Random Forest model proved to be the best performing and achieving the highest accuracy and robustness. The comparison study helped determine which model was optimal for deployment by illuminating the advantages and disadvantages of each model. The results and analysis demonstrated the capabilities of algorithms for machine learning in improving COPD risk prediction and highlighted areas for future improvement.

CHAPTER 6

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

6.1 Impact on Life

A progressive lung illness marked by lasting respiratory symptoms and reduced airflow, encompassing disorders like emphysema and chronic bronchitis are called Chronic Obstructive Pulmonary illness. In our day of sophisticated technology, continuous exposure to irritants like cigarette smoke is a common cause of COPD, which is rising. The research on "Predictive Modeling and Risk Assessment of COPD Using a Machine Learning Approach" promises substantial consequences on society and healthcare. By applying machine learning algorithms, this research intends to better the accuracy and efficiency of COPD diagnosis and risk assessment, enabling healthcare practitioners to make more informed decisions.

Accurate and prompt prediction of COPD can transform medical diagnostics, leading to improved patient outcomes and societal benefits. This research can improve patient care, eliminate misdiagnoses, and speed treatment interventions, ultimately decreasing COPD-related problems, hospitalizations, and healthcare expenditures. The focus on data-driven models and comparative analysis of diverse techniques boosts the practicality of the study, supporting the creation of optimum diagnostic and risk assessment solutions for real-world healthcare settings. Moreover, enhanced COPD prediction approaches might assist underprivileged populations and places with limited healthcare resources, making sophisticated diagnostic capabilities more accessible and inexpensive. This coincides with global health programs targeted at lowering health inequalities and enhancing healthcare equity.

In summary, our discovery offers the potential to usher in a new era of precision medicine for COPD diagnosis and management, expanding scientific understanding and boosting public health outcomes, healthcare accessibility, and general well-being globally.

6.2 Impact on Society and Environment

While the major focus of the research on "Predictive Modeling and Risk Assessment of COPD Using a Machine Learning Approach" is on healthcare and technology breakthroughs, its environmental impact is equally notable. The integration of machine learning algorithms, vital for advancements in medical diagnosis, demands intense computer operations. These operations, notably the training and optimization of sophisticated models, demand enormous computer power, typically depending on energy-intensive technology like graphics processing units (GPUs).

The environmental effect derives from the energy consumption required with training and deploying these sophisticated models. The high computing needs, especially with huge datasets and complicated model designs, lead to higher energy usage, therefore leaving a carbon footprint. This research stimulates a consideration on the environmental consequences of adopting state-of-the-art technologies in healthcare.

However, breakthroughs in hardware efficiency, cloud computing tactics, and sustainable computing practices are always improving. Researchers and developers are becoming aware of the environmental risks connected with AI applications, leading to efforts to improve algorithms and hardware for energy efficiency. The incorporation of green computing methods and the utilization of renewable energy sources in data centers can lessen the environmental effect, integrating technical advancement with eco-friendly concerns.

In summary, while the fundamental contributions of this study lay in enhancing medical diagnoses, its environmental effect is a vital concern. As the area evolves, attempts to combine technical innovation with environmental sustainability will be crucial in ensuring that healthcare innovations remain socially responsible and ecologically mindful.

6.3 Ethical Aspects

The research on "Predictive Modeling and Risk Assessment of COPD Using a Machine Learning Approach" is anchored by a dedication to ethical issues throughout its approach. The use of patient data implies a responsibility to maintain privacy and confidentiality, stressing adherence to ethical rules and regulatory criteria. Informed permission, a cornerstone of ethical research methods, is adequately acquired when necessary. The potential social impact of the study on healthcare underlines the significance of ethical deployment, ensuring that improvements in predictive diagnostics translate to enhanced patient care without sacrificing human rights. The ethical dimension is further expressed in the thorough recording of the research process, enabling examination, reproducibility, and responsible information distribution. Additionally, it is vital to examine and reduce any biases in the machine learning models to ensure fair and equitable healthcare results. Overall, ethical issues are important to the research's commitment to societal well-being and the proper deployment of cutting-edge technology in the healthcare area.

