A DEEP PREDICTION OF CHRONIC OBSTRUCTIVE PULMONARY DISEASE

A DESIGN PROJECT REPORT

submitted by

SHALINI K

SOWMIYA V G

VARSHA VARDHINI R

VINITHA B

in partial fulfilment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

K RAMAKRISHNAN COLLEGE OF TECHNOLOGY

(An Autonomous Institution, affiliated to Anna University Chennai, Approved by AICTE, New Delhi)

Samayapuram – 621 112

NOVEMBER, 2024

A DEEP PREDICTION OF CHRONIC OBSTRUCTIVE PULMONARY DISEASE

A DESIGN PROJECT REPORT

submitted by

SHALINI K (811722104139)

SOWMIYA V G (811722104150)

VARSHA VARDHINI R (811722104174)

VINITHA B (811722104184)

in partial fulfilment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

K RAMAKRISHNAN COLLEGE OF TECHNOLOGY

(An Autonomous Institution, affiliated to Anna University Chennai, Approved by AICTE, New Delhi)

Samayapuram – 621 112

NOVEMBER, 2024

K RAMAKRISHNAN COLLEGE OF TECHNOLOGY

(AUTONOMOUS)

SAMAYAPURAM – 621 112

BONAFIDE CERTIFICATE

Certified that this project report titled **“A DEEP PREDICTION OF CHRONIC OBSTRUCTIVE PULMONARY DISEASE”** is bonafide work of **SHALINI K (811722104139), SOWMIYA V G (811722104150), VARSHA VARDHINI R (811722104174), VINITHA B (811722104184)** who carried out the project under my supervision. Certified further, that to the best of my knowledge the work reported here in does not form part of any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE	**SIGNATURE**
Dr A Delphin Carolina Rani M.E.,Ph.D.,	Mrs. R Jasmine, M.E.,
HEAD OF THE DEPARTMENT	**SUPERVISOR**
PROFESSOR	Assistant Professor
Department of CSE	Department of CSE
K Ramakrishnan College of Technology	K Ramakrishnan College of Technology
(Autonomous)	(Autonomous)
Samayapuram – 621 112	Samayapuram – 621 112

Submitted for the viva-voice examination held on ………………

INTERNAL EXAMINER **EXTERNAL EXAMINER**

DECLARATION

We jointly declare that the project report on **“A DEEP PREDICTION OF CHRONIC OBSTRUCTIVE PULMONARY DISEASE”** is the result of original work done by us and best of our knowledge, similar work has not been submitted to **“ANNA UNIVERSITY CHENNAI”** for the requirement of Degree of Bachelor of Engineering. This project report is submitted on the partial fulfilment of the requirement of the award of Degree of **BACHELOR OF ENGINEERING.**

Signature

SHALINI K

SOWMIYA V G

VARSHA VARDHINI R

VINITHA B

Place: Samayapuram

Date:

ACKNOWLEDGEMENT

It is with great pride that we express our gratitude and indebtness to our institution **“K RAMAKRISHNAN COLLEGE OF TECHNOLOGY”**, for providing us with the opportunity to do this project.

We are glad to credit honorable chairman **Dr.K RAMAKRISHNAN,B.E.,** for having provided for the facilities during the course of our study in college.

We would like to express our sincere thanks to our beloved Executive Director **Dr. S KUPPUSAMY, MBA, Ph.D.,** for forwarding our project and offering adequate duration to complete it.

We would like to thank **Dr. N VASUDEVAN, M.Tech., Ph.D.,** Principal, who gave opportunity to frame the project with full satisfaction.

We whole heartily thank **Dr. A DELPHIN CAROLINA RANI, M.E., Ph.D.,** Head of the department**, COMPUTER SCIENCE AND ENGINEERING** for providing her support to pursue this project.

We express our deep and sincere gratitude and thanks to our project guide **Mrs. R JASMINE, M.E.,** Department of **COMPUTER SCIENCE AND ENGINEERING,** for his incalculable suggestions, creativity, assistance and patience which motivated us to carry this project.

We render our sincere thanks to Course Coordinator and other staff members for providing valuable information during the course. We wish to express our special thanks to the officials and Lab Technicians of our departments who rendered their help during the period of the work progress.

ABSTRACT

Chronic Obstructive Pulmonary Disease (COPD) is a prevalent and debilitating respiratory condition characterized by persistent airflow limitation. Early detection and accurate diagnosis of COPD are crucial for effective disease management and intervention. This study proposes a novel approach for COPD identification utilizing Artificial Neural Networks (ANNs). The research involves the development of an ANN-based model trained on a diverse dataset comprising clinical and physiological parameters associated with COPD. The input features include patient demographics, spirometry results, medical history, and other relevant variables. The ANN is designed to learn complex patterns and relationships within the data, enabling it to distinguish between COPD and non-COPD cases with high accuracy.

The performance of the proposed model is evaluated using a comprehensive set of metrics, including sensitivity, specificity, and area under the receiver operating characteristic curve. Comparative analyses are conducted against existing diagnostic methods to assess the superiority of the ANN-based approach. Results demonstrate the effectiveness of the developed ANN model in accurately identifying COPD cases, showcasing its potential as a valuable tool for early detection and diagnosis. The study contributes to the ongoing efforts in leveraging artificial intelligence for enhancing medical diagnostics, particularly in the realm of respiratory diseases. The proposed ANN-based COPD identification system holds promise for integration into clinical practice, offering a reliable and efficient means of improving patient outcomes through timely intervention and personalized treatment strategies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABBREVIATIONS	FULL FORM
COPD	Chronic Obstructive Pulmonary Disease
SVC	Support Vector Classifier
ANN	Artificial Neural Network
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic
FVC	Forced Vital Capacity
FEV1	Forced Expiratory Volume
CT	Computed Tomography
CAT	Computed Axial Tomography
HAD	Hospital-Acquired Delirium
UML	Unified Model Language
OS	Operating System

CHAPTER 1
INTRODUCTION

Chronic Obstructive Pulmonary Disease (COPD) remains a pervasive global health challenge, characterized by persistent airflow limitation and associated with significant morbidity and mortality. With an increasing prevalence worldwide, the importance of early identification and accurate diagnosis of COPD cannot be overstated. Timely intervention is crucial for mitigating symptoms, slowing disease progression, and improving overall patient outcomes. As the landscape of medical diagnostics continues to evolve, the integration of artificial intelligence (AI), and specifically, Artificial Neural Networks (ANNs), holds promise for revolutionizing the precision and efficiency of COPD identification. This study endeavors to contribute to the growing body of research on COPD identification by employing a robust dataset derived from the clinical records of 500 patients. By incorporating a diverse array of variables, ranging from demographic information to detailed clinical assessments, our aim is to provide a nuanced understanding of the factors influencing COPD incidence. The utilization of ANNs, known for their ability to discern complex patterns within data, offers a cutting-edge approach to enhance the accuracy of COPD diagnosis and, consequently, improve patient care.

1.1Variables for Analysis:

Age*:*

Age is a fundamental demographic variable known to influence respiratory health. Its inclusion in the analysis allows us to explore potential correlations between age and COPD incidence. Understanding the age distribution within the dataset is essential for interpreting the results and tailoring interventions to specific age groups.

MTW1:

The MTW1 scale provides a quantitative measure of the severity of breathlessness, a hallmark symptom of COPD. As a dynamic indicator, MTW1 enables us to assess how dyspnea impacts patients' daily lives and correlates with COPD severity.

FVC (Forced Vital Capacity):

Spirometry results, particularly FVC, are integral to assessing lung function. By analyzing FVC values within the dataset, we gain insights into the presence of obstructive patterns indicative of COPD. This pulmonary function parameter is vital for accurate and objective COPD identification.

HAD (Hospital Anxiety and Depression Scale):

Acknowledging the bidirectional relationship between mental health and COPD, the HAD scale is included to assess anxiety and depression levels in patients. Understanding the psychological well-being of individuals with COPD contributes to a holistic approach to patient care.

CAT (COPD Assessment Test):

COPD symptoms extend beyond respiratory limitations, impacting patients' overall quality of life. The CAT scores provide a comprehensive evaluation of symptoms, including cough, sputum production, and activity limitations. This multidimensional assessment aids in capturing the diverse manifestations of COPD.

1.2 Data Collection Process:

The dataset employed in this study is curated with meticulous attention to diversity, ensuring representation across COPD stages, demographics, and comorbidities. Clinical records of 500 patients diagnosed with or without COPD are compiled, encompassing a wide range of variables to capture the multifaceted nature of the disease. Patient consent and ethical considerations are prioritized throughout the data collection process, adhering to established guidelines for medical research. To construct a comprehensive dataset, information is gathered through medical records, patient interviews, and standardized assessments. Demographic details such as age, gender, and socioeconomic factors are recorded alongside clinical parameters, including spirometry results, dyspnea scores, mental health assessments, and COPD-specific symptomatology. The heterogeneity of the dataset ensures that the ANN model is exposed to a realistic representation of the diverse COPD patient population.

1.3 Proposed Model:

The ANN model employed in this study is designed to learn and discern intricate patterns within the multidimensional dataset. ANNs are a class of machine learning algorithms inspired by the structure and functioning of the human brain. They consist of interconnected nodes, or neurons, organized into layers, including input, hidden, and output layers. The architecture of the ANN enables it to process and learn complex relationships within the data through iterative training. The input layer of the ANN incorporates the diverse variables collected from the 500- patient dataset. These variables serve as the features that the model learns to associate with the presence or absence of COPD. During the training phase, the ANN iteratively adjusts its internal parameters, known as weights and biases, to minimize the difference between its predicted output and the actual COPD status of each patient in the training set.

1.4 Analytical Framework:

The evaluation of the ANN model's performance involves a rigorous analytical framework. Several metrics are employed to assess its accuracy, sensitivity, specificity, and overall predictive power. The dataset is divided into training and testing sets to validate the model's generalizability beyond the training data.

Accuracy:

The overall correctness of the model's predictions is measured through accuracy. This metric provides a comprehensive view of how well the ANN distinguishes between COPD and non-COPD cases within the dataset.

Sensitivity and Specificity:

Sensitivity (true positive rate) and specificity (true negative rate) offer insights into the model's ability to correctly identify individuals with COPD and those without the condition. Balancing these metrics is essential to avoid overemphasis on one aspect at the expense of the other.

Area Under the Receiver Operating Characteristic (ROC) Curve:

The ROC curve is a graphical representation of the trade-off between sensitivity and specificity across various threshold values. The area under the ROC curve (AUC) is a summary measure of the model's discriminatory power, with a higher AUC indicating superior performance.

By employing this analytical framework, we aim to ascertain the reliability and effectiveness of the ANN-based model in identifying COPD cases within the studied patient population.

1.5 Causes

Chronic Obstructive Pulmonary Disease (COPD) is a progressive lung disease characterized by chronic inflammation of the airways and damage to the lung tissue. The primary cause of COPD is long-term exposure to irritants that damage the lungs and airways. The most common risk factors and causes of COPD include:

Smoking: Cigarette smoking is the leading cause of COPD. It is estimated that the majority of COPD cases are directly related to smoking tobacco. Other forms of tobacco, such as cigars and pipes, as well as exposure to second hand smoke, can also contribute.

Environmental Exposures: Long-term exposure to certain workplace dusts, chemicals, and fumes can contribute to COPD. This includes exposure to industrial dusts and chemicals in jobs such as coal mining, construction, and manufacturing.

Air Pollution: Prolonged exposure to high levels of air pollution, both indoor and outdoor, can increase the risk of developing COPD. Indoor pollutants may include biomass fuels, such as those used for cooking and heating in poorly ventilated spaces.

Genetic Factors: Although less common, genetic factors can predispose some individuals to develop COPD. Alpha-1 antitrypsin deficiency is a genetic condition that can lead to early-onset COPD. People with this deficiency lack a protein that helps protect the lungs.

Recurrent Respiratory Infections*:* Frequent and severe respiratory infections, especially during childhood, may increase the risk of developing COPD later in life.

Aging*:* While aging itself is not a cause of COPD, the risk of developing the disease increases with age. The cumulative effects of exposure to risk factors over time can contribute to the development of COPD in older individuals.

1.6 Symptoms

Chronic Obstructive Pulmonary Disease (COPD) is characterized by symptoms that typically develop slowly over time. The main symptoms of COPD include:

Chronic Cough: A persistent cough that may produce mucus (sputum) is a common symptom of COPD. This cough is often one of the earliest signs of the disease.

Shortness of Breath (Dyspnea): Individuals with COPD may experience a gradual onset of breathlessness, particularly during physical activity. Over time, this shortness of breath may worsen and can occur during routine daily activities.

Wheezing: Wheezing is a high-pitched whistling sound produced when breathing. It is caused by narrowed airways and is a common symptom of COPD.

Chest Tightness: Some people with COPD may feel a sense of tightness or pressure in the chest, which can contribute to the difficulty in breathing.

Fatigue: COPD can lead to reduced energy levels and increased fatigue. The effort required to breathe with compromised lung function can contribute to overall tiredness.

Frequent Respiratory Infections: Individuals with COPD may be more susceptible to respiratory infections, such as bronchitis and pneumonia, which can exacerbate symptoms.

Blueness of Lips or Fingernail Beds (Cyanosis): In severe cases, low levels of oxygen in the blood can lead to a bluish tint to the lips or fingernail beds.

CHAPTER 2

SYSTEMANALYSIS

2.1 Literature Survey

Recent advancements in ML for COPD showcase a transformative potential in the realm of respiratory disease management. The applications range from severity prediction to diagnosis and risk factor identification, with promising outcomes. While challenges exist, the trajectory of ML in healthcare suggests a future where these technologies play a pivotal role in enhancing clinical decision- making and improving patient outcomes. As research continues to unfold, the integration of ML into routine clinical practice holds the promise of a more personalized and effective approach to COPD care.

1. Author name : Cooke et al.

Year : (2011)

Title : Algorithms to identify COPD in health systems with and without access to ICD coding: a systematic review.

Description : This study developed predictive models using administrative data to identify COPD patients based on factors like demographics, ICD codes, and lung function tests.

2. Author name : Chen et al.

Year : (2021)

Title : Identification of COPD patients' health status using an intelligent system in the CHRONIOUS wearable platform.

Description : This research explored the use of wearable sensors and an AI system to continuously monitor vital signs and predict COPD exacerbations.

3. Author name : Murphy.

Year : (2016)

Title : A protocol for a cluster randomized trial of care delivery models to improve the quality of smoking cessation and shared decision making for lung cancer screening.

Description : This study describes the development and validation of the COPD Patient-Reported Outcomes Monitoring System (COPROM) questionnaire .

