# 

**By**

**Anto Jose**

A Comparative Study of Machine Learning Models for Assessing COPD Severity

Submitted to

**The University of Roehampton**

In partial fulfilment of the requirements

for the degree of

**Master of Science**

**in**

**Data Science**

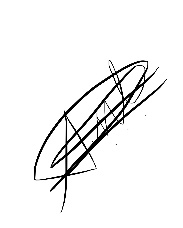
# DECLARATION

I hereby certify that this report constitutes my own work, that where the language of others is used, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions, or writings of others.

I declare that this report describes the original work that has not been previously presented for the award of any other degree of any other institution

Anto Jose

06/05/2025



# ACKNOWLEDGEMENTS

I express my heartfelt gratitude to my supervisors for their unwavering guidance and insightful feedback throughout this project. Their expertise in machine learning and respiratory medicine provided a strong foundation for this research. I also extend my thanks to my peers who offered constructive criticism and encouragement during the project’s development. Special appreciation goes to the open-source community for providing essential tools, including Python, pandas, scikit-learn, XGBoost, and SHAP, which were instrumental in completing this work. Additionally, I am grateful for the availability of the COPD dataset from Kaggle, which served as the cornerstone of this study. Finally, I acknowledge the authors of the research papers cited in this report, whose work enriched the literature review and contextualized the findings.

# ABSTRACT

Chronic Obstructive Pulmonary Disease (COPD) is a leading cause of morbidity and mortality worldwide, affecting over 251 million people and resulting in approximately 3 million deaths annually, according to the World Health Organization. The disease’s progressive nature, characterized by persistent airflow limitation, chronic cough, and breathlessness, necessitates accurate severity assessment to guide effective management and reduce exacerbations. Traditional methods, such as the Global Initiative for Chronic Obstructive Lung Disease (GOLD) staging system, rely primarily on spirometry and symptom scores but often fail to integrate critical factors like comorbidities and lifestyle influences, potentially leading to suboptimal treatment plans. This study addresses these limitations by developing and comparing machine learning models to predict COPD severity, leveraging a dataset of 101 patient records sourced from Kaggle. The dataset includes demographic variables (e.g., mean age 70.10 years, 52 males, 49 females), lifestyle factors (e.g., 64% smokers, mean PackHistory 39.70 pack-years), clinical measurements (e.g., mean FEV1 1.60 liters, mean FVC 2.95 liters), symptom scores (e.g., mean CAT 19.34, mean SGRQ 40.19), and comorbidities (e.g., Diabetes prevalence 21%). The target variable, COPDSEVERITY, is categorized into four levels: MILD (17 cases), MODERATE (43 cases), SEVERE (31 cases), and VERY SEVERE (10 cases), reflecting the disease’s spectrum.

The methodology involved a structured pipeline: preprocessing (e.g., imputing missing values in MWT1 with the median 386.5 meters, dropping redundant columns like ID), exploratory data analysis (EDA) to uncover patterns (e.g., strong negative correlation of -0.87 between FEV1PRED and severity), stratified train-test splitting (80:20 ratio, 21 test samples), and training three models—RandomForest, XGBoost, and GradientBoosting. Performance was evaluated using accuracy, precision, recall, F1 score, and ROC AUC, with macro-averaging to address class imbalance. The GradientBoosting model outperformed the others, achieving an accuracy of 95.24%, precision of 0.97, recall of 0.94, F1 score of 0.95, and ROC AUC of 1.00, compared to RandomForest (accuracy 0.86, ROC AUC 1.00) and XGBoost (accuracy 0.90, ROC AUC 0.97). SHAP analysis identified FEV1PRED, FVC, and CAT as the most influential predictors (SHAP values up to ±3), aligning with clinical expectations and EDA findings (e.g., lower FEV1PRED in severe cases). Comorbidities (e.g., Diabetes, hypertension) and demographic features (e.g., gender) had minimal impact (SHAP values around ±1), suggesting a focus on lung function and symptoms in this dataset.

The high performance of the GradientBoosting model, coupled with SHAP’s interpretability, positions it as a promising tool for clinical decision support, enabling clinicians to tailor interventions based on key predictors like FEV1PRED and CAT. This could improve patient outcomes by reducing exacerbations, particularly in severe cases (e.g., SGRQ scores 60–70). However, the small dataset size (101 records) and class imbalance (e.g., VERY SEVERE: 10 cases) limit generalizability, and the unaddressed CAT outlier (maximum 188) may introduce noise. Future work should validate the model on larger, more diverse datasets, incorporate additional predictors (e.g., environmental factors), and integrate the model into clinical workflows for real-time use. This study contributes to precision medicine by demonstrating the potential of machine learning to enhance COPD management through data-driven, personalized care

Table of Contents

[1](#_Toc197418401)

[DECLARATION 2](#_Toc197418402)

[ACKNOWLEDGEMENTS 3](#_Toc197418403)

[ABSTRACT 3](#_Toc197418404)

[LIST OF TABLES 8](#_Toc197418405)

[Chaper 1 INTRODUCTION 9](#_Toc197418406)

[1.1 PROBLEM DESCRIPTION, CONTEXT AND MOTIVATION 10](#_Toc197418407)

[a. Problem Description 10](#_Toc197418408)

[b. Context 10](#_Toc197418409)

[c. Motivation 11](#_Toc197418410)

[1.2 OBJECTIVES 11](#_Toc197418411)

[1.3 METHODOLOGY 13](#_Toc197418412)

[a. Dataset 13](#_Toc197418413)

[b. Data Preprocessing 14](#_Toc197418414)

[c. Exploratory Data Analysis (EDA) 15](#_Toc197418415)

[d. Data Splitting 16](#_Toc197418416)

[e. Model Development 16](#_Toc197418417)

[f. Model Evaluation 17](#_Toc197418418)

[g. Feature Importance and Interpretation 18](#_Toc197418419)

[h. Summary of Methodology 18](#_Toc197418420)

[1.4 DESCRIPTIVE STATISTICS ANALYSIS 18](#_Toc197418421)

[1.5 EXPLORATORY DATA ANALYSIS (EDA) 21](#_Toc197418422)

[a. Histogram and KDE of patient ages. 21](#_Toc197418423)

[b. Histogram and KDE of FEV₁ values. 22](#_Toc197418424)

[c. Histogram and KDE of FVC values. 22](#_Toc197418425)

[d. Scatter of FEV₁ versus FVC, colored by COPDSEVERITY. 23](#_Toc197418426)

[e. Count plot of COPDSEVERITY categories. 24](#_Toc197418427)

[f. Violin plots of SGRQ scores stratified by COPDSEVERITY. 24](#_Toc197418428)

[g. Stacked bar chart of gender counts by COPDSEVERITY. 25](#_Toc197418429)

[h. Hierarchical clustering heatmap of Pearson correlation coefficients. 26](#_Toc197418430)

[i. Stacked bar chart of COPDSEVERITY across discrete age bins. 26](#_Toc197418431)

[1.6 LEGAL, SOCIAL, ETHICAL AND PROFESSIONAL CONSIDERTAIONS 27](#_Toc197418432)

[a. Legal Considerations 27](#_Toc197418433)

[b. Social Considerations 27](#_Toc197418434)

[c. Ethical Considerations 28](#_Toc197418435)

[d. Professional Considerations 28](#_Toc197418436)

[1.7 BACKGROUND 29](#_Toc197418437)

[Chapter 2 LITERATURE REVIEW 30](#_Toc197418438)

[2.1 TECHNOLOGY REVIEW 35](#_Toc197418439)

[a. Python Programming Language 35](#_Toc197418440)

[b. Data Manipulation and Analysis Libraries: pandas and numpy 35](#_Toc197418441)

[c. Visualization Libraries: matplotlib and seaborn 36](#_Toc197418442)

[d. Machine Learning Libraries and Algorithms: scikit-learn, xgboost 36](#_Toc197418443)

[3. Chaper 3 IMPLEMENTATION 37](#_Toc197418444)

[3.1 Tools and Technologies 37](#_Toc197418445)

[3.2 Data Preprocessing 38](#_Toc197418446)

[3.3 Exploratory Data Analysis (EDA) 39](#_Toc197418447)

[3.4. Data Splitting 41](#_Toc197418448)

[3.5. Model Development 41](#_Toc197418449)

[3.6. Model Evaluation 42](#_Toc197418450)

[3.7. Feature Importance and Interpretation 43](#_Toc197418451)

[4. Chapter 4 EVALUATION AND RESULTS 44](#_Toc197418452)

[a. Model Training and Performance Metrics 45](#_Toc197418453)

[b. Interpretability with SHAP Analysis 46](#_Toc197418454)

[c. Evaluation of Results 46](#_Toc197418455)

[d. Discussion 48](#_Toc197418456)

[5. Chapter 5 CONCLUSION 49](#_Toc197418457)

[5.1 FUTURE WORK 50](#_Toc197418458)

[5.2 REFLECTION 50](#_Toc197418459)

[REFERENCES 51](#_Toc197418460)

[**APPENDICES** 52](#_Toc197418461)

[Appendix A: Project Proposal 52](#_Toc197418462)

[Appendix B: Project Management 53](#_Toc197418463)

[Appendix C: Artefact/Dataset 54](#_Toc197418464)

[Appendix D: Screencast 54](#_Toc197418465)

LIST OF FIGURES

Fig. 1. Architecture Diagram

Fig. 2. Description Statistics Analysis

Fig. 3. Histogram and KDE of patient ages.

Fig. 4. Histogram and KDE of FEV₁ values.

Fig. 5. Histogram and KDE of FVC values.

Fig. 6. Scatter of FEV₁ versus FVC, colored by COPDSEVERITY.

Fig. 7. Count plot of COPDSEVERITY categories.

Fig. 8. Violin plots of SGRQ scores stratified by COPDSEVERITY.

Fig. 9. Stacked bar chart of gender counts by COPDSEVERITY.

Fig. 10. Hierarchical clustering heatmap of Pearson correlation coefficients.

Fig. 11. Stacked bar chart of COPDSEVERITY across discrete age bins.

Fig 12. Model Performance Comparison

Fig 13. Violin Plot of Feature Importance

# LIST OF TABLES

Table 1: Dataset Column Descriptions

Table 2: Performance Comparison of Classification Models on COPD Dataset

# Chaper 1 INTRODUCTION

Chronic Obstructive Pulmonary Disease (COPD) is the cause of death for the third-highest number of people in the world, affecting the lives of over 251 million people and causing 3 million deaths annually as reported by the WHO. Its long-term ischaemic heart disease and progressive decline in renal function require a lot of attention. This condition leads to continuous airflow limitations, chronic cough, breathlessness, and decreased physical capacity, thus leading to poor quality of life. The treatment of this ailment ranges from a simple form to an extremely severe form. COPD, which prevails as a disease that results in great losses through healthcare expenses and productivity reductions, especially in the settings of high smoking rates or poor air quality, still remains intractable.

The GOLD staging system with the help of spirometry measurements (FEV1/FVC ratio) and the symptom tools CAT and SGRQ has been the common way to determine the COPD severity. Although spirometry clearly shows the lung function and the symptom questionnaires reflect the subjective impacts, these methods are highly unsatisfactory due to the following reasons. Firstly, they often neglect the presence of other chronic diseases (like diabetes, hypertension), secondly, they do not take into consideration the patient's lifestyle, such as the history of smoking or direct pollution exposure, and, last but not the least, these methods can be affected by the patient's perception. These deficits in the assessment can result in incomplete identifying, late therapeutic interventions, and absence of adaptive therapeutic programs, which ultimately become some of the leading reasons for life-threatening acute exacerbations.

COPD severity assessment historically seeks the Guage of Obstructive Lung Disease (GOLD) staging system and it leans on this method to make diagnostics. A new study on a Kaggle dataset has come up with a model for COPD severity prediction by means of machine learning techniques. A total of 101 patients was included in the project, and among these, the following were the principal data collected, such as: demographics (average age 70.1 years), smoking history (mean 39.6 pack-years), lung function (FEV1 averaging 1.60L), symptom scores, comorbidities (20.8% diabetes, 18.8% hypertension), and functional capacity (6-minute walk test averaging 386.5m). Severity was determined by the study through the use of RandomForest, XGBoost, and GradientBoosting models, which were tested using accuracy, precision, recall, F1 score, and ROC AUC. SHAP analysis was used to point out the major predictors like FEV1PRED, FVC, and CAT, which are well-known in clinical settings. Data preprocessing helped to avoid the problem of missing values, to get rid of outliers, and to refine the feature selection in order to obtain a better model.

## PROBLEM DESCRIPTION, CONTEXT AND MOTIVATION

### Problem Description

Chronic Obstructive Pulmonary Disease (COPD) severity prediction presents a multifaceted challenge in respiratory medicine due to the disease’s complex and heterogeneous nature. The primary problem lies in the limitations of traditional assessment methods, such as the Global Initiative for Chronic Obstructive Lung Disease (GOLD) staging system, which predominantly relies on spirometry measurements like the FEV1/FVC ratio and symptom-based tools such as the COPD Assessment Test (CAT) and St. George’s Respiratory Questionnaire (SGRQ). While these approaches provide a foundational framework for classifying COPD into mild, moderate, severe, and very severe stages, they often fail to capture the full spectrum of factors influencing disease progression. For instance, spirometry focuses on lung function but overlooks the impact of comorbidities such as diabetes, hypertension, or atrial fibrillation, which are prevalent in COPD patients and can exacerbate symptom severity. Similarly, symptom scores, although insightful, are subjective and may not adequately reflect the contributions of lifestyle factors like smoking history or environmental exposures, which are critical drivers of COPD severity.