6.4 Sustainability Plan

The sustainability strategy for the research on "Predictive Modeling and Risk Assessment of COPD Using a Machine Learning Approach" is meant to reduce environmental effect while assuring continuous beneficial social benefits. Efforts will be aimed towards enhancing computational efficiency, adopting energy-efficient techniques, and researching cloud computing alternatives to minimize total energy consumption related with model training and deployment. Furthermore, the research team is devoted to adopting eco-friendly approaches in data management and storage. The approach promotes continuing awareness and incorporation of green computing concepts, combining technical breakthroughs with environmental sustainability. Continuous review and adaption of sustainable strategies will be included in the research process, showing a dedication to responsible and eco-conscious scientific innovation. This strategy not only helps the development of better diagnostic tools but also ensures that such progress is done in an ecologically friendly way.

6.5 Summary

The research on "Predictive Modeling and Risk Assessment of COPD Using a Machine Learning Approach" has significant implications for healthcare, society, and the environment. By leveraging advanced machine learning techniques, it intends to increase the accuracy and efficiency of COPD diagnosis and risk assessment, leading to better patient outcomes and less load on healthcare systems. These AI-driven tools can facilitate early detection and ensure equitable access to diagnostics, particularly in underserved communities, while also potentially reducing healthcare costs and stimulating innovation and job creation.

Ethical considerations are integral, focusing on patient privacy, informed consent, transparency, and mitigating biases. The research also addresses environmental impacts by optimizing computational efficiency, adopting energy-efficient algorithms, and implementing eco-friendly practices in data management. Overall, this research promises to revolutionize healthcare delivery, promote health equity, stimulate socio-economic development, and ensure environmentally responsible scientific innovation.

CHAPTER 7

CONCLUSION AND FUTURE WORK

7.1 Conclusion

The goal of the research is to develop predictive models for COPD risk assessment using machine learning algorithms. The study successfully identified and implemented several models, utilizing the Random Forest algorithm having the maximum accuracy and robustness. The models will be integrated into a user-friendly risk assessment tool, meant to aid healthcare professionals and patients in earlier detection and individualized management of COPD.

The research highlighted the potential of machine learning in enhancing COPD risk prediction and management, particularly in resource-limited settings like Bangladesh. By leveraging comprehensive datasets and advanced algorithms, the study provided valuable insights into key risk factors and their interactions.

7.2 Further Suggested Works

While this research achieved significant milestones, several areas for future work were identified. Expanding the dataset to include more diverse and larger populations can improve model generalizability and accuracy. Incorporating additional data sources such as wearable device data and real-time monitoring can enhance the predictive capabilities of the models.

Future research can also explore the integration of deep learning techniques, which may provide even higher accuracy and deeper insights into complex patterns within the data. Additionally, developing more sophisticated and interpretable models can address ethical concerns and promote trust and adoption among healthcare professionals.

7.3 Limitations/ Conflict of Interests

The research faced limitations such as the availability and quality of data, particularly from rural areas. Potential biases in the dataset due to socio-economic and cultural differences were acknowledged. Computational resource constraints also limited the extent of model training and evaluation.

No conflicts of interest were identified in the conduct of this research. The study was conducted with the sole aim of advancing knowledge and improving COPD risk assessment.

References:

[1] J. L. M. Amaral, A. J. Lopes, J. M. Jansen, A. C. D. Faria, and P. L. Melo, "Machine learning algorithms and forced oscillation measurements applied to the automatic identification of chronic obstructive pulmonary disease," *Computer Methods and Programs in Biomedicine*, vol. 105, no. 3, pp. 183–193, Mar. 2012, doi: 10.1016/j.cmpb.2011.09.009.

[2] J. Peng, C. Chen, M. Zhou, X. Xie, Y. Zhou, and C.-H. Luo, "A Machine-learning Approach to Forecast Aggravation Risk in Patients with Acute Exacerbation of Chronic Obstructive Pulmonary Disease with Clinical Indicators," Scientific Reports, vol. 10, no. 1, Feb. 2020, doi: 10.1038/s41598-020-60042-1.

[3] D. S. Alam, M. A. Chowdhury, A. T. Siddiquee, S. Ahmed, and J. D. Clemens, "Prevalence and determinants of Chronic Obstructive Pulmonary Disease (COPD) in Bangladesh.," PubMed, vol. 12, no. 6, pp. 658–67, Jan. 2015, doi: 10.3109/15412555.2015.1041101.

[4] Mridul Das Joshe, Nazmul Hassan Emon, M. Islam, Nushrat Jahan Ria, M. Masum, and Sheikh, "Symptoms Analysis Based Chronic Obstructive Pulmonary Disease Prediction In Bangladesh Using Machine Learning Approach," Jul. 2021, doi: https://doi.org/10.1109/icccnt51525.2021.9580078.

