

Integrated Machine Learning Framework for the Determination of Respiratory Diseases by Utilizing the Metagenomics Approaches



By

Sehar Hifsa

(Fall-2022-MSBI-00000400261)

Department of Sciences (MS Bioinformatics)

School of Interdisciplinary Engineering and Sciences (SINES)

National University of Sciences and Technology (NUST)

Islamabad, Pakistan

(2024)

Integrated Machine Learning Framework for the Determination of Respiratory Diseases by Utilizing the Metagenomics Approaches



By

Sehar Hifsa

(Fall-2022-MSBI-00000400261)

A thesis submitted to the National University of Sciences and Technology, Islamabad,

in partial fulfillment of the requirements for the degree of

Master of Science in

Bioinformatics

Supervisor: Dr.Rehan Zafar Paracha

School of Interdisciplinary Engineering and Sciences (SINES)

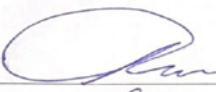
National University of Sciences and Technology (NUST)


Islamabad, Pakistan

(2024)

THESIS ACCEPTANCE CERTIFICATE


Certified that final copy of MS/MPhil thesis written by Ms. Sehar Hifsa Registration No. 00000400261 of SINES has been vetted by undersigned, found complete in all aspects as per NUST Statutes/Regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature with stamp:  Associate Professor
SINES - NUST, Sector H-12
Islamabad
Name of Supervisor: Dr. Lehan Saif
Date: 19/08/24

Signature of HoD with stamp:  Dr. Mir Hiyas Ahmad
HoD Engineering
Professor
SINES - NUST, Sector H-12
Islamabad
Date: 20/08/2024

12

Countersign by

Signature (Dean/Principal):  Principal SINES
Date: 20/08/2024

AUTHOR'S DECLARATION

I...Sehar Hifsa....hereby state that my MS thesis titled “...Integrated Machine Learning Framework for the Determination of Respiratory Diseases by Utilizing the Metagenomics Approaches...” is my own work and has not been submitted previously by me for taking any degree from National University of Sciences and Technology, Islamabad or anywhere else in the country/ world.

At any time if my statement is found to be incorrect even after I graduate, the university has the right to withdraw my MS degree.

Name:—Sehar Hifsa—————

Date: —August,2024—————

I would like to dedicate this thesis to my loving Father Raziq Mehmood, my mother Munaza Bashir, my sister Hazeefa Raziq and my Brother Muhammad Kaleem for their support and love.

ACKNOWLEDGEMENTS

All praise for Almighty ALLAH Who is the ultimate source of all knowledge. Almighty Allah has made me reach this present pedestal of knowledge with quality of doing something novel, stimulating and path bearing. All thanks to Allah Who gave me wisdom, strength and hope to accomplish my dreams and made me successful. All respects are for Holy Prophet Hazrat Muhammad (PBUH) who is the symbol of guidance and fountain of knowledge.

I am deeply thankful to my parents, whose constant love and support enabled me to reach this point. They were a continuous source of guidance during my educational course. Their diligence and struggles motivated me to pursue my goals during this difficult time.

I would like to express my sincere gratitude to my supervisor, Dr. Rehan Zafar Paracha, for all of his assistance and moral support during this research, including his recommendations and outstanding leadership. His deep knowledge, skills, proficiency, and valuable consultations are the basis of this successful research. His humble behavior and providing a healthy environment for work have enabled me to overcome the problems more effectively. I am thankful to my GEC members Dr. Zamir Hussain and Dr. Uzma Habib for their guidelines and assistance throughout this duration. I am grateful to the Principal, HOD, and all faculty members for providing an academic environment and assistance at the SINES.

I would like to extend my heartfelt thanks to my dear brother and sister, their unwavering support has motivated me to tackle obstacles and seek success. Their patience and prayers helped me a lot during my academic phase. Finally, I acknowledge my dear Uncle Muhammad Mehmood, Muhammad Zahoor, my dear cousin and my friend Haleema Parveen for their immense care and help throughout the research phase.

Contents

LIST OF TABLES	VI
LIST OF FIGURES	VIII
LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS	IX
ABSTRACT	XII
1 INTRODUCTION	1
1.1 Respiratory Microbiome	1
1.1.1 Lung Microbiome	1
1.1.2 Composition of Lung Microbiome	2
1.1.3 Association of Lung Microbiome	2
1.1.3.1 Microbiome of Upper Respiratory Tract	2
1.1.3.2 Microbiome of Nasopharynx	3
1.1.4 Factors Affecting Lung Microbiome	3
1.1.4.1 Genetic Factors	3
1.1.4.2 Environmental Factors	3
1.1.4.3 Antibiotic Usage	4
1.2 Respiratory Diseases	4
1.2.1 Chronic Obstructive Pulmonary Disease	4
1.2.1.1 Stages of COPD	4
1.2.1.2 Symptoms of COPD	5
1.2.1.3 Prevalence of COPD	5
1.2.1.4 Lung Microbes in COPD	6
1.2.2 COVID-19	6
1.2.2.1 Symptoms of COVID-19	6
1.2.2.2 Prevalence of COVID-19	7

1.2.2.3	Lung Microbes in COVID-19	7
1.2.3	Lung Cancer	7
1.2.3.1	Symptoms of Lung Cancer	7
1.2.3.2	Prevalence of Lung Cancer	8
1.2.3.3	Lung Microbiome in Lung Cancer	8
1.2.4	Tuberculosis	8
1.2.4.1	Symptoms of TB	8
1.2.4.2	Prevalence	9
1.2.4.3	Lung Microbiome in TB	9
1.3	Metagenome	9
1.3.1	Types of Metagenome	9
1.3.1.1	Amplicon Sequencing	10
1.3.1.2	Shotgun Sequencing	10
1.4	Machine Learning	10
1.4.1	Different ML Models	11
1.4.1.1	Random Forest	11
1.4.1.2	Support Vector Machine	11
2	LITERATURE REVIEW	12
2.1	Chronic Obstructive Pulmonary Disease	12
2.1.1	Metagenomics on COPD	13
2.2	COVID-19	14
2.2.1	Metagenomics on COVID-19	15
2.3	Lung Cancer	16
2.3.1	Metagenomics on Lung Cancer	17
2.4	Tuberculosis	18
2.4.1	Metagenomics on TB	19
2.5	ML in Metagenomics	20
2.6	Study Rationale	21
2.7	Objectives	21
3	MATERIALS AND METHODS	22
3.1	Dataset Collection	23
3.2	Metagenomics	23
3.2.1	Amplicon Sequencing	23
3.2.2	In-Silico Analysis of Amplicons	24

3.2.2.1	Quantitative Insights Into Microbial Ecology 2	25
3.2.2.2	Demultiplexing	25
3.2.2.3	Data Input	25
3.2.2.4	Denoising	26
3.2.2.5	Amplicon Sequence Variants and Operational Taxonomic Units	26
3.2.2.6	Taxonomical Classification	26
3.2.2.7	Visualization	27
3.2.3	Shotgun Sequencing	27
3.2.4	In-Silico Shotgun Analysis	27
3.2.4.1	Galaxy	29
3.2.4.2	Importing Datasets	29
3.2.4.3	Quality Control	29
3.2.4.4	Taxonomical Assignment	30
3.2.4.5	MetaphlAn	30
3.3	Machine Learning	31
3.3.1	Input Data	32
3.3.2	Splitting of Data	33
3.3.3	Feature Selection	33
3.3.4	Model Building	33
3.3.5	Model Training	33
3.3.6	Model Evaluation	34
4	RESULTS	35
4.1	Amplicon Analysis Results	35
4.1.1	COPD Amplicon Dataset	35
4.1.2	COVID-19 Amplicon Results	40
4.2	Shotgun Analysis	45
4.2.1	In-Silico Shotgun Analysis of COPD	46
4.2.2	In-Silico Shotgun Analysis of COVID-19	48
4.2.3	In-Silico Shotgun Analysis of Lung Cancer	49
4.2.4	In-Silico Shotgun Analysis of TB	49
4.3	Machine Learning	50
4.3.1	Random Forest	50
4.3.1.1	Chronic Obstructive Pulmonary Disease Dataset	51
4.3.1.2	COVID-19 Dataset	52

4.3.1.3	Lung Cancer Dataset	53
4.3.1.4	Tuberculosis Dataset	54
4.3.1.5	All Datasets	55
4.3.2	Support Vector Machine	56
4.3.2.1	Chronic Obstructive Pulmonary Disease Dataset . . .	56
4.3.2.2	COVID-19 Dataset	57
4.3.2.3	Lung Cancer Dataset	58
4.3.2.4	Tuberculosis Dataset	59
4.3.2.5	All Datasets	60
5	DISCUSSION	62
6	CONCLUSIONS AND FUTURE RECOMMENDATIONS	65

List of Tables

3.1	Datasets used for the Study	23
4.1	Summary of Demultiplexed Sequences of COPD Dataset	36
4.2	Statistical Aspects after denoising of COPD Dataset	38
4.3	Summary of COPD Dataset Features	39
4.4	Taxonomy Classification on the basis of Feature Id's	39
4.5	Summary of Demultiplexed Sequences of COVID-19 Dataset	41
4.6	Statistical Aspects after denoising of COVID-19 Dataset	43
4.7	Summary of COPD Dataset Features	43
4.8	Taxonomy Classification on the basis of Feature Id's	44
4.9	Datasets Used for ML	50
4.10	Summary of RF Model Scores	56
4.11	Summary of SVM Model Scores	61

List of Figures

3.1	General Workflow for Methodology	22
3.2	In-Silico Analysis steps used for Amplicon Analysis	24
3.3	In-Silico Analysis steps used for Metagenome Analysis	28
3.4	Steps used for ML	32
4.1	Forward Reads Frequency Diagram of COPD	36
4.2	Reverse Reads Frequency Diagram of COPD	37
4.3	Quality Score Graph for Forward Reads of COPD	37
4.4	Quality Score Graph for Reverse Reads of COPD	38
4.5	Bar Graph for Taxonomy Classification of COPD	40
4.6	Forward Reads Frequency Diagram for COVID-19	41
4.7	Reverse Reads Frequency Diagram for COVID-19	42
4.8	Quality Score Graph for Forward Reads of COVID-19	42
4.9	Quality Score Graph for Reverse Reads of COVID-19	43
4.10	Bar Graph for Taxonomy Classification of COVID-19	45
4.11	Graph representing the Phred Score per base for COPD Before Fastp .	47
4.12	Graph representing the Adaptor Content in COPD Dataset Before Fastp	47
4.13	Graph representing the Phred Score per base for COPD After Fastp . .	48
4.14	Graph representing the Adaptor Content in COPD Dataset After Fastp	48
4.15	Feature Selected Using Lasso for COPD	51
4.16	Confusion Matrix for COPD through RF	52
4.17	Feature Selected Using Lasso for COVID-19	52
4.18	Confusion Matrix for COVID-19 through RF	53
4.19	Feature Selected Using Lasso for Lung Cancer	53
4.20	Confusion Matrix for Lung Cancer through RF	54
4.21	Feature Selected Using Lasso for TB	54
4.22	Confusion Matrix for TB through RF	55
4.23	Feature Selected Using Lasso for All Datasets	55

4.24	Confusion Matrix for All Datasets through RF	56
4.25	Feature Selected Using Lasso for COPD	57
4.26	Confusion Matrix for COPD through SVM	57
4.27	Feature Selected Using Lasso for COVID-19	58
4.28	Confusion Matrix for COVID-19 through SVM	58
4.29	Feature Selected Using Lasso for Lung Cancer	59
4.30	Confusion Matrix for Lung Cancer through RF	59
4.31	Feature Selected Using Lasso for TB	60
4.32	Confusion Matrix for TB through SVM	60
4.33	Feature Selected Using Lasso for All Datasets	61
4.34	Confusion Matrix for All Datasets through SVM	61