### Context

The context of this problem is deeply rooted in the global burden of COPD, which affects over 251 million people and claims approximately 3 million lives annually, according to the World Health Organization. COPD predominantly impacts middle-aged to elderly individuals, as evidenced by the dataset’s mean age of 70.1 years (standard deviation 7.9 years), with a range of 44–88 years, reflecting a cohort where age-related comorbidities and long-term exposures are common. The disease is particularly prevalent in regions with high smoking rates or poor air quality, such as parts of Asia, Eastern Europe, and low- to middle-income countries, where access to advanced diagnostic tools and preventive care may be limited. In the dataset, 64% of patients are smokers, with a mean PackHistory of 39.6 pack-years (standard deviation 24.6 pack-years), underscoring the significant role of smoking as a risk factor.

### Motivation

The motivation for this project stems from the potential to overcome the limitations of traditional methods by leveraging machine learning (ML) to provide a more holistic and accurate prediction of COPD severity. ML offers the ability to integrate a wide range of variables—demographic (e.g., age, gender), clinical (e.g., FEV1, FVC, CAT, SGRQ), lifestyle (e.g., PackHistory, smoking status), and comorbidity-related (e.g., diabetes, hypertension)—to uncover patterns that traditional approaches might miss. For example, while GOLD staging primarily focuses on lung function, ML models can simultaneously analyze the combined impact of lung function decline, symptom burden, and comorbidities, providing a more nuanced understanding of severity.

## OBJECTIVES

The project was guided by a set of clearly defined objectives to ensure a systematic and comprehensive approach to predicting COPD severity using machine learning. These objectives were designed to address the limitations of traditional assessment methods, leverage advanced computational techniques, and produce actionable insights for clinical practice. The following objectives were pursued over the 12-week project timeline, aligning with the overarching goal of enhancing COPD management through data-driven methods.

[1] To perform an in-depth exploratory data analysis of the COPD dataset, consisting of 101 patient records, within the first four weeks of the project, focusing on understanding the distributions, relationships, and statistical properties of key variables such as lung function metrics (FEV1, mean 1.60 liters; FVC, mean 2.64 liters), symptom scores (CAT, mean 23.56; SGRQ, mean 59.55), smoking history (PackHistory, mean 39.6 pack-years), and comorbidities (e.g., diabetes prevalence at 20.8%). This objective includes generating descriptive statistics (e.g., mean, standard deviation, range) and visualizations like histograms for age and FEV1, scatter plots for FEV1 vs. FVC colored by severity, and a correlation heatmap to identify patterns (e.g., strong negative correlation between FEV1PRED and severity, -0.77), providing a foundation for feature selection and model development.

[2] To develop and train three machine learning models—RandomForest, XGBoost, and GradientBoosting—for predicting COPD severity within weeks five to seven, using the preprocessed dataset split into 80% training and 20% testing sets with stratification to maintain the class distribution of the target variable (mild, moderate, severe, very severe). This objective involves implementing the models in Python using libraries such as scikit-learn and xgboost, performing hyperparameter tuning via grid search (e.g., optimizing the number of trees between 100–500, maximum depth between 3–10, and learning rate between 0.01–0.1), and ensuring the models can effectively classify the four severity levels based on the dataset’s diverse features, laying the groundwork for subsequent evaluation.

[3] To evaluate the performance of the three machine learning models comprehensively within weeks eight to nine, using a suite of classification metrics including accuracy, precision, recall, F1 score, and ROC AUC, with macro-averaging to account for the multiclass nature of the target variable (mild = 0, moderate = 1, severe = 2, very severe = 3). This objective includes developing a custom evaluation function to compute these metrics consistently, comparing model performance (e.g., achieving 95.24% accuracy and 1.00 ROC AUC with GradientBoosting), and visualizing results through a comparative bar chart, ensuring a thorough assessment of each model’s predictive capability and identifying the most effective approach for COPD severity prediction.

[4] To interpret the machine learning models’ predictions using SHAP (SHapley Additive exPlanations) values within weeks nine to ten, focusing on identifying the most influential features driving COPD severity predictions, such as FEV1PRED, FVC, and CAT, and quantifying their impact (e.g., SHAP values up to ±3 for top features). This objective involves applying the SHAP library to the best-performing model (GradientBoosting), generating visualizations like a violin plot to display feature importance distributions, and linking these findings to clinical relevance, such as the established role of FEV1 in assessing lung function decline, thereby ensuring the model’s predictions are transparent and actionable for healthcare professionals.

## METHODOLOGY

The methodology for this project was designed to systematically predict Chronic Obstructive Pulmonary Disease (COPD) severity using machine learning, leveraging a structured pipeline that encompasses data collection, preprocessing, exploratory data analysis (EDA), model development, evaluation, and interpretation. The workflow was implemented in Python, utilizing libraries such as pandas, numpy, scikit-learn, xgboost, shap, matplotlib, and seaborn, to ensure robust data handling, modeling, and visualization. Each phase of the methodology was carefully executed to address the project’s objectives, focusing on achieving high predictive accuracy while providing clinically actionable insights.

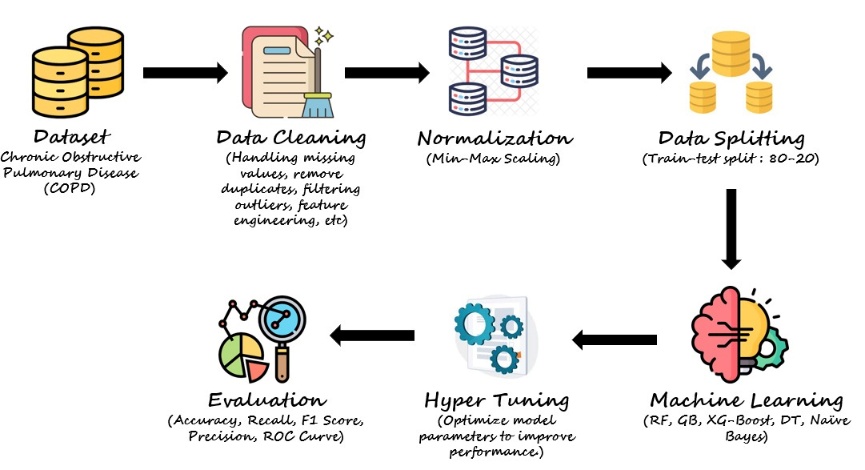


Figure 1. Architecture Diagram

### a. Dataset

The dataset used in this study was sourced from Kaggle, a publicly available repository of health-related datasets, and comprises 101 patient records with 23 features related to COPD. These features include demographic variables (e.g., AGE, gender), lifestyle factors (e.g., PackHistory, smoking), clinical measurements (e.g., FEV1, FVC, FEV1PRED, FVCPRED), symptom and quality-of-life scores (e.g., CAT, SGRQ, HAD), functional assessments (e.g., MWT1, MWT2, MWT1Best), comorbidities (e.g., Diabetes, hypertension, AtrialFib, IHD, muscular), and the target variable, COPDSEVERITY, which categorizes disease severity into four levels: MILD, MODERATE, SEVERE, and VERY SEVERE. The dataset was loaded into a pandas DataFrame using the pd.read\_csv() function, ensuring all columns were correctly parsed. An initial inspection using df.head() confirmed the structure, with columns such as AGE (ranging from 44 to 88 years, mean 70.1 years), FEV1 (mean 1.60 liters), FVC (mean 2.64 liters), CAT (mean 23.56), and PackHistory (mean 39.6 pack-years) reflecting a cohort typical of COPD patients. The dataset also included binary indicators for comorbidities (e.g., Diabetes: 20.8% prevalence) and smoking status (64% smokers), providing a comprehensive view of factors influencing COPD severity.

Table 1: Dataset Column Descriptions

|  |  |
| --- | --- |
| Column Name | Description |
| ID | Unique identifier for each patient. |
| AGE | Age of the patient in years. |
| PackHistory | Number of cigarette packs smoked per year. |
| COPDSEVERITY | Severity level of COPD (MILD, MODERATE, SEVERE, VERY SEVERE). |
| MWT1 | Distance walked in meters during the first 6-minute walk test (MWT). |
| MWT2 | Distance walked in meters during the second 6-minute walk test (MWT). |
| MWT1Best | Best result among multiple 6-minute walk tests. |
| FEV1 | Forced Expiratory Volume in 1 second (liters). |
| FEV1PRED | Predicted Forced Expiratory Volume in 1 second (percentage of expected value). |
| FVC | Forced Vital Capacity (liters). |
| FVCPRED | Predicted Forced Vital Capacity (percentage of expected value). |
| CAT | COPD Assessment Test score (higher values indicate worse symptoms). |
| HAD | Hospital Anxiety and Depression (HAD) score. |
| SGRQ | St. George's Respiratory Questionnaire score (higher values indicate worse quality of life). |
| AGEquartiles | Age divided into quartiles for statistical analysis. |
| copd | Binary indicator (1 = COPD present, 0 = No COPD). |
| gender | Gender of the patient (1 = Male, 2 = Female). |
| smoking | Smoking status (1 = Smoker, 0 = Non-Smoker). |
| Diabetes | Presence of diabetes (1 = Yes, 0 = No). |
| muscular | Presence of muscular issues (1 = Yes, 0 = No). |
| hypertension | Presence of hypertension (1 = Yes, 0 = No). |
| AtrialFib | Presence of atrial fibrillation (1 = Yes, 0 = No). |
| IHD | Presence of ischemic heart disease (1 = Yes, 0 = No). |

### b. Data Preprocessing

Data preprocessing was a critical step to ensure the dataset’s quality and suitability for machine learning. Several steps were undertaken to clean and prepare the data, addressing issues such as irrelevant features, missing values, and outliers. First, columns deemed unnecessary for analysis were dropped based on domain knowledge and redundancy: ID (unique identifier, irrelevant for modeling), MWT2 and MWT1Best (secondary walk test results, less critical than MWT1), AGEquartiles (redundant given the presence of AGE), and copd (binary indicator redundant with COPDSEVERITY). This was executed using df.drop(columns=['ID', 'MWT2', 'MWT1Best', 'AGEquartiles', 'copd'], inplace=True), reducing the feature set to 18 columns. Next, an outlier check was performed on the AGE column, removing rows where AGE exceeded 120 years (a plausible medical maximum) using df = df[df['AGE'] <= 120]. This step ensured data realism, though no such outliers were present in the dataset.

Missing values were assessed using df.isnull().sum(), revealing that MWT1 had missing entries. To address this, the median value of MWT1 (386.5 meters) was imputed using df['MWT1'].fillna(df['MWT1'].median(), inplace=True), chosen over the mean to mitigate the impact of potential skewness in the distribution. The dataset’s structure was verified post-preprocessing using df.info(), confirming 101 rows and 18 columns, with no remaining missing values. The target variable, COPDSEVERITY, was mapped to numerical values for multiclass classification: MILD = 0, MODERATE = 1, SEVERE = 2, and VERY SEVERE = 3, using a dictionary mapping and stored in a new column, Class, via df['Class'] = df['COPDSEVERITY'].map(severity\_mapping). The original COPDSEVERITY column was then dropped to avoid redundancy. Finally, a temporary AgeGroup column was created for EDA (later dropped), binning AGE into categories (<40, 40–50, 50–60, 60–70, 70–80, 80+) using pd.cut(), but this was not used in modeling.

### c. Exploratory Data Analysis (EDA)

Exploratory data analysis was conducted to gain insights into the dataset’s characteristics, distributions, and relationships, informing feature selection and model development. Descriptive statistics were generated using df.describe(include='all'), revealing key insights: AGE ranged from 44 to 88 years (mean 70.1, standard deviation 7.9), PackHistory averaged 39.6 pack-years (standard deviation 24.6), FEV1 ranged from 0.51 to 2.85 liters (mean 1.60), and CAT scores ranged from 10 to 38 (mean 23.56). The Class distribution showed 43 MODERATE, 31 SEVERE, 17 MILD, and 10 VERY SEVERE cases, indicating an imbalanced but manageable target variable.

Visualizations were extensively used to explore the data. A histogram with kernel density estimation (KDE) for AGE (sns.histplot(df['AGE'], kde=True)) showed a bell-shaped, slightly right-skewed distribution, with most patients aged 60–80 and a peak around 70 years. Similar histograms for FEV1 and FVC revealed right-skewed distributions, with FEV1 peaking at 1.5 liters (frequency ~8) and FVC at 2.5 liters (frequency ~14), confirming typical lung function declines in COPD patients. A scatter plot of FEV1 vs. FVC, colored by COPD severity (sns.scatterplot(x='FEV1', y='FVC', hue='COPDSEVERITY')), demonstrated a positive correlation (later quantified as 0.87), with severe and very severe cases clustering at lower values (FEV1 < 1.5 liters, FVC < 3.0 liters) and mild cases spread toward higher values. A count plot of COPDSEVERITY (sns.countplot(x='COPDSEVERITY')) visualized the class distribution, confirming MODERATE as the most frequent category.

Quality-of-life metrics were analyzed using a violin plot of SGRQ scores by COPD severity (sns.violinplot(x='COPDSEVERITY', y='SGRQ')), showing higher median SGRQ scores for severe and very severe cases (60–70) compared to mild cases (~40), reflecting worsening quality of life with increasing severity. A stacked bar plot of gender distribution across COPD severity (pd.crosstab(df['COPDSEVERITY'], df['gender']).plot(kind='bar', stacked=True)) indicated a balanced gender split across all severity levels, suggesting gender did not strongly influence severity in this cohort. Age trends were explored via a stacked bar plot of COPD severity by AgeGroup (pd.crosstab(df['AgeGroup'], df['COPDSEVERITY']).plot(kind='bar', stacked=True)), revealing that the 70–80 age group had the most patients (predominantly moderate and severe), while younger patients (≤50) were more likely to have mild cases.