[5] R. S. R. Biswas and F. K. Chowdhury, "Risk factors of chronic obstructive pulmonary disease among rural women, Chittagong, Bangladesh," ˜ the œEgyptian Journal of Bronchology, vol. 11, no. 3, pp. 188–192, Jul. 2017, doi: 10.4103/1687-8426.211395.

[6] I. Sutradhar, R. D. Gupta, M. Hasan, A. Wazib, and M. Sarker, "Prevalence and risk factors of Chronic Obstructive pulmonary disease in Bangladesh: a systematic review," Curēus, Jan. 2019, doi: 10.7759/cureus.3970.

[7] S. Swaminathan et al., "A machine learning approach to triaging patients with chronic obstructive pulmonary disease," PloS One, vol. 12, no. 11, p. e0188532, Nov. 2017, doi: 10.1371/journal.pone.0188532.

[8] I. Mohamed, M. M. Fouda, and K. M. Hosny, "Machine Learning Algorithms for COPD Patients Readmission Prediction: A Data Analytics approach," IEEE Access, vol. 10, pp. 15279–15287, Jan. 2022, doi: 10.1109/access.2022.3148600.

[9] A. Badnjevic, L. Gurbeta, and E. Custovic, "An expert diagnostic system to automatically identify asthma and chronic obstructive pulmonary disease in clinical settings," Scientific Reports, vol. 8, no. 1, Aug. 2018, doi: 10.1038/s41598-018-30116-2.

[10] D. Sanchez-Morillo, M. A. Fernandez-Granero, and A. Leon-Jimenez, "Use of predictive algorithms in-home monitoring of chronic obstructive pulmonary disease and asthma," Chronic Respiratory Disease, vol. 13, no. 3, pp. 264–283, Jun. 2016, doi: 10.1177/1479972316642365.

[11] S. Alamgir, Md. A. S. Talukder, D. S. Hossain, F. Afroz, Mst. S. A. Begom, and M. Mehjabin, “The clinical and spirometric profiles and staging of COPD, asthma in smokers and nonsmokers, Bangladesh,” Scholars Journal of Applied Medical Sciences, vol. 9, no. 9, pp. 1332–1335, Sep. 2021, doi: 10.36347/sjams.2021.v09i09.002.
[12] D. Spathis and P. Vlamos, “Diagnosing asthma and chronic obstructive pulmonary disease with machine learning,” Health Informatics Journal, vol. 25, no. 3, pp. 811–827, Aug. 2017, doi: 10.1177/1460458217723169.
[13] T. Goto, C. A. Camargo, M. K. Faridi, B. J. Yun, and K. Hasegawa, “Machine learning approaches for predicting disposition of asthma and COPD exacerbations in the ED,” ˜ the œAmerican Journal of Emergency Medicine, vol. 36, no. 9, pp. 1650–1654, Sep. 2018, doi: 10.1016/j.ajem.2018.06.062.
[14] C.-T. Wu et al., “Acute exacerbation of a chronic obstructive pulmonary disease Prediction system using wearable device data, machine learning, and deep learning: Development and cohort study,” JMIR Mhealth and Uhealth, vol. 9, no. 5, p. e22591, May 2021, doi: 10.2196/22591.
[15] S. Zeng, M. Arjomandi, Y. Tong, Z. C. Liao, and G. Luo, “Developing a machine learning model to predict severe chronic obstructive pulmonary disease exacerbations: Retrospective cohort study,” JMIR. Journal of Medical Internet Research/Journal of Medical Internet Research, vol. 24, no. 1, p. e28953, Jan. 2022, doi: 10.2196/28953.
[16] World Health Organization: WHO and World Health Organization: WHO, “Chronic obstructive pulmonary disease (COPD),” Mar. 16, 2023. https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd)
[17] “Global Initiative for Chronic Obstructive Lung Disease - Global Initiative for Chronic Obstructive Lung Disease - GOLD,” Global Initiative for Chronic Obstructive Lung Disease - GOLD, May 07, 2024. https://goldcopd.org/

[18] “What is COPD? | NHLBI, NIH,” NHLBI, NIH, Oct. 25, 2023. https://www.nhlbi.nih.gov/health/copd
[19] “Global health estimates.” https://www.who.int/data/global-health-estimates
[20] “GHO.” https://www.who.int/data/gho
[21] A. Rajkomar, J. Dean, and I. Kohane, “Machine learning in medicine,” New England Journal of Medicine/˜ the œNew England Journal of Medicine, vol. 380, no. 14, pp. 1347–1358, Apr. 2019, doi: 10.1056/nejmra1814259.
[22] Md. Kamruzzaman Khan, et al. (2021). Prevalence and Risk Factors of Chronic Obstructive Pulmonary Disease in Bangladesh: A Systematic Review. Cureus, 13(11), e18322. PubMed
[23] Bangladesh Pulmonary & Critical Care Society. (2020). Bangladesh guidelines for chronic obstructive pulmonary disease. Bangladesh Pulmonary & Critical Care Society