Nomenclature

16S rDNA	16S Ribosomal Deoxyribonucleic Acid
18S rRNA	18S Ribosomal Ribonucleic Acid
ACE2	Angiotensin-Converting Enzyme 2
ANNs	Artificial Neural Networks
ASVs	Amplicon Sequence Variant
BAL	Bronchoalveolar Lavage
BALF	Bronchoalveolar Lavage Fluid
CMV	Cytomegalovirus
COPD	Chronic Obstructive Pulmonary Disease
COVID-19	Corona Virus Disease
DADA2	Divisive Amplicon Denoising Algorithm 2
EBV	Epstein-Barr virus
ENA	European Nucleotide Archive
ERK	Signal-Regulated Kinase
GOLD	Global Initiative for Chronic Obstructive Lung Disease
HIV	Human Immunodeficiency Virus
HPV	Human Papillomavirus
HPVB19	Human Parvovirus B19
HRE	High-Risk Exacerbators

HSV	Herpes Simplex Virus
lncRNAs	Long Noncoding RNAs
LRT	Lower Respiratory Tract
LUSC	Lung Squamous Cell Carcinoma
MAGs	Metagenome Assembled Genomes
MAPK	Mitogen-Activated Protein Kinase
ML	Machine Learning
MUC5AC	Mucin 5AC
MUC5B	Mucin 5B
NETs	Neutrophil Extracellular Traps
NGS	Next Generation Sequencing
NIs	Nosocomial Infections
NSCLC	Non-Small Cell Lung Cancer
OTU	Operational Taxonomic Unit
PHEIC	Public Health Emergency of International Concern
PI3K	Phosphatidylinositol 3-Kinase
PM	Particulate matter
PRRs	Primary Recognition Receptors
QC	Quality Control
QIIME2	Quantitative Insights Into Microbial Ecology
RF	Random Forest
SARS-CoV	Severe Acute Respiratory Syndrome
SCLC	Small-Cell Lung Cancer
SRA	Sequence Read Archive
SVM	Support Vector Machine

TB	Tuberculosis
UBT	Upper Bronchial Tract
URT	Upper Respiratory Tract
V3	Variable Region 3 of 16S rRNA gene
V4	Variable Region 4 of 16S rRNA gene
WHO	World Health Organization

ABSTRACT

A respiratory disease refers to a condition that affects the organs involved in respiration. The lungs are the central part of the respiratory system. Advancements in sequencing technology have revealed that respiratory tract primarily lungs consist of unique and diverse microbes. The alteration in the balance of these microbes have been found in different respiratory diseases e.g. Chronic Obstructive Pulmonary Disease (COPD), asthma, Tuberculosis (TB), lung cancer, pneumonia and SARS-CoV. Globally, pulmonary infections continue to be a leading source of mortality with COPD about 3 million deaths annually, asthma affecting about 334 million people, lung cancer killing 1.6 million people and TB infecting 10 million cases worldwide acquiring great concern to understand the composition and role of microbial communities in respiratory diseases. This study utilized the two secondary 16S amplicon datasets to analyze and identify the bacterial compositions mainly in COPD and SARS-CoV. The outcome has given the taxonomic classification and relative abundance of micro-organisms in lungs during different disease states. Moreover, the four shotgun datasets are also analyzed to provide detail insight into microbial content, identifying different bacterial species and assessing the observed differences during diseased and healthy states in COPD and SARS-CoV, lung cancer and TB disorders. The key bacterial phylum determined by the analysis are Bacteroidetes, Firmicutes and Proteobacteria with different families and crucial bacterial species with their relative abundances. The discriminating phylum in lung cancer are Actinobacteria and Fusobacteria are distinct phylum in lung cancer and crucial species are *Halomonas-sp-LBP4*, *Campylobacter-jejuni* and *Haemophilus-influenzae* while *Candidatus-Saccharibacteria* is discriminated with significant species are *Neisseria-subflava* and *Prevotella-melaninogenica* in TB sequences after performing metagenome analysis. In COPD the important species are *Haemophilus-influenzae* and *Staphylococcus-aureus* and COVID-19 has *Staphylococcus-epidermidis*, *Malassezia-restricta* and *Corynebacterium-propinquum*. Amplicon analysis shows that *Fusobacteria* is also present and *Staphylococcaceae*, *Pseudomonaceae* and *Flavobacteriaceae* are identified in both amplicons and shotgun analysis. Additionally, using the taxonomic classification tables of bacterial species with their relative abundances in specific disorders two sophisticated machine learning models are generated to classify

the bacteria into diseased and control also providing information about the relative abundances present in the feature data. These models are trained on each disease dataset and then one model with all combined datasets. All models gives the good accuracies and potential to categorize the microbial species precisely. The COVID-19 datasets has high accuracies among all datasets with 94% in RF and 88% in SVM. These procedures give valuable insight into the understanding of respiratory microbe's composition and patterns in infected and control states. These findings lead to the basis of further comprehensive and valuable studies focusing on composition and functional profiling of respiratory microbiota and investigating a better way to cure these painful disorders.

Keywords: Respiratory microbiota, COPD, COVID-19, lung cancer, TB, amplicon, shotgun, taxonomical classification, relative abundance, ML.

Chapter 1

INTRODUCTION

The respiratory system is the basis of life. Human respiratory system consists of different organs, muscles and various bones which are helpful in breathing process. Respiration involves two main mechanisms, inhalation (air goes in) and exhalation (air moves out) from the lungs. The respiratory tract is further divided into the upper respiratory tract (nose, pharynx, larynx) and lower respiratory tract (trachea, bronchi, lungs). Humans and microbes are correlated. These microbes include bacteria, viruses, fungi, archaea, etc. The human body also has a variety of microbes at different body sites e.g. gut, skin, respiratory and urogenital tract etc. There are various critical processes maintained by these micro-organisms including digestion, prevention of diseases and maintaining immune balance [1].

1.1 Respiratory Microbiome

Respiratory tract contains the complex communities of bacteria known as 'respiratory microbiota'. High throughput sequencing techniques were used for the verification of micro-organisms in the respiratory tract within the lower respiratory tract. Most of the research is on the dysbiosis of the microflora of the lower respiratory tract in chronic respiratory disorders. Alteration of airway microbes occurs in persistent airway disorders including severe asthma, bronchiectasis, and COPD [2].

1.1.1 Lung Microbiome

The microbial community that resides particularly in the lungs is collectively considered as lung microbiota. The lower respiratory tract is comprised of the transfer of micro-organisms from the upper respiratory tract through processes of inhalation, mucosal dispersion, and microaspiration [3]. After entering into the lungs there are certain mechanisms involved that are responsible for the removal of these micro-organisms

from the lungs to maintain a balanced and healthy environment. These mechanisms include host defense system or immune system responses, mucociliary clearance, and cough which are helpful for the elimination of microbial content from the respiratory tract and lungs. In this way, the lungs perform their central role of respiration efficiently [4].

1.1.2 Composition of Lung Microbiome

The composition of the microbiome of the lungs is very diverse as it includes different viral, fungal, and bacterial colonies. The mycobiota (fungal species) include *Penicillium*, *Candida*, and *Saccharomyces* belonging to the division of Ascomycetes. The viral community of the lungs includes phages. *Haemophilus*, *Streptococcus*, and *Prevotella* are among the bacterial species present in the lungs. All these micro-organisms play a crucial role in maintaining a healthy state in the lungs and help to perform the normal functions of the lungs [5].

1.1.3 Association of Lung Microbiome

The microbiome of lung is affected by several environmental and microflora present in the human body. The microflora of Upper respiratory tract also has a significant effect on the lung micro-organisms. The effect of upper pulmonary tract and micro-organisms of nasopharynx is discussed below.

1.1.3.1 Microbiome of Upper Respiratory Tract

The Upper Respiratory Tract (URT) consists of different structures including the nasal cavity, nasopharynx, larynx and oropharynx. All these structures have specific microbial ecosystems and alteration leads to various contagious infectious disorders. The URT has a dense amount of microbes as compared to the lower respiratory tract. URT is typically composed of *Corynebacterium*, *Streptococcus*, *Moraxella*, *Dolosigranulum* and *Bifidobacterium*. These genera maintain the control environment in the nasopharynx and oropharynx. Moreover, the URT is persistently in association with the surrounding biosphere it is home to various micro-organisms. Therefore a dynamic microbial ecosystem is exhibited by the upper airway. In the same way, various factors are involved in the composition of the microbial environment, disruption leads to different respiratory disorders [6]. The respiratory system is also engaged in wide-ranging functions such as humidification, respiratory airway passage, warming, immunological defense, olfactory sensation, and filtering of inhaled air.

1.1.3.2 Microbiome of Nasopharynx

The nasopharynx plays a vital role in respiration and using the oropharynx helps in the passage of air from the nasal cavity to the windpipe [7]. The different bacterial genera exhibit the nasopharynx such as non-typeable *Haemophilus influenzae*, *Streptococcus pneumoniae*, and *Moraxella catarrhalis* [8]. There are other dominant bacterial genera such as *Corynebacterium* and *Dolosigranulum* which are also helpful in the functioning of the nasal cavity [9]. Therefore alteration in bacterial community results in severe diseases and affects human health badly. The common diseases of the nasopharynx are nasopharyngeal carcinoma, tonsillitis, and nasopharyngitis.

The microbiome of the upper respiratory system plays a significant role in the development of lung microbiome. The microorganisms present in the nasal cavity, nasopharynx, and oropharynx are eventually transferred to the lungs, forming the foundation of lung flora. Therefore, studying these microbes is essential to gain a better understanding of lung microbiota [10]. It is essential to comprehend the microbial compositions present in the upper respiratory tract, as they have a significant influence on the lung's primary microbial community where proper gaseous exchange takes place. Understanding these microbial compositions can help us gain valuable insights into the lung functioning and its impact on overall health.

1.1.4 Factors Affecting Lung Microbiome

Numerous factors can alter the microbial composition of the lungs, including genetics, environmental aspects, and the use of antibiotics.

1.1.4.1 Genetic Factors

The respiratory tract features cilia-like structures that aid in the removal of pathogenic microorganisms through a process called mucociliary clearance. Additionally, a protective mucus layer lines the airway passage, with MUC5AC and MUC5B being the mucin-encoding genes that contribute to the elimination of harmful bacteria from the lungs [11].

1.1.4.2 Environmental Factors

Humans are often exposed to various habitats, and the micro-organisms present in these habitats can significantly impact the composition of their lung microbiome. Previous studies have demonstrated that exposure to dust can increase the risk of developing asthma, and similarly, exposure to cigarette smoke can also alter the communities of micro-organisms in the lungs. Dietary fibers impact the lung microbiota, particularly the proportion of bacterial phyla, such as Bacteroidetes and Firmicutes [11].

1.1.4.3 Antibiotic Usage

Antibiotic use in the antenatal period affects the intestinal microbiota and also disturbs the lung microbiome. Some studies showed that there are perturbations in the lung microbiome after antibiotic use. Vancomycin plus neomycin was used in a study that enhanced T-cell and natural killer cell activation leading to the reduction of melanoma B16 lung metastasis in their mouse lungs. Therefore, medication alters the microbial content in the respiratory system [11].

Consequently, all these factors alter the microflora of the pulmonary system and these alterations result in chronic and acute disorders. There are various breathing diseases e.g. pulmonary fibrosis, asthma, lung cancer, chronic pulmonary respiratory disease (COPD), COVID-19, pneumonia cystic fibrosis and TB etc.

1.2 Respiratory Diseases

There are several respiratory diseases which are responsible for millions of deaths across the world. Most common respiratory diseases are COPD (COPD), Asthma, TB , Lung cancer, Pneumonia, Cystic Fibrosis etc. Now non-culture techniques e.g. Next Generation Sequencing (NGS) contributes for analyzing the role of microbial content in these disorders. Some pulmonary diseases and the micro-organisms role in progression of diseases is discussed below.