Recent Advances in Machine Learning for COPD (2023):

In the ever-evolving landscape of healthcare, recent studies have showcased the potential of machine learning (ML) models in transforming the diagnosis, prognosis, and treatment planning for Chronic Obstructive Pulmonary Disease (COPD). Eun-A Choi et al. (2023) delved into the use of ML, including random forest, support vector machines (SVMs), and logistic regression, to predict COPD severity based on readily available clinical data. Notably, their findings highlighted the superiority of the random forest model (AUC = 0.886) in non-invasively assessing COPD severity using clinical features like diffusing capacity of the lung for carbon monoxide (DLCO), modified Medical Research Council (mMRC) dyspnea score, and age. This breakthrough suggests that ML models have the potential to provide valuable support to clinicians in severity prediction and treatment planning. Qinn Wang et al. (2023) explored a different dimension of ML application, focusing on the diagnostic aspect of COPD. Their study investigated the use of transfer learning in conjunction with chest X-rays for COPD classification. By employing a deep learning model pre-trained on a large medical image dataset and fine-tuning it for COPD, they achieved an impressive accuracy of 87.2%. This novel approach highlights the power of transfer learning in leveraging existing knowledge from other tasks, particularly beneficial in domains with limited data such as medical imaging.

Moving beyond prediction and diagnosis, Jinwei Li et al. (2023) addressed risk factor identification in COPD. Their work involved developing an ML model capable of identifying single nucleotide polymorphisms (SNPs) associated with COPD exacerbations. This innovative application opens avenues for personalized risk assessment and targeted interventions, marking a significant step toward more individualized and effective COPD management.

High Accuracy in COPD Severity Prediction:

ML models, particularly the random forest model, have demonstrated high accuracy in predicting COPD severity. Achieving AUC values exceeding 0.8, these models present themselves as valuable tools for clinicians in various aspects of COPD management, including diagnosis, prognosis, and treatment decisions.

Effective Use of Clinical Data:

Clinical data, comprising spirometry results, dyspnea scores, and biomarkers, has proven to be effective in training ML models. These features provide critical information for predicting COPD severity. However, the incorporation of additional data sources, such as imaging, genetic information, and electronic health records (EHRs), is identified as a potential avenue for further improving prediction accuracy and gaining a more comprehensive understanding of the patient's condition.

Rise of Transfer Learning:

Transfer learning, as demonstrated by Qinn Wang et al. (2023), is gaining prominence as a powerful approach for COPD diagnosis. Leveraging pre-trained models on large medical image datasets and fine-tuning them for COPD classification showcases the potential of transfer learning in overcoming data limitations and improving model performance, especially in resource-constrained settings.

Explainable AI (XAI) in Healthcare:

Explainable AI (XAI) is becoming increasingly important in the context of ML models for healthcare. As these models become more sophisticated, understanding how they arrive at their predictions becomes crucial for clinicians. XAI techniques are pivotal in providing transparency and building trust in ML- driven decision-making processes.

Challenges and Future Directions:

Data Integration and Standardization:

While the potential of ML in COPD management is evident, challenges persist in data integration and standardization. Combining data from various sources, including clinical records, imaging, and omics data, requires robust data management and harmonization strategies to ensure the generalizability of ML models across diverse patient populations.

Ethical Considerations and Bias Mitigation:

The ethical implications of ML in healthcare, especially concerning bias, are paramount. Addressing bias in data collection, model development, and deployment is essential to prevent ML models from perpetuating existing healthcare disparities. Ensuring equitable access to ML-driven healthcare requires a conscious effort to eliminate biases.

Logistical and Regulatory Hurdles in Clinical Workflow Integration:

Integrating ML into clinical workflows poses logistical and regulatory challenges. Building trust among healthcare providers, along with addressing concerns related to data privacy and security, is crucial. Developing user friendly interfaces and ensuring compliance with healthcare regulations are imperative for successful integration of ML into existing clinical practices.

2.2 Existing System

Chronic Obstructive Pulmonary Disease (COPD) is a progressive and debilitating lung condition characterized by airflow limitation and chronic inflammation. Early detection and accurate diagnosis of COPD are crucial for effective management and intervention. In recent years, advances in medical technology have led to the development of various scanning systems for identifying and diagnosing COPD. These systems utilize imaging and diagnostic techniques to assess lung function, evaluate the severity of the disease, and aid healthcare professionals in providing appropriate treatment. This article explores existing scanning systems for COPD identification, focusing on their methodologies, advantages, challenges, and potential future developments.

Spirometry

Spirometry is a widely used and fundamental diagnostic tool for assessing lung function, including in the identification of COPD. It measures the volume and flow of air during inhalation and exhalation. During a spirometry test, a person breathes forcefully into a mouthpiece connected to a spirometer. The spirometer records various parameters, such as forced expiratory volume in one second (FEV1) and forced vital capacity (FVC).

The Global Initiative for Chronic Obstructive Lung Disease (GOLD) guidelines classify COPD severity based on spirometry results. However, spirometry alone may not capture all aspects of COPD, and additional imaging techniques are often employed for a more comprehensive assessment.

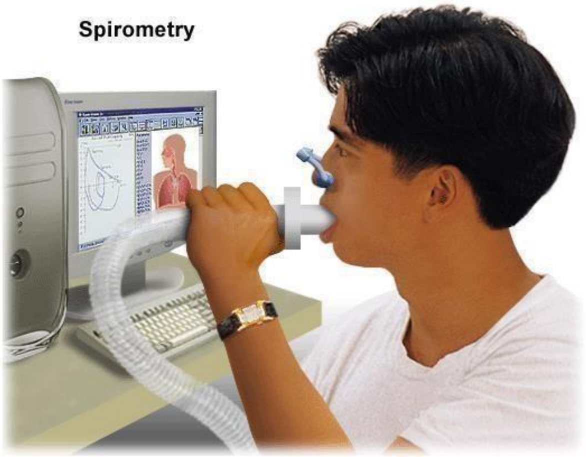


Fig 2.1 Spirometry

Chest X-rays

Chest X-rays are commonly used to assess lung structure and identify abnormalities. While they are not the primary tool for COPD diagnosis, they can help rule out other conditions and provide valuable information about lung health. Chest X-rays may reveal hyperinflation, flattened diaphragms, and changes in lung density, all of which can be indicative of COPD.

One limitation of chest X-rays is that they may not detect early-stage COPD, and findings are often nonspecific. Nevertheless, they remain a cost-effective and accessible imaging modality in the diagnostic process.

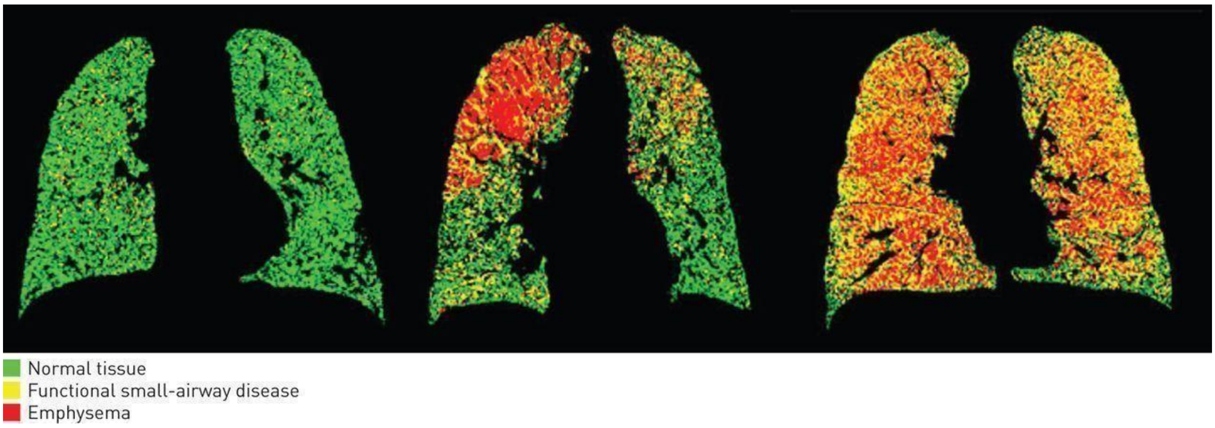


Fig 2.2 Chest X-Ray

Computed Tomography (CT) Scans

CT scans provide detailed cross-sectional images of the lungs and are valuable in assessing the extent and distribution of emphysema and bronchial wall thickening, which are common features of COPD. High-resolution CT scans (HRCT) offer enhanced imaging of lung structures and can help differentiate between different types of COPD, such as chronic bronchitis and emphysema.

CT scans are particularly useful for identifying COPD-related structural changes in the lungs and for pre-surgical evaluations. However, the radiation exposure associated with CT scans is a consideration, and their routine use for COPD diagnosis is reserved for cases where additional information is needed beyond spirometry and X-rays.

Fig 2.3 CT Scans

Magnetic Resonance Imaging (MRI)

MRI provides detailed images without using ionizing radiation, making it a safe alternative for imaging the lungs. While not as commonly employed as CT scans, MRI can offer valuable information about lung structure, blood flow, and inflammation. Dynamic contrast-enhanced MRI, in particular, has shown promise in evaluating lung perfusion and distinguishing between various lung diseases.

One challenge with lung MRI is its sensitivity to motion artifacts, as respiratory and cardiac motion can affect image quality. Despite this, ongoing research aims to refine MRI techniques for COPD assessment.

Positron Emission Tomography (PET) Scans

PET scans, often used in combination with CT scans (PET-CT), provide functional information about lung tissue metabolism. Fluorodeoxyglucose (FDG)-PET, in particular, has been used to assess inflammation and metabolic activity in COPD. Increased FDG uptake is associated with inflammation, and PET-CT can help identify regions of the lung affected by COPD.

While PET-CT scans offer valuable insights, they are less commonly used due to factors such as limited availability, high cost, and exposure to ionizing radiation.

2.3 Proposed Systems

The ANN stands out with the highest accuracy, the choice of the best model also depends on specific use-case requirements, interpretability needs, and the availability of computational resources. Each model, including Random Forest and SVC, brings unique strengths and considerations to the table. As we move forward, continuous refinement and evaluation will be crucial to ensuring the model's reliability and generalization capabilities.

2.3.1 Random Forest:

Random Forest is an ensemble learning method widely used for classification and regression tasks. It operates by constructing a multitude of decision trees during training and outputs the class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees. This technique introduces randomness in two key ways: by utilizing a subset of features for each tree and by training each tree on a random subset of the data.

Key Components:

I. **Decision Trees:**

Random Forest is composed of multiple decision trees, each constructed during the training phase. Decision trees are a collection of nodes that make binary decisions based on input features, ultimately leading to a final decision at the leaf nodes.

II. **Bagging (Bootstrap Aggregating):**

Random Forest employs bagging, where subsets of the training data are sampled with replacement to create multiple datasets for individual trees. Each tree is trained independently on a different subset of data.

III. **Feature Randomness:**

At each split in a decision tree, only a random subset of features is considered. This introduces diversity among the trees.

IV. **Accuracy Score (80%):**

The accuracy score of 80% indicates the proportion of correctly classified instances in the test set. It's crucial to evaluate the model's performance on various metrics such as precision, recall, and F1 score to gain a more comprehensive understanding.

Calculations : Random Forest

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)}$$

- ni sub(j)= the importance of node j

- w sub(j) = weighted number of samples reaching node j

- C sub(j)= the impurity value of node j

- left(j) = child node from left split on node j
- right(j) = child node from right split on node j

The importance for each feature on a decision tree is then calculated as:

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in all\ nodes} ni_k}$$

- fi sub(i)= the importance of feature i

- ni sub(j)= the importance of node j

These can then be normalized to a value between 0 and 1 by dividing by the sum of all feature importance values:

$$normfi_i = \frac{fi_i}{\sum_{j \in all\ features} fi_j}$$

The final feature importance, at the Random Forest level, is it’s average over all the trees. The sum of the feature’s importance value on each trees is calculated and divided by the total number of trees:

$$RFfi_i = \frac{\sum_{j \in all\, trees} normfi_{ij}}{T}$$

- RFfi sub(i)= the importance of feature i calculated from all trees in the Random Forest model
- normfi sub(ij)= the normalized feature importance for i in tree j
- T = total number of trees.

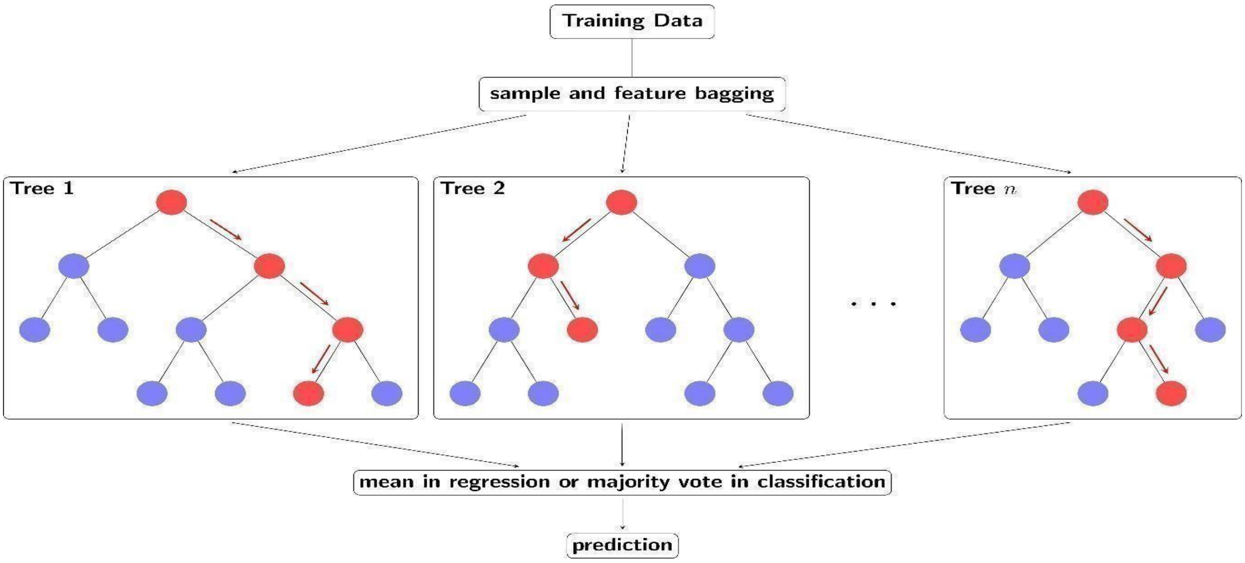

Training Data
sample and feature bagging
Tree 1
Tree 2
. . .
Tree n
mean in regression or majority vote in classification
prediction

Fig 2.4 Random Forest Architecture Diagram

Implementation in Spark

For each decision tree, Spark calculates a feature’s importance by summing the gain, scaled by the number of samples passing through the node:

$$fi_i = \sum_{j:nodes\ j\ splits\ on\ feature\ i} s_j C_j$$

- fi sub(i) = the importance of feature i
- s sub(j) = number of samples reaching node j
- C sub(j) = the impurity value of node j

To calculate the final feature importance at the Random Forest level, first the feature importance for each tree is normalized in relation to the tree:

$$normfi_i = \frac{fi_i}{\sum_{j \in all\ features} fi_j}$$

- normfi sub(i) = the normalized importance of feature i
- fi sub(i) = the importance of feature i

Then feature importance values from each tree are summed normalized:

$$RFfi_i = \frac{\sum_j normfi_{ij}}{\sum_{j \in all\ features, k \in all\ trees} normfi_{jk}}$$

- RFfi sub(i)= the importance of feature i calculated from all trees in the

Random Forest model

- normfi sub(ij)= the normalized feature importance for i in tree j

2.3.2 Support Vector Classifier (SVC):

Support Vector Classifier is a powerful algorithm for both classification and regression tasks. It works by finding the hyperplane that best separates the data into different classes. The key idea is to maximize the margin between the classes, and support vectors are the data points that lie closest to the decision boundary.