[24] Mohammad Shahedul Alam, et al. (2016). Prevalence and Determinants of Chronic Obstructive Pulmonary Disease (COPD) in Bangladesh. COPD, 13(4), 382-389.
[25] S. Ter Stal, "Look who's talking: appearance of embodied conversational agents in eHealth," 2021. doi: 10.3990/1.9789036551267.

Appendix A
Complex Engineering Problems (EP) and Complex Engineering Activities (EA) Analysis

Title: Predictive Modeling and Risk Assessment For COPD Using A Machine Learning Approach

Students ID: 202-15-3831
202-15-3846

CO Description for FYDP

CO	CO Descriptions	PO
Phase -I		
CO1	Integrate recently gained and previously acquired knowledge to identify a **COPD prediction** problem for the Final Year Design Project (FYDP)	**PO1**
CO2	Analyze different aspects of the goals in designing a solution for this FYDP	**PO2**
CO3	Explore diverse problem domains through a literature review, delineate the issues, and establish these goals for the FYDP	**PO4**
CO4	Perform economic evaluation and cost estimation and employ suitable project management procedures throughout the development life cycle of the FYDP	**PO11**
Phase -II		
CO5	Design and develop technical solutions and system components or processes that meet specified requirements, ensuring compliance with public health and safety standards, as well as considering cultural, socioeconomic, and environmental factors in this FYDP	**PO3**
CO6	Choose and apply appropriate methodologies, resources, and contemporary engineering and IT technologies to address complex engineering processes, encompassing prediction and modeling, while adhering to relevant constraints in this FYDP	**PO5**
CO7	Analyze societal, health, safety, legal, and cultural considerations, along with associated responsibilities, in the context of professional engineering practice and the resolution of this problem, employing logical reasoning guided by contextual understanding.	**PO6**
CO8	Comprehend and evaluate the enduring sustainability and impact of professional engineering endeavors in addressing intricate engineering challenges within social and environmental frameworks.	**PO7**
CO9	Implement ethical principles and adhere to professional standards and norms in this FYDP	**PO8**

CO10	Capable of operating proficiently both individually and as a team member or leader across diverse teams and interdisciplinary settings in this FYDP.	**PO9**
CO11	Proficiently communicate with the engineering community and broader society regarding complex engineering endeavors, including the ability to comprehend and generate comprehensive reports and design documentation, as well as provide and receive clear instructions throughout this FYDP.	**PO10**
CO12	Acknowledge the importance of self-directed and life-long learning within the evolving landscape of technology and possess the readiness and capability to engage in lifelong learning endeavors.	**PO12**

Addressing CO (1 to 12), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP), and Attainment of Complex Engineering Activities (EA)

Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP):

SN	EP Definition	Attain-ment	CO	Justification (with Knowledge Profile)	References
1.	**EP1: Depth of Knowledge Required**	Yes	CO1, CO2, CO3, CO5, CO6, CO7 and CO8	This research illustrates basic engineering principles by applying data augmentation techniques and a range of Machine Learning models for data analysis and classification tasks. The research displays **knowledge (K4)** by using transfer learning approaches employing complex machine learning models such as Logistic Regression, Decision Tree Classifier, Random Forest, Linear Discriminant Analyzer, and Ada Boost. The research employs **engineering concepts and design (K5)** through the execution of a systematic technique of experimentation. The research focuses on the application of machine learning models to engineering practice and technology in the K6 sector.	**Page no:** [9-21] **Section:** [3] **Page no:** [22-29] **Section:** [4] **Page no:** [30-43] **Section:** [5]
				The project uses **engineering practice & design (K5**) using the figure of process of experimentation. The project covers **engineering practice & technology (K6)** by utilizing Machine Learning methodology.	**Page no:** [9-11] **Section:** [3.2(B)] **Page no:** [20] **Section:** [3.4]
				This project ensures to **K8 (Research Literature)** by synthesizing insights from recent studies, to predict COPD using machine learning, showcasing a comprehensive understanding of current methodologies.	**Page no:** [4-7] **Section:** [2.2, 2.3]