1.2.1 Chronic Obstructive Pulmonary Disease

COPD is a medical condition marked by inflammation of the lungs, sputum production with cough and breathing difficulties [10]. Various factors result in COPD including inflammation in the lungs, causes of exacerbation rates and response to different therapies. Other characteristics such as pathogenic bacteria, viruses and environmental factors, can also increase the likelihood of exacerbation. Contiguous Diseases are among the primary causes of exacerbations, with viruses such as Coronavirus, Influenza virus as well as bacteria like *Pseudomonas aeruginosa*, and *Morxella* being frequently implicated factors. Thus microbiome contributes to the development, advancement and medication of COPD [12].

1.2.1.1 Stages of COPD

For a better understanding of the severity and classification of the disease, scientists introduced the system named Global Initiative for Obstructive Lung Disease (GOLD) introduced a system for the classification of different stages of COPD based on Forced Expiratory volume in 1s (FEV1). Mainly, the four stages are according to the GOLD standard used globally for COPD severity checks.

- Stage 1 (Mild)
- Stage 2 (Moderate)
- Stage 3 (severe)
- Stage 4 (very severe)

Stage one has few or no symptoms, while the moderate shows fewer symptoms with cough. After moderate the condition becomes severe. The last stage with very severe symptoms and extreme difficulty in respiration [13]

1.2.1.2 Symptoms of COPD

The symptoms of a disease can vary depending on the intensity and duration of the illness. Additionally, patients experience difficulty in doing their daily activities. Various symptoms influence the regular work and lifestyle of the patient. Cough appears as the most prevailing symptom among the others. The mild symptoms includes:

- Sputum production
- Cough
- Dyspnea

The most severe and uncommon includes:

- Respiratory discomfort
- Acute bronchitis
- Wheezing [14].

1.2.1.3 Prevalence of COPD

Prevalence, another important factor of any disease, indicates the extent to which people are affected by it. COPD is a highly prevalent disease worldwide, including in Pakistan. According to research, Two-hundred fifty-one million people are affected by COPD around the world, causing three million deaths. In middle or low-income countries COPD is responsible for a 90 % death rate. In Pakistan, 2.1 % of individuals are affected by COPD. The research estimated that the number of deaths resulting from COPD will reach 4.5 million worldwide by 2030 [15].

1.2.1.4 Lung Microbes in COPD

Microbes play a significant role in human health and disease. Beneficial bacteria help maintain a healthy environment and protect against harmful microbes. On the other hand, harmful microorganisms lead to infectious and non-infectious diseases. The lungs contain a significant amount of microbiota that aids in their normal functioning. When there is dysbiosis (an imbalance in the microbiota), it can lead to various disorders. In individuals with COPD, there is a decrease in the diversity of beneficial bacteria, and an increase in disease-causing bacterial species and phyla such as Brevundimonas, Neisseria, Moraxella and species from the Cyanobacteria phylum. Additionally, viruses such as Rhinovirus and Influenza virus can contribute to the progression of COPD. In cases of acute exacerbations of COPD, there is an increase in the amount of Klebsiella and Acinetobacter species [10].

1.2.2 COVID-19

COVID-19 is a pulmonary virus that mainly affects the lungs causing pneumonia. The virus badly affects the respiratory system but it also damages the kidneys, circulatory system, digestive and central nervous system leading to multi-organ system failure. The main receptor for COVID-19 in our body is Angiotensin-Converting Enzyme 2 (ACE2) which plays an important role in providing a favorable environment for the growth of the virus [16].

1.2.2.1 Symptoms of COVID-19

The research carried on COVID-19 symptoms shows that symptoms can appear after a little while of virus encounter generally between 2-14 days. The symptoms of COVID-19 may vary a little from person to person but the most common symptoms of COVID-19 includes the following:

- Cough
- Fever
- Fatigue
- Dyspnea
- Headache
- Abdominal pain
- Sputum production
- Loss of smell and taste [17].

1.2.2.2 Prevalence of COVID-19

After the emergence of the first case in China, the world encountered horrible COVID-19 virus waves one after the other. In the COVID-19 era, there are four to five strains of COVID-19 emerged and it badly affects human lives all over the world. The prevalence of the coronavirus is too much high as it has affected 40 million people across the world according to a study done in 2021. The dangerous waves also affect low-income countries and Pakistan and according to research, its prevalence rate in Pakistan is 0.019 %. Although it seems less, the figure resulted in 1,580,631 positive cases and 30,656 deaths in Pakistan [18].

1.2.2.3 Lung Microbes in COVID-19

Although there is little known about the role of lung microbiota in COVID-19. Some studies show that dysbiosis in lung micro-organisms has occurred in SARS-Cov-2. The pathogenic bacteria in COVID-19 have increased according to evidence from a few studies. In research on COVID-19 patients, biopsies and Broncho Alveolar Lavage Fluid (BALF) samples are used to focus on lung microbiome. The fungal genera consist of Cladosporium, Dipodascus, Aspergillus, and Candida with the dominant one being Cutaneotrichosporon. The bacterial genera involved are Brevundimonas, Enterobacteriaceae and Acinetobacter [19]. The abundance of Enterobacter cloacae, Klebsiella oxytoca, and Lactobacillus reuteri are also seen in SARS-Cov-2 patients [20].

1.2.3 Lung Cancer

Lung cancer is a cancer in the tissues of the lungs. Lung cancer is among the most prevalent cancers. Tobacco smoking is among the top factors that lead to severe lung cancer stages. The types of lung cancer are broadly categorized into Small-Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC). Non-Small Cell Lung Cancer is the most prevalent cancer among the others. Non-small cell Lung Cancer is a cancer that starts in the lining of lung cells. The most common types of NSCLC are adenocarcinoma and squamous cell carcinoma [21].

1.2.3.1 Symptoms of Lung Cancer

Although most cancers are diagnosed at later stages there are a few symptoms that patients of lung cancer encounter. The symptoms are listed below:

- Weakness
- Weight loss

- Hemoptysis
- Severe Chest Pain
- Cough with wheezing [22].

1.2.3.2 Prevalence of Lung Cancer

The world has been facing lethal cancer disease from a long ago. Lung cancer is among the top spreading cancers. Smoking is the top factor for respiratory disease specifically lung cancer. The death rate of lung cancer estimated was at 1.8 million and the diagnosed cases were 2.0 million worldwide. The 19 million cases of cancer are diagnosed in Pakistan showing the critical situation of disease prevalence [23].

1.2.3.3 Lung Microbiome in Lung Cancer

The imbalance of microbes is also related to lung cancer. The concentration of microorganisms decreased in lung cancer. The strains of *Pseudomonas* and *Escherichia coli* are found in cancer patients. The Cytomegalovirus (CMV) in lung cancer tissues, *Thermus* bacteria richness increased in the last stages of cancer. The patients with Lung Squamous Cell Carcinoma (LUSC) and adenocarcinoma majorly occupy *Actinomyces*, *Veillonel*, *Rothia*, *Capnocytophaga*, *Arthobacter* and *Proteobacteria* [24].

1.2.4 Tuberculosis

TB is also a lethal disease which is caused by a bacterium called *Mycobacterium TB*. TB affects the respiratory passage and lungs badly resulting in pulmonary fibrosis, bronchiectasis respiratory cavitation and restriction of breath. *Mycobacterium TB* interacts with different host processes and adapts according to them. Alterations in oxygen tension and, the emergence of granulomas also adapt their metabolism mechanism are among the modification processes [25].

1.2.4.1 Symptoms of TB

There are different recognizable symptoms shown by a TB patient. The dominant symptoms are:

- Fatigue
- Breathlessness
- Shivering and fever

- Discomfort in the chest
- Chronic cough with mucus

The other accessory symptoms might include hematological manifestations, low blood sugar level, more oxidative stress, inadequate Vitamin D, and modified host microbial community are correlated to TB [26].

1.2.4.2 Prevalence

According to estimates from the World Health Organization , 25 % of people in the world are Mtb-positive and consequently at risk of developing TB. There are a couple of contributors that give rise to mycobacterium infection including HIV infection, malnutrition, and smoking. African countries also had a high prevalence of TB specifically East African countries. The reported fatalities from TB in 2015 were 1.4 million and increased to 1.49 million in 2018 and 10 million new cases diagnosed [27].

1.2.4.3 Lung Microbiome in TB

The decrease in diversity of micro-organisms is shown by TB individuals. Some bacterial species of Coprococcus, Treponema, Prevotella and Leptotrichia. While there is also a rise in an abundance of pathogenic bacteria Pseudomonas, Streptococcus, Gramulicatella, Moryella and Mogibacterium. The relative abundance of Mycobacterium also increased [28].

1.3 Metagenome

A metagenome corresponds to the sequencing of all the microbes present in a particular environment [29]. The term "Metagenomics" was used first time by Jo Handelsman in 1998. She provides a way to identify and study the non-culturable microbes through the metagenomic technique. Currently, microbes of different environments are studied through metagenomics [30]. In terms of overall biological material and number of cells, microorganisms make up the majority of life forms on Earth and play crucial roles in worldwide component processes [31].

1.3.1 Types of Metagenome

The two most used techniques of high-throughput sequencing are amplicon and shotgun sequencing.

1.3.1.1 Amplicon Sequencing

Amplicon sequencing is the targeted approach in which only the marker genes are sequenced instead of the whole genome of organisms. Prokaryotes utilized 16S rDNA, while eukaryotes utilized 18S rDNA, primarily fungi. The pipeline in amplicon includes various tools to accomplish the microbial classification. The most common tool used for amplicon analysis is Quantitative Insights into Microbial Ecology Version 2. The ancient method used for classification is the Operational Taxonomic unit (OTU) while the recent and most used is Amplicon Sequence Variants (ASV). ASV represents the unique sequences and high accuracy and closely related sequences are clustered. After ASV formation they are used to identify the microbes present in the specimen. In short, amplicon sequencing can be used for bacteria or archaea identification from the samples used for analysis [32].

1.3.1.2 Shotgun Sequencing

Whole genome or Shotgun sequencing is used for the complete sequencing of the genome of microbes.. Shotgun gives detailed information on genomes to the species level of the organism. The functional and various pathways details are highlighted through the whole genome sequencing of microbes. The method also gives the knowledge about the other micro-organisms present in the environment from the focused one only. The pipeline of shotgun sequencing involves different tools at different steps. The FASTQC, Bowtie, Metaphlan, Kraken, Bracken and LefSe are the frequently used tools in whole genome sequencing. The method is of great importance in understanding the microbe's role in multiple human diseases encountering multiple body sites [32].

1.4 Machine Learning

Machine Learning (ML), trains machines to process information with greater efficiency. Understanding from evidence is the aim of ML. Numerous research investigations have been conducted on the subject of teaching computers to operate on their own. Several methods are used in ML to address data-related issues. The sort of model that works effectively for a specific problem, and the choice of algorithm used relies on the type of query you want to answer. The ML process to teach a function that translates a command to an outcome using sample combinations of inputs and outputs is known as supervised learning. The train and test datasets are separated from the input dataset. From training data, the ML model gets trained, and the model is evaluated through test data.

Unsupervised learning is the type in which there is no sample data learning. It is up to the algorithms to find and display the intriguing structure within the data.

Few features were identified from the data by the unsupervised learning algorithms. It recognizes the class of the data when it is introduced by using the previously learned features. Its primary applications are in decreasing attribute and clustering [33].

There are various ways in which we can use ML and deep learning in metagenomics to handle and manage microbiome data. We can use ML models for taxonomic classification of the microbiome, biomarker bacteria in any disease, functional annotation of micro-organisms, sequence or sample classification, disease identification, and for therapeutic efficacy [34].

1.4.1 Different ML Models

There are various ML models which we use for different purposes. Support Vector Machine, Random Forest, Naive Bayes and Logistic Regression models. Some of them are discussed in the below section.