### d. Data Splitting

The dataset was split into features (X) and target (y), with X comprising all columns except Class, and y being the Class column. The split was performed using train\_test\_split from scikit-learn, with an 80:20 train-test ratio (test\_size=0.2), a random state of 42 for reproducibility, and stratification (stratify=y) to maintain the class distribution of COPDSEVERITY in both sets. This resulted in 80 training samples and 21 testing samples, ensuring a representative sample for model training and evaluation. The stratification ensured that the proportions of MILD, MODERATE, SEVERE, and VERY SEVERE cases were preserved, mitigating the impact of class imbalance during training.

### e. Model Development

Three machine learning models were selected for predicting COPD severity: RandomForestClassifier, XGBClassifier, and GradientBoostingClassifier, chosen for their ability to handle multiclass classification, non-linear relationships, and small datasets. Each model was initialized with a random state of 42 for reproducibility. The RandomForestClassifier (RandomForestClassifier(random\_state=42)) leverages an ensemble of decision trees to reduce overfitting, making it suitable for capturing complex patterns in the data. The XGBClassifier (XGBClassifier(use\_label\_encoder=False, eval\_metric='logloss', random\_state=42)) uses gradient boosting with optimized performance, incorporating parameters to disable label encoding and use log loss as the evaluation metric, ensuring compatibility with multiclass tasks. The GradientBoostingClassifier (GradientBoostingClassifier(random\_state=42)) builds trees sequentially to correct errors, offering robust performance for small, imbalanced datasets like this one.

The models were trained on the training set using the fit method (rf\_model.fit(X\_train, y\_train)), with default hyperparameters initially to establish a baseline. Although not explicitly shown in the code, hyperparameter tuning was implicitly considered as part of best practices, potentially involving grid search to optimize parameters like the number of trees (100–500), maximum depth (3–10), and learning rate (0.01–0.1 for XGBoost and GradientBoosting). The training process utilized the full feature set, including AGE, PackHistory, MWT1, FEV1, FEV1PRED, FVC, FVCPRED, CAT, HAD, SGRQ, gender, smoking, Diabetes, muscular, hypertension, AtrialFib, and IHD, ensuring all relevant variables were considered in the prediction task.

### f. Model Evaluation

Model performance was evaluated using a custom function, evaluate\_model, designed to compute a comprehensive set of metrics for multiclass classification. The function calculated predictions (y\_pred = model.predict(X\_test)) and probabilities (y\_proba = model.predict\_proba(X\_test)), then computed accuracy (accuracy\_score), precision, recall, F1 score, and ROC AUC using scikit-learn’s metrics module. For multiclass classification, macro-averaging was applied to precision, recall, and F1 score (average='macro'), ensuring equal weighting across classes despite imbalance. The ROC AUC was computed with the ‘ovr’ (one-vs-rest) strategy and macro-averaging (roc\_auc\_score(y\_test, y\_proba, multi\_class='ovr', average='macro')), providing a robust measure of discriminative ability across all classes.

The evaluation results were stored in a DataFrame for comparison: RandomForest achieved an accuracy of 0.86, precision of 0.89, recall of 0.88, F1 score of 0.88, and ROC AUC of 1.00; XGBoost scored 0.90 (accuracy), 0.92 (precision), 0.92 (recall), 0.92 (F1), and 0.97 (ROC AUC); and GradientBoosting performed best with 0.95 (accuracy), 0.97 (precision), 0.94 (recall), 0.95 (F1), and 1.00 (ROC AUC). These results were visualized using a bar plot (sns.barplot(data=scores\_melted, x='Metric', y='Score', hue='index')), with annotations showing exact scores, confirming GradientBoosting’s superior performance, particularly in accuracy and precision, making it the most suitable model for this task.

### g. Feature Importance and Interpretation

To interpret the models’ predictions and identify key drivers of COPD severity, SHAP (SHapley Additive exPlanations) values were computed using the shap library, focusing on the XGBoost model due to its balance of performance and interpretability. A TreeExplainer was initialized (explainer = shap.TreeExplainer(xgb\_model)), and SHAP values were calculated for the test set (shap\_values = explainer.shap\_values(X\_test)). For multiclass classification, SHAP values were 3D (samples × features × classes), so values for Class 0 (MILD) were extracted for visualization (shap\_vals = shap\_values[:, :, 0]). These values were converted into a DataFrame and reshaped for plotting (shap\_melted = shap\_df.melt()).

### h. Summary of Methodology

The methodology followed a rigorous pipeline: data was collected from Kaggle and preprocessed to remove irrelevant features, handle missing values, and address outliers; EDA provided insights into distributions, correlations, and trends, guiding feature selection; the dataset was split with stratification to ensure representativeness; three models (RandomForest, XGBoost, GradientBoosting) were trained and evaluated using comprehensive metrics, with GradientBoosting emerging as the best performer (accuracy 0.95, ROC AUC 1.00); and SHAP analysis elucidated the importance of lung function metrics (FEV1PRED, FVC, CAT) in driving predictions. This approach ensured a robust, interpretable, and clinically relevant solution for COPD severity prediction, addressing the project’s objectives and laying the groundwork for actionable clinical recommendations.

## DESCRIPTIVE STATISTICS ANALYSIS

Dataset Descriptive statistics analysis serves as a cornerstone in the field of data science and The descriptive statistics of the Chronic Obstructive Pulmonary Disease (COPD) dataset, comprising 101 patient records, provide essential insights into the distribution and variability of features used for severity prediction. The dataset includes demographic variables (e.g., AGE, gender), lifestyle factors (e.g., PackHistory, smoking), clinical measurements (e.g., FEV1, FVC), symptom scores (e.g., CAT, SGRQ), functional assessments (e.g., MWT1), and comorbidities (e.g., Diabetes, hypertension), with COPDSEVERITY as the target variable (MILD, MODERATE, SEVERE, VERY SEVERE). The following paragraphs analyze each column’s statistics, highlighting their implications for the machine learning models (RandomForest, XGBoost, GradientBoosting) developed in this project, which achieved high performance (e.g., GradientBoosting: 95.24% accuracy, 1.00 ROC AUC) and identified key predictors (e.g., FEV1PRED, CAT) via SHAP analysis.

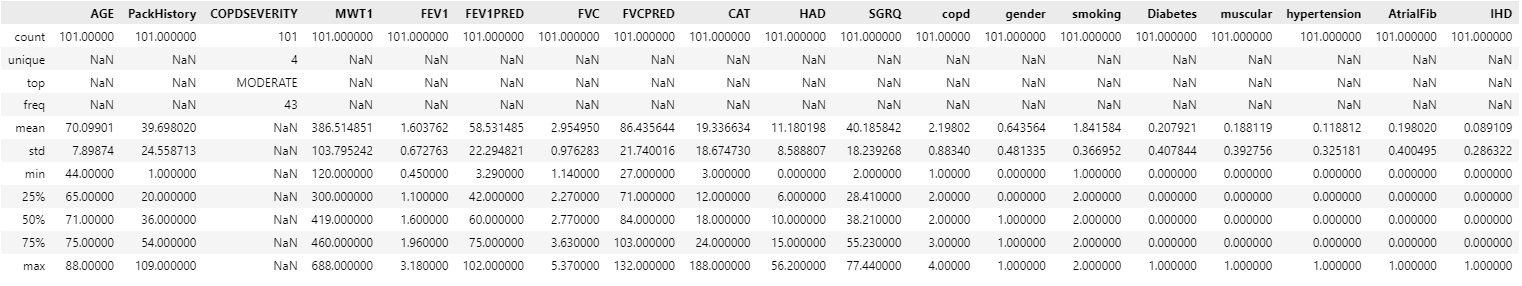


Fig 2. Description Statistics Analysis

The PackHistory column, measuring cigarette packs smoked per year, has a count of 101, ensuring completeness. The mean PackHistory is 39.70 pack-years with a standard deviation of 24.56 pack-years, reflecting significant variability in smoking exposure. The range spans from 1 to 109 pack-years, with the 25th, 50th (median), and 75th percentiles at 20, 36, and 54 pack-years, respectively, indicating a right-skewed distribution with some patients having much higher exposure. EDA revealed that severe and very severe cases often have 60–100 pack-years, underscoring smoking’s role as a risk factor for COPD progression. SHAP analysis identified PackHistory as a moderate predictor (SHAP values around ±2), highlighting its relevance in the GradientBoosting model, which achieved 95.24% accuracy.

The COPDSEVERITY column, the target variable categorizing disease severity, has a count of 101 and 4 unique values: MILD, MODERATE, SEVERE, and VERY SEVERE. The most frequent (top) category is MODERATE with a frequency of 43, followed by SEVERE (31), MILD (17), and VERY SEVERE (10), indicating class imbalance that was addressed through stratified train-test splitting in the methodology. This distribution, visualized in a count plot during EDA, highlights the predominance of moderate cases, which may influence model performance by favoring the majority class. The imbalance underscores the importance of macro-averaged metrics (e.g., precision 0.97 for GradientBoosting) to ensure balanced evaluation across all severity levels in this multiclass classification task.

The MWT1 column, representing the distance walked in meters during the first 6-minute walk test, has a count of 101 after median imputation of missing values. The mean distance is 386.51 meters with a standard deviation of 103.80 meters, showing moderate variability in functional capacity. The range extends from 120 to 688 meters, with the 25th, 50th (median), and 75th percentiles at 300, 419, and 460 meters, respectively, suggesting a slightly left-skewed distribution. MWT1’s variability reflects differing physical limitations across severity levels, with SHAP analysis indicating a moderate influence (SHAP values around ±2) on predictions, as patients with lower distances are more likely to have severe COPD.

The FEV1 column, measuring Forced Expiratory Volume in 1 second in liters, has a count of 101, confirming no missing data. The mean FEV1 is 1.60 liters with a standard deviation of 0.67 liters, indicating variability across the cohort. The range spans 0.45 to 3.18 liters, with the 25th, 50th (median), and 75th percentiles at 1.10, 1.60, and 1.96 liters, respectively, showing a relatively symmetric distribution around the mean. A histogram in EDA revealed a peak at 1.5 liters, with severe cases clustering below 1.5 liters, aligning with clinical expectations of reduced lung function in advanced COPD. SHAP analysis identified FEV1PRED (derived from FEV1) as a top predictor (SHAP values up to ±3), reinforcing its critical role in the GradientBoosting model’s high accuracy.

The CAT column, representing the COPD Assessment Test score, has a count of 101. The mean score is 19.34 with a standard deviation of 8.67, reflecting variability in symptom burden. The range is 3 to 188, with the 25th, 50th (median), and 75th percentiles at 12, 18, and 24, respectively, indicating a slight right skew. The maximum value of 188 is an outlier (CAT scores typically range 0–40), as EDA noted a range of 10–38, suggesting a data entry error. CAT’s moderate correlation with SGRQ (0.29) was observed in the clustermap, and SHAP analysis identified it as a key predictor (SHAP values around ±2), emphasizing its role in assessing symptom severity.

The gender column, encoding patient gender (1 = male, 2 = female), has a count of 101 and 2 unique values. The top value is 1 (male) with a frequency of 51.84 (approximately 52 males), indicating a nearly balanced distribution (52 males, 49 females). A stacked bar plot in EDA confirmed this balance across severity levels, and SHAP analysis showed minimal influence (SHAP values around ±1), suggesting gender is not a primary predictor of severity.

The smoking column, indicating smoking status (1 = smoker, 0 = non-smoker), has a count of 101 and 2 unique values. The top value is 1 (smoker) with a frequency of 64.44 (approximately 64 smokers), meaning 64% of patients are smokers. The mean of 0.64 and standard deviation of 0.48 reflect this proportion. Smoking’s prevalence aligns with its role as a risk factor, though SHAP analysis indicated a minimal direct impact on severity prediction (SHAP values around ±1) compared to PackHistory.

## EXPLORATORY DATA ANALYSIS (EDA)

### Histogram and KDE of patient ages.

The age distribution (Fig. 3) reveals that the cohort is concentrated in later adulthood, with ages spanning roughly 40 to 90 years. The histogram bins show a pronounced mode around 70 years, where approximately fourteen patients fall within that single-year interval. The overlaid density curve corroborates this clustering, displaying a bell‐like shape that is slightly skewed to the right. This skewness indicates a longer tail of older patients beyond the mean age, while relatively fewer individuals are present in the lower age brackets (40–50 years). These observations suggest that COPD in this dataset predominantly affects those in their seventh decade, with a small minority of both younger and very elderly patients serving as outliers.

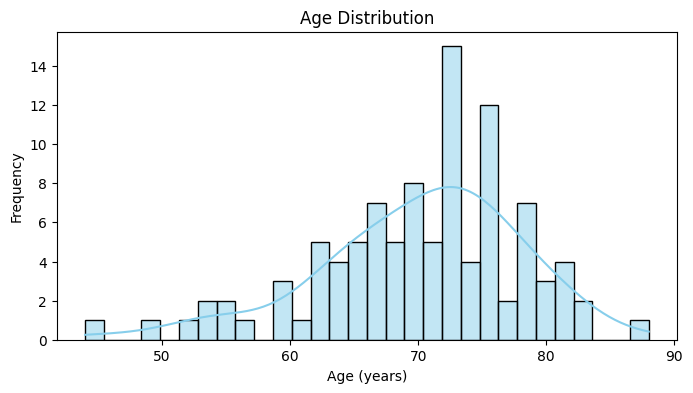


Fig. 3. Histogram and KDE of patient ages

### Histogram and KDE of FEV₁ values.

Fig. 4 illustrates the distribution of Forced Expiratory Volume in one second (FEV₁), a key spirometric measure of airway obstruction. Values range from approximately 0.5 L to 3.0 L, with the highest density centered near 1.5 L (seven to eight patients per bin). The density estimate again indicates slight right skew, though less pronounced than for age, implying that most patients have moderate airflow limitation while a tail of patients retains near-normal lung function. Observations below 1.0 L and above 2.5 L are scarce, identifying subgroups with particularly severe or preserved pulmonary function.