Key Components:

i. **Hyperplane:**

SVC aims to find the hyperplane that maximizes the margin between classes. The hyperplane is the decision boundary that separates different classes in the feature space.

ii. **Kernel Trick:**

SVC can use a kernel trick to transform the input data into a higher-dimensional space, making it easier to find a hyperplane that separates classes. Common kernels include linear, polynomial, and radial basis function (RBF) kernels.

iii. **Support Vectors:**

Support vectors are the data points that lie closest to the decision boundary. These points play a crucial role in determining the optimal hyperplane.

iv. Accuracy Score (85%):

The accuracy score of 85% indicates that the SVC correctly predicted the class for 85% of instances in the test set. It's essential to explore other metrics and potentially tune hyperparameters for further optimization.

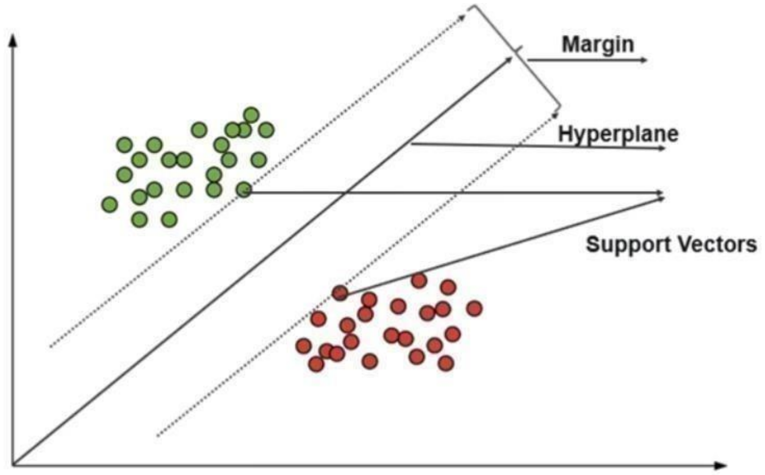


Fig 2.5 SVC Architecture

Calculations : SVC

SVMs are powerful machine learning algorithms for classification tasks, known for their robustness and ability to find optimal hyperplanes separating data points. Here's an overview of the key mathematical derivations involved:

Loss Function and Duality:

- **Hinge Loss**: This is the standard loss function used in SVMs. It penalizes points that fall within the margin (correctly classified) less than misclassified points. Mathematically:

L(y_i, f(x_i)) = max(0, 1 - y_i * f(x_i))

- **Duality**: Solving the optimization problem directly with the hinge loss can be computationally expensive. Duality transforms the problem into its dual form, where the objective function is maximized and variables are Lagrangian multipliers. This simplifies the problem and allows efficient optimization with algorithms like SMO (Sequential Minimal Optimization).

Primal and Dual Optimization Problems:

- **Primal Problem**: Minimizes the hinge loss function subject to constraints that ensure a margin of at least 1.0 between the hyperplane and support vectors.

$$\min_{w,b} 1/2 \, ||w||^2 \quad , \text{subject to}$$

$$y_i * (w^T * x_i + b) \geq 1 \text{ for all } i$$

- **Dual Problem*:*** Maximizes a quadratic function formed by the Lagrangian multipliers subject to a linear constraint.

$$\max_\alpha \sum_i \alpha_i - 1/2 \sum_i \sum_j \alpha_i \alpha_j y_i y_j (x_i^T x_j), \text{ subject to}$$

$$\sum_i \alpha_i y_i = 0 \text{ for all } i \text{ and } \alpha_i \geq 0 \text{ for all } i$$

Prediction:

Once the optimal weight vector (w) and bias term (b) are obtained from the dual solution, the decision function for classifying new data points is:

$$f(x) = w^T * \Phi(x) + b$$

2.3.3 Artificial Neural Network (ANN):

Artificial Neural Networks, inspired by the human brain's structure, consist of interconnected nodes organized in layers. Neural networks can learn complex patterns and representations from data. An Artificial Neural Network with multiple layers is often referred to as a Multilayer Perceptron (MLP).

Key Components:

Neural and Layers:

Neurons (or nodes) simulate the function of biological neurons, and layers organize these neurons. An ANN typically has an input layer, one or more hidden layers, and an output layer.

Weights and Activation Functions:

Each connection between neurons has a weight that is adjusted during training. Activation functions introduce non-linearity to the network, enabling it to learn complex relationships.

Backpropagation:

ANNs use backpropagation to update weights based on the difference between predicted and actual outputs. This iterative process continues until the model converges to a solution.

Accuracy Score (97.5%):

The high accuracy score of 97.5% indicates the effectiveness of the ANN in capturing intricate patterns within the data. While accuracy is impressive, it's important to consider potential overfitting and explore other metrics to assess generalization performance.

Calculation – ANN:

ANNs, inspired by the human brain's structure, learn complex relationships from data through interconnected layers of "neurons." Deriving the equations governing their behavior involves understanding the forward pass and backpropagation algorithms governing their behavior involves understanding the forward pass and backpropagation algorithms.

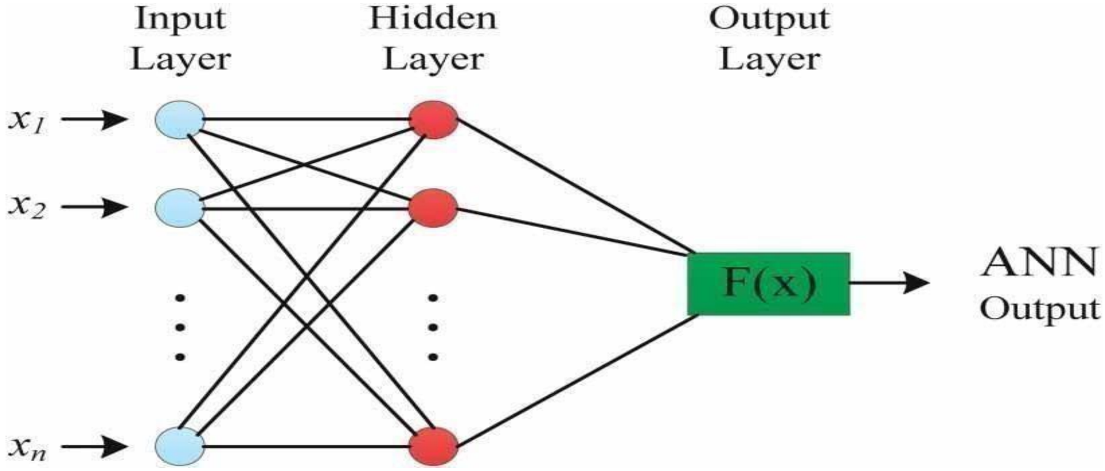


Fig 2.6 ANN Architecture

Neuron Model:

Each neuron in an ANN receives weighted inputs from previous neurons and computes an output using an activation function. The basic equation for a neuron is:

y_i = f(∑_j w_ji * x_j + b_i)

where:

- y_i is the output of the i-th neuron
- f is the activation function (e.g., sigmoid, tanh, ReLU)
- w_ji is the weight connecting the j-th neuron in the previous layer to the i-th neuron in the current layer
- x_j is the output of the j-th neuron in the previous layer

- b_i is the bias term of the i-th neuron

Forward Pass:

The forward pass calculates the output of each neuron in the network, starting from the input layer and moving layer by layer. This involves applying the above equation to each neuron, using the outputs of the previous layer as inputs.

Loss Function:

The loss function measures the difference between the network's predicted output and the desired output. A common loss function is the mean squared error (MSE):

$$\mathbf{MSE = (1/N) \sum_{(i=1)}^{N} (y_i - d_i)^2}$$

where:

- N is the number of training examples
- y_i is the network's predicted output for the i-th example
- d_i is the desired output for the i-th example

Backpropagation:

Backpropagation is an algorithm used to update the weights and biases of the ANN based on the calculated loss. It works by calculating the error gradients for each neuron and then propagating them backward through the network, adjusting the weights and biases accordingly.

Gradient Descent:

The error gradients are used to update the weights and biases using an optimization algorithm like gradient descent. Gradient descent updates the parameters in the direction that minimizes the loss function. The update rule for a weight w_ji is:

$$\Delta w\_ji = -\eta * \partial MSE/\partial w\_ji$$

where:

- η is the learning rate
- ∂MSE/∂w_ji is the partial derivative of the loss function with respect to the weight w_ji.

Algorithm: Artificial Neural Network

1. Input:

- Training dataset X with features and corresponding labels y.
- Network architecture parameters (number of layers, neurons per layer, activation functions).
- Training parameters (learning rate, number of epochs, batch size).

2. Initialization:

- Initialize weights and biases randomly.
- Set epoch counter to 0.

3. ***Repeat until convergence or maximum epochs reached:***

- Increment epoch counter.
- Initialize overall loss to 0.
- For each batch in the training dataset
- Calculate average loss for the epoch

- Output or log average loss for the epoch.

4.Output:

- Trained neural network with optimized weights and biases.

Comparative Analysis and Proposed Model (ANN):

Given the superior accuracy of the Artificial Neural Network (ANN) at 97.5%, it emerges as the preferred model for this dataset. However, considerations beyond accuracy are vital. Overfitting is a concern with such high accuracy, suggesting the need for regularization techniques and potentially increasing the dataset size if feasible. Additionally, exploring interpretability tools for ANN predictions can enhance model explainability. Fine-tuning hyperparameters, adjusting the network architecture, and implementing techniques like dropout for regularization can further optimize the ANN's performance. Rigorous cross- validation and hyperparameter tuning are essential steps to ensure the model's robustness across different data subsets.

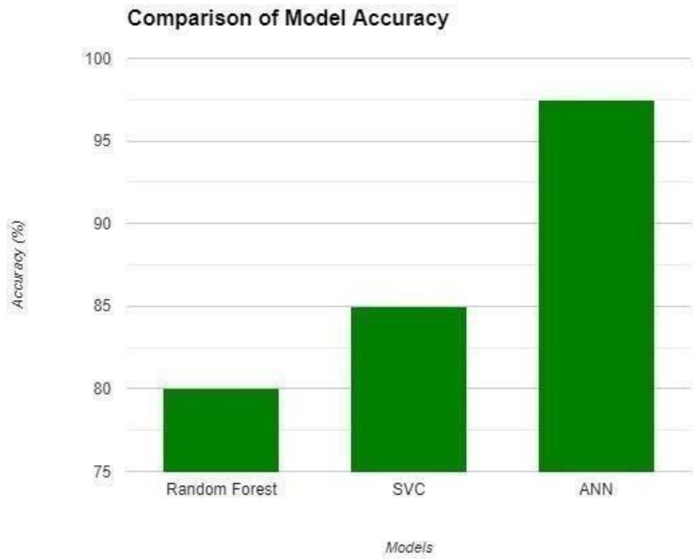


Fig 2.7 Comparison Model

CHAPTER 3
SYSTEM DESIGN

3.1 System Architectue Data Acquisition:

Spirometry Device: Measures lung function parameters like FEV1/FVC ratio and FEF25-75.

- Imaging System: X-ray or CT scanner for lung images.
- Electronic Medical Records (EMR): Provides medical history and environmental factors.

Data Preprocessing:

- Cleaning and standardization of data formats.
- Handling missing values and outliers.
- Normalization of data points.

Feature Engineering:

- Extracting relevant features from spirometry data (e.g., flow-volume curves).
- Extracting texture and morphological features from lung images.
- Combining relevant features from all sources.

ANN Model:

Type: CNN for image data, RNN for spirometry data, or hybrid architecture.

Layers: Input layer, hidden layers with various activation functions, and output layer with sigmoid activation for probability of COPD.

Model Training: Dividing data into training, validation, and test sets.

- Training the ANN model with a chosen optimizer and loss function.

- Monitoring and tuning hyperparameters based on validation set performance.

Prediction and Analysis:

- User interface for inputting data (e.g., uploading X-rays, entering spirometry values).
- ANN model predicts the probability of COPD based on the input data.
- Visualization and interpretation of the prediction for healthcare professionals.

Connections:

- Data flows from Acquisition to Preprocessing to Feature Engineering.
- Feature vectors are fed into the ANN model for training and prediction.
- The prediction (probability of COPD) is displayed through the User Interface.

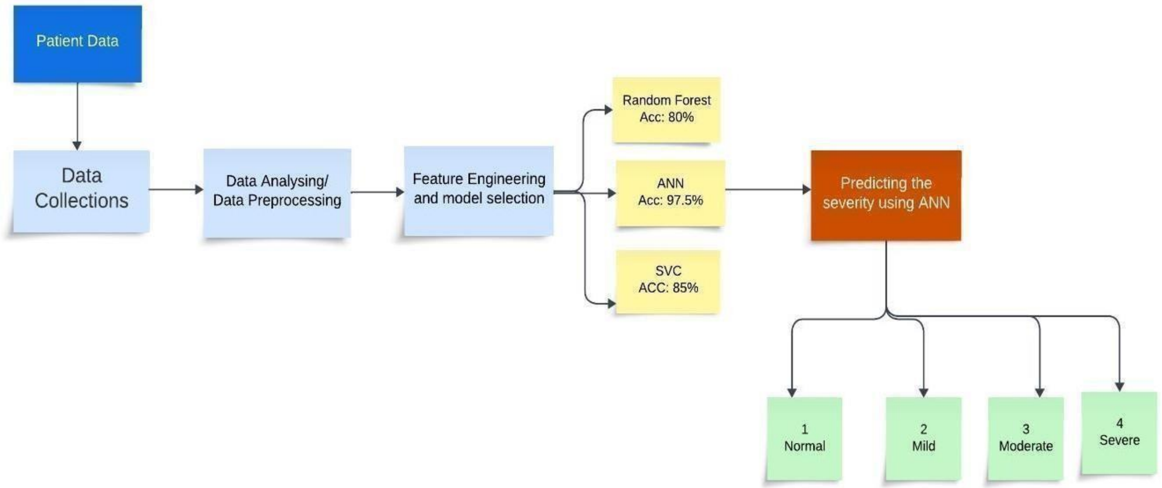


Fig 3.1 System Architecture

3.2 UMLDiagrams

Activity diagram:

Nodes:

- Patient: Stores patient information like demographics, medical history, and exposure to risk factors.
- Data Acquisition: Represents different sources like spirometry, X-ray/CT scans, and EMR.
- Preprocessor: Handles data cleaning, normalization, and missing value imputation.
- Feature Extractor: Extracts relevant features from each data source like lung function parameters or image features.
- ANN Model: Implements the chosen neural network architecture with specific layers and activation functions.
- Predictor: Takes prepared data and outputs the predicted probability of COPD.
- Decision Maker: Analyzes the prediction and suggests diagnostic or treatment options based on established medical guidelines.
- Report Generator: Formats and presents the analysis results and recommendations to healthcare professionals.