1.4.1.1 Random Forest

Decision tree methods, one of the most widely used supervised learning techniques, are quite adaptable because they are non-parametric. Another recursive split technique that uses a group of distinct decision trees for prediction is called a Random Forest (RF). Since numerous models are integrated to create a single RF, as well as a combined approach. To accomplish this, the procedure makes copies of the original data that are bootstrapped, projecting one tree for every bootstrap. As numerous trees are averaged and combined to produce reliable and precise predictions, RF are less likely to over-fit than a single tree.

1.4.1.2 Support Vector Machine

Support Vector Machine are designed for classification tasks where multiple variables are involved and there are two classes (e.g., cases or controls) in a multi-layered model. SVM use a multidimensional, border that ideally divides classes, to distinguish between different classes (i.e., outcome categories). To locate an independent selection surface, SVM use a conversion of data that propagates the input into a space with greater dimensions represented as a kernel operator. SVM are hampered by being a "black box" technique, meaning that metrics for how predictors are integrated to optimize the hyperplane are not supplied, even though they can produce predictions that are highly precise when utilizing adaptive nonlinear kernels [35].

Chapter 2

LITERATURE REVIEW

This chapter is a review about the previous studies on pulmonary diseases and microorganisms role in the development of the disorder. Non-culture techniques e.g. metagenomics approaches includes primarily the 16S rRNA and metagenome sequencing.

2.1 Chronic Obstructive Pulmonary Disease

COPD is a chronic disorder and multifactor disease with a large number of patients. The characteristics of COPD patients are diminished lung function, shortness of breath and low-quality air. Advanced techniques like Next Generation Sequencing have clarified the role of respiratory tract microbiota, especially lung microbes encountering different disorders [36]

The respiratory tract is a gateway for the entrance of many bacteria and viruses. The intake and elimination of microbes are disturbed by pulmonary infection and illness. The structure of the lung microbiota varies as the disease progresses. As a result, lung dysbiosis that causes COPD exacerbations has a considerable influence on lung function and overall health. A growing research indicates that lung microbiome dysbiosis is a contributing factor in the onset and intensification of COPD [37].

Globally, COPD is the third leading cause of death. The underlying cause of acute flares-up involves both bacterial and viral contagions. Around 50 % of COPD acute exacerbations may be caused by respiratory viral pathogens. The aberrant inflammation is triggered by inhaled toxic gases and proteases are essential for the etiology of COPD [38].

Smoking-related inflammation is characterized by proteolytic enzymes and excessive neutrophilic infiltration. The Protein Phosphatase 2A (PP2A) activity is reduced in COPD which is a modulator in the process of inflammation. Similarly, the spread of pulmonary diseases is also related to the receptor signaling of Formylated peptides and formyl peptide receptors (FPR). These peptides are produced by pathogens as a

result of tissue damage. Hence, an endless loop is established whereby inflammation stimulates various signaling pathways in cells, protease-antiprotease imbalances, and mucous elevated levels [24].

Particulate matter (PM) is an intricate mixture consisting of fluids, pollutants, and heavy metals and differs in terms of quantity and surface area. Metals like Fe, Pb, Cu, and organic compounds phenols, volatile organic compounds constitute the PM. PM considerably raises symptoms of COPD, acting as a stimulator factor for acute flare-ups of the disease. According to research, exposure to PM is responsible for the increase in pulmonary inflammation in COPD. Additionally, when exposed to PM, COPD patients are more vulnerable to contagious diseases. The reason is that PM contains pathogens and these pathogens cause alterations in the microbes in COPD patients, resulting in the deterioration of the condition of the disease [39].

2.1.1 Metagenomics on COPD

The respiratory microbiota is more poorly studied than the gut microbiome. The latest research shows that chronic respiratory conditions such as COPD and bronchial asthma can result in a changed microflora concerning its different bacterial species. The respiratory bacterial composition and diversity were analyzed in mild COPD patients with different pathological features of the disease. The bacterial drivers in COPD were *Prevotella*, *Streptococcus*, *Staphylococcus*, *Veillonella* and *Pseudomonas*. The research shows there is a relation between imbalanced microflora and respiratory tract malfunctioning over the whole extent of COPD. However, more harmful pathogens are present in patients with acute COPD. The COPD patients were treated with antibiotics that reduce the Proteobacterial and increase the family of Firmicutes [40].

The common bacterial phylum in COPD are Firmicutes, Bacteroidetes and Proteobacteria. The bacterial species dominant in COPD are *Pseudomonas*, *Staphylococcus*, *Veillonella* and *Streptococcus*. The fungi specifically, there was a large increase in *Corynebacterium*, *Actinomyces*, *Actinobacillus* and *Selenomonas* in COPD subjects. The more abundant species reported were *Streptococcus* and *Staphylococcus* in COPD patients [37]. In a study meta-transcriptomic technique was used on the lungs of COPD patient to identify the pathogenic bacterial species and their role in the infection. Pathogenic bacteria such as *Rothia*, *Streptococcus*, *Pseudomonas* and other bacterial colonization were more common in COPD patients. However, the flare-ups were more common in patients where the cause of infection was *Pseudomonas* [41].

There was another study in which shotgun metagenome sequencing was performed on sputum samples of COPD patients. The metagenome classifies the sequences on a specie-level and functional annotation can be possible. Metagenomic sequencing of Upper Respiratory Tract (URT) was conducted to improve understanding of microbial content concerning disease. The bacterial composition of COPD was *Prevotella*,

Pseudomonas and *Fusobacteria*. The other genera include *Neisseria*, *Veillonella*, *Lactobacillus* and *Streptococcus* [42].

Another study utilized the 16s rRNA V3-V4 in which sputum samples of COPD patients were sequenced. The sputum sampling of patients was sequenced at different stages of patients. The microbial heterogeneity was reduced in High-Risk Exacerbators (HRE) e.g. *Streptococcus*, *Prevotella* and *Gemella morbillorum*. Proteobacteria, Firmicutes, and Actinobacteria were same among the both groups. The network analysis involves the highest rank genera using SparCC for interaction analysis. *Actinomyces* exhibit a negative association with *Moraxella* in HRE. *Haemophilus*, *Streptococcus*, *Lactobacillus* and *Canocytophage* also show negative association in HRE [43].

Research was conducted on sputum samples of COPD individuals to better understand the species diversity of the Upper Bronchial Tract (UBT) microbiome and identify physiological alterations linked to the illness, we used metagenomic sequencing of the UBT microorganism. Ten healthy and eight COPD patients' bacterial metagenomes were sequenced. The intensity of the disease was connected with variations in *Streptococcus pneumoniae* abundance and functional categories associated with a decreased ability for bacterial metabolism activities. *Stenotrophomonas*, *Streptococcus* and *Brucella* were only in diseased samples while *Pseudomonas*, *Fusobacterium*, and *Veionella* were in all samples [42].

2.2 COVID-19

A single, positive-strand RNA virus called SARS-CoV-2 is responsible for the severe respiratory illness that affects people. The World Health Organization (WHO) declared a Public Health Emergency of International Concern (PHEIC) in January 2020. SARS-CoV-2 is a contagious disease that spreads to nearly every continent. These viruses can mutate and adapt to various environments, so the risks are persistent and long-lasting [44].

The Primary Recognition Receptors (PRRs) triggered the cytokines production. Virus chemicals and cell deterioration lead to cytokine storms. Other factors such as host responses, spread mechanism, and metabolic changes also contribute to cytokines formation. Neutrophil Extracellular Traps (NETs) are produced by neutrophils that induce microthrombosis, inflammation and cell death. All of these factors result in difficulty in breathing [45].

There were multiple factors for the deaths of patients infected with COVID-19. Nosocomial infections mainly in the lower bronchial tract are also a major risk factor for COVID-19 patients. Multiple factors were responsible for the deaths of patients infected with COVID-19. A lot of bacterial species are responsible for nosocomial infections (NIs). The identified bacterial species through research and exploration were *Escherichia coli*, *Staphylococcus species*, *Pseudomonas*, *Klebsiella pneumoniae* and *Acine-*

tobacter species. The viral agents and bacterial infections or co-infections that can happen during hospitalization can result in the death of many patients with respiratory disorders. After co-infections, there was severe lung tissue damage which induced more vulnerability to bacterial infectious disease. Thus, following a viral infection, bacterial superinfection can arise which leads to more severe and worse conditions [46].

Various reasons are responsible for the intensity of COVID-19 mainly patients with previous medical conditions and weak immune system also contributes significantly. There is also hyperactivation of NF- κ B and viral load is also more in intense conditions. The intensity of infection contributes to malfunction and lung damage. All these conditions lead to pneumonia, COPD, and other lung diseases [47]

2.2.1 Metagenomics on COVID-19

The research on the upper bronchial tract of COVID-19 patients was carried out for a better understanding of the correlation between micro-organisms and the intensity of the disease. There are only a few studies that are focused on the lower bronchial tract but it is also significant in intensity of disease [48].

Inflammation and Immune system response can aggravate the infection. Several Cytokines MIP-1 alpha, IL-6, CCL2 and CXCL10/IP10 were released in the host body. Microorganisms can alter and worsen the infected condition. *Veillonella* and *Haemophilus* had a high concentration in COVID-19 patients in the oropharynx region. The nasopharynx had a high burden of *Streptococcus*, *Staphylococcus*, and *Corynebacterium*. The dominant fungi species was *Candida* while the bacterial species were *Enterococcus faecium*, *Klebsiella* and *Staphylococcus* [49].

Metagenome Sequencing was used on COVID-19 patients to analyze the respiratory microbiome and co-infection pathogens. The frequent bacterial species were *Klebsiella*, *Staphylococcus*, and *Enterococcus*. A high frequency of viruses was detected including Epstein-Barr Virus (EBV), Herpes Simplex Virus (HSV) and Human Parvovirus B19 (HPVB19). The fungal species that were also detected most frequently were *Cryptococcus*, candidemia and *Aspergillus*. Hence co-infections occurred and worsened the condition [50].

Analyzing the upper bronchial tract and COVID-19 is of great interest. The research was carried out on metagenomic RNA-seq sequences to investigate the microbial metabolic processes. The micro-organism's composition differs considerably with different micro-organism divisions consisting of archaea, viruses, fungal and bacterial species. Several pathways are linked to COVID-19 disease severity such as virulence, cell death, membrane transporters, and immune system. There was notable variation in the microbiome of healthy and diseased groups. The primary phyla were Bacteroidetes, Firmicutes, Actinobacteria, Cyanobacteria and Proteobacteria which were dominant in infected individuals. Additionally, the abundant genera in infected individuals were

Actinomyces, *Streptomyces*, *Pseudomonas*, *Staphylococcus*, *Corynebacterium* and *Mycobacterium*. There was a negative correlation between COVID-19 and microbiome diversity in the upper bronchial tract [51].

There was research carried out on BALF samples compilation of metatranscriptomes to understand the lower respiratory micro-organisms. For the understanding of functions, pathways analysis was performed. There were various affected pathways some of them were recruited here Carbapenem biosynthesis, N-Glycan biosynthesis, Glycolysis / Gluconeogenesis and D-Alanine metabolism [52]. Significant research was carried out to analyze the difference in lung microbiome between the infected and healthy individuals. Bronchoalveolar lavage sampling was used to carry out 16s rRNA sequencing. A notable change was observed in the richness of micro-organisms richness. *Acinetobacter*, *Enterobacteriaceae*, *Pseudomonas*, and *Sphingobacterium* were prominent among the infected persons [53].

In a study, together with assembly and binning techniques, whole-metagenome shotgun sequencing data were used to recreate Metagenome-Assembled Genomes (MAGs) of five hundred fourteen COVID-19-related nasopharyngeal and fecal samples across 6 distinct groups. The MetaWRAP (v1.3.2) and MaxBin2 (v2.2.6) are used for generating metagenome assemblies and binning. However, a few numbers of nrMAGs were found in the nasopharynx samples due to contamination [54].