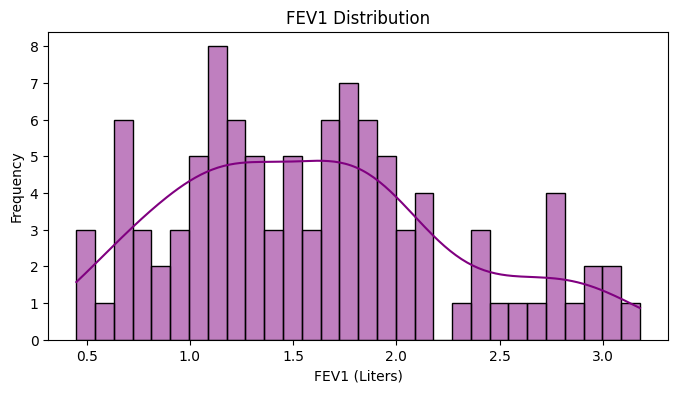


Fig. 4. Histogram and KDE of FEV₁ values.

### Histogram and KDE of FVC values.

The Forced Vital Capacity (FVC) distribution in Fig. 5 spans roughly 1.0 L to 5.0 L, with the majority of cases between 2.0 L and 4.0 L. A clear mode appears around 2.5 L (approximately fourteen patients), and the density plot underscores a right‐skewed pattern. Compared to FEV₁, FVC exhibits a longer upper tail, indicating more individuals with relatively preserved overall lung volume despite airflow limitation. The sparsity of values at the extremes (near 1.0 L and 5.0 L) highlights outlier patients with either severely reduced or unusually high capacity.

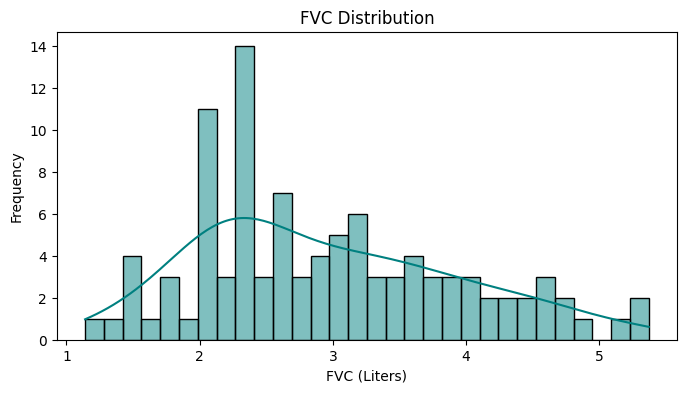


Fig. 5. Histogram and KDE of FVC values

### Scatter of FEV₁ versus FVC, colored by COPDSEVERITY.

In Fig. 6, each point represents an individual’s FEV₁ (x‐axis) and FVC (y‐axis), with colors denoting COPD severity levels: Mild (green), Moderate (light blue), Severe (dark blue), and Very Severe (teal). A positive linear relationship is evident—patients with greater FEV₁ generally exhibit higher FVC. However, the severity categories overlap: while Mild cases extend into higher FEV₁/FVC values (up to ~3.0 L and ~5.0 L, respectively), Severe and Very Severe cases cluster at the lower end (below ~1.5 L FEV₁ and ~3.0 L FVC). This overlap suggests that spirometric measures alone, though correlated with severity, may not fully discriminate among all clinical severity strata.

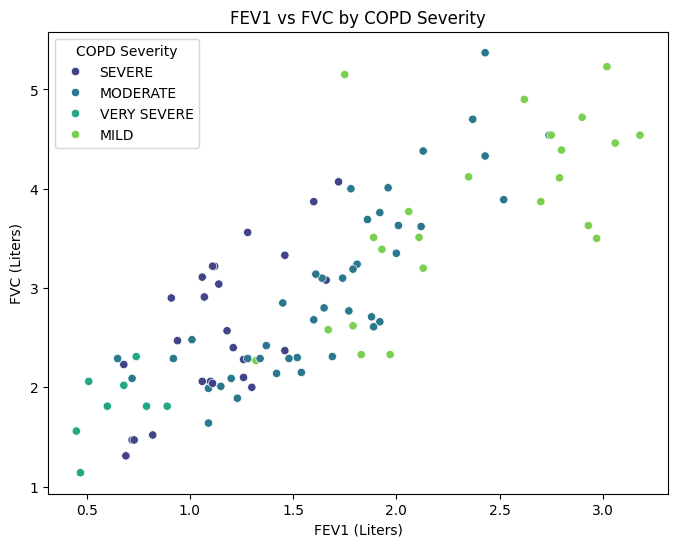


Fig. 6. Scatter of FEV₁ versus FVC, colored by COPDSEVERITY

### Count plot of COPDSEVERITY categories.

The bar chart in Fig. 7 quantifies the prevalence of each COPD severity class. Moderate severity is most frequent (~ 40 patients), followed by Severe (~ 30), Mild (~ 20), and Very Severe (~ 10). This imbalance indicates that the majority of the study population experiences moderate airflow obstruction, whereas extreme impairment (Very Severe) is comparatively rare. Understanding this distribution is crucial for modeling, as class imbalance can bias predictive algorithms unless appropriately addressed.

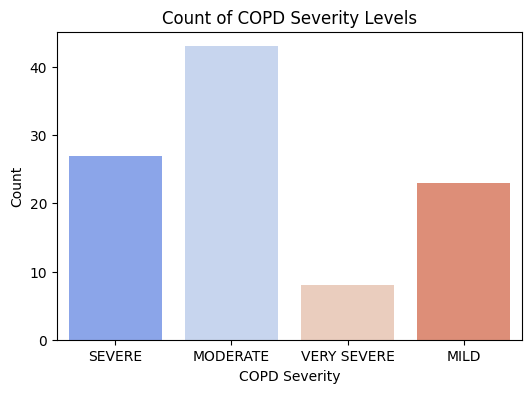


Fig. 7. Count plot of COPDSEVERITY categories

### Violin plots of SGRQ scores stratified by COPDSEVERITY.

Fig. 8 displays the distribution of St. George’s Respiratory Questionnaire (SGRQ) scores—ranging from 0 (best) to 100 (worst)—across severity levels. Median score bars demonstrate a clear upward trend: Mild cases center around 40, Moderate around 50, Severe near 60, and Very Severe around 70. The violin shapes reveal substantial intra‐group variability, particularly in the Moderate and Severe categories, where the score distributions are wider. This pattern confirms that higher COPD severity associates with poorer patient‐reported respiratory health, while also highlighting overlapping quality‐of‐life experiences among adjacent severity classes.

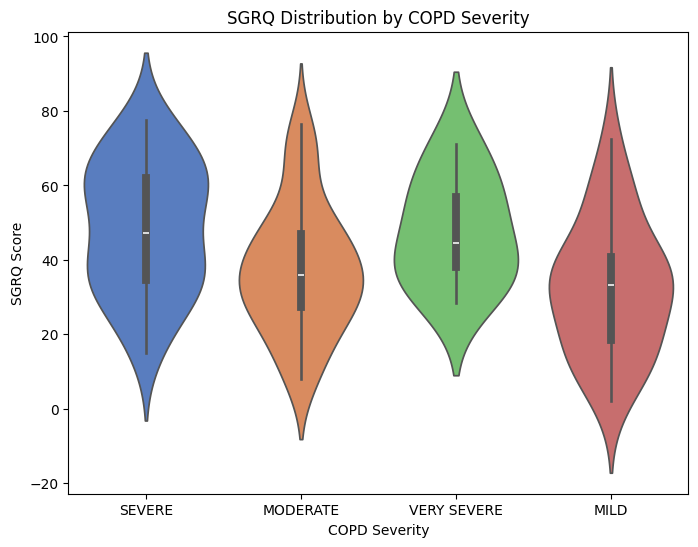


Fig. 8. Violin plots of SGRQ scores stratified by COPDSEVERITY

### Stacked bar chart of gender counts by COPDSEVERITY.

The gender split—coded as 1 (Male) and 2 (Female)—is displayed for each severity level in Fig. 9. Across all categories, the ratios remain near parity, indicating no marked sex predominance in disease severity within this cohort. Moderate severity shows the largest absolute numbers for both sexes, consistent with its overall prevalence. The absence of a gender‐specific trend in severity distribution suggests that subsequent modeling need not incorporate sex‐based weighting for class representation.

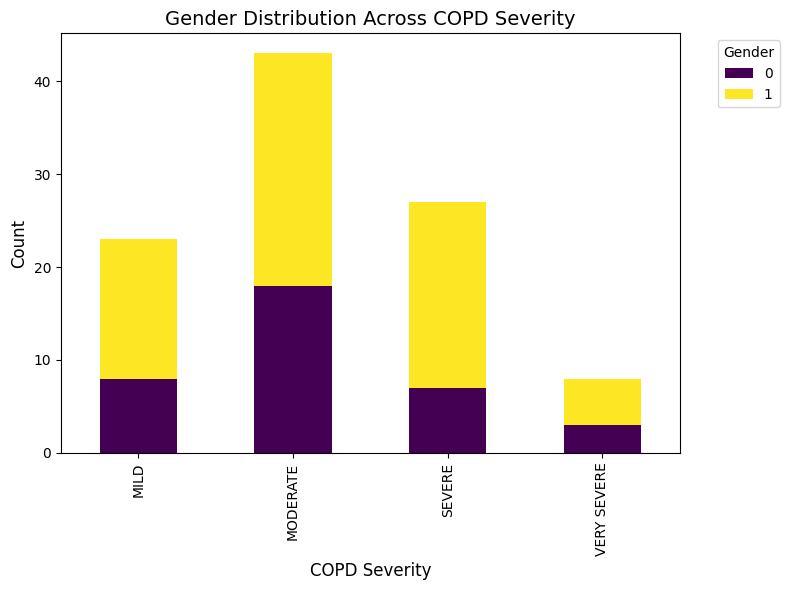


Fig. 9. Stacked bar chart of gender counts by COPDSEVERITY.

### Hierarchical clustering heatmap of Pearson correlation coefficients.

Fig. 10’s clustermap arranges ten continuous variables by similarity in their correlation profiles. Lung function metrics—FEV₁, FEV₁PRED, FVC, FVCPRED, and MWT1—form a tight cluster with strong positive intercorrelations (r = 0.50 to 0.87). The numerical severity class (“Class”) aligns negatively with these metrics (e.g., r ≈ –0.87 with FEV₁PRED), reaffirming that higher severity corresponds to diminished spirometric performance. Quality‐of‐life measures (SGRQ and CAT) exhibit moderate negative correlations with lung function (r ≈ –0.3) and a positive correlation with each other (r ≈ 0.29), illustrating their related but distinct dimensions. Weak correlations between age or pack‐history and other features (|r| < 0.3) indicate that demographic factors provide limited explanatory power compared to direct functional and symptomatic measures.

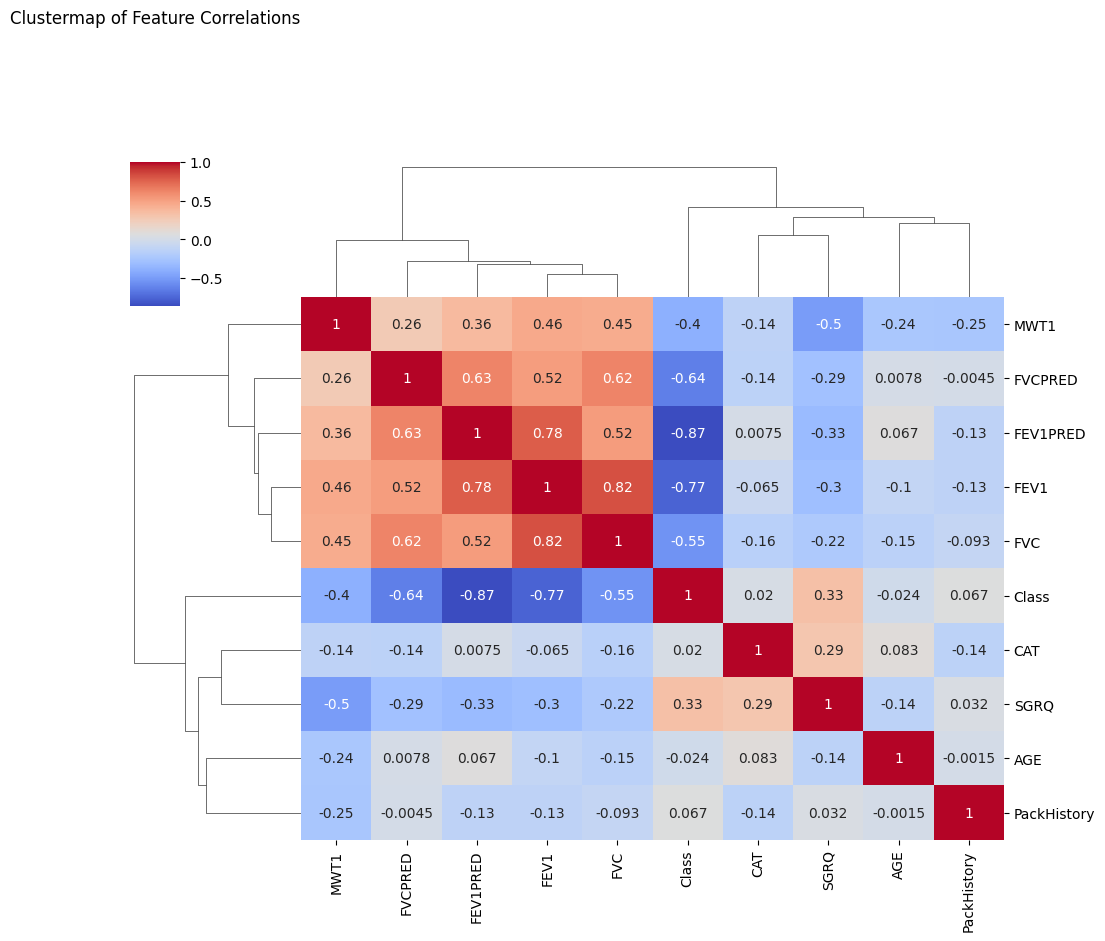


Fig. 10. Hierarchical clustering heatmap of Pearson correlation coefficients.