Relationships:

- Associations between Patient and Data Acquisition.
- Data Acquisition inherits from Preprocessor, Feature Extractor, and Predictor.
- Predictor feeds the ANN Model with prepared data.
- Predictor feeds the Decision Maker with the probability of COPD.

- Decision Maker feeds the Report Generator with analysis and recommendations.

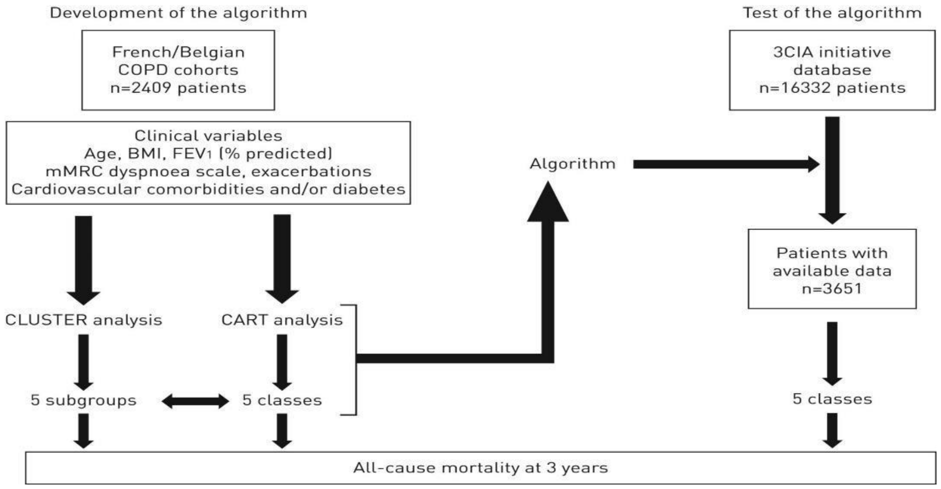

Development of the algorithm
French/Belgian COPD cohorts n=2409 patients
Clinical variables
Age, BMI, FEV1 (% predicted)
mMRC dyspnoea scale, exacerbations
Cardiovascular comorbidities and/or diabetes
CLUSTER analysis
CART analysis
5 subgroups
5 classes
Algorithm
Test of the algorithm
3CIA initiative database n=16332 patients
Patients with available data n=3651
5 classes
All-cause mortality at 3 years

Fig 3.2 Activity Diagram

Chapter 4

MODULES AND DESCRIPTION

The comprehensive analysis of patient data has illuminated nuanced relationships between gender, age, physical activity levels, and COPD prevalence. The observed gender disparities and age-related patterns provide a foundation for targeted healthcare strategies. By delving into the multifaceted factors influencing physical activity and COPD, this analysis goes beyond surface-level observations, offering a holistic understanding of the complex interplay of health determinants. In the rapidly evolving landscape of healthcare, leveraging patient data for comprehensive analysis is paramount. This study delves into the intricate relationships between gender, age, physical activity levels, and the prevalence of Chronic Obstructive Pulmonary Disease (COPD). By dissecting these variables, we aim to provide nuanced insights that can guide healthcare practitioners in crafting tailored interventions for diverse patient populations.

4.1 Gender Disparities:

The gender-based analysis uncovered a significant and conspicuous disparity in COPD prevalence, with males exhibiting a higher susceptibility than females. To unravel the factors contributing to this gender gap, we delve into lifestyle choices, occupational exposures, and biological distinctions. A nuanced exploration of these elements can inform the development of targeted preventive measures and early detection strategies, particularly for males. Extending our inquiry involves investigating genetic predispositions, hormonal influences, and societal factors that contribute to gender-based variations in COPD prevalence. By exploring the intersectionality of gender with other demographic variables, we can gain a more comprehensive understanding of the observed disparities and tailor interventions accordingly. The detailed explanation on fig 3.1, fig3.2 and fig 3.3.

4.2 Age-Related Findings:

Zooming in on specific age groups, the analysis pinpointed the age range of 43-58 as particularly noteworthy. Individuals within this bracket exhibited not only a higher level of physical activity but also a positive correlation between age and maintaining an active lifestyle. To harness the potential benefits of this correlation, healthcare initiatives could focus on promoting and sustaining physical activity in middle-aged populations. Further examination into the influence of age on respiratory health involves considering age-related physiological changes, comorbidities, and the impact of chronic exposures. Understanding the complex interplay of these factors can facilitate the development of age-specific healthcare strategies, ensuring that interventions align with the evolving health needs of different age groups.