2.3 Lung Cancer

Lung cancer is a widespread disease affecting millions of people across the world. Various factors are responsible for the disease's development. There were various genes and transcripts involved in the pathogenesis of the disease. The primary risk factor considered is tobacco while radioactive gas called radon also contributes significantly to lung cancer. Nuclear radiation Exposure to asbestos, chromium some viruses such as Human papillomavirus (HPV), Mycobacterium TB, inflammation due to various reasons, coal and biomass used for cooking plays an effective role in disease [55].

It appears that there is widespread dysregulation of Long Non-Coding RNA (lncRNA) expression in cancer, where these molecules function as tumor suppressors or oncogenes. The lncRNA plays a crucial role in the expression of genes and the disruption of the expression of lncRNA has a critical role in cancer pathogenesis. For the diagnostics indicators, lncRNA is of great importance. There was a difference in expression between healthy and cancerous cells. There was research conducted in which the expression level of lncRNAs e.g LINC00968, PCAT19, ADAMTS9-AS2, SVIL-AS1 was downregulated, and NUTM2A-AS1, PCAT6, LINC01214, LINC00673, FEZF1-AS1 show upregulated expression in cancerous cells [56]. The human body contains a variety of microbes that are essential to preserving the host body's microbiome. Disruption of micro-organisms brings about various disorders. Production of cancerous substances, pathogenic factors,

and inflammatory responses can be reasons of disrupted micro-organism communities to lead to cancer [57].

2.3.1 Metagenomics on Lung Cancer

The micro-environment containing the viruses, bacteria and other microbes in the lower bronchial tract is considered as microbiome of the lung. The source of lung microbial content is the upper respiratory tract. A healthy lung has bacterial phyla Proteobacteria, Actinobacteria and Fusobacteria [58]. The microbial communities of the lung are unique from other body sites. Microaspiration and oral microbial content serve as the primary source of the lower respiratory tract. Moreover, coughing, ciliary transport, and immunological responses were responsible for microbial transportation. Furthermore, repeated usage of antimicrobials supports the infection link between pathogens imbalance and cancer development [59].

The lung microbiome can impact immune mechanisms related to cancerous cells. Respiratory flora has an impact on tumor-promoting pathways since several compounds produced by bacteria were linked to carcinogenesis. According to an analysis, respiratory carcinoma had a high regulation of pathways e.g. amino acid metabolism, lipid metabolism, and xenobiotic biodegradation. The gene expressions in pulmonary cells may be further influenced by this changed microorganism biochemical composition. The Veillonella had enhanced the expression of the Extracellular Signal-Regulated Kinase (ERK), Phosphatidylinositol 3-Kinase (PI3K), and Mitogen-Activated Protein Kinase (MAPK) networks in gene expression profiles of the pulmonary tissues [60].

Through infectious substances, genetically harmful pathways, and host autoimmune processes, the microbial community can contribute to the progression of cancer. A study was conducted on patients with lung cancer using 16S sequencing. The analysis shows that in later stages of lung carcinoma, Legionella and Thermus were found providing insight into the pathology of tumor formation. The cancer group exhibited a considerable over-representation of the genera e.g. *Novosphingobium*, *Acinetobacter*, *Rhizobium*, *Prevotella*, and *Christensenellaceae R-7 group*. There was an enrichment of *Haemophilus influenzae* (*H. influenzae*) in the SCC group. *Novosphingobium*, *Parasutterella*, and *Micrococcus* were a few ASVs that had low concentrations [61].

DNA degradation can be caused by *Bacteroides fragilis*. The low concentrations of anti-cancer p53 gene cause aggression of bacteria and inflammation of NF- κ B promotes carcinogenesis. For an improved picture, an amplicon sequencing analysis was performed on lung carcinoma patients. A different taxa Acidovorax was detected with a mutation in the TP53 gene. *Rhodofera*, *Klebsiella*, *Pseudomonas*, and *Anaerococcus* were among the prominent ones. *Polaromonas* and *Comamonas* were also present although not prominent according to biometrics [62].

A lot of studies bring to light that pathogens and carcinoma of the lung are closely related, some researchers have used the latest innovations in metagenomics sequencing

to discover germs that contribute to lung cancer as biomarkers for tumor detection. In metatranscriptomic research conducted on lung cancer patients. *Prevotella* with different species, *Haemophilus*, *Mycobacterium*, *Moraxella*, *Rothia*, and *Streptococcus* were among the dominant bacterial genera in tumor cells. Furthermore, statistical analysis showed that *Pedobacter*, *Prevotella*, *Klebsiella* and *Mycobacterium* were among the putative biomarkers [63].

There was research that focused on the microflora of lungs during the cancer. Sampling was carried out in two ways. Shotgun analysis was carried out on the cancer and control samples. Distance-based RDA analysis is used to analyze the role of other factors in lung cancer progression. Analysis shows that there is an association between the factors and microflora. Actinobacteria, Firmicutes, and Proteobacteria had high abundances while *Mycobacterium* and *Streptococcus* were among the prevalent genera in the cancer patients [64].

2.4 Tuberculosis

The primary cause of TB is the bacterial species *Mycobacterium TB*. The disease cycle starts with the source of disease, contagious particle formation, and entrance of these pathogens into the host and then transferred to another prone person. TB patients have aerosols that are transmitted through sniffing, screaming, and coughing the effective one [65].

In a report, WHO declared that TB was the cause of 1.4 1 million deaths and around 10 million were infected in 2019. Co-infections were the most prominent cause of developing TB. The SARS-CoV-2 and HIV were the dominant disease agents. Other health issues e.g. nutritional deficiency, organ implantation, and hyperglycemia were also the reasons for the development and intensity of the disease. There were some genetic factors, variation of A4 hydrolase, gene and protein expression patterns of *mycobacterium*, immune defense reaction and capability of epigenetics. All these factors play a potential role in disease progression [66].

Bacterial adaptation in the pathology of TB also plays a major role. There was research that summarizes that host reaction during infection varies as T-cell activation and differentiation occur. MTB strain Rostov had a less intense response than the H37Rv strain in the mouse models that were used for the research. The Quiescent bacteria become intensified and result in the formation of cavities in immuno-suppressed individuals. Lung impairment happens as the pathogen spreads in diseased persons [67].

Drug resistance in TB infection is a serious issue and biofilm formation has its role in the resistance and *Mycobacterium* pathogenesis. Previous studies have provided evidence that various genes and about 115 proteins play their parts significantly in the creation of biofilms. DevR, Rv0097, Rv1996, and RegX3 were the common proteins

that perform their activity in bacterial adhesion. A few genes were also involved lpqY-sugABC, Rv0195, secA2 and Rv1176c. Genetic regulators e.g. Rv0199, nirB and pks1 also had crucial regulation in microbial colonization. Gene ontology analysis revealed a few transcription factors, such as RegX3 and NarL while some kinases as PknG and PknB were also involved in pathogenesis [68].

2.4.1 Metagenomics on TB

There were very few studies on the lung microbiota of TB patients. Here are the findings of a study in which they performed 16S rRNA sequencing on the lung microbiota of TB patients. The *Anoxybacillus* which is a rod-shaped bacteria was significantly high in diseased persons. The relative abundance of Firmicutes, Bacillaceae and Mycobacterial communities was higher in diseased patients than the normal ones. Overall study shows that the lung communities differ potentially between the infected and healthy individuals [69].

The micro-organisms of the pulmonary tract have an important impact on the host defense system and in shaping the breathing system. Lung flora affects pulmonary infections such as TB, COPD, asthma, and cystic fibrosis. The lower respiratory tract has a lower microbial community than the URT. The nasal communities of microbes include *Propionibacterium*, *Corynebacterium*, and *Staphylococcus* while *Neisseria*, *Haemophilus* and *Prevotella* were enriched in the oropharynx. Amplicon rRNA sequencing revealed unique symbiotic bacteria during persistent pulmonary infection. Usually, Pulmonary flora has few sample studies and little knowledge due to difficulties in sampling. A shotgun metagenomic sequencing revealed that *Actinomyces*, *Caulobacter* and *M.tuberculosis* were enriched in TB patients [70].

In a study, a shotgun metagenomic analysis was carried out on TB-infected individuals. *Rothia* was prominent in throat culture while *Pasteurella* and *Klebsiella* were dominant in BALF. In Untreated TB patients, there was a high abundance of *Staphylococcus aureus*, *E.coli* and *Pasteurella multocida*. *Staphylococcus aureus* can cause multiple organ damage and various illnesses. TB patients had a low ratio of *Prevotella melaninogenica* which leads to aggravating the inflammation and deteriorates the pulmonary system [71].

The pulmonary microflora fluctuate during several disease states and these fluctuations encounter inflammations which results in lung function alterations. To evaluate the microbial communities in the respiratory system both amplicon and shotgun techniques were used. Amplicon shows various taxa had high relative abundance in TB patients e.g. Absconditabacteria, Tenericutes, Proteobacteria, Spirochaetes, TM7 and Gracilibacteria while at the genera level TB patients had *Neisseria*, *Fusobacterium*, *Lautropia*, *Streptococcus* and *Campylobacter*. Diversity relationships between different genera abundance in TB patients, their association robustness, and effects on host responses were also analyzed [72].

Very little is known about the lung microbiome of TB patients with dysbiosis and mostly analyzed using amplicon sequencing. For a better understanding of lung microbes, shotgun analysis was performed on bronchoalveolar lavage samples between TB patients and healthy persons. Alpha diversity of TB patients decreased dramatically than control ones. Mycobacterium was on the top of the list in TB-infected persons while Actinobacteria was dominant in the perspective of different phylums. Pseudomonas, Selenomonas, Streptococcus salivarius, and Neisseria were present in healthy individuals. Finally, system biology techniques were used to find the hub bacteria. Pseudomonas, Rothia and Mycobacterium were among the hub microbes [73].

2.5 ML in Metagenomics

BALF samples with lower pulmonary infections were extracted for metagenome sequencing. Furthermore, the RF model was built to check the influence of bacteria on hospitalization. Patients were divided into two groups based on their residence period. The model was trained with various factors but it performed better when bacterial content was integrated with diagnostic parameters [74].

ELasticNet is used for linear models, with the metadata of Cystic fibrosis patients. ElasticNet is a regularization technique like lasso which extracts the significant features for model training. The model also predicts lung functioning. 16S rRNA sequencing was also done on the patient samples. Models trained on microbial information give significantly better prediction than models incorporated with patients' metadata. The model accurately predicts the diseased bacterial species e.g. Achromobacter and Pseudomonas [75].

To understand the association between bacterial extracellular vesicles and pulmonary disorders. ANN ML model was built that can differentiate between asthma, lung cancer and COPD patients based on bacterial taxonomy. The taxonomical clustering was derived from the OTU created by the 16S amplicon sequencing. The model trained on COPD shows the Firmicutes on top while the model for lung cancer and asthma shows the Proteobacteria was the most prominent [76].

A study was conducted that uses the RF model to align the important features using metadata and micro-organism information of Cystic Fibrosis patients and normal individuals. To reduce the dimensionality of the data, Brouta wrapper technique was used. The diversity of microflora also plays an effective role in the categorization of bacterial species. Using different variables and parameters, the cross-validated error measures were duplicated one hundred times. In contrast, the model does not perform well in a grouping of cystic fibrosis and control samples taken from the oral swabs. It also can't differentiate between the Pancreatic sufficient (PS) and pancreatic insufficient (PI) groups of patients. The R software was used for the ML analysis [77].

2.6 Study Rationale

Lungs have diverse and low microbial content. The identification and classification of microbes role in respiratory disorders requires more research and identification techniques. There is no comparative research on respiratory diseases in context of metagenomics even the metagenome of diseased individuals are not compared with healthy persons in most of previous studies. To bring into consideration these limitations and gaps we are performing in-silico analysis of respiratory microbiome. The analysis will improves our understanding on the role and relative abundance of microbes in diseased and control conditions. We will identify the different significant patterns of diseased data using different ML algorithms.