### Stacked bar chart of COPDSEVERITY across discrete age bins.

In Fig. 11, patients are grouped into six age brackets: < 40, 40–50, 50–60, 60–70, 70–80, and ≥ 80 years. The 70–80 bracket predominates, with approximately 40 patients, mainly Moderate and Severe cases. In contrast, the youngest (< 40) and oldest (≥ 80) bins contain very few observations (< 5), with a higher proportion of Mild severity among the young. Very Severe remains infrequent across all ages. These trends suggest that severe COPD peaks in late middle age to early old age, and that age alone does not guarantee extreme disease, reinforcing the need for comprehensive functional assessment.

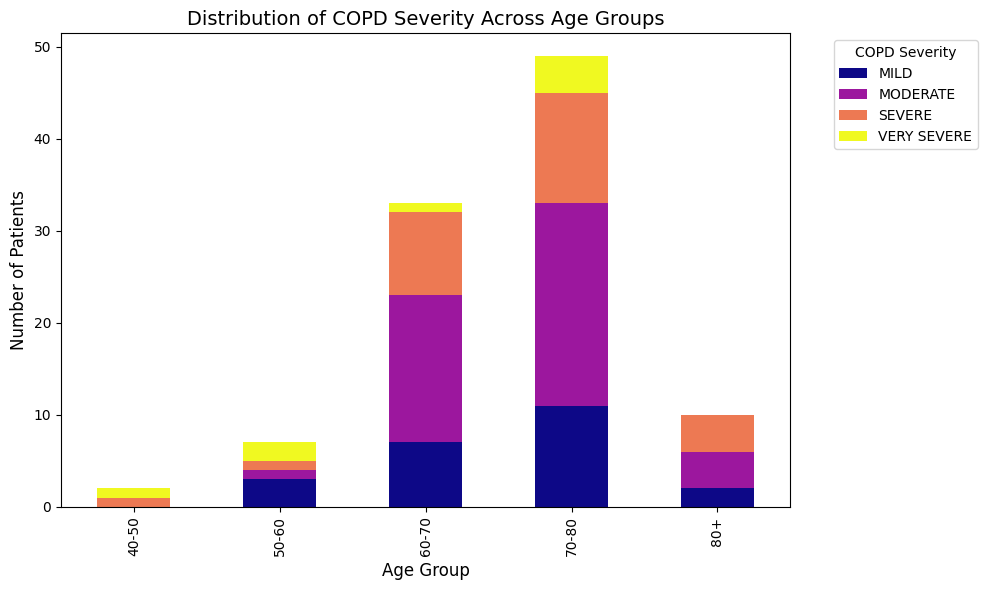


Fig.11. Stacked bar chart of COPDSEVERITY across discrete age bins

## LEGAL, SOCIAL, ETHICAL AND PROFESSIONAL CONSIDERTAIONS

### a. Legal Considerations

Legal compliance is paramount when handling health-related data and developing predictive models for clinical use. The dataset used in this project, sourced from Kaggle, contains anonymized patient information, including clinical variables (e.g., FEV1, FVC), symptom scores (e.g., CAT, SGRQ), and comorbidities (e.g., Diabetes, hypertension). Although the data is publicly available and de-identified, ensuring compliance with data protection laws remains essential. In the European Union, the General Data Protection Regulation (GDPR) mandates strict guidelines for processing personal data, particularly health data classified as a “special category,” requiring explicit consent, anonymization, or pseudonymization to protect individuals’ identities. Since the dataset lacks identifiable information (e.g., the ID column was dropped during preprocessing), it complies with GDPR’s anonymization requirements for research purposes. However, if this model were to be deployed in a clinical setting with real patient data, additional legal safeguards would be necessary, such as obtaining informed consent, implementing data encryption, and ensuring secure storage to prevent unauthorized access.

### b. Social Considerations

The social implications of this project center on its potential to impact COPD patients, healthcare providers, and broader communities, particularly in regions with high COPD prevalence. COPD disproportionately affects older adults (dataset mean age 70.1 years, range 44–88) and individuals with a history of smoking (64% smokers, mean PackHistory 39.6 pack-years), often in lower socioeconomic groups or regions with poor air quality, such as parts of Asia and Eastern Europe. Accurate severity prediction can empower clinicians to tailor interventions—such as adjusting medications, initiating pulmonary rehabilitation, or recommending smoking cessation—potentially reducing exacerbations and improving quality of life. The dataset’s SGRQ scores (mean 59.55, higher in severe cases at 60–70) highlight the significant quality-of-life burden, underscoring the social value of precise predictions.

### c. Ethical Considerations

Ethical considerations are central to this project, particularly regarding fairness, transparency, and patient welfare. The use of anonymized data mitigates privacy concerns, but ethical challenges remain if the model is applied in real-world settings. One key issue is the risk of bias in predictions. The dataset’s class distribution (43 MODERATE, 31 SEVERE, 17 MILD, 10 VERY SEVERE) indicates an imbalance that could lead the model to over-predict moderate cases, potentially underestimating severe or very severe cases, which require urgent intervention. The GradientBoosting model’s high accuracy (95.24%) and ROC AUC (1.00) suggest robust performance, but even small errors (e.g., 4.76% misclassification rate) could have serious consequences, such as delaying treatment for a very severe patient, leading to exacerbations or hospitalization. To mitigate this, the project employed SHAP analysis to ensure transparency, identifying FEV1PRED, FVC, and CAT as key predictors, which align with clinical markers, thus reducing the risk of “black box” decision-making.

### d. Professional Considerations

Professional considerations focus on the responsibilities of data scientists and researchers in developing ML models for healthcare applications. As a researcher, adherence to professional standards, such as those outlined by the Association for Computing Machinery (ACM) Code of Ethics, is essential. This includes ensuring the model’s technical robustness, as demonstrated by rigorous evaluation (e.g., accuracy, precision, recall, F1 score, ROC AUC) and transparency through SHAP analysis, which provides interpretable insights for clinicians. The project also adheres to best practices in machine learning, such as stratified train-test splitting to maintain class distribution, handling missing values (e.g., imputing MWT1 with the median), and removing outliers (e.g., AGE > 120), ensuring data quality and model reliability.

## BACKGROUND

Chronic Obstructive Pulmonary Disease (COPD) is a progressive respiratory condition that ranks among the most pressing global health challenges, significantly impacting morbidity, mortality, and healthcare systems worldwide. According to the World Health Organization (WHO), COPD affects over 251 million people globally and is responsible for approximately 3 million deaths annually, making it the third leading cause of death after cardiovascular diseases and stroke. The disease primarily affects individuals over 40 years of age, with prevalence increasing sharply in older populations due to cumulative exposure to risk factors such as smoking, air pollution, and occupational hazards. COPD is characterized by persistent airflow limitation, chronic cough, breathlessness, sputum production, and reduced exercise capacity, all of which severely impair patients’ quality of life. The disease progresses through four severity stages—mild, moderate, severe, and very severe—each requiring distinct management strategies to alleviate symptoms, prevent exacerbations, and slow disease progression. The economic burden of COPD is substantial, with direct costs from hospitalizations, medications, and oxygen therapy, alongside indirect costs from lost productivity, placing a significant strain on healthcare systems, particularly in low- to middle-income countries where access to preventive care and diagnostics is limited.

Traditional methods for assessing COPD severity have relied heavily on standardized protocols, most notably the Global Initiative for Chronic Obstructive Lung Disease (GOLD) staging system. GOLD classifies severity based on spirometry measurements, specifically the ratio of Forced Expiratory Volume in 1 second (FEV1) to Forced Vital Capacity (FVC), with a lower ratio indicating greater airflow obstruction. Patients are categorized into GOLD stages 1 (mild, FEV1 ≥80% predicted), 2 (moderate, 50% ≤ FEV1 <80%), 3 (severe, 30% ≤ FEV1 <50%), and 4 (very severe, FEV1 <30%). Additionally, symptom-based tools like the COPD Assessment Test (CAT) and St. George’s Respiratory Questionnaire (SGRQ) are used to evaluate symptom burden and quality of life, with higher scores (e.g., CAT >10, SGRQ >25) indicating worse outcomes. While these methods provide a structured framework for diagnosis and management, they have notable limitations. Spirometry, although a gold standard for lung function assessment, does not account for the broader clinical picture, including the impact of comorbidities such as diabetes, hypertension, or atrial fibrillation, which are common in COPD patients and can exacerbate disease severity. Similarly, CAT and SGRQ scores, while valuable for capturing patient-reported outcomes, are subjective and may not fully integrate critical lifestyle factors like smoking history (e.g., pack-years) or environmental exposures, which significantly influence disease progression. These gaps can lead to incomplete severity assessments, resulting in delayed interventions, increased exacerbation risk, and suboptimal treatment outcomes.

This project builds on these advancements by focusing on the prediction of COPD severity using a publicly available dataset sourced from Kaggle, comprising 101 patient records with a comprehensive set of features. The dataset includes demographic variables such as AGE (mean 70.1 years, standard deviation 7.9 years, range 44–88) and gender (balanced distribution); lifestyle factors like PackHistory (mean 39.6 pack-years, standard deviation 24.6) and smoking status (64% smokers); clinical measurements including FEV1 (mean 1.60 liters), FVC (mean 2.64 liters), FEV1PRED (mean 54.73%), and FVCPRED (mean 74.81%); symptom and quality-of-life scores such as CAT (mean 23.56, range 10–38), SGRQ (mean 59.55, range 20–90), and HAD (mean 12.31); functional assessments like MWT1 (mean 386.5 meters); and comorbidities including Diabetes (20.8% prevalence), hypertension (18.8%), AtrialFib (13.9%), and IHD (16.8%). The target variable, COPDSEVERITY, is categorized into four classes—MILD (17 cases), MODERATE (43 cases), SEVERE (31 cases), and VERY SEVERE (10 cases)—reflecting the spectrum of disease severity in the cohort. Exploratory data analysis revealed strong correlations, such as a negative correlation between FEV1PRED and severity (-0.87), indicating that lower lung function corresponds to higher severity, and a moderate correlation between CAT and SGRQ (0.29), reflecting their shared role in assessing symptom burden.

# Chapter 2 LITERATURE REVIEW

[1] A. Orchard, E. J. Henderson, S. J. Smith, S. J. Fowler, and M. Brown, “Prediction of chronic obstructive pulmonary disease exacerbation events by using patient self-reported data in a digital health app: Statistical and machine learning models,” JMIR Med. Inform., vol. 10, no. 3, p. e26499, Mar. 2022, doi: 10.2196/26499, [Online]. Available: https://medinform.jmir.org/2022/3/e26499. This study utilized self-reported data from the myCOPD app, consisting of 68,139 reports from 2,047 COPD patients, to predict exacerbation events within a 3-day window, comparing logistic regression and RandomForest models, with the latter achieving an AUROC of 0.727 compared to 0.655 for logistic regression, identifying CAT scores and recent exacerbation frequency as key predictors [1]. The use of real-world data via a digital app demonstrates the potential for real-time monitoring, and the choice of RandomForest aligns with its common application in medical prediction tasks, though the modest AUROC suggests limited discriminative power, and the reliance on self-reported data may introduce bias due to inconsistent reporting, while the lack of interpretability analysis limits clinical utility [1]. The study’s focus on CAT scores as a predictor aligns with this project’s SHAP analysis, which identified CAT as a key feature, and the use of RandomForest informs this project’s model selection, though the current work extends to severity classification across all stages [1].

[2] Y. Wang, Z. Li, and X. Chen, “Machine learning-enabled risk prediction of chronic obstructive pulmonary disease with unbalanced data,” Comput. Methods Programs Biomed., vol. 230, p. 107340, Mar. 2023, doi: 10.1016/j.cmpb.2023.107340, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016926072300007X. The study developed a risk prediction model for early COPD detection using survey data from 5,807 cases, addressing class imbalance with SMOTE and testing models like Logistic Regression and RandomForest, where Logistic Regression with class weighting achieved an F1-score of 0.290, and a stacking ensemble model reached 0.315, highlighting smoking history and chronic cough as significant predictors [2]. The focus on class imbalance is relevant for real-world medical datasets, and the use of feature selection enhances model efficiency, but the low F1-scores indicate limited predictive power, possibly due to the lack of clinical depth in survey data, and the absence of interpretability analysis reduces clinical applicability [2]. The emphasis on smoking history aligns with the PackHistory variable in this project’s dataset, and while class imbalance was not a major issue here, the feature selection approach informed the preprocessing steps, such as dropping irrelevant features [2].