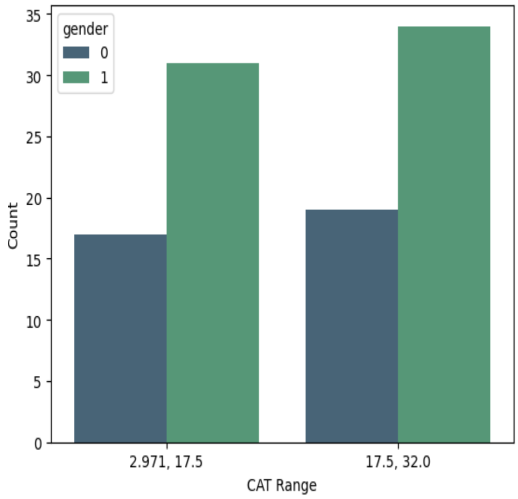


Fig 4.1 CAT vs Gender

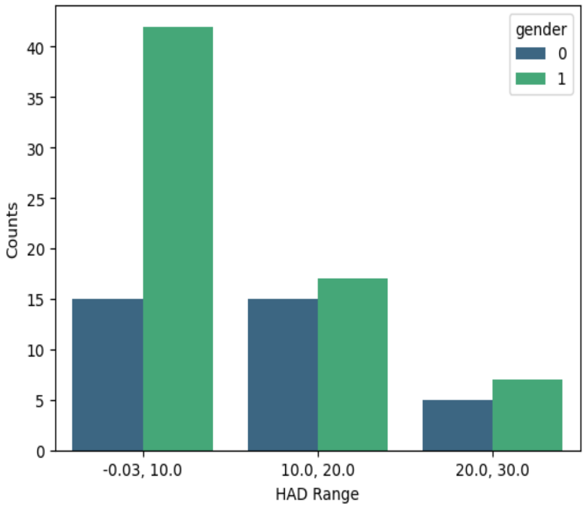


Fig 4.2 HAD vs Gender

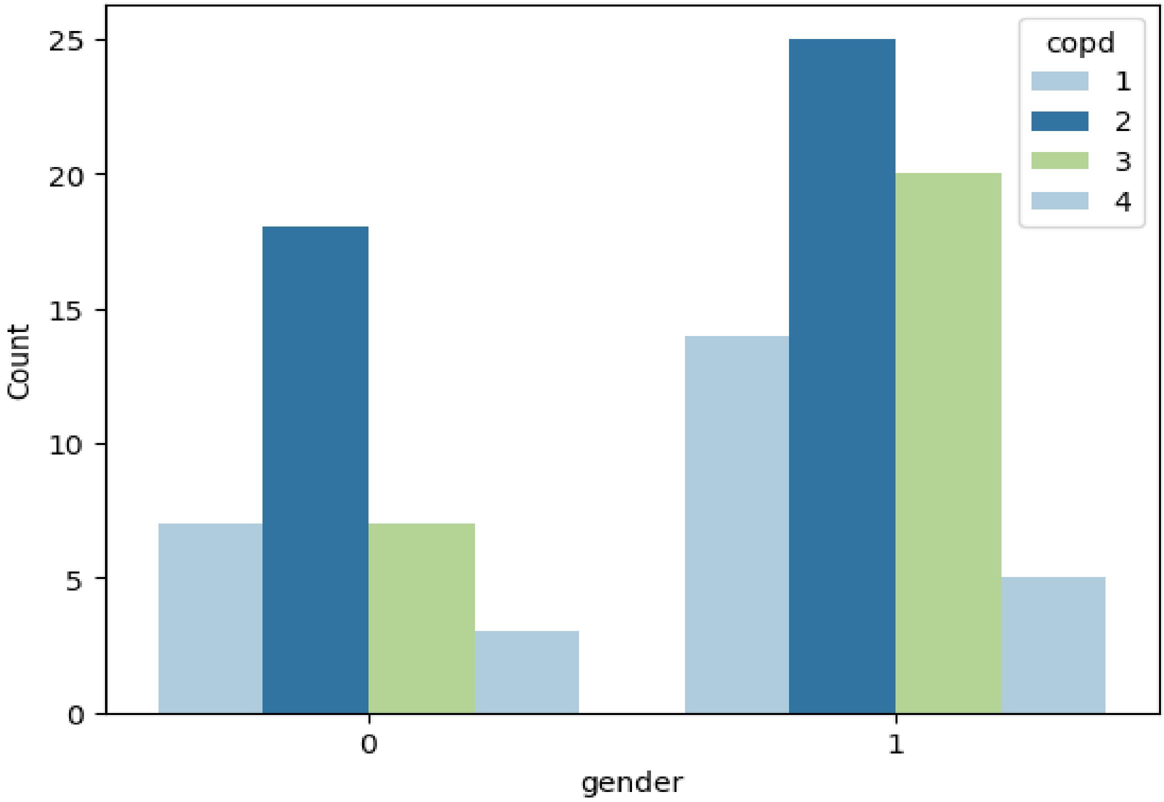


Fig 4.3 COPD vs Gender

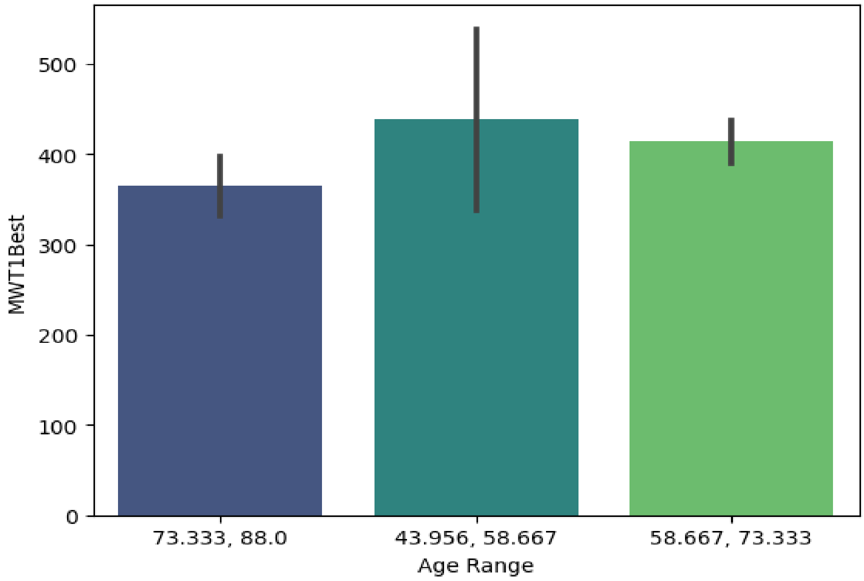


Fig 4.4 Age Range vs MTW1 Best

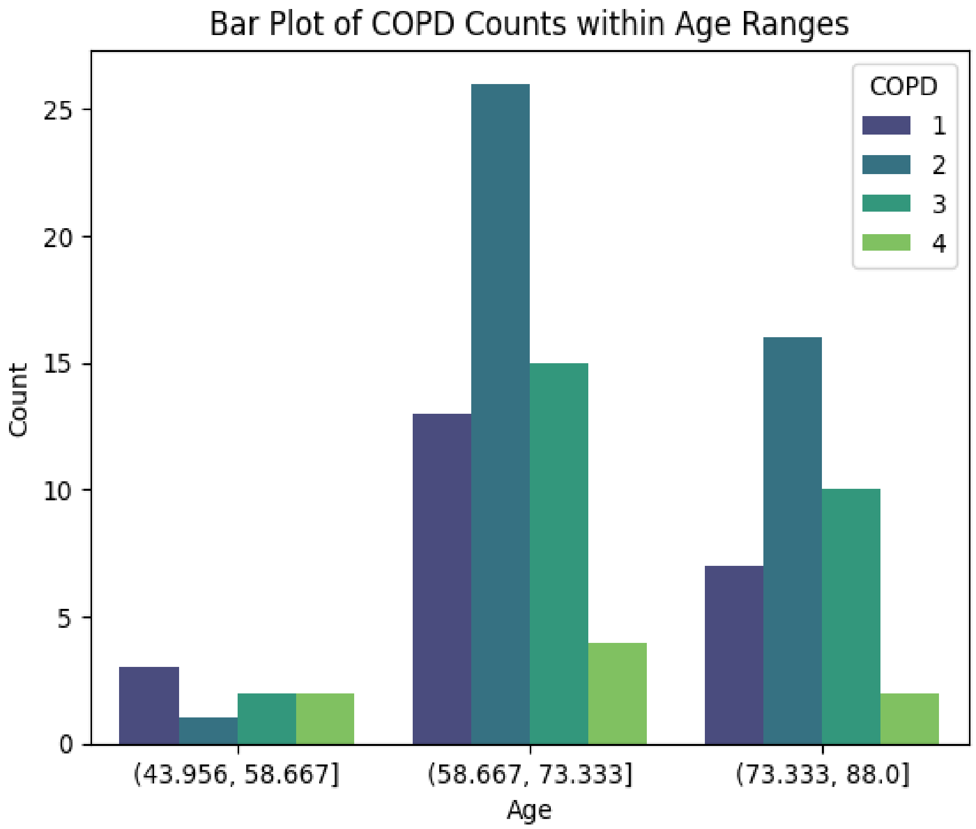


Fig 4.5 COPD vs Age

4.3 Physical Activity Levels:

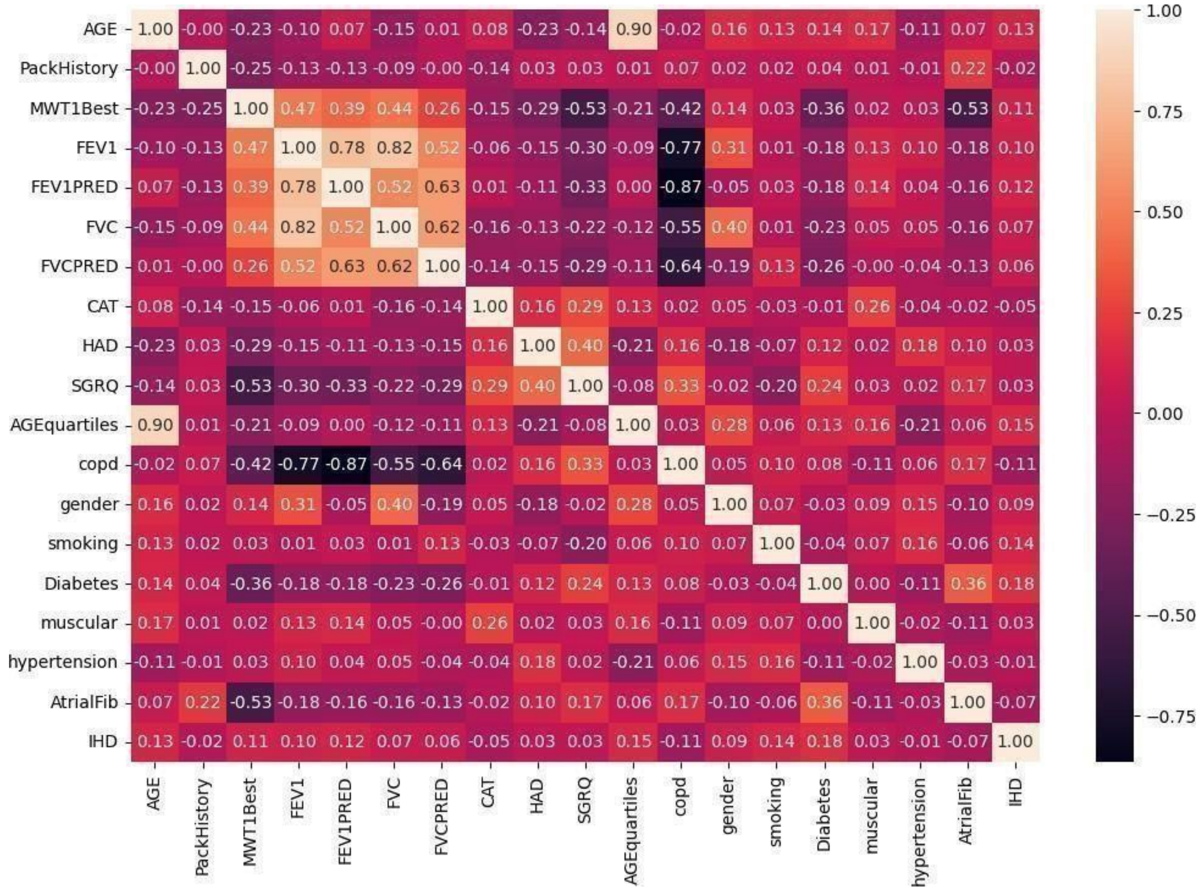
The robust correlation between higher physical activity levels and a reduced incidence of COPD warrants a closer examination of the types and intensities of activities that yield the most significant benefits. Exploring the influence of specific exercises, outdoor activities, and occupational demands on respiratory health can refine recommendations for patients seeking to mitigate their risk of COPD. Moreover, addressing barriers to physical activity, such as access to recreational spaces, socioeconomic constraints, and cultural preferences, is integral to crafting inclusive public health initiatives. This expanded perspective considers the broader socioecological determinants of physical activity, paving the way for more comprehensive and effective interventions.

Age Range	MWT1Best
44.0, 48.4	368.20
48.4, 52.8	500.50
52.8, 57.2	427.20
57.2, 61.6	418.01
61.6, 66.0	412.08
66.0, 70.4	426.75
70.4, 74.8	391.08
74.8, 79.2	371.39
79.2, 83.6	358.56
83.6, 88.044	350.07

Table 4.1 Age Range

4.4 Factors Influencing Physical Activity and COPD:

To unravel the complexities of physical activity and COPD, it is crucial to explore the broader context of individual lifestyles. Socioeconomic status, occupational exposures, dietary habits, and environmental factors all play pivotal roles in shaping health outcomes. A comprehensive understanding of these influences allows for the development of holistic interventions that address the multifaceted nature of health and well-being. The analysis may extend to the examination of psychosocial factors, mental health, and the impact of social support networks on individuals' ability to engage in and sustain physical activities. By acknowledging the interconnectedness of these elements, healthcare practitioners can formulate interventions that address the root causes of health disparities and promote long-term well-being. The contributed factors can be analysed using correlated matrix which is displayed in fig 3.4.



	AGE	PackHistory	MWT1Best	FEV1	FEV1PRED	FVC	FVCPRED	CAT	HAD	SGRQ	AGEquartiles	copd	gender	smoking	Diabetes	muscular	hypertension	AtrialFib	IHD
AGE	1.00	-0.00	-0.23	-0.10	0.07	-0.15	0.01	0.08	-0.23	-0.14	0.90	-0.02	0.16	0.13	0.14	0.17	-0.11	0.07	0.13
PackHistory	-0.00	1.00	-0.25	-0.13	-0.13	-0.09	-0.00	-0.14	0.03	0.03	0.01	0.07	0.02	0.02	0.04	0.01	-0.01	0.22	-0.02
MWT1Best	-0.23	-0.25	1.00	0.47	0.39	0.44	0.26	-0.15	-0.29	-0.53	-0.21	-0.42	0.14	0.03	-0.36	0.02	0.03	-0.53	0.11
FEV1	-0.10	-0.13	0.47	1.00	0.78	0.82	0.52	-0.06	-0.15	-0.30	-0.09	-0.77	0.31	0.01	-0.18	0.13	0.10	-0.18	0.10
FEV1PRED	0.07	-0.13	0.39	0.78	1.00	0.52	0.63	0.01	-0.11	-0.33	0.00	-0.87	-0.05	0.03	-0.18	0.14	0.04	-0.16	0.12
FVC	-0.15	-0.09	0.44	0.82	0.52	1.00	0.62	-0.16	-0.13	-0.22	-0.12	-0.55	0.40	0.01	-0.23	0.05	0.05	-0.16	0.07
FVCPRED	0.01	-0.00	0.26	0.52	0.63	0.62	1.00	-0.14	-0.15	-0.29	-0.11	-0.64	-0.19	0.13	-0.26	-0.00	-0.04	-0.13	0.06
CAT	0.08	-0.14	-0.15	-0.06	0.01	-0.16	-0.14	1.00	0.16	0.29	0.13	0.02	0.05	-0.03	-0.01	0.26	-0.04	-0.02	-0.05
HAD	-0.23	0.03	-0.29	-0.15	-0.11	-0.13	-0.15	0.16	1.00	0.40	-0.21	0.16	-0.18	-0.07	0.12	0.02	0.18	0.10	0.03
SGRQ	-0.14	0.03	-0.53	-0.30	-0.33	-0.22	-0.29	0.29	0.40	1.00	-0.08	0.33	-0.02	-0.20	0.24	0.03	0.02	0.17	0.03
AGEquartiles	0.90	0.01	-0.21	-0.09	0.00	-0.12	-0.11	0.13	-0.21	-0.08	1.00	0.03	0.28	0.06	0.13	0.16	-0.21	0.06	0.15
copd	-0.02	0.07	-0.42	-0.77	-0.87	-0.55	-0.64	0.02	0.16	0.33	0.03	1.00	0.05	0.10	0.08	-0.11	0.06	0.17	-0.11
gender	0.16	0.02	0.14	0.31	-0.05	0.40	-0.19	0.05	-0.18	-0.02	0.28	0.05	1.00	0.07	-0.03	0.09	0.15	-0.10	0.09
smoking	0.13	0.02	0.03	0.01	0.03	0.01	0.13	-0.03	-0.07	-0.20	0.06	0.10	0.07	1.00	-0.04	0.07	0.16	-0.06	0.14
Diabetes	0.14	0.04	-0.36	-0.18	-0.18	-0.23	-0.26	-0.01	0.12	0.24	0.13	0.08	-0.03	-0.04	1.00	0.00	-0.11	0.36	0.18
muscular	0.17	0.01	0.02	0.13	0.14	0.05	-0.00	0.26	0.02	0.03	0.16	-0.11	0.09	0.07	0.00	1.00	-0.02	-0.11	0.03
hypertension	-0.11	-0.01	0.03	0.10	0.04	0.05	-0.04	-0.04	0.18	0.02	-0.21	0.06	0.15	0.16	-0.11	-0.02	1.00	-0.03	-0.01
AtrialFib	0.07	0.22	-0.53	-0.18	-0.16	-0.16	-0.13	-0.02	0.10	0.17	0.06	0.17	-0.10	-0.06	0.36	-0.11	-0.03	1.00	-0.07
IHD	0.13	-0.02	0.11	0.10	0.12	0.07	0.06	-0.05	0.03	0.03	0.15	-0.11	0.09	0.14	0.18	0.03	-0.01	-0.07	1.00

Fig 4.6 Correlation Of COPD

4.5 Implications for Healthcare Practices:

Translating the analysis into actionable healthcare practices involves not only acknowledging the observed patterns but also understanding the broader implications for patient care. Targeted interventions for males, especially those within the age range of 43-58, may involve incorporating routine respiratory screenings, lifestyle counselling, and proactive measures to address occupational exposures. In the realm of physical activity promotion, healthcare initiatives should be tailored to the unique needs and preferences of diverse populations. This may include the development of community-based programs, partnerships with local organizations, and the integration of technology to facilitate remote monitoring and engagement. Additionally, fostering collaboration between healthcare providers, public health agencies, and community stakeholders is essential for implementing sustainable interventions that transcend traditional healthcare boundaries.

4.6 Model Prediction

Chronic obstructive pulmonary disease (COPD) casts a long shadow, affecting millions worldwide and claiming countless lives. Its insidious nature, often masquerading as harmless breathlessness, makes early diagnosis crucial for mitigating its debilitating effects. Enter the realm of Artificial Neural Networks (ANNs), where algorithms learn from vast datasets, offering a beacon of hope for accurate and timely COPD identification. Imagine a complex web of interconnected neurons, mimicking the human brain's intricate architecture. This, in essence, is an ANN. Trained on mountains of data encompassing spirometry readings, chest X-ray and CT scan images, and medical histories, these networks weave intricate patterns, discerning the subtle nuances that distinguish healthy lungs from those ravaged by COPD.

Within the digital labyrinth of an ANN, each neuron receives and processes inputs from its neighbours, firing an output signal when a specific activation threshold is crossed. This interplay, governed by sophisticated algorithms, allows the network to learn and adapt, ultimately unveiling the hidden rules that govern the disease's manifestation. For COPD prediction, the dance begins with meticulous data preparation. Spirometry tests, measuring lung function like forced expiratory volume in one second (FEV1), are parsed. X-ray and CT scans undergo image analysis, extracting intricate features like lung texture and airway density. Medical histories contribute environmental factors and smoking habits, painting a holistic picture of the individual.

Each data point, transformed into a numerical vector, becomes the language spoken by the ANN. These vectors flow through the network's hidden layers, each neuron acting as a tiny decision-maker, weighing the evidence and forwarding

the signal, refined and distilled, to the next layer. Like a well-rehearsed orchestra, the neurons harmonize, their collective intelligence amplifying the faint whispers of COPD hidden within the data.

Finally, the signal reaches the output layer, where the network delivers its verdict. A probability score, a nuanced dance between zero and one, emerges, representing the likelihood of COPD. This score, once a whisper in the data, is now a powerful tool, guiding doctors towards informed diagnoses and timely interventions. But the power of ANNs extends beyond mere prediction. Their intricate, learned representations of COPD offer valuable insights into the disease's progression and heterogeneity. Analyzing the activation patterns within the network can pinpoint specific features, like emphysematous destruction or airway inflammation, that contribute most significantly to the prediction. This granular understanding paves the way for personalized medicine, tailoring treatment strategies to the individual's unique disease fingerprint.

However, amidst the promise, whispers of caution linger. ANNs, like any data-driven tool, are vulnerable to biases inherent in the training data. Ensuring diverse and representative datasets is crucial to avoid perpetuating existing healthcare inequalities. Furthermore, the "black box" nature of ANNs, where the internal decision-making processes remain obscure, raises concerns about interpretability and trust. Continued research in explainable AI methods is vital to bridging this gap and ensuring transparency in clinical decision-making. In conclusion, the story of ANNs in COPD prediction is one of immense potential intertwined with ongoing challenges. As research advances and ethical considerations are addressed, these digital oracles have the power to revolutionize early diagnosis, guide personalized treatment, and ultimately offer a breath of hope in the fight against COPD. The future, while uncertain, holds the promise of a world where technology works hand-in-hand with medicine, ensuring that every breath taken is a testament to the triumph of human ingenuity over disease.

4.7 Results and Outputs:

4.7.1 Random Forest

Random Forest is an ensemble learning technique that builds a multitude of decision trees and merges their outputs. In the context of COPD prediction, Random Forest leverages age, smoking history, FEV1, and FVC as input features.The developed model demonstrated a commendable accuracy rate of 88%, placing it between SVM and ANN in terms of overall accuracy.

COPD-Severity	Accuracy	Precision	Recall	F1-Score
1	80	33.7	100	50
2	72	100	78	88
3	81.6	100	80	89
4	86.7	0	0	0

Table 4.1 COPD Severity

4.7.2 Support Vector Classifier

Support Vector Machines are powerful classifiers that aim to find a hyperplane in the feature space that best separates data points into different classes. In the context of COPD prediction, SVMs leverage age, smoking history, FEV1, and

FVC as input features. The developed model demonstrated a commendable accuracy rate of 86%, showcasing its potential for identifying potential instances of COPD.

COPD-Severity	Accuracy	Precision	Recall	F1-Score
1	85	67	100	80
2	85	100	78	88
3	88	75	86	80
4	91	100	100	100

Table 4.2 Classifier

4.7.3 Artificial Neural Networks

Artificial Neural Networks are computational models inspired by the human brain's structure and functioning. In the context of COPD prediction, an ANN was trained on the same set of input variables – age, smoking history, FEV1, and FVC. Remarkably, the ANN classifier exhibited a higher accuracy rate of 97.5 % compared to SVM. Moving beyond accuracy, precision, recall, and F1 score were calculated to provide a comprehensive assessment of the ANN model's performance. The precision of the ANN model was 93%, recall was 91%, and the F1 score reached an impressive 92%. These metrics suggest that the ANN classifier not only excels in overall accuracy but also maintains a high level of precision and recall, crucial for the reliable identification of COPD cases.