2.7 Objectives

- To identify the micro-biome content present in respiratory disorders analyzing in-silico amplicon and shotgun sequencing data.
- To build a ML model that can differentiate between different respiratory diseases based on metagenomics data.

Chapter 3

MATERIALS AND METHODS

The proposed methodology is for the identification and better understanding of the microbial communities in pulmonary disorders primarily the lung microbiota. Metagenomics datasets of sequence reads are used in the aspect of lung and nasopharynx samples. Additionally, different ML models are also built that can differentiate between the diseased and healthy samples based on bacterial abundance and metagenomics data. The general workflow for methodology is represented in the Figure 3.1.

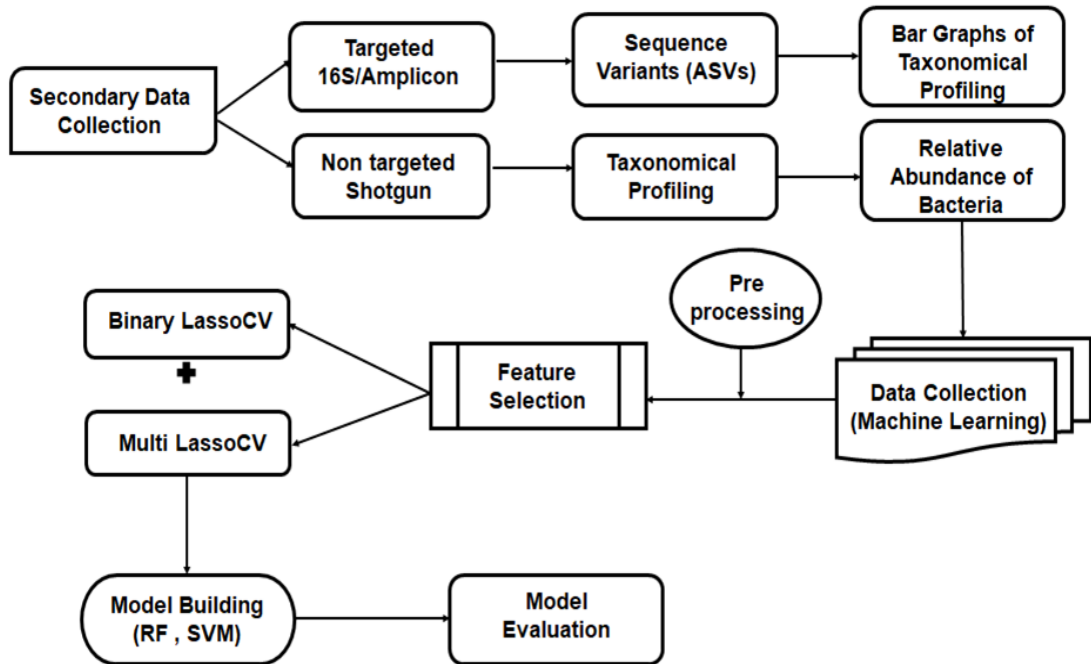


Figure 3.1: General Workflow for Methodology

3.1 Dataset Collection

The datasets are selected on some attributes that plays a significant role on the overall research and its outcomes. Datasets of 16S rRNA and shotgun sequencing are selected for the computational analysis of reads. Two datasets of amplicon and four datasets of shotgun are used. The datasets are downloaded from publicly available repositories e.g. European Nucleotide Archive (ENA), Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA). Reads should be pair-end and have both normal and diseased samples. Reads should be of Homo Sapiens. The sampling sites should be respiratory tract primarily the lungs. The details of datasets used for analysis are represented in the Table 3.1.

Table 3.1. Datasets used for the Study

Accession No.	Disease	Sampling Site	No. of Samples	Study Type
PRJNA636302	COPD	Lung	78	Amplicon
PRJNA687143	COVID-19	Lung	52	Amplicon
PRJEB9034	COPD	Lung	18	Shotgun
PRJNA743981	COVID-19	Nasopharynx	99	Shotgun
PRJNA714488	Lung Cancer	Lung	47	Shotgun
PRJNA655567	TB	Lung	61	Shotgun

3.2 Metagenomics

Metagenomics is the collective study of microbial communities in a specified environment. Several methods are utilized for the identification and understanding of microbial communities. Culture-based methods provide evidence with few limitations while non-culture techniques have brought a good shift for detailed analysis. Non-culture-based techniques e.g. NGS are more appropriate for analyzing the microbiome. Amplicon and shotgun are widely used techniques for analyzing the microbiota in different environments. Metatranscriptome, culturome and virome are also used for micro-organisms analysis. All these techniques work with both DNA and RNA extraction and sequencing [32].

3.2.1 Amplicon Sequencing

The bacterial gene has both conserved and variable regions. The conserved regions are similar among all bacteria and variable regions vary in different bacterial species. Since variable regions are mostly sequenced to identify the bacterial species in various samples. The variable parts of a gene range from one to nine represented as V1-V9.

The commonly used regions for sequencing are V4, V3-V4 and V1-V2. The base pairs range is approximately 1500 bp of any particular gene. Full-length sequencing of the 16S rRNA gene is possible through the PacBIOs Sequel and Nanopore sequencing while 10,000bp can be sequenced. The primary steps for amplicon sequencing involve DNA Extraction, PCR Amplification, Library Preparation and Sequencing [78].

3.2.2 In-Silico Analysis of Amplicons

In-silico platforms commonly used for amplicon analysis are Mothur and Quantitative Insights Into Microbial Ecology(QIIME). After sequencing, the quality of raw reads is analyzed. The tools used for quality control are the Deblur and Divisive Amplicon Denoising Algorithm 2. The well-known methods used for read clustering are OTUs and ASVs. An important step is a taxonomical classification, which involves the databases e.g. Greengenes2 and Silva databases. Visualization is mostly in bar graphs with bacterial species classification [79]. The graphical representation of methodology used for amplicon analysis is represented in the Figure 3.2.

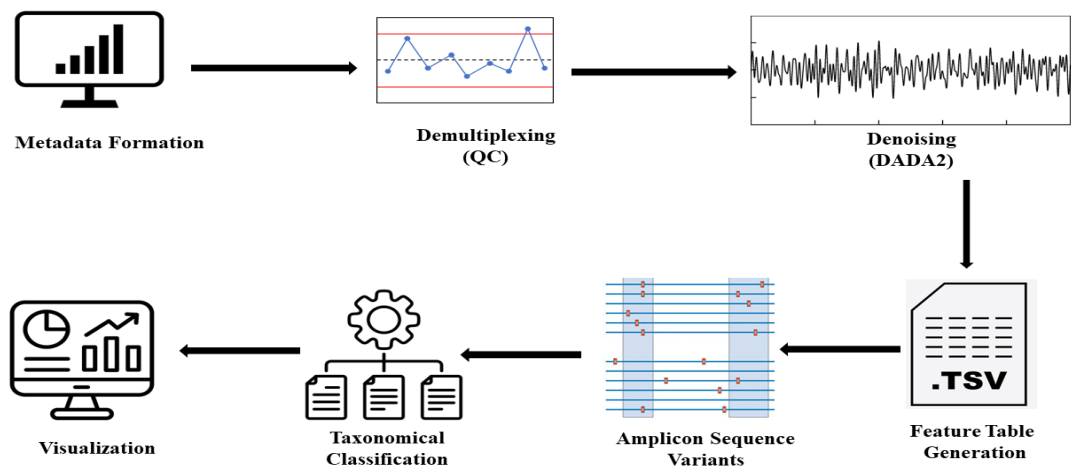


Figure 3.2: In-Silico Analysis steps used for Amplicon Analysis

The primary steps used here for amplicon sequencing utilizing Quantitative Insights Into Microbial Ecology (QIIME2) are:

- Metadata formation
- Demultiplexing
- Denoising
- Feature Table Generation
- Amplicon Sequence Variants(ASVs)
- Taxonomical Classification
- Visualization

3.2.2.1 Quantitative Insights Into Microbial Ecology 2

QIIME2 version 2023.2 (<https://qiime2.org/>) is a powerful tool used for the analysis of amplicon sequences. There are mainly three main interfaces of QIIME2, q2cli is the command line, q2galaxy is the graphical user, and the Artifact API interface uses Python. Here we used the command line interface using QIIME2. QIIME2 is installed through WSL. Linux commands are used for installation of the latest version of Miniconda in Windows from the Anaconda website (<https://docs.anaconda.com/free/miniconda/>). After installation of miniconda QIIME2 is installed within conda environment. QIIME2 2023.2 distribution is installed for analysis.

3.2.2.2 Demultiplexing

After installation, the datasets are downloaded from the ENA (<https://www.ebi.ac.uk/ena/browser/>) and SRA (<https://www.ncbi.nlm.nih.gov/sra>). Data should be demultiplexed, pooled samples can't be used for analysis. The datasets used for analysis are already in separate paired fastq files for each sample. Demultiplexing also gives additional information about the data e.g. sequence count summary and quality plots.

3.2.2.3 Data Input

The first step is the creation of path files for input and metadata files. For path files, three columns are created with SampleID, forward absolute file path, and reverse absolute file path as paired sequences are used. The metadata files contain sample IDs, sample names, their SRA-IDs, body sites year of publication, etc. The fastq files of all samples are collectively input through the path file and their format is converted into qza. QIIME2 can't work with any other format of files.

3.2.2.4 Denoising

After data input, the denoising of data is carried out through Divisive Amplicon Denoising Algorithm 2 (DADA2). For denoising, qza sequence files along with metadata are input for Quality Control (QC). DADA2 separates the chimeric sequences. The tool uses two main parameters trim for trimming at the beginning and end of sequence reads and trunc for truncation at any specific position of sequence. As the quality plots are generated through demultiplexing then the low-quality reads are trimmed or truncated from the sequences. The output files are, a file showing the statistics with filtered, denoise parameters, etc, representative sequences with feature information, and a feature table showing frequency per sample, sample details, and feature details. These features uses the ASVs as the DADA2 uses the ASVs for representation of sequences in the data.

3.2.2.5 Amplicon Sequence Variants and Operational Taxonomic Units

There are two methods used for species classification. Operational Taxonomic Units use a 97% identity index for grouping similar sequences. OTUs reduce the computational power and mistakes to some extent from the analysis. Clustering was in wide-ranging groups leading to more general predictions. Different OTUs picking methods were used including reference-based and de-novo. OTUs were widely used techniques before the era of denoising methods. Recently, the trend shifted towards denoising methods primarily the ASVs. ASVs are the most advanced technique for sequence classification. There are no identity thresholds instead matched biological variants for grouping. The results are more accurate and different sequence variants between highly correlated bacterial species. It results in the exact match rather than a general way of grouping as in OTUs. Sometimes ASVs lead to strain-level classification providing more insights into the microbial communities. In simple terms, denoised OTU is ASV.

3.2.2.6 Taxonomical Classification

There are different taxonomical classification methods after the creation of ASVs. Reference-based methods using databases and naive Bayes pre-trained classifiers classifiers. Reference-based uses the Greengenes Database and Silva database for the grouping of microbial communities. Both can be used for classification but as Silva is updated regularly hence Silva is preferred over the Greengenes database. As pre-trained classifiers work smoothly so we select the silva-138-99-nb-classifier.qza naive Bayes classifier with 99 % sequence similarity classifier used for classification of microbial species.