2.	**EP2: Range of Conflicting Requirements**	Yes	CO2, and CO7	This project addresses **EP-2** by recognizing challenges in COPD prediction, such as the limitations of traditional diagnostics and the complexities of integrating advanced machine learning algorithms. Through comparative analysis, it refines predictive models, improving diagnostic accuracy and healthcare interventions for better patient outcomes in COPD management.	**Page no:** [30-43] **Section:** [5]
3.	**EP3: Depth of analysis requir ed**	Yes	CO2, and CO6	This project addresses **EP-3** by comparing experimental outcomes, highlighting advanced machine learning algorithms as the key solution for improving COPD prediction among various approaches.	**Page no:** [22-25] **Section:** [4.2] **Page no:** [30-38] **Section:** [5.2, 5.4]
4.	**EP4: Familiarity of Issues**	Yes	CO8	This project's multidisciplinary approach extends beyond computer science and engineering, influencing medical diagnoses in respiratory disorders like COPD, contributing to breakthroughs in public health and healthcare practices which suggests **EP-4.**	**Page no:** [22-29] **Section:** [4] **Page no:** [30-43] **Section:** [5]
5.	**EP5: Extends of application codes**	No	CO5	N/A	N/A
6.	**EP6: Extends of stakeholders involved and conflicting requirements**	No	CO8	N/A	N/A

7.	**EP7: Interdependence**	Yes	CO5	This project's comprehensive approach addresses high-level concerns by integrating numerous components across data collecting, statistical analysis, and proposed technique, guaranteeing a holistic solution to challenging challenges in deepfake detection which ensures **EP-7.**	**Page no:** [22-29] **Section:** [4]

Addressing CO11 with Complex Engineering Activities (EA) [Some or all of the following]:

SN	EA Definition	Attainment	CO	Justification	References
1.	**EA1: Range of resources**	Yes	CO11	Our study employs varied resources such as high-performance computing infrastructure, GPUs, machine learning techniques, annotated datasets, and ethical concerns to ensure systematic research and contribute to advancements in COPD Prediction through machine learning methods.	**Page no:** [20-21] **Section:** [3.5]
2.	**EA2: Level of interaction**	No		N/A	N/A
3.	**EA3: Innovation**	No		N/A	N/A

4.	**EA4: Consequences for society and the environment**	Yes		This project enhances healthcare by improving COPD prediction with advanced machine learning, promoting sustainability through efficient computation and adhering to patient data privacy guidelines.	**Page no:** [30-43] **Section:** [5.1, 5.2, 5.4]
5.	**EA-5: Familiarity**	Yes		This project expands existing research by exploring advanced machine learning for COPD prediction, demonstrated through comprehensive analysis and novel approaches, offering new insights into the field.	**Page no:** [4-7] **Section:** [2.1, 2.3]

Addressing CO (4, 9, 10, and 12):

SN	COs	Attainment	Justification	References
1	CO4	Yes	This project addresses **CO4** by combining good project management and financial control, assuring thorough planning, resource allocation, and budget estimation for optimal resource use throughout the study lifecycle.	**Page no:** [3] **Section:** [1.6]
2	CO9	Yes	The project upholds ethical standards by prioritizing patient privacy, securing informed consent, and transparently documenting the research process. This ensures responsible knowledge dissemination and societal well-being through the ethical application of advanced healthcare technologies, meeting CO9 requirements.	**Page no:** [39-43] **Section:** [5.3]
3	CO10	No	N/A	N/A
4	CO12	Yes	The project exemplifies CO12 by embracing continuous learning and adaptation in a dynamic technological landscape. Its extensive data collection, rigorous statistical analysis, meticulous methodology, and thorough experimental results showcase a commitment to staying current and refining techniques to tackle modern challenges.	**Page no:** [9-21, 22-27] **Section:** [3.2, 3.3, 3.4, 3.5, 4.2]

PREDICTIVE MODELING AND RISK ASSESSMENT FOR COPD USING A MACHINE LEARNING APPROACH.

ORIGINALITY REPORT

19%	14%	4%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	9%
2	Submitted to Huntington Beach Union High School District Student Paper	3%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
4	era.library.ualberta.ca Internet Source	<1%
5	worldwidescience.org Internet Source	<1%
6	www.researchgate.net Internet Source	<1%
7	docshare.tips Internet Source	<1%
8	Melissa N. Galea Holmes, John A. Weinman, Lindsay M. Bearne. "Are Walking Treatment Beliefs and Illness Perceptions Associated	<1%