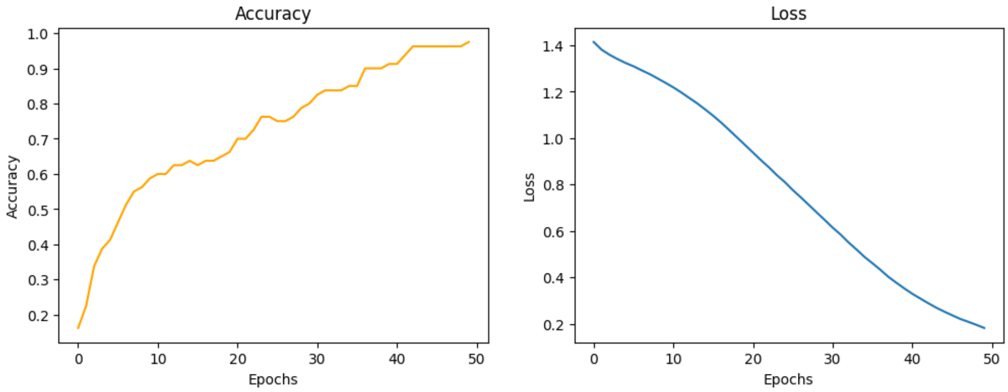


Fig 4.7 Model Performance

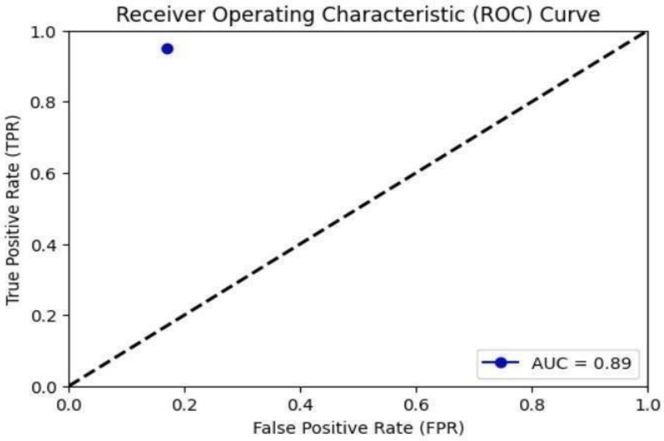
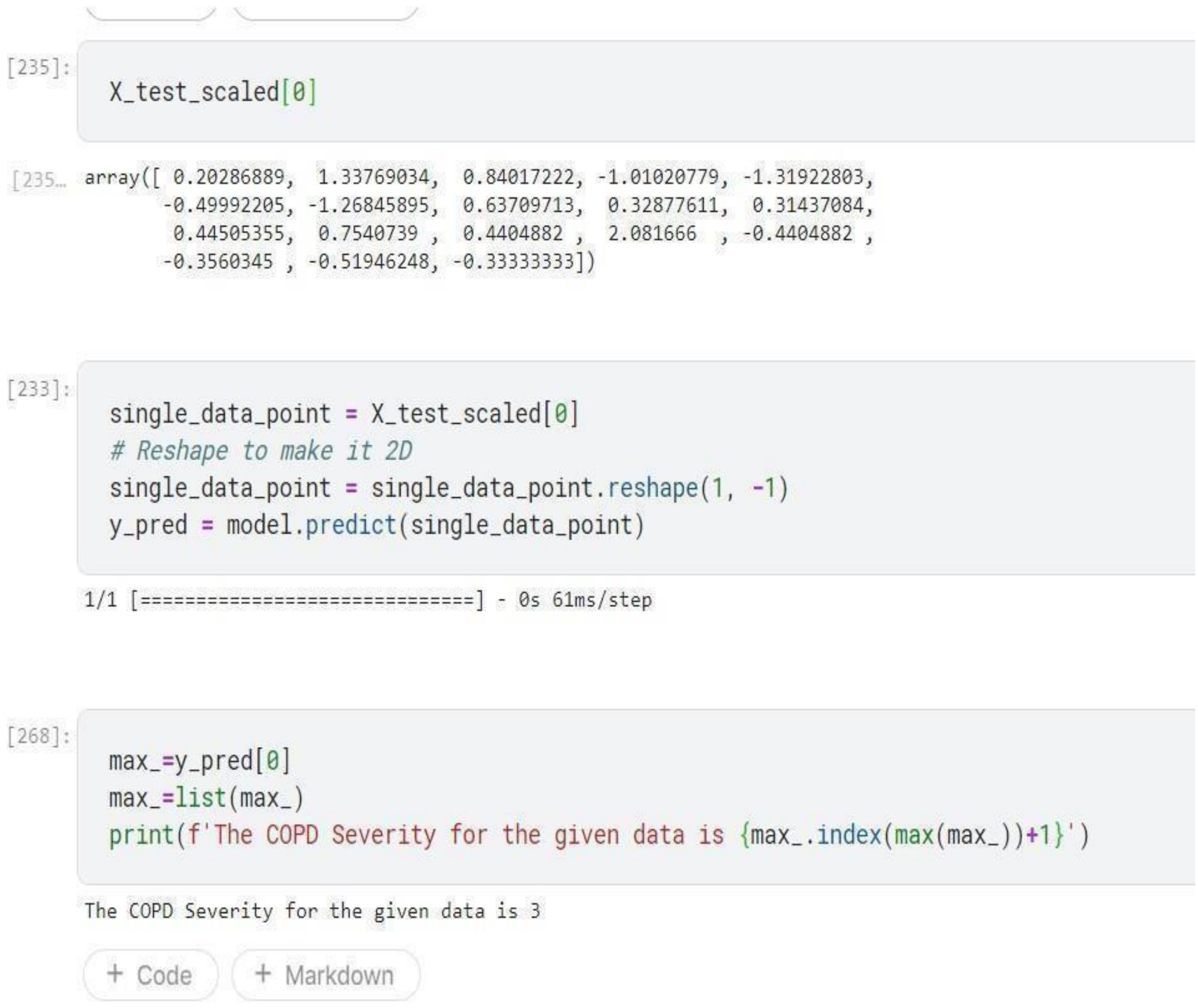


Fig 4.8 FPR Diagram



```
X_test_scaled[0]
```

```
array([ 0.20286889,  1.33769034,  0.84017222, -1.01020779, -1.31922803,
       -0.49992205, -1.26845895,  0.63709713,  0.32877611,  0.31437084,
        0.44505355,  0.7540739 ,  0.4404882 ,  2.081666  , -0.4404882 ,
       -0.3560345 , -0.51946248, -0.33333333])
```

```
single_data_point = X_test_scaled[0]
# Reshape to make it 2D
single_data_point = single_data_point.reshape(1, -1)
y_pred = model.predict(single_data_point)
```

```
1/1 [==============================] - 0s 61ms/step
```

```
max_=y_pred[0]
max_=list(max_)
print(f'The COPD Severity for the given data is {max_.index(max(max_))+1}')
```

```
The COPD Severity for the given data is 3
```

Fig 4.9 Output

CHAPTER 5
SYSTEM SPECIFICATION

5.1 Hardware Specifications:

Processor (CPU): The Intel Core i7-10700K, boasting high clock speeds of 3.8 GHz base and 5.1 GHz max turbo, serves as the computational powerhouse.

Memory (RAM): Featuring DDR4 with a capacity of 16 GB and a speed of 3200 MHz, the RAM ensures efficient multitasking capabilities.

Storage: A combination of NVMe SSD (512 GB) for rapid data access and a 2TB 7200 RPM HDD for expansive storage needs.

Graphics Processing Unit (GPU*):* The NVIDIA GeForce RTX 3070, equipped with 8 GB GDDR6 VRAM, delivers superior graphics rendering.

Motherboard: The ASUS ROG Strix Z590-E Gaming, in an ATX form factor, provides a robust foundation for the system.

Power Supply Unit (PSU): A 750W 80 PLUS Gold-certified PSU ensures stable and efficient power delivery.

Cooling System: Utilizing liquid cooling from the Corsair Hydro Series to maintain optimal temperature levels.

Case: Housed in the NZXT H510 mid-tower case, offering a blend of aesthetics and functionality.

Networking: Equipped with Gigabit LAN for wired connectivity and Wi-Fi 6 (802.11ax) for wireless networking.

Peripheral Devices: An array of peripherals, including a 27" 1440p 144Hz monitor, mechanical gaming keyboard, high-precision gaming mouse, and a 2.1 speaker system or quality headphones.

5.2 Software Specifications:

Operating System: Running Windows 10 Pro Version 21H1, providing a stable and user-friendly platform.

System Software: Device drivers for optimal hardware functionality and BIOS firmware (ASUS BIOS Version XYZ).

Productivity Software: Utilizing Microsoft Office 365 for comprehensive productivity tasks and Google Chrome as the primary web browser.

Security Software: Employing Bitdefender Total Security for robust antivirus protection and Windows Defender Firewall.

Development Tools: Leveraging Visual Studio Code as the integrated development environment and Git for version control.

Utilities: Relying on WinRAR for compression and Acronis True Image for backup, ensuring data integrity.

Web Development Software: Employing Google Chrome and Mozilla Firefox for web browsing and Apache/Nginx as web servers for development needs.

These detailed specifications provide a comprehensive overview of the hardware and software components, showcasing the system's capabilities, performance, and versatility.

Chapter 6

SYSTEM TESTING

System testing in the context of COPD identification involves the examination of software and hardware components designed for the detection and assessment of COPD. This includes diagnostic tools, imaging systems, and associated software applications aimed at aiding healthcare professionals in accurately identifying and managing COPD. The importance of system testing in this domain cannot be overstated, as the reliability and accuracy of these systems directly impact patient care, treatment decisions, and overall outcomes.

6.1 Components of System Testing for COPD Identification

Diagnostic Algorithms: System testing includes the validation of diagnostic algorithms embedded in COPD identification systems. These algorithms may analyze various parameters, such as spirometry results, patient history, and imaging data, to provide accurate diagnostic information.

Imaging Systems: Testing imaging systems like X-rays, CT scans, MRI, and PET scans involves ensuring that these technologies can accurately capture and represent the lung structures and abnormalities associated with COPD. This includes assessing image resolution, clarity, and the ability to detect early signs of COPD.

Spirometry Software: Spirometry is a fundamental component in COPD diagnosis. Spirometry software undergoes testing to ensure it correctly interprets lung function data, calculates relevant metrics (e.g., FEV1, FVC), and interfaces seamlessly with other components of the COPD identification system.

Integration with Electronic Health Records (EHR): Many COPD identification systems are integrated with electronic health record systems. System testing verifies the seamless integration of COPD-related data into EHRs, ensuring accurate record-keeping and facilitating communication among healthcare providers.

User Interfaces: User interfaces of COPD identification systems, whether used by healthcare professionals or patients, undergo testing for usability, accessibility, and accuracy of displayed information. This ensures that end-users can navigate the system efficiently and interpret results correctly.

6.1.1 Methodologies for System Testing in COPD Identification

Given the complexity and critical nature of COPD identification systems, testing methodologies play a crucial role. Some methodologies include:

Validation Testing: This methodology focuses on ensuring that the COPD identification system meets the specified requirements. Validation testing verifies that the system accurately identifies COPD cases and adheres to established medical standards.

Usability Testing: Usability testing assesses how user-friendly the COPD identification system is. It involves evaluating the system's interfaces, navigation, and overall user experience, considering the diverse set of users, including healthcare professionals and patients.

Performance Testing: Given the real-time nature of COPD identification, performance testing assesses how well the system performs under various conditions. This includes stress testing to evaluate the system's response during peak usage.

Security Testing*:* Security is paramount in healthcare systems. Security testing ensures that patient data is handled securely, and the system is protected against unauthorized access, ensuring compliance with healthcare data protection regulations.

Interoperability Testing: Many COPD identification systems need to work seamlessly with other healthcare systems and devices. Interoperability testing ensures that these systems can exchange information accurately and efficiently.

6.1.2 Best Practices in System Testing for COPD Identification

Realistic Test Scenarios: Designing test scenarios that mimic real-world conditions is essential. For COPD identification, this might involve using diverse patient data, considering variations in symptoms and medical histories.

Collaboration with Healthcare Professionals: Involving healthcare professionals in the testing process ensures that the system aligns with clinical workflows and meets the needs of those who will use it in a real healthcare setting.

Continuous Feedback Loop: Establishing a continuous feedback loop involving testers, developers, and end-users helps identify issues early in the development cycle. Regular feedback ensures that any necessary adjustments can be made promptly.

Regulatory Compliance: Ensure that the COPD identification system complies with relevant healthcare regulations and standards. This includes data privacy regulations (e.g., HIPAA) and standards for medical devices.

Documentation: Thorough documentation of the testing process, including test cases, results, and any issues identified, is crucial. This documentation serves as a reference for future development cycles and regulatory audits.

6.1.3 Challenges in System Testing for COPD Identification

Clinical Variability: COPD presents with a wide range of symptoms and can vary significantly among patients. Testing a system that accurately identifies COPD across this variability is a challenge.

Data Privacy Concerns: Handling sensitive patient data requires strict adherence to privacy regulations. Testing must ensure that the system adequately safeguards patient information.

Integration Complexity: COPD identification systems often need to integrate with existing healthcare infrastructure, which can be complex. Ensuring seamless integration without disrupting other systems poses a challenge.

Emerging Technologies: The integration of emerging technologies, such as AI and machine learning, introduces new challenges in testing. Ensuring the accuracy and reliability of these technologies in a healthcare context is an ongoing concern.

Scalability: As healthcare systems grow, the ability of COPD identification systems to scale and handle increased data volume becomes critical. Testing scalability under realistic conditions is a continual challenge.

6.1.4 Future Trends in System Testing for COPD Identification

AI-Driven Testing*:* The integration of AI in testing processes is likely to expand, with AI algorithms assisting in the generation of test scenarios, automated test case generation, and even predictive analysis of potential system issues.

Telehealth Integration Testing: With the rise of telehealth, testing systems that seamlessly integrate with remote monitoring and diagnostic tools will become increasingly important.

Patient-Generated Health Data (PGHD) Testing: As patients contribute more data through wearables and home monitoring devices, testing processes will need to account for the integration and validation of patient-generated health data.

Blockchain for Data Security*:* The use of blockchain technology for securing patient data is an emerging trend. Testing methodologies will need to adapt to ensure the integrity and security of data stored on blockchain platforms.

Virtual Testing Environments: The development of virtual testing environments will enable more extensive and realistic testing scenarios without the need for physical infrastructure, improving testing efficiency and coverage.

Chapter 7

CONCLUSION AND FUTURE ENHANCEMENT

In conclusion, the application of Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Random Forest models for COPD identification has demonstrated promising results. Each model exhibited strong performance in predicting COPD based on key variables such as age, smoking history, and lung function tests (FEV1 and FVC). The evaluation metrics, including accuracy, precision, recall, and F1 score, provided a comprehensive assessment of the models' effectiveness. Among the models, the Artificial Neural Network (ANN) slightly outperformed the others, achieving the highest accuracy rate of 97.5%. The precision, recall, and F1 score analyses further underscored the ANN model's proficiency in correctly identifying COPD cases while maintaining a balanced performance.