3.2.2.7 Visualization

After taxonomical classification, the last step is visualization. The taxonomy files are input and bar graphs are generated from them. Bar graphs show the identified bacterial communities at different levels within each sample. As we view the bar plot in QIIME2 view, there are several parameters e.g. bar width, taxonomic levels and color palette and download options etc. Taxonomical levels with their description are the following:

- Level 1 represents the domain level as either the identified is bacteria or archaea or some other domain
- Level 2 represents the Phylum of bacteria or archaea
- Level 3 representing the class of identified microbes
- Level 4 represent the order of bacteria
- Level 5 represents the family of microbes
- Level 6 shows the genus of bacteria or archaea
- Level 7 represents the species level of microbes

3.2.3 Shotgun Sequencing

There are different High-throughput Sequencing technologies. The base pair length, expense, range of inaccuracy, and speed also play important roles. Various factors e.g. sequencing accuracy, the relative abundance of the sample, and the type of sequencing are considered during the choice of sequencing. Shotgun is preferred among all because of its accuracy. The commonly used technique for sequencing around the world is Illumina. The laboratory steps for metagenome sequencing involve microbial sample collection, DNA extraction from the samples, QC, DNA fragmentation, Metagenome library preparation and finally sequencing [30].

3.2.4 In-Silico Shotgun Analysis

After metagenome sequencing, raw reads are obtained. The QC of raw reads is analyzed and low reads are trimmed by various software e.g., MultiQC, FastQC, Cutadapt, and Trimmomatic are among the commonly used tools. FastQC is widely used for QC of raw reads with ten QC points. Trimmomatic and Cutadapt are used for trimming low-quality or for short reads. After QC, alignment is done using Bowtie or BWA tools. The reads are aligned to the human genome to remove the host reads and focus only on the microbial content. Taxonomical Classification is carried out in two

ways. The use of Reference genome and metagenome assemblies while reference-based genome classification is using Metaphlan tool. The MetaWrap, is used for MAGs for taxonomical classification. Functional Analysis is the additional information provided by the metagenome sequencing and HUMAnN2 is used for functional annotation. The graphical representation of methodology used for shotgun analysis is represented in the Figure 3.3.

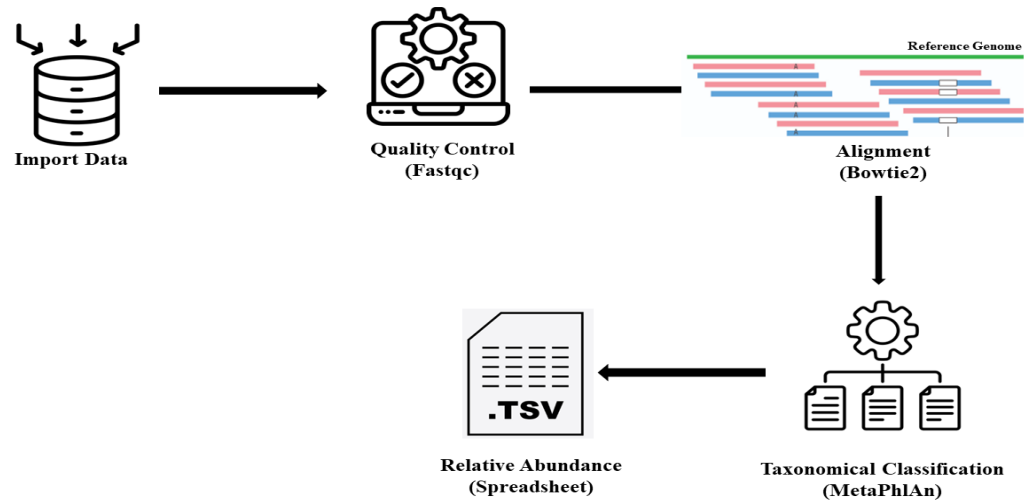


Figure 3.3: In-Silico Analysis steps used for Metagenome Analysis

The methodology used for the selected datasets are:

- Import Data
- QC
- Alignment
- Taxonomical Classification
- Relative Abundances

3.2.4.1 Galaxy

Galaxy is a well-known framework for the computational processing of high-throughput sequencing data. It is an open-source platform. An HP ML350pT08 LFF CTO server with 24 cores (2 CPUs) and 288GB of RAM running Debian OS (Jessie v7) hosted the Galaxy instance allocated for detection. Two additional RAID5 systems with five 4TB disks each were used for storage. Galaxy has two significant dimensions:

- A command line interface (CLI)
- An intuitive Web-based graphic user interface (GUI)

The Graphical user interface is easy to use while the command-line requires expertise and an understanding of Linux. A major reason for using the galaxy project is extensive memory which handles the large amount of data generated through high throughput techniques. Additionally, providing a vast variety of tools for different analyses and provides a user-friendly environment. [80].

3.2.4.2 Importing Datasets

The datasets are present in the public repository of ENA (<https://www.ebi.ac.uk/ena/browser/>). The format for input files is fastq and pair-end sequences are used. After checking the desired parameters raw reads are uploaded from ENA to the Galaxy project through file links.

3.2.4.3 Quality Control

After uploading into the Galaxy platform, the first step is quality checking of fastq files. The sequence files contain low-quality reads, GC content, over-represented sequences, adapter content, and duplicates. A tool called FastQC is intended to identify any issues in high-throughput sequencing datasets. The tool checks all these parameters and makes an extensive report on them. If there is adapter content, and reads with low-quality phred scores, high level of duplication, or any error the pre-processing is performed to improve the quality of data. Fastp is a program made to quickly and easily prepare Fastq files all at once. Base correction, polyA tail trimming, adapter trimming and duplicates can be removed using Fastp. Both FastQC and Fastp generate the html reports through which we can analyze all these parameters for further processing.

3.2.4.4 Taxonomical Assignment

Taxonomical profiling is the identification and measuring of the relative abundances of the micro-organisms content in the collected samples. It is a crucial step in shotgun sequencing as it classifies the microflora which is required for further analysis. There are two popular ways of profiling bacterial communities in samples. There are two popular ways of profiling bacterial communities in samples.

- Reference Based Profiling
- Metagenome Assembled Genomes (MAGs)

In contrast to assembly, reference-based computational methods properly recognize and assess the identified taxa and current genes in a microbial flora by homology, based on annotated reference sequence data. The capacity of reference-based metagenomic profiling to identify rare and difficult-to-assemble genomes is a significant benefit over assembly. This makes it possible to produce reliable ecological statistics for common and uncommon taxa, which are challenging to measure precisely due to numerous technical omissions in data derived only from metagenome assemblies.

All the microbial communities in sampling have their individual and unique genomes. Reference-based genome techniques can't properly align and lack some information. Reconstructing a genome effectively from the provided collection of metagenomic sequences is therefore one of the primary objectives in the processing of metagenomic data. Pre-processing, assembly, and binning is the pipeline for MAGs. A hybrid of short and extensive reads is the perfect way to create assemblies. The sample-wise sequencing depth information can help binning methods decide which contigs should form a MAG. Metagenome assembly and binning have been accomplished through the development of many processes. ATLAS, Muffin, and MetaWRAP are used for the formation of metagenome assemblies [81].

3.2.4.5 Metaphlan

MetaPhlan 4, is a method that utilizes a combined additional of microorganism genomes and MAGs to generate a wider range of species-level genome bins⁴² (SGBs) and precisely identify their existence and abundance in metagenomes, thereby leveraging the best aspects of both reference- and assembly-based metagenome profiling. SGBs are defined exclusively by the MAGs and comprise both recognized species (kSGBs) and non-characterized species (uSGBs). MetaPhlan4 is the most updated version of Metaphlan. The tool uses the read homology to estimate the relative abundance of samples. The input can be of three ways, single or multiple fastq files, a SAM file and an alignment file generated by the Metaphlan. The first step in the MetaPhlan pipeline is to use Bowtie2 to align the raw reads against the database of SGB-specific

markers. SGB-specific markers are specific to the species-level classification of bacterial species. After alignment then microbial categorization and finding their relative abundances in the samples is the important step in downstream analysis. The tool uses reference-based genome alignment and various databases to attain the purpose. In output, four files are generated:

- A tab-separated file with prediction and relative abundances of sample
- A BIOM file contains taxonomic profiles
- A Bowtie2 contains information on alignment quality and information
- A SAM with aligner, mapping score, and all alignment positions, etc.

[82].

In our analysis for taxonomical grouping their relative abundances are calculated using Metaphlan4. The input files are pair-end fasta files. The '-very-sensitive' and reads shorter than 70 bp '-read-min-len-70' parameter is used by the default. In this way, short reads and very high accuracy can be achieved. In SampleID, the sample-ids provided in the metadata of datasets are replaced for each sample. Mostly the CHOCOPhlan databases were used for microbial profiling. The mpa-vOct2022-CHOCOPhlanSGB-202212 is the most updated database that we used for the taxonomical categorization of microbiome.

3.3 Machine Learning

The goal of computer science's ML field is to use data to learn how to perform better across a range of tasks. ML is commonly employed in applied healthcare research to denote automated, highly adaptable, and computationally demanding methods for detecting patterns in intricate data structures, such as underlying dimensions, nonlinear connections, and interactions. Generally, there are two types of learning. Supervised learning learns from the labeled data while Unsupervised learning learns on without labeled data and learns few features [35]. The graphical representation of methodology used for ML models is represented in the Figure 3.4.

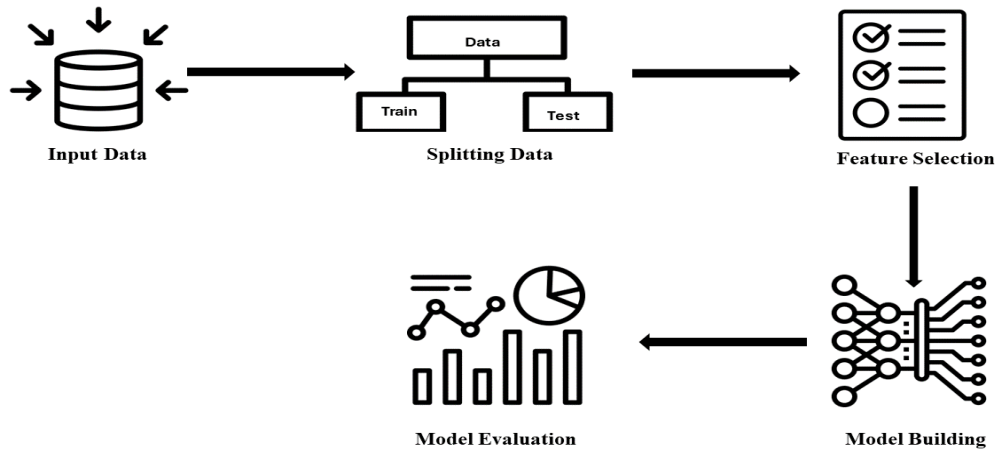


Figure 3.4: Steps used for ML

The ML analysis comprises the following steps:

- Input Data
- Splitting of Data
- Feature Selection
- Model Building
- Training of Model
- Model Evaluation

3.3.1 Input Data

Data is in the form of the relative abundance of the species calculated from the Metaphlan4. After the identification and abundance calculation, all the Metaphlan tables are merged. Metaphlan can differentiate where the particular bacterial species is present or not and itself places the zero in the missing value rows. After merging, information on all the classification levels is discarded as we are interested in the bacterial species hence only species-level information is available in rows. The merged relative abundance table has both normal and diseased samples of four respiratory disorders e.g. TB, COPD, COVID-19, and lung cancer. Hence the merged table is input in the model.

3.3.2 Splitting of Data

The splitting of data is a crucial step in ML models. There are different splitting criteria for splitting of data e.g. 70/30 and 80/20. Mostly 80/20 ratio is used for splitting into train and test sets. Train data is used for the training or learning of the model while test data is used for the evaluation of the model. We have used the 80/20 ratio for splitting of data. The splitting is helpful for evaluating and measuring the performance of the model.

3.3.3 Feature Selection

Extracting the significant and appropriate attributes from the initial feature collection is known as feature selection. It results in a reduction of dimensions of data as duplicates and unnecessary attributes are removed. Using feature selection techniques, learning algorithms can be trained on the data. Effective feature selection can enhance learning outcomes, boost accuracy, and shorten learning times [83]. We have used the Least absolute shrinkage and selection operator (Lasso) or L1 regularization method in RF and SVM for significant feature selection for the training of the models. The most significant bacterial species relevant to a particular disease based on their abundance are selected and used for the learning of the model.