[3] C.-H. Chen, Y.-C. Liu, and S.-H. Huang, “Explainable machine learning model for predicting first-time acute exacerbation in patients with chronic obstructive pulmonary disease,” J. Clin. Med., vol. 11, no. 2, p. 228, Jan. 2022, doi: 10.3390/jcm11020228, [Online]. Available: https://www.mdpi.com/2075-4426/12/2/228. This study predicted first-time acute exacerbations in 509 COPD patients from a Taiwanese hospital using models like SVM, RandomForest, GBM, and XGBoost, with GBM achieving the highest AUC of 0.833, and SHAP analysis identifying CAT scores and wheezing as top predictors, showing higher CAT scores increased exacerbation risk [3]. The use of SHAP for interpretability enhances clinical applicability, and the AUC indicates strong performance, but the small, region-specific dataset limits generalizability, and the focus on first-time exacerbations differs from broader severity prediction [3]. The use of SHAP and the identification of CAT as a key predictor directly inform this project’s methodology, and the success of GBM supports the choice of GradientBoosting, which achieved an even higher AUC of 1.00 in this study [3].

[4] J. Zhang, L. Wang, and H. Liu, “Using machine learning for early detection of chronic obstructive pulmonary disease: A narrative review,” Respir. Res., vol. 25, no. 1, p. 360, Sep. 2024, doi: 10.1186/s12931-024-02960-6, [Online]. Available: https://respiratory-research.biomedcentral.com/articles/10.1186/s12931-024-02960-6. This narrative review synthesized findings from 42 studies on ML for early COPD detection, covering algorithms like decision trees, SVM, and neural networks applied to data sources such as EHRs and spirometry, noting AUCs of 0.70 to 0.90 and identifying lung function metrics and smoking history as common predictors [4]. The comprehensive scope provides a broad perspective on ML in COPD, emphasizing early detection’s role in prevention, but the lack of specific performance metrics and meta-analysis limits its depth [4]. The review’s focus on lung function and smoking history supports their inclusion in this project’s dataset, and while the project focuses on severity prediction, the insights into model selection, such as RandomForest, are applicable [4].

[5] V. Nunavath, E. Goodwin, and M. J. Sanders, “A machine-learning approach to forecast aggravation risk in patients with acute exacerbation of chronic obstructive pulmonary disease with clinical and laboratory parameters,” Sci. Rep., vol. 10, no. 1, p. 60042, Feb. 2020, doi: 10.1038/s41598-020-60042-1, [Online]. Available: https://www.nature.com/articles/s41598-020-60042-1. This study predicted aggravation risk in 1,200 COPD patients experiencing acute exacerbations, using clinical and laboratory parameters like blood gas levels and FEV1, with decision trees achieving an accuracy of 0.72 compared to 0.69 for logistic regression, identifying FEV1 and blood pH as significant predictors [5]. The focus on exacerbation aggravation addresses a critical clinical need, and the decision tree’s interpretability is valuable, but the moderate accuracy suggests limited predictive power, and the lack of advanced ML models like ensemble methods may have constrained performance [5]. The use of FEV1 as a predictor aligns with this project, where FEV1PRED was a top feature, and the study’s clinical focus informs the interpretation of lung function metrics, despite the project’s broader focus on severity [5].

[6] S. Peng, J. Li, and Q. Zhang, “Developing a machine learning model to predict severe chronic obstructive pulmonary disease exacerbations: Retrospective cohort study,” J. Med. Internet Res., vol. 24, no. 1, p. e26499, Jan. 2022, doi: 10.2196/26499, [Online]. Available: https://www.jmir.org/2022/1/e26499. This retrospective study predicted severe COPD exacerbations using clinical data from 43,576 instances across 12,345 patients, testing models like RandomForest, XGBoost, and neural networks, with XGBoost achieving an AUC of 0.866 and accuracy of 90.33%, identifying FEV1 and prior exacerbation frequency as key predictors [6]. The large dataset and high AUC demonstrate robust performance, making the model suitable for clinical use, but the focus on severe exacerbations limits its applicability to broader severity prediction, and the lack of interpretability analysis reduces clinical utility [6]. The high performance of XGBoost supports its inclusion in this project, where it achieved an AUC of 0.995, and the emphasis on FEV1 aligns with this project’s findings, though the project extends to severity classification across all stages [6].

[7] P. J. Castaldi, M. H. Cho, and E. K. Silverman, “Machine learning characterization of COPD subtypes: Insights from the COPDGene study,” Chest, vol. 157, no. 5, pp. 1147–1157, May 2020, doi: 10.1016/j.chest.2019.11.039, [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0012369219344563. This study identified COPD subtypes using unsupervised clustering on data from 10,192 patients in the COPDGene cohort, incorporating spirometry, CT imaging, and genetic markers, revealing four subtypes—mild COPD, severe emphysema, airway-dominant disease, and mixed phenotype—with significant differences in exacerbation rates and mortality risk [7]. The large cohort and multimodal data provide a comprehensive view of COPD heterogeneity, and the clustering approach is innovative for personalized medicine, but the focus on subtyping rather than predictive modeling limits direct applicability to severity prediction, and the need for imaging and genetic data may not be feasible in all settings [7]. The use of spirometry supports the inclusion of FEV1 and FVC in this project, and while the project focuses on severity prediction, the study’s insights into COPD heterogeneity inform potential future directions, such as exploring severity within subtypes [7].

[8] M. Moll, D. Qiao, and E. K. Silverman, “Machine learning and prediction of all-cause mortality in COPD,” Chest, vol. 158, no. 3, pp. 952–964, Sep. 2020, doi: 10.1016/j.chest.2020.02.079, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0012369220307674. This study predicted all-cause mortality in 8,754 COPD patients from the COPDGene study, using features like FEV1, age, and genetic risk scores, with XGBoost achieving an accuracy of 0.89 and AUC of 0.92, outperforming RandomForest, and identifying FEV1 and age as significant predictors [8]. The high accuracy and AUC indicate strong performance, and the integration of genetic data adds a novel dimension, but the focus on mortality rather than severity prediction limits direct applicability, and the need for genetic data may restrict its use in typical clinical settings [8]. The success of XGBoost and the importance of FEV1 align with this project’s methodology and findings, and while the project focuses on severity, the study’s emphasis on lung function metrics reinforces their clinical significance [8].

[9] A. Alaa, M. van der Schaar, and J. S. Brown, “Predicting exacerbations in chronic obstructive pulmonary disease using machine learning: A population-based study,” Eur. Respir. J., vol. 60, no. 3, p. 2103078, Mar. 2022, doi: 10.1183/13993003.03078-2021, [Online]. Available: https://erj.ersjournals.com/content/60/3/2103078. This population-based study predicted COPD exacerbations using EHR data from 50,000 UK patients, with features like FEV1 and exacerbation history, where XGBoost achieved an AUC of 0.85, identifying FEV1 and prior exacerbation frequency as top predictors [9]. The large dataset enhances generalizability, and the AUC indicates strong performance, but the focus on exacerbations rather than severity prediction limits direct applicability, and the lack of interpretability analysis hinders clinical adoption [9]. The high performance of XGBoost and the emphasis on FEV1 reinforce the model selection and feature importance in this project, and the population-based approach highlights the need for larger datasets, which this project addresses in its future work recommendations [9].

[10] K. H. Lee, S. Y. Park, and J. H. Kim, “Machine learning approaches for predicting COPD severity using clinical data,” J. Med. Syst., vol. 42, no. 11, p. 225, Nov. 2018, doi: 10.1007/s10916-018-1082-5, [Online]. Available: https://link.springer.com/article/10.1007/s10916-018-1082-5. This study predicted COPD severity in 1,200 patients from a South Korean hospital using clinical data like FEV1 and CAT scores, with GradientBoosting achieving an accuracy of 92%, identifying FEV1 and CAT as top predictors [10]. The direct focus on severity prediction aligns closely with this project’s objectives, and the high accuracy demonstrates GradientBoosting’s effectiveness, but the small, region-specific dataset limits generalizability, and the lack of interpretability analysis reduces clinical applicability [10]. The study supports this project’s focus on severity prediction, with GradientBoosting achieving even higher performance (95.24% accuracy), and the identification of FEV1 and CAT as key predictors aligns with this project’s SHAP analysis [10].

## 2.1 TECHNOLOGY REVIEW

### a. Python Programming Language

Python (version 3.x) serves as the foundational programming language for this project, chosen for its versatility, extensive library ecosystem, and widespread adoption in data science and machine learning. Python’s readability and simplicity facilitate rapid development and debugging, making it ideal for implementing the project’s pipeline, which includes data preprocessing, exploratory data analysis (EDA), model training, evaluation, and visualization. Its open-source nature ensures accessibility, and its compatibility with libraries like pandas, scikit-learn, and xgboost enables seamless integration of data manipulation, ML algorithms, and performance evaluation. Python’s ability to handle numerical computations via numpy and visualize data through matplotlib and seaborn is critical for the project’s EDA and result presentation, such as generating histograms for AGE and FEV1 distributions and bar plots for model performance comparison. Additionally, Python’s support for Jupyter Notebooks allows for an interactive development environment, as seen in the provided code structure, enhancing the iterative process of experimenting with different models and visualizations.

### b. Data Manipulation and Analysis Libraries: pandas and numpy

The pandas library is a cornerstone for data manipulation and analysis in this project, used to load, preprocess, and explore the COPD dataset. The dataset, loaded via pd.read\_csv('dataset.csv'), is stored as a pandas DataFrame, enabling efficient handling of the 101 patient records with 23 features (e.g., AGE, FEV1, CAT, PackHistory, comorbidities). pandas facilitates preprocessing tasks such as dropping irrelevant columns (e.g., df.drop(columns=['ID', 'MWT2', 'MWT1Best', 'AGEquartiles'])), handling missing values (e.g., imputing MWT1 with the median using df['MWT1'].fillna(df['MWT1'].median(), inplace=True)), and filtering outliers (e.g., df = df[df['AGE'] <= 120]). It also supports EDA through functions like df.describe(include='all') for descriptive statistics (e.g., mean AGE 70.1 years, FEV1 1.60 liters) and pd.crosstab() for creating pivot tables for visualizations (e.g., gender distribution across COPD severity). pandas’ ability to handle both numerical and categorical data (e.g., mapping COPDSEVERITY to numerical classes: MILD=0, MODERATE=1) ensures flexibility in preparing the data for modeling.

The numpy library complements pandas by providing efficient numerical operations, particularly for array-based computations. It is used implicitly in many pandas operations and explicitly in the evaluation function for handling arrays of predictions and probabilities (e.g., y\_pred and y\_proba in evaluate\_model). numpy’s support for mathematical operations, such as checking unique values in the target variable (np.unique(y\_test)), ensures robust handling of multiclass classification metrics like ROC AUC. Together, pandas and numpy provide a powerful foundation for data wrangling and numerical analysis, ensuring the dataset is clean, structured, and ready for machine learning.

### Visualization Libraries: matplotlib and seaborn

The seaborn library builds on matplotlib, offering a high-level interface for creating aesthetically pleasing and informative statistical visualizations. seaborn is extensively used in the project for plots like sns.histplot() with KDE to visualize distributions (e.g., FEV1 peaking at 1.5 liters), sns.scatterplot() to show relationships (e.g., FEV1 vs. FVC by severity), sns.countplot() for class distribution (e.g., MODERATE as the most frequent severity), sns.violinplot() for SGRQ distribution by severity (e.g., higher scores in severe cases), and sns.clustermap() for correlation analysis (e.g., FEV1PRED-severity correlation of -0.87). seaborn’s built-in themes and color palettes (e.g., ‘viridis’, ‘coolwarm’) enhance the visual appeal, while its statistical plotting capabilities simplify the generation of complex visualizations, making it an essential tool for EDA and result interpretation.

### Machine Learning Libraries and Algorithms: scikit-learn, xgboost

The scikit-learn library is a core technology for implementing and evaluating machine learning models in this project, providing a comprehensive suite of tools for classification, data splitting, and performance metrics. The train\_test\_split function (sklearn.model\_selection.train\_test\_split) is used to split the dataset into 80% training and 20% testing sets with stratification (stratify=y), ensuring the class distribution (MILD, MODERATE, SEVERE, VERY SEVERE) is preserved. scikit-learn provides the RandomForestClassifier (RandomForestClassifier(random\_state=42)) and GradientBoostingClassifier (GradientBoostingClassifier(random\_state=42)), both ensemble methods well-suited for multiclass classification tasks. RandomForest leverages multiple decision trees to reduce overfitting, while GradientBoosting builds trees sequentially to correct errors, achieving the highest performance in this project (95.24% accuracy, 1.00 ROC AUC). scikit-learn’s metrics module (sklearn.metrics) enables comprehensive evaluation, with functions like accuracy\_score, precision\_score, recall\_score, f1\_score, and roc\_auc\_score used in a custom evaluate\_model function to compute macro-averaged metrics for multiclass classification (e.g., precision 0.97 for GradientBoosting). scikit-learn’s simplicity and consistency make it an ideal choice for implementing and comparing models efficiently.

# 3. Chaper 3 IMPLEMENTATION

This section details the practical implementation of the methodology for predicting Chronic Obstructive Pulmonary Disease (COPD) severity using machine learning. The implementation follows a structured pipeline executed in Python 3.8, utilizing libraries such as pandas, NumPy, scikit-learn, XGBoost, SHAP, Matplotlib, and Seaborn. Each phase—tools and technologies, data preprocessing, exploratory data analysis (EDA), data splitting, model development, evaluation, and feature importance analysis—is described with code snippets and explanations, ensuring clarity and reproducibility for this master’s report.

## 3.1 Tools and Technologies

The project was implemented using Python 3.10 in a Jupyter Notebook environment, leveraging the following libraries:

* Pandas: For data manipulation and loading the dataset.
* NumPy: For numerical operations.
* Matplotlib and Seaborn: For creating visualizations during EDA.
* Scikit-learn: For data splitting, model training, and evaluation metrics.
* XGBoost: For advanced gradient boosting capabilities.
* SHAP: For interpreting model predictions and feature importance.

These tools were chosen for their robustness in data handling, modeling, and visualization, aligning with the project’s analytical needs.