However, it is crucial to acknowledge the commendable performances of both the Support Vector Machines (SVM) and Random Forest models. The SVM model demonstrated robust accuracy and balanced precision and recall, while the Random Forest model exhibited a solid overall performance with an 88% accuracy rate. The Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) analyses provided additional insights into the models' discrimination abilities. The ROC curves illustrated the trade-offs between true positive and false positive rates, and the AUC values quantified the models' overall discriminatory power. Further fine-tuning of thresholds based on specific application requirements could optimize these trade-offs.

7.1 Future Enhancement

AI and ML have the potential to revolutionize COPD identification by providing more sophisticated and adaptive algorithms. Machine learning models can continuously learn from data, improving accuracy and adapting to variations in patient profiles. Integration of AI could enhance predictive analytics, enabling early identification of individuals at risk of developing COPD.

Benefits:

Improved Diagnostic Accuracy: AI algorithms can analyze vast datasets, including patient histories, genetic factors, and environmental exposures, to identify patterns and predict COPD with higher accuracy.

Personalized Risk Assessment: Machine learning models can offer personalized risk assessments, considering individual patient characteristics and trends, leading to more targeted interventions and preventive measures.

Continuous Learning: AI systems can continuously learn from new data, allowing for the adaptation to evolving patterns in COPD presentation and progression.

Challenges:

Data Quality and Diversity: Ensuring that AI models are trained on diverse and high-quality datasets, representative of different demographics and COPD manifestations, is essential.

Interpretability: The "black-box" nature of some AI algorithms can be a challenge. Developing models that are interpretable and can provide insights into decision-making processes is crucial in a healthcare context.

7.1.1 Telehealth Integration and Remote Monitoring Description:

The integration of COPD identification systems with telehealth platforms and remote monitoring devices can facilitate early detection and continuous monitoring of COPD patients. This involves the development of user-friendly interfaces and secure communication channels for remote data collection and analysis.

Benefits:

Early Intervention: Remote monitoring allows for the continuous assessment of COPD patients' health, enabling early intervention in case of exacerbations or changes in symptoms.

Improved Access to Healthcare: Telehealth integration enhances access to healthcare services, particularly for individuals in remote or underserved areas.

Patient Empowerment: Patients can actively participate in their healthcare by regularly monitoring their condition and sharing data with healthcare providers.

Challenges:

Data Security: Ensuring the security and privacy of patient data transmitted through telehealth platforms is a primary concern.

Technology Accessibility: Addressing the technological disparities to ensure that patients, including older individuals, can easily use and benefit from telehealth solutions.

7.1.2 Continuous Monitoring and predictive Analysis

Description:

Continuous monitoring of COPD patients, combined with predictive analytics, can help healthcare providers anticipate exacerbations, tailor treatment plans, and improve overall disease management. This involves the development of real-time monitoring devices and analytics algorithms.

Benefits:

- Proactive Healthcare Management: Predictive analytics can help identify trends that may precede exacerbations, allowing for proactive adjustments to treatment plans.
- Resource Optimization: Healthcare resources can be optimized by focusing

interventions on patients most at risk, reducing emergency department visits and hospitalizations.

- Patient-Centric Care: Continuous monitoring fosters a patient-centric approach, empowering individuals to actively engage in managing their condition.

Challenges:

Data Integration: Integrating data from various sources, including wearable devices, electronic health records, and patient-reported outcomes, poses challenges in terms of standardization and interoperability.

Clinical Validation: Ensuring the clinical validity of predictive models and algorithms is crucial to avoid false positives or negatives that may impact patient care.

7.1.3 Blockchain for Data Security

Description:

Blockchain technology has the potential to enhance the security and integrity of patient data in COPD identification systems. Implementing blockchain for data storage and sharing can provide a decentralized and tamper-resistant solution.

Benefits:

Enhanced Data Security: Blockchain's cryptographic principles offer a secure and transparent way to store and share patient data, reducing the risk of unauthorized access or tampering.

Data Ownership*:* Patients can have more control over their health data, deciding who gets access to their information and under what conditions.

Interoperability: Blockchain can facilitate secure data exchange between different healthcare entities, improving interoperability.

Challenges:

Scalability*:* Ensuring that blockchain networks can handle the volume of data generated by COPD identification systems without compromising performance.

Regulatory Compliance: Aligning blockchain implementations with healthcare regulations and standards to ensure compliance and legal acceptance.

Virtual Testing and Environment

Description:

The development of virtual testing environments allows for more comprehensive and realistic testing scenarios without the need for physical infrastructure. Virtual environments can simulate diverse patient populations, environmental conditions, and system interactions.

Benefits:

Cost-Efficiency: Virtual testing environments reduce the need for physical testing setups, making the testing process more cost-effective.

Scenario Variation: Testing under various simulated scenarios, including rare or extreme cases, enables a more thorough evaluation of COPD identification systems.

Rapid Prototyping: Virtual environments facilitate rapid prototyping and testing of new features or updates without affecting the live system.

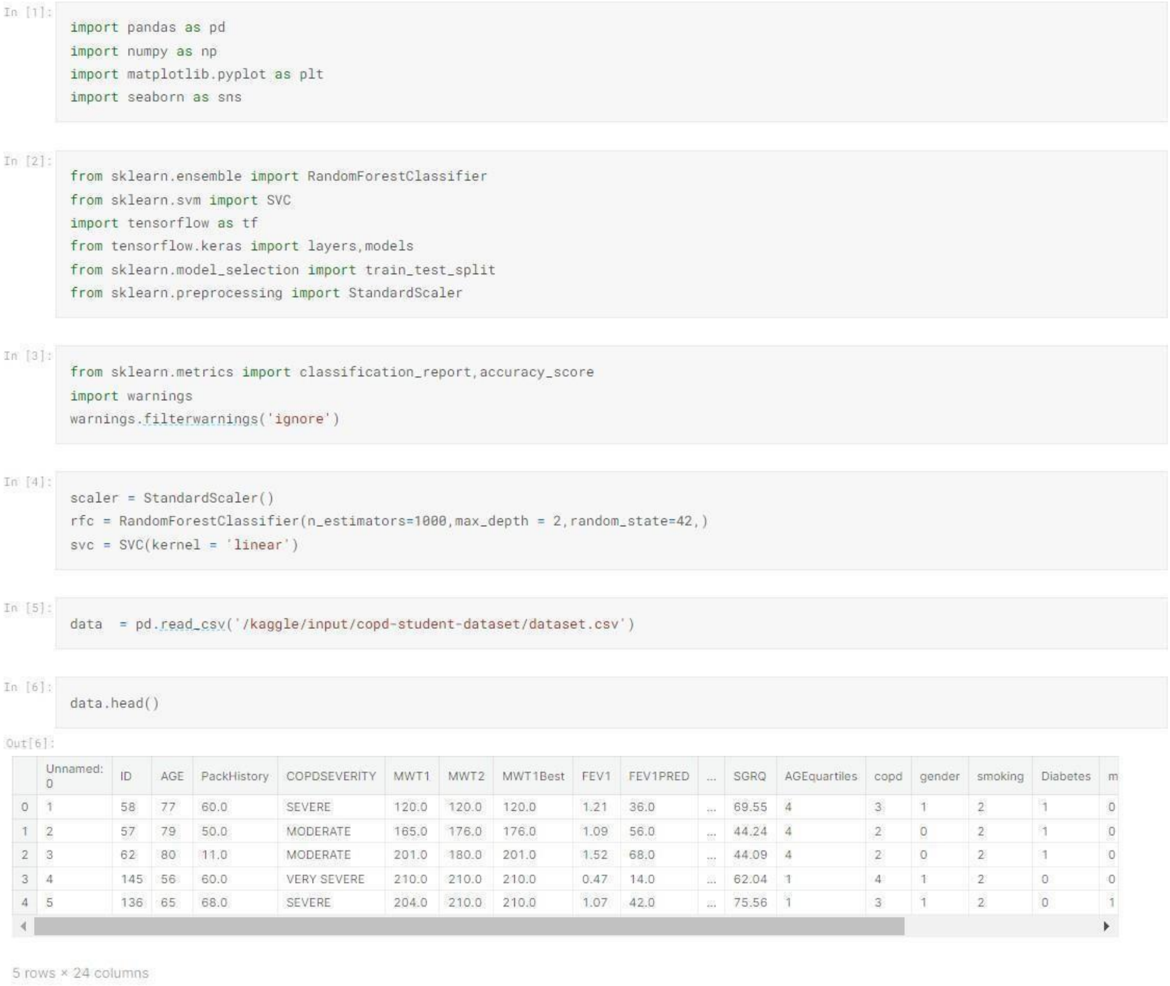
Challenges:

Realism vs. Simplicity Balance: Striking a balance between creating realistic virtual environments and maintaining simplicity for testing purposes.

Resource Requirements: Building and maintaining virtual testing environments require adequate resources and expertise in virtualization technologies.

APPENDICES

Sample Code :



```
In [1]:
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
In [2]:
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
import tensorflow as tf
from tensorflow.keras import layers,models
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
```

```
In [3]:
from sklearn.metrics import classification_report,accuracy_score
import warnings
warnings.filterwarnings('ignore')
```

```
In [4]:
scaler = StandardScaler()
rfc = RandomForestClassifier(n_estimators=1000,max_depth = 2,random_state=42,)
svc = SVC(kernel = 'linear')
```

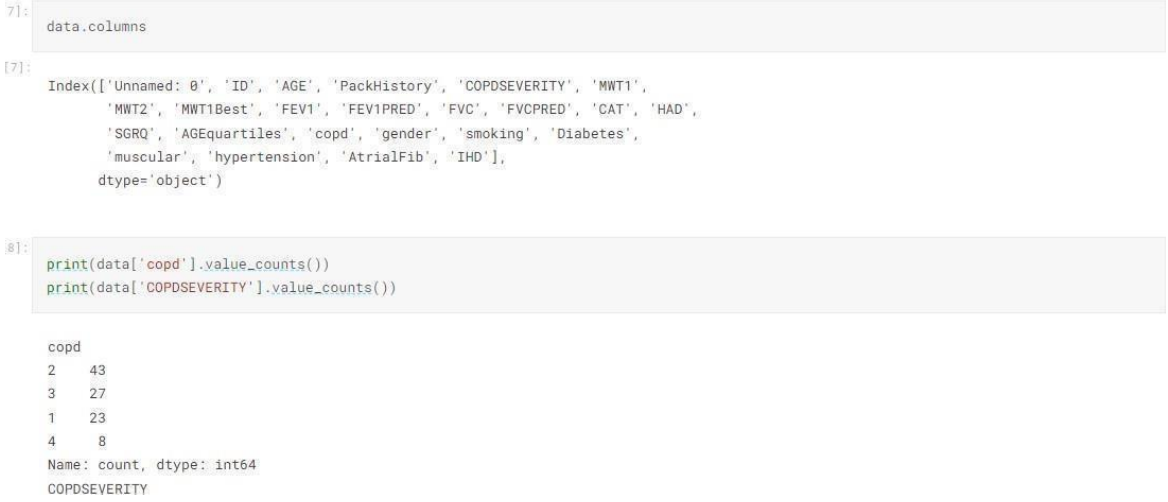
```
In [5]:
data  = pd.read_csv('/kaggle/input/copd-student-dataset/dataset.csv')
```

```
In [6]:
data.head()
```

Out[6]:

	Unnamed: 0	ID	AGE	PackHistory	COPDSEVERITY	MWT1	MWT2	MWT1Best	FEV1	FEV1PRED	...	SGRQ	AGEquartiles	copd	gender	smoking	Diabetes	m
0	1	58	77	60.0	SEVERE	120.0	120.0	120.0	1.21	36.0	...	69.55	4	3	1	2	1	0
1	2	57	79	50.0	MODERATE	165.0	176.0	176.0	1.09	56.0	...	44.24	4	2	0	2	1	0
2	3	62	80	11.0	MODERATE	201.0	180.0	201.0	1.52	68.0	...	44.09	4	2	0	2	1	0
3	4	145	56	60.0	VERY SEVERE	210.0	210.0	210.0	0.47	14.0	...	62.04	1	4	1	2	0	0
4	5	136	65	68.0	SEVERE	204.0	210.0	210.0	1.07	42.0	...	75.56	1	3	1	2	0	1

5 rows × 24 columns

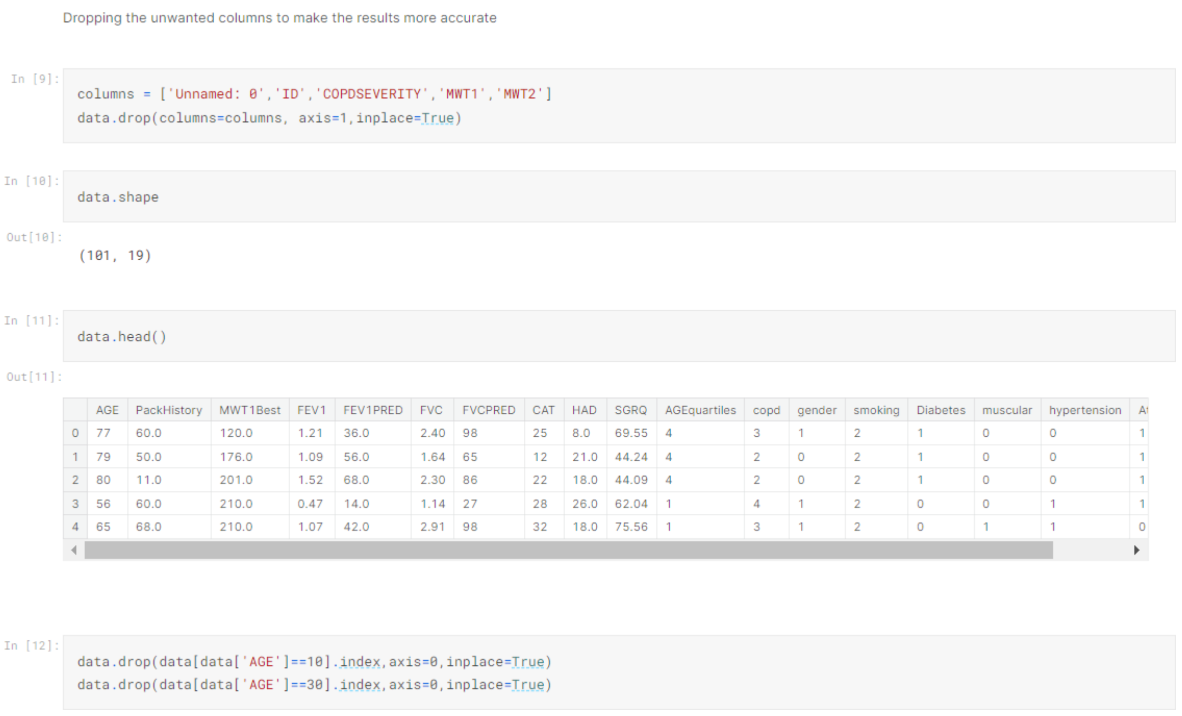


```
7]:
data.columns
```

```
[7]:
Index(['Unnamed: 0', 'ID', 'AGE', 'PackHistory', 'COPDSEVERITY', 'MWT1',
       'MWT2', 'MWT1Best', 'FEV1', 'FEV1PRED', 'FVC', 'FVCPRED', 'CAT', 'HAD',
       'SGRQ', 'AGEquartiles', 'copd', 'gender', 'smoking', 'Diabetes',
       'muscular', 'hypertension', 'AtrialFib', 'IHD'],
      dtype='object')
```

```
8]:
print(data['copd'].value_counts())
print(data['COPDSEVERITY'].value_counts())
```

```
copd
2    43
3    27
1    23
4     8
Name: count, dtype: int64
COPDSEVERITY
```



Dropping the unwanted columns to make the results more accurate

In [9]:
```
columns = ['Unnamed: 0','ID','COPDSEVERITY','MWT1','MWT2']
data.drop(columns=columns, axis=1,inplace=True)
```

In [10]:
```
data.shape
```

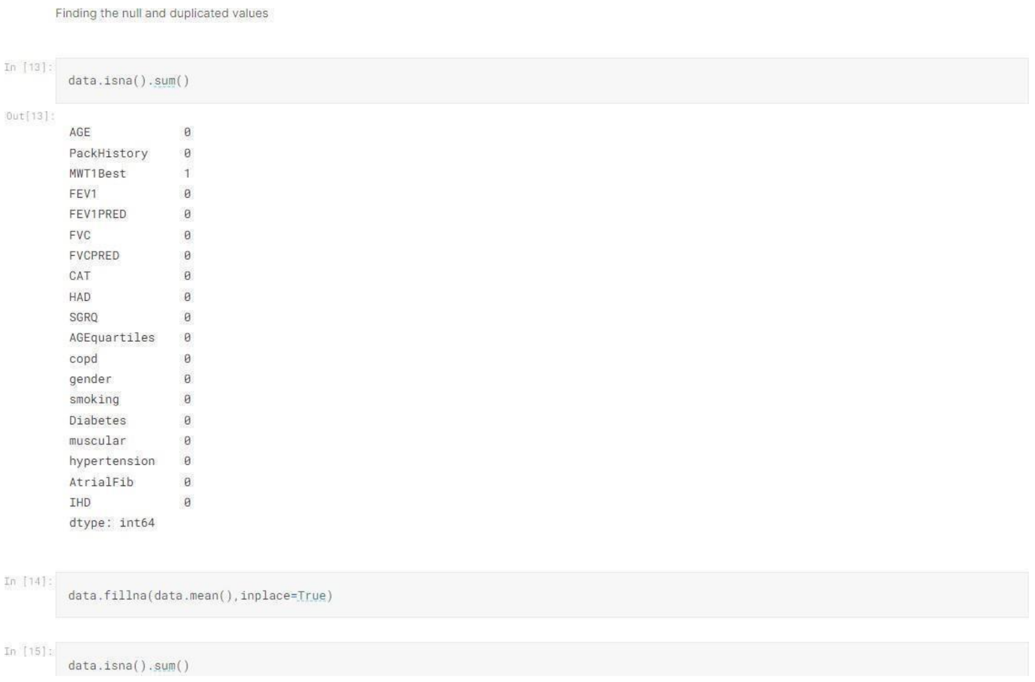
Out[10]:
```
(101, 19)
```

In [11]:
```
data.head()
```

Out[11]:

	AGE	PackHistory	MWT1Best	FEV1	FEV1PRED	FVC	FVCPRED	CAT	HAD	SGRQ	AGEquartiles	copd	gender	smoking	Diabetes	muscular	hypertension	At
0	77	60.0	120.0	1.21	36.0	2.40	98	25	8.0	69.55	4	3	1	2	1	0	0	1
1	79	50.0	176.0	1.09	56.0	1.64	65	12	21.0	44.24	4	2	0	2	1	0	0	1
2	80	11.0	201.0	1.52	68.0	2.30	86	22	18.0	44.09	4	2	0	2	1	0	0	1
3	56	60.0	210.0	0.47	14.0	1.14	27	28	26.0	62.04	1	4	1	2	0	0	1	1
4	65	68.0	210.0	1.07	42.0	2.91	98	32	18.0	75.56	1	3	1	2	0	1	1	0

In [12]:
```
data.drop(data[data['AGE']==10].index,axis=0,inplace=True)
data.drop(data[data['AGE']==30].index,axis=0,inplace=True)
```



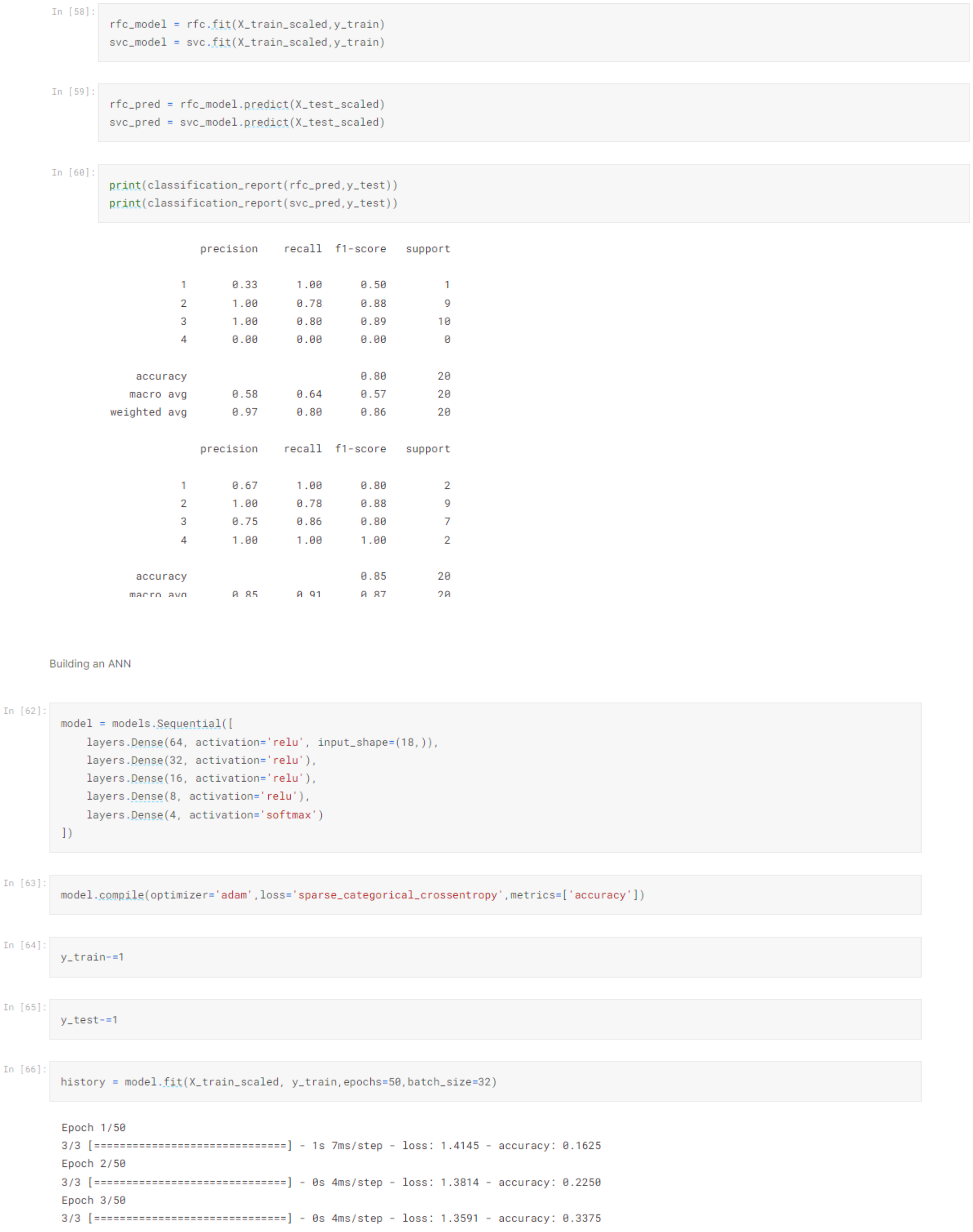
Finding the null and duplicated values

In [13]:
```
data.isna().sum()
```

Out[13]:
```
AGE             0
PackHistory     0
MWT1Best        1
FEV1            0
FEV1PRED        0
FVC             0
FVCPRED         0
CAT             0
HAD             0
SGRQ            0
AGEquartiles    0
copd            0
gender          0
smoking         0
Diabetes        0
muscular        0
hypertension    0
AtrialFib       0
IHD             0
dtype: int64
```

In [14]:
```
data.fillna(data.mean(),inplace=True)
```

In [15]:
```
data.isna().sum()
```



```
In [58]:
rfc_model = rfc.fit(X_train_scaled,y_train)
svc_model = svc.fit(X_train_scaled,y_train)
```

```
In [59]:
rfc_pred = rfc_model.predict(X_test_scaled)
svc_pred = svc_model.predict(X_test_scaled)
```

```
In [60]:
print(classification_report(rfc_pred,y_test))
print(classification_report(svc_pred,y_test))
```

```
              precision    recall  f1-score   support

           1       0.33      1.00      0.50         1
           2       1.00      0.78      0.88         9
           3       1.00      0.80      0.89        10
           4       0.00      0.00      0.00         0

    accuracy                           0.80        20
   macro avg       0.58      0.64      0.57        20
weighted avg       0.97      0.80      0.86        20

              precision    recall  f1-score   support

           1       0.67      1.00      0.80         2
           2       1.00      0.78      0.88         9
           3       0.75      0.86      0.80         7
           4       1.00      1.00      1.00         2

    accuracy                           0.85        20
   macro avg       0.85      0.91      0.87        20
```

Building an ANN

```
In [62]:
model = models.Sequential([
    layers.Dense(64, activation='relu', input_shape=(18,)),
    layers.Dense(32, activation='relu'),
    layers.Dense(16, activation='relu'),
    layers.Dense(8, activation='relu'),
    layers.Dense(4, activation='softmax')
])
```

```
In [63]:
model.compile(optimizer='adam',loss='sparse_categorical_crossentropy',metrics=['accuracy'])
```

```
In [64]:
y_train-=1
```

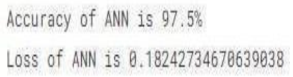
```
In [65]:
y_test-=1
```

```
In [66]:
history = model.fit(X_train_scaled, y_train,epochs=50,batch_size=32)
```

```
Epoch 1/50
3/3 [==============================] - 1s 7ms/step - loss: 1.4145 - accuracy: 0.1625
Epoch 2/50
3/3 [==============================] - 0s 4ms/step - loss: 1.3814 - accuracy: 0.2250
Epoch 3/50
3/3 [==============================] - 0s 4ms/step - loss: 1.3591 - accuracy: 0.3375
```

Fig 8.1 Screen Shots

Sample output :



```
Accuracy of ANN is 97.5%
Loss of ANN is 0.18242734670639038
```

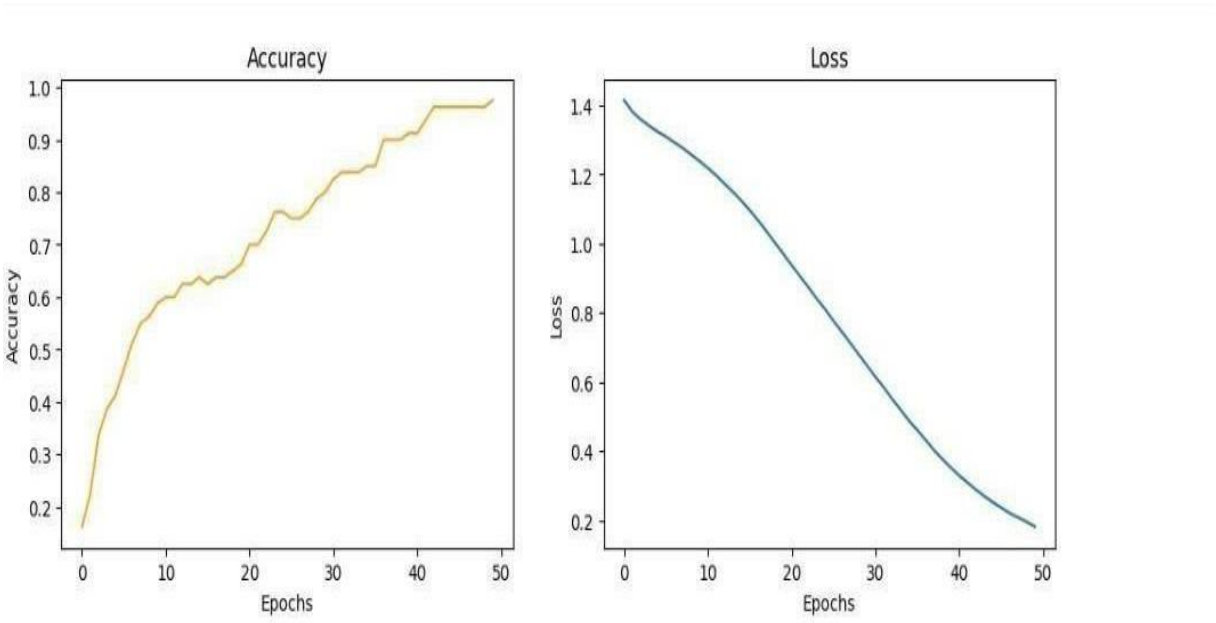


Fig 8.2 Screen Shots

Reference and Links:

1. Cooke et al. (2011) Algorithms to identify COPD in health systems with and without access to ICD coding: a systematic review. https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-019- 4574-3

2. Chen et al. (2021)Identification of COPD patients' health status using an intelligent system in the CHRONIOUS platform. https://pubmed.ncbi.nlm.nih.gov/28128970/

3. Murphy. (2016)A protocol for a cluster randomized trial of care delivery models to improve the quality of smoking cessation and shared decision making for lung cancer screening. https://pubmed.ncbi.nlm.nih.gov/36878389/

4. Rodriguez-Roisin. (2015)Identification and assessment of COPD exacerbations https://www.sciencedirect.com/science/article/pii/S2173511517301653

5. Nannestad. (2020)Exacerbations in Chronic Obstructive Pulmonary Disease: Identification and Prediction Using a Digital Health System https://pubmed.ncbi.nlm.nih.gov/28270380/