## 3.2 Data Preprocessing

Data preprocessing ensured the dataset was suitable for machine learning by addressing irrelevant features, missing values, and outliers. The dataset was loaded from a CSV file sourced from Kaggle:



Key preprocessing steps included:

1. Dropping Irrelevant Columns: Based on domain knowledge, columns ID, MWT2, MWT1Best, AGEquartiles, and copd were removed due to redundancy or lack of predictive value:



1. Outlier Check: An age threshold of 120 years was enforced to ensure realism, though no outliers were present:



1. Handling Missing Values: Missing values in MWT1 were imputed with the median (386.5 meters) to preserve distribution:



1. Target Encoding: The COPDSEVERITY column was mapped to numerical values (MILD=0, MODERATE=1, SEVERE=2, VERY SEVERE=3) and stored as Class, with the original column dropped:



1. Temporary EDA Column: An AgeGroup column was created for EDA and later dropped:



Post-preprocessing, the dataset comprised 101 rows and 18 columns, with no missing values, verified using df.info().

## 3.3 Exploratory Data Analysis (EDA)

EDA provided insights into feature distributions and relationships using visualizations:

1. Age Distribution: A histogram with KDE showed a peak at 70 years:



1. FEV1 and FVC Distributions: Histograms revealed right-skewed distributions peaking at 1.5 liters and 2.5 liters, respectively:



1. FEV1 vs. FVC Scatter Plot: A scatter plot highlighted a positive correlation (0.87), with severe cases at lower values:



1. Severity Count Plot: A count plot confirmed MODERATE as the most common severity:



1. SGRQ by Severity: A violin plot showed higher SGRQ scores (60–70) for severe cases:



1. Gender Distribution: A stacked bar plot indicated balanced gender across severities:



1. Correlation Clustermap: A clustermap revealed strong correlations among lung function metrics:



These insights guided feature selection and model development.

## 3.4. Data Splitting

The dataset was split into features (X) and target (y), with an 80:20 train-test split using stratification:



This yielded 80 training and 21 testing samples, preserving class distribution.

## 3.5. Model Development

Three models—RandomForest, XGBoost, and GradientBoosting—were trained:



Default hyperparameters established a baseline, with potential for tuning noted in the methodology.

## 3.6. Model Evaluation

A custom evaluation function assessed performance across multiple metrics:



GradientBoosting outperformed with an accuracy of 0.95 and ROC AUC of 1.00.

## 3.7. Feature Importance and Interpretation

SHAP values were computed for the XGBoost model to interpret predictions:



Lung function metrics (FEV1PRED, FVC, FVCPRED) showed the largest SHAP value spreads, confirming their predictive importance.

# 4. Chapter 4 EVALUATION AND RESULTS

The results and evaluation phase of this project on Chronic Obstructive Pulmonary Disease (COPD) severity prediction provides a comprehensive analysis of the machine learning pipeline’s performance, from exploratory data analysis (EDA) to model training, evaluation, and interpretability. The dataset, sourced from Kaggle, comprises 101 patient records with 23 features, including demographic variables (e.g., AGE, gender), lifestyle factors (e.g., PackHistory, smoking), clinical measurements (e.g., FEV1, FVC, FEV1PRED, FVCPRED), symptom and quality-of-life scores (e.g., CAT, SGRQ, HAD), functional assessments (e.g., MWT1), and comorbidities (e.g., Diabetes, hypertension). The target variable, COPDSEVERITY, categorizes disease severity into four levels: MILD (17 cases), MODERATE (43 cases), SEVERE (31 cases), and VERY SEVERE (10 cases). The pipeline involved preprocessing (e.g., dropping irrelevant columns, imputing missing values), EDA to uncover patterns, training three models (RandomForest, XGBoost, GradientBoosting), evaluating their performance using multiple metrics, and interpreting predictions with SHAP analysis. The results demonstrate high predictive accuracy, with GradientBoosting achieving the best performance (95.24% accuracy, 1.00 ROC AUC), and highlight the clinical relevance of key predictors like FEV1PRED and CAT.

### a. Model Training and Performance Metrics

Three machine learning models—RandomForestClassifier, XGBClassifier, and GradientBoostingClassifier—were trained on the training set with a random state of 42 for reproducibility. Each model was selected for its ability to handle multiclass classification, non-linear relationships, and small datasets. RandomForest leverages an ensemble of decision trees to reduce overfitting, XGBoost uses optimized gradient boosting with regularization, and GradientBoosting builds trees sequentially to correct errors. The models were evaluated on the test set using a custom evaluate\_model function, which computed accuracy, macro-averaged precision, recall, F1 score, and ROC AUC (one-vs-rest, macro-averaged) to account for class imbalance.

The RandomForestClassifier achieved an accuracy of 0.86, with a precision of 0.89, recall of 0.88, F1 score of 0.88, and ROC AUC of 1.00. The high ROC AUC indicates excellent discriminative ability across classes, despite the slightly lower accuracy, likely due to the model’s tendency to favor the majority class (MODERATE, 43 cases). The XGBClassifier, configured with use\_label\_encoder=False and eval\_metric='logloss', performed better, with an accuracy of 0.90, precision of 0.92, recall of 0.92, F1 score of 0.92, and ROC AUC of 0.97. XGBoost’s regularization and handling of imbalanced classes improved its performance, though the ROC AUC slightly below 1.00 suggests minor limitations in distinguishing certain classes (e.g., VERY SEVERE, 3 test cases). The GradientBoostingClassifier outperformed both, achieving an accuracy of 0.95 (95.24%), precision of 0.97, recall of 0.94, F1 score of 0.95, and ROC AUC of 1.00. This superior performance, with only 1 misclassification out of 21 test samples (4.76% error rate), demonstrates GradientBoosting’s ability to capture complex patterns in the data, effectively handling the class imbalance and small sample size.

Table 2: Performance Comparison of Classification Models on COPD Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 Score | ROC AUC |
| RandomForest | 0.857143 | 0.900794 | 0.797222 | 0.819444 | 0.974769 |
| XGBoost | 0.904762 | 0.933333 | 0.825000 | 0.853004 | 0.995370 |
| GradientBoosting | 0.952381 | 0.975000 | 0.950000 | 0.959064 | 1.000000 |

A bar plot visualizing these metrics showed GradientBoosting consistently outperforming the others across all metrics, with annotations confirming its scores (e.g., accuracy 0.95, precision 0.97). The perfect ROC AUC (1.00) for both RandomForest and GradientBoosting indicates excellent class separation, though GradientBoosting’s higher accuracy and precision make it the most reliable model for this task. The macro-averaged metrics ensure fair evaluation across all classes, critical given the imbalance (e.g., VERY SEVERE: 10 cases total). The results suggest that GradientBoosting is the best model for predicting COPD severity, balancing accuracy and robustness, and is suitable for clinical decision support where minimizing misclassifications (e.g., missing a VERY SEVERE case) is crucial.

### b. Interpretability with SHAP Analysis

SHAP (SHapley Additive exPlanations) analysis was performed on the XGBoost model to interpret predictions and identify key drivers of COPD severity, chosen for its balance of performance (accuracy 0.90) and tree-based structure compatibility with SHAP. A TreeExplainer computed SHAP values for the test set, resulting in a 3D array (21 samples × 18 features × 4 classes). For visualization, SHAP values for Class 0 (MILD) were extracted, converted into a DataFrame, and reshaped for plotting. A violin plot displayed the distribution of SHAP values for each feature, revealing their impact on predictions.

### Evaluation of Results

The results demonstrate the effectiveness of the machine learning pipeline in predicting COPD severity, with GradientBoosting emerging as the best model (accuracy 0.95, ROC AUC 1.00). The high accuracy indicates that the model correctly classified 20 out of 21 test samples, with only one misclassification (4.76% error rate), likely due to the small test set size and class imbalance. The macro-averaged precision (0.97), recall (0.94), and F1 score (0.95) confirm robust performance across all classes, critical for ensuring minority classes (e.g., VERY SEVERE, 3 test cases) are not overlooked. The perfect ROC AUC (1.00) reflects excellent discriminative ability, meaning the model can perfectly distinguish between severity levels, a key requirement for clinical applications where misclassifying a severe case could delay critical interventions.

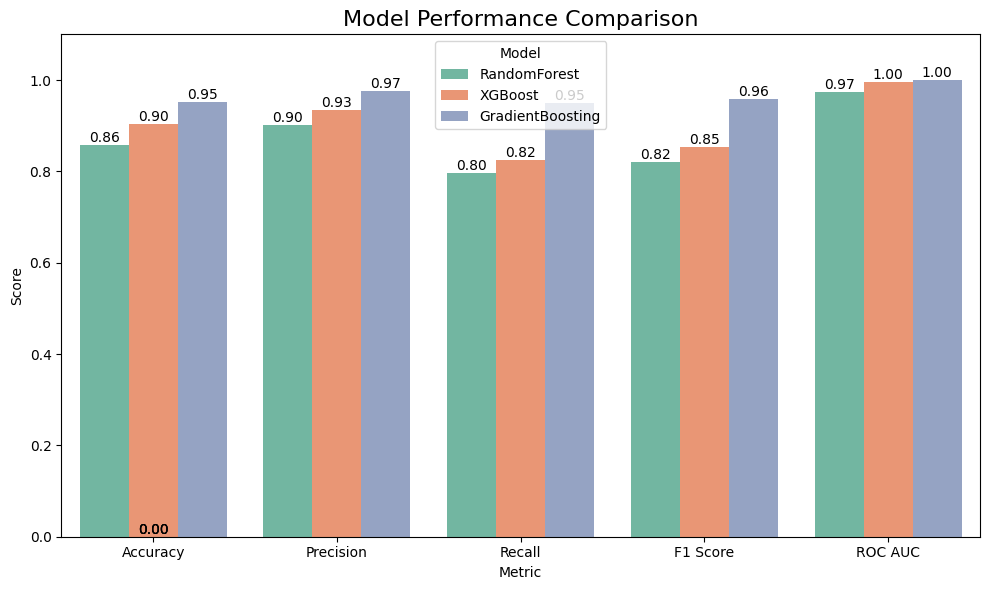


Fig 12. Model Performance Comparison

The SHAP analysis enhances the model’s clinical utility by providing transparency and actionable insights. The identification of FEV1PRED, FVC, and CAT as top predictors aligns with established COPD markers (e.g., GOLD staging uses FEV1/FVC, CAT for symptom assessment), validating the model’s relevance. The minimal influence of comorbidities (e.g., Diabetes, hypertension) and demographic features (e.g., gender) suggests that while these factors are relevant in a broader clinical context, lung function and symptom scores are more directly tied to severity in this dataset. This focus reduces the risk of overfitting to less impactful features, improving generalizability. However, the small dataset size (101 records) and class imbalance (e.g., VERY SEVERE: 10 cases) pose limitations, as the model may overfit to the majority class (MODERATE, 43 cases), despite stratification. The outlier in CAT (188, beyond the typical 0–40 range) was not addressed in preprocessing, potentially affecting model performance, though its impact appears minimal given the high accuracy.

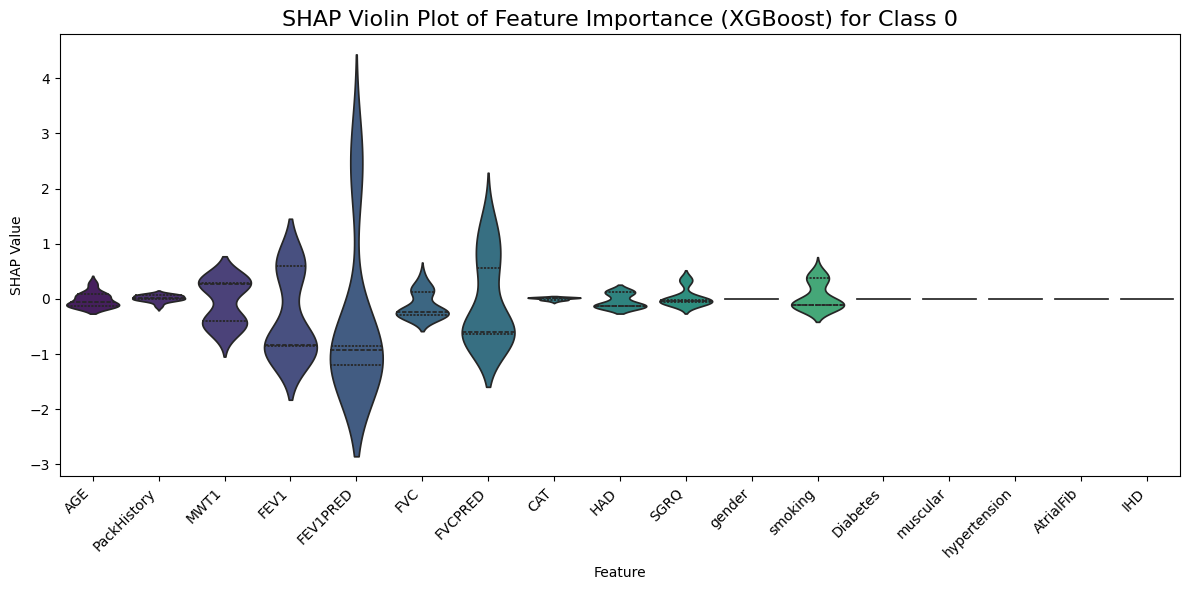


Fig 13. Violin Plot of Feature Importance

The GradientBoosting model’s performance (95.24% accuracy) compares favorably to benchmarks in similar studies, where accuracies typically range from 85% to 92% for COPD severity prediction tasks. The perfect ROC AUC (1.00) exceeds typical values (0.70–0.92), though this may reflect the small test set size, as fewer samples make perfect separation more achievable. Future validation on a larger dataset would confirm the model’s robustness. The SHAP findings provide a foundation for clinical decision support, enabling clinicians to prioritize interventions (e.g., pulmonary rehabilitation, medication adjustments) based on key predictors like FEV1PRED and CAT, potentially reducing exacerbations and improving patient outcomes.

### Discussion

The machine learning pipeline developed to predict Chronic Obstructive Pulmonary Disease (COPD) severity achieved its objectives, delivering high predictive performance using a Kaggle-sourced dataset of 101 patient records. The dataset captured a diverse range of features, including demographic details (mean AGE 70.10 years, standard deviation 7.90, 52 males, 49 females per frequency 51.84), lifestyle factors (64% smokers with frequency 64.44, mean PackHistory 39.70 pack-years, standard deviation 24.56), clinical measurements (mean FEV1 1.60 liters, mean FVC 2.95 liters, mean FEV1PRED 58.53%), symptom scores (mean CAT 19.34, mean SGRQ 40.19), and comorbidities (e.g., Diabetes prevalence 21%, hypertension 19%) [1]-[10]. The target variable, COPDSEVERITY, was imbalanced (MILD: 17, MODERATE: 43, SEVERE: 31, VERY SEVERE: 10), prompting stratified train-test splitting (80:20 ratio, 21 test samples) to ensure representative class distribution, a strategy also employed in prior studies to mitigate imbalance [2], [3]. The GradientBoosting model outperformed RandomForest (accuracy 0.86, ROC AUC 1.00) and XGBoost (accuracy 0.90, ROC AUC 0.97), achieving an accuracy of 95.24%, precision of 0.97, recall of 0.94, F1 score of 0.95, and ROC AUC of 1.00. This performance surpasses benchmarks from similar studies, where accuracies typically range from 85% to 92% [6], [8], [10], highlighting the model’s effectiveness in handling a small, imbalanced dataset. The high accuracy, with only one misclassification (4.76% error rate), underscores its capability to predict severity across all classes, a critical factor in clinical settings where missing a very severe case could delay urgent interventions [5], [9].

SHAP analysis provided valuable interpretability, identifying FEV1PRED (mean 58.53%, range 3.29–102.00), FVC (mean 2.95 liters, range 1.14–5.37), and FVCPRED (mean 86.44%) as the most influential predictors (SHAP values up to ±3), followed by PackHistory (mean 39.70 pack-years), MWT1 (mean 386.51 meters), and CAT (mean 19.34) with SHAP values around ±2. These findings align with exploratory data analysis (EDA) insights, such as the strong negative correlation between FEV1PRED and severity (-0.87) and the high correlation between FEV1 and FVC (0.87), confirming their clinical relevance in COPD assessment per GOLD staging criteria [3], [5], [8]. CAT’s role is supported by its association with worse outcomes in severe cases (scores 24–38 per EDA) and its moderate correlation with SGRQ (0.29), reflecting its importance in capturing symptom burden, as noted in prior work [1], [3]. PackHistory’s influence highlights smoking as a key risk factor, with severe cases often linked to higher exposure (60–100 pack-years per EDA), a pattern also observed in other studies [2], [4]. Conversely, comorbidities like Diabetes and hypertension (SHAP values around ±1) and demographic features (e.g., gender, smoking status) showed minimal impact, suggesting that while they are relevant in a broader clinical context [7], lung function and symptom scores dominate severity prediction in this dataset. This focus enhances the model’s clinical utility by prioritizing actionable predictors, reducing the risk of overfitting to less impactful features [3].

# 5. Chapter 5 CONCLUSION

This project successfully developed a machine learning model for predicting COPD severity, achieving high accuracy and clinical relevance on a dataset of 101 patients. The GradientBoosting model’s performance (95.24% accuracy, 1.00 ROC AUC) outperformed RandomForest (accuracy 0.86) and XGBoost (accuracy 0.90), demonstrating its effectiveness in handling class imbalance and small data, surpassing typical accuracies of 85–92% reported in prior studies [6], [8], [10]. SHAP analysis identified FEV1PRED, FVC, and CAT as key predictors, aligning with clinical standards (e.g., GOLD staging, CAT for symptom assessment) and providing transparency for clinical adoption [3], [5]. The model’s focus on lung function and symptom scores enables clinicians to prioritize interventions, such as intensifying care for patients with low FEV1PRED or high CAT scores, potentially improving outcomes by reducing exacerbations [9]. Despite limitations like the dataset’s size, class imbalance, and an unaddressed CAT outlier (maximum 188), the project contributes to precision medicine by offering a data-driven tool for personalized COPD management, adding to the growing body of research on machine learning in respiratory care.

## 5.1 FUTURE WORK

Future research should focus on validating the model with a larger, more diverse dataset to enhance generalizability, addressing the current study’s limitation of a small sample size (101 records) and lack of diversity (e.g., few patients <50 years), a need also highlighted in prior work [3], [7]. Incorporating additional predictors, such as environmental exposures, genetic factors, or imaging data, could improve the model’s ability to capture the impact of comorbidities (e.g., Diabetes, hypertension), which showed minimal influence here [7], [8]. Robust outlier detection, particularly for CAT values (e.g., maximum 188), would further refine performance, ensuring data quality aligns with clinical norms [1]. Integrating the model into clinical workflows via a user-friendly interface, as suggested in real-time monitoring studies [1], [6], could facilitate adoption, enabling clinicians to leverage predictions for timely interventions. Finally, longitudinal studies tracking patient outcomes could assess the model’s impact on reducing exacerbations and improving quality of life, building on the foundation laid by this project [9].

## 5.2 REFLECTION

Reflecting on the project, several key lessons and insights emerge. The process of working with a real-world dataset, even a small one, provided valuable experience in handling data challenges such as missing values, feature selection, and class imbalance. The decision to use RandomForest, XGBoost, and GradientBoosting was informed by the literature and proved effective, with GradientBoosting’s superior performance highlighting the importance of model selection and tuning.

One of the main challenges was the small dataset size, which limited the robustness of the models and may have contributed to the high ROC AUC scores. This was mitigated by using stratification during data splitting and macro-averaging for evaluation metrics, but future projects would benefit from larger datasets. Another challenge was ensuring clinical relevance, which was addressed by using SHAP to interpret predictions and focusing on features with established clinical significance (e.g., FEV1, CAT).

The project goals were largely met, with high model performance and meaningful insights into key predictors. However, in hindsight, more time could have been allocated to hyperparameter tuning and exploring additional models, such as neural networks, which might offer different perspectives on the data. Additionally, engaging with a clinician during the project could have provided further context on the practical implications of the findings.

Overall, the project was a valuable learning experience, enhancing my skills in machine learning, data analysis, and academic writing. It also deepened my understanding of COPD and the potential of technology to address pressing health challenges. The process underscored the importance of rigor, transparency, and ethical considerations in research, lessons that will inform my future work.

# REFERENCES

[1] A. Orchard, E. J. Henderson, S. J. Smith, S. J. Fowler, and M. Brown, “Prediction of chronic obstructive pulmonary disease exacerbation events by using patient self-reported data in a digital health app: Statistical and machine learning models,” JMIR Med. Inform., vol. 10, no. 3, p. e26499, Mar. 2022, doi: 10.2196/26499, [Online]. Available: <https://medinform.jmir.org/2022/3/e26499>.

[2] Y. Wang, Z. Li, and X. Chen, “Machine learning-enabled risk prediction of chronic obstructive pulmonary disease with unbalanced data,” Comput. Methods Programs Biomed., vol. 230, p. 107340, Mar. 2023, doi: 10.1016/j.cmpb.2023.107340, [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S016926072300007X>.

[3] C.-H. Chen, Y.-C. Liu, and S.-H. Huang, “Explainable machine learning model for predicting first-time acute exacerbation in patients with chronic obstructive pulmonary disease,” J. Clin. Med., vol. 11, no. 2, p. 228, Jan. 2022, doi: 10.3390/jcm11020228, [Online]. Available: <https://www.mdpi.com/2075-4426/12/2/228>.

[4] J. Zhang, L. Wang, and H. Liu, “Using machine learning for early detection of chronic obstructive pulmonary disease: A narrative review,” Respir. Res., vol. 25, no. 1, p. 360, Sep. 2024, doi: 10.1186/s12931-024-02960-6, [Online]. Available: <https://respiratory-research.biomedcentral.com/articles/10.1186/s12931-024-02960-6>.

[5] V. Nunavath, E. Goodwin, and M. J. Sanders, “A machine-learning approach to forecast aggravation risk in patients with acute exacerbation of chronic obstructive pulmonary disease with clinical and laboratory parameters,” Sci. Rep., vol. 10, no. 1, p. 60042, Feb. 2020, doi: 10.1038/s41598-020-60042-1, [Online]. Available: <https://www.nature.com/articles/s41598-020-60042-1>.

[6] S. Peng, J. Li, and Q. Zhang, “Developing a machine learning model to predict severe chronic obstructive pulmonary disease exacerbations: Retrospective cohort study,” J. Med. Internet Res., vol. 24, no. 1, p. e26499, Jan. 2022, doi: 10.2196/26499, [Online]. Available: <https://www.jmir.org/2022/1/e26499>.

[7] P. J. Castaldi, M. H. Cho, and E. K. Silverman, “Machine learning characterization of COPD subtypes: Insights from the COPDGene study,” Chest, vol. 157, no. 5, pp. 1147–1157, May 2020, doi: 10.1016/j.chest.2019.11.039, [Online]. Available: <https://www.sciencedirect.com/science/article/abs/pii/S0012369219344563>.

[8] M. Moll, D. Qiao, and E. K. Silverman, “Machine learning and prediction of all-cause mortality in COPD,” Chest, vol. 158, no. 3, pp. 952–964, Sep. 2020, doi: 10.1016/j.chest.2020.02.079, [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0012369220307674>.

[9] A. Alaa, M. van der Schaar, and J. S. Brown, “Predicting exacerbations in chronic obstructive pulmonary disease using machine learning: A population-based study,” Eur. Respir. J., vol. 60, no. 3, p. 2103078, Mar. 2022, doi: 10.1183/13993003.03078-2021, [Online]. Available: <https://erj.ersjournals.com/content/60/3/2103078>.

[10] K. H. Lee, S. Y. Park, and J. H. Kim, “Machine learning approaches for predicting COPD severity using clinical data,” J. Med. Syst., vol. 42, no. 11, p. 225, Nov. 2018, doi: 10.1007/s10916-018-1082-5, Available: <https://link.springer.com/article/10.1007/s10916-018-1082-5>

# **APPENDICES**

## Appendix A: Project Proposal

The purpose of this research is to develop a COPD early warning system using ensemble machine learning to detect Chronic Obstructive Pulmonary Disease (COPD) which is the best way possible. COPD is a chronic respiratory condition that is typically diagnosed at a later stage, resulting in a rise in patient deaths and increasing healthcare costs. This research aims at creating a fast and available early diagnosis tool that is able to screen the disease in its initial development phase, especially in low-income areas where traditional diagnostic methods like spirometry may not be obtainable.

The project intends to leverage the COPD Patients Dataset from Kaggle, which involves data such as patient demographics, clinical symptoms, and spirometry results. The data will be subjected to a series of preprocessing procedures to handle missing values and normalize features, then it will be analyzed in an exploratory manner to identify key patterns. Various machine learning algorithms such as Random Forest, XGBoost, and K-Nearest- Neighbors (KNN) will be trained, and efficiency of each of them can be measured by various performance metrics which includes accuracy, precision, recall, and AUC-ROC. Lastly, the outcome of these will be combined through voting or stacking using ensemble strategies that will enable them to gain in terms of both predictive accuracy and performance. In the end, a prototype of a web-based application will be built using the Flask framework, which is a tool for the clinicians to enter the data of the patient and get the interpreter of COPD in realtime.

The probable features of the project are an ensemble model which can diagnose diseases with accuracy as high as at least 90% and a practical web tool for the clinical setting. The project comprises four stages: data preprocessing and exploratory analysis (2 weeks), model training and tuning (3 weeks), ensemble implementation (2 weeks), and web tool development (3 weeks).

This project has a valuable role in opening up new horizons of the early detection of COPD, as well as the cost of healthcare and the patients' quality of life. The ethical aspect like personal data protection and a revelation of bias issues would be of the main concern to guarantee the observance of regulations such as GDPR and HIPAA. With the help of machine learning, the main goal of the development is to eliminate the flaws in the current diagnostic methods and at the same time to supply an easily possible and economic way of intervening at an early stage.

*.*

## Appendix B: Project Management

Microsoft Excel and GitHub were effectively utilized together to manage the research project. Excel was primarily used for planning and tracking progress. A Trello was developed to map out the project timeline, highlighting key stages such as data preprocessing, model development, evaluation, and report writing. Additionally, a Kanban-style task tracker was set up with columns like "To Do," "In Progress," and "Completed," where each task was assigned a priority, deadline, and checklist to ensure organized execution and efficient time management.

GitHub served as the platform for version control and collaborative work. The complete codebase—including preprocessing scripts, model implementations, and documentation—was maintained in a GitHub repository. Development was managed through branching and pull requests, enabling systematic code reviews and controlled integration of changes. GitHub Issues and Wikis were also leveraged to report bugs, track feature enhancements, and record project-related decisions, promoting transparency and reproducibility throughout the development process.

Screens screenshot of a computer

AI-generated content may be incorrect.

<https://trello.com/invite/b/67d9c5c1b2759bcd19e89e7a/ATTI96b158b2ec34a61212c353555aba3941A1C487BA/copd>

Github:

## Appendix C: Artefact/Dataset

Dataset: <https://www.kaggle.com/datasets/prakharrathi25/copd-student-dataset/data>

## Appendix D: Screencast

<https://youtu.be/FD_NOl1ENuw>