3.3.4 Model Building

After significant feature selection, the RF and SVM classifiers are selected for building a model. Sklearn is a widely used library for ML models. We also used Scikit learn for training and building our models. Hence two ML models e.g. RF and SVM are built for the better identification of microbial species in the respiratory disorders.

3.3.5 Model Training

After model building, the important step is the training of the model. The training set of data is used for the training of the model. The labeled data is used for the learning of the model so that it effectively learns and enhances its capacity to go well with the test data. Mostly, 80 % of data is utilized for the training of models.

3.3.6 Model Evaluation

Model Evaluation is to analyze the performance of the build model. In this step, we can categorize how well the model has learned and how it works with unseen and real data. The 20 % of data that is not used for the training is used for the testing of the build models. In this way we can analyze the reliability of the model.

Chapter 4

RESULTS

The primary purpose of the executed research is to explore the pathogenic microbes in pulmonary disorders. The important techniques used to achieve the objectives are amplicon and shotgun analysis. After that ML algorithms are used for the classification of samples into diseased and healthy based on relative abundances of Bacteria generated after shotgun analysis.

4.1 Amplicon Analysis Results

After raw data retrieval from ENA, amplicon analysis is performed. The steps for analysis are the creation of path and metadata files, demultiplexing and denoising of the raw reads, feature table, ASVs and taxonomical classification.

4.1.1 COPD Amplicon Dataset

In-silico analysis of amplicon sequences is carried out to analyze the microbial community and diversity in respiratory diseases. The analysis is carried out in the most familiar tool for amplicon analysis, QIIME2. First, the pair-end fastq files are downloaded from the ENA for COPD. The ENA Id for COPD is PRJNA636302. The datasets contains 78 samples of different stages of disease. To input the files for the analysis, path files are created with the ENA links of fastq files. The path file and fastq files are placed in the directory. The metadata files are also created for the analysis.

To analyze the quality of the raw reads, demultiplexing of the sequences is done. The demultiplexing step will result in two sections. The first section shows the overview of demultiplexed sequence count summary of forward and reverse reads and histograms of forward and reverse reads frequency histograms. The summary of demultiplexed summary of sequence count for COPD is shown in Table 4.1. The histograms for forward reads are shown in Figure 4.1 and histogram for reverse reads in Figure 4.2.

The second section shows the interactive quality plots for forward and reverse reads and demultiplexed sequence length summary table. The interactive quality plots for forward reads are shown in Figure 4.3 and for reverse reads for shown in Figure 4.4.

Table 4.1. Summary of Demultiplexed Sequences of COPD Dataset

Statistics	Forward Reads	Reverse Reads
Minimum	56784	56784
Median	200026.0	200026.0
Mean	228741.012821	228741.012821
Maximum	450590	450590
Total	17841799	17841799

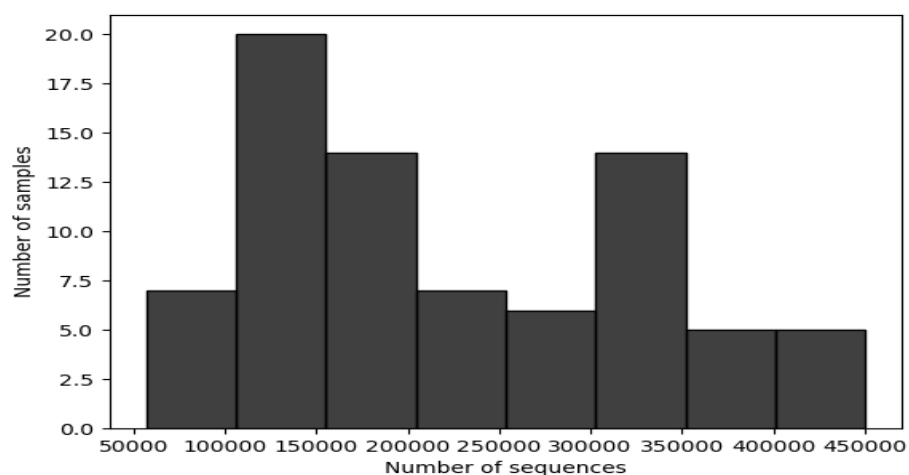


Figure 4.1: Forward Reads Frequency Diagram of COPD

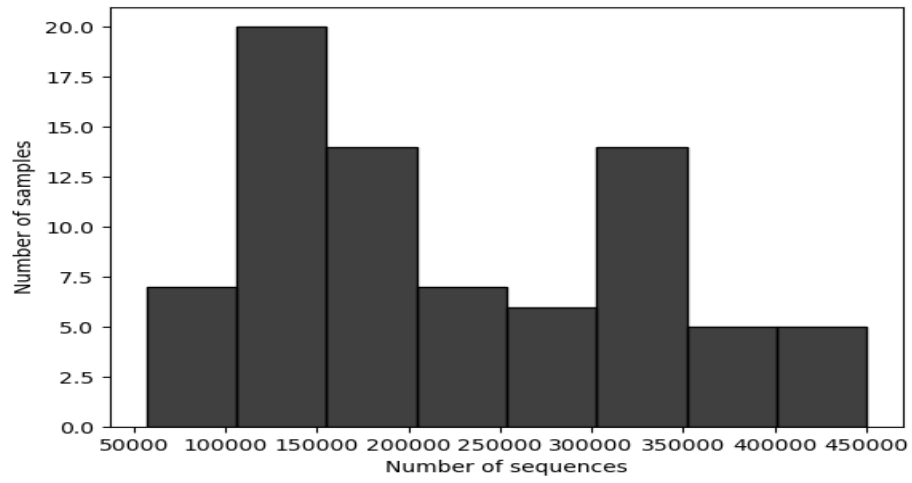


Figure 4.2: Reverse Reads Frequency Diagram of COPD

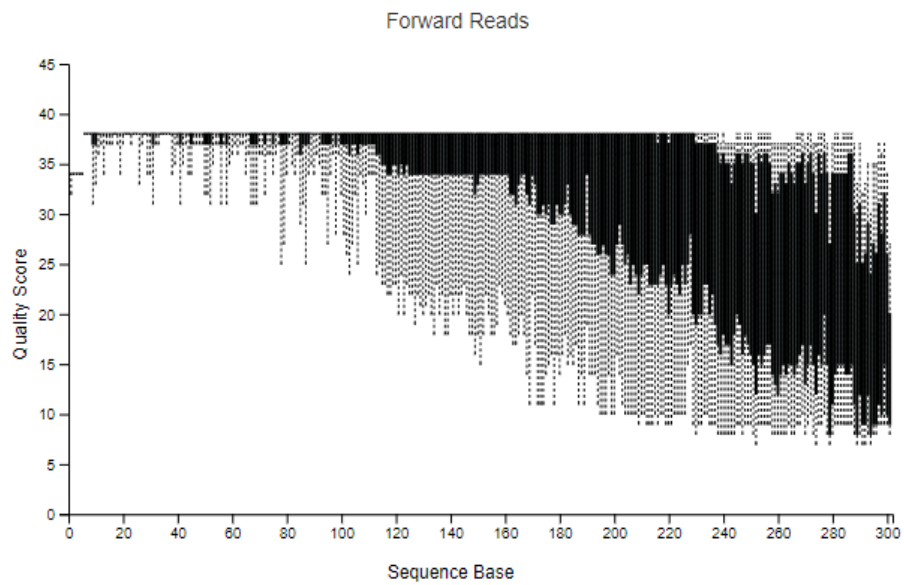


Figure 4.3: Quality Score Graph for Forward Reads of COPD

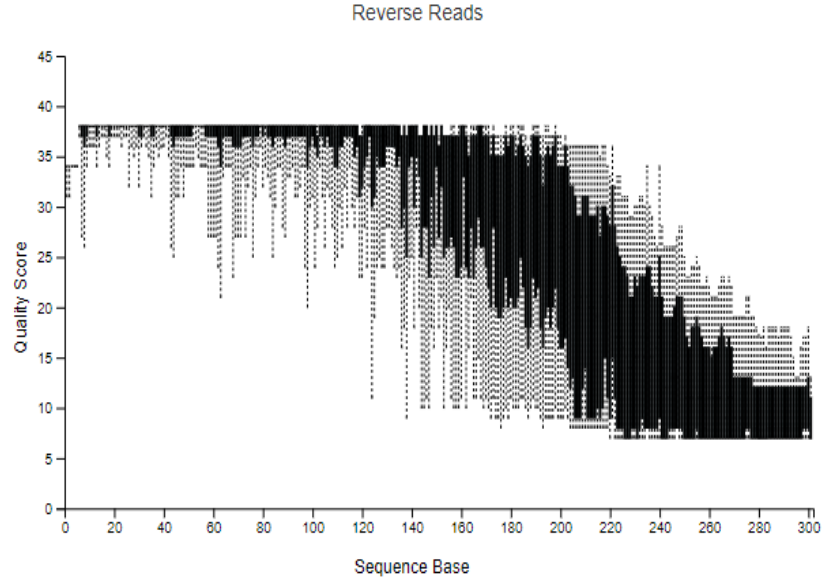


Figure 4.4: Quality Score Graph for Reverse Reads of COPD

The DADA2 tool is further used to denoise and improve the quality of the sequences. As shown in the quality plots for forward reads the quality of reads from 260 onwards have low Phred scores less than 20 so we truncated sequences from 260 onwards. Similarly, the Phred scores for forward reads are low from 220 sequence bases. Hence the reads are truncated from the 220 sequence bases onwards and only good quality reads are preserved for future use. The tool outputs three files; one showing the full statistical information of the reads, representative sequences from the input reads and a feature table. The statistical aspects of the COPD read are displayed in Table 4.2. Representative sequences or ASVs are shown with feature Id and sequence length. The feature table consists of frequency per sample and features, interactive sample details with sample depth and metadata and finally the complete feature details table. The summary of frequency estimates are shown in Table ??.

Table 4.2. Statistical Aspects after denoising of COPD Dataset

Sample-id	Input	Filtered	Passed Filter %	Denoised	Merged	Merged Filters %	Non-Chimeric	Non-Chimeric %
Sample1	210501	104581	49.68	104097	101588	48.26	16996	8.07
Sample10	171481	89081	51.95	88609	87238	50.87	10015	5.84
Sample11	176480	76604	43.41	76245	75332	42.69	17769	10.07
Sample12	215058	116606	54.22	116064	114508	53.25	14470	6.73
Sample13	127802	68706	53.76	68420	67741	53	17426	13.64
Sample14	176401	70257	39.83	69429	67278	38.14	15058	8.54
Sample15	144061	81368	56.48	80336	78077	54.2	17793	12.35
Sample16	104199	57923	55.59	57205	55755	53.51	14755	14.16
Sample17	396779	208566	52.56	205808	192378	48.48	27610	6.96
Sample18	131099	71700	54.69	71230	70223	53.56	14571	11.11

Table 4.3. Summary of COPD Dataset Features

Metric	Sample
Number of Samples	78
Number of Features	3412
Total Frequency	1,527,651

For taxonomic analysis we have used silva-138-99-nb-classifier which is naive Bayes classifier with 99 % sequence similarity used for classification of the sequences in the dataset. This step results in taxonomic file which has data in tab separated values with three columns; feature id, taxon and confidence level for classification. The table for taxonomic classification is shown in Table 4.4. Then the taxonomy file is represented in bar graphs for the better visualization of classification. The bar chart for taxonomy is displayed in Figure 4.5. In COPD samples, Proteobacteria, Firmicutes, Actinobacteriota and Fusobacteriota were the most abundant phylum of bacteria. The abundant families of bacteria are Enterobacteriaceae, Moraxellaceae, Pasteurellaceae, Lactobacillaceae, Micrococcaceae, Neisseriaceae, Streptococcaceae, Actinomycetaceae and Leptotrichiaceae which plays significant impact in disorder. At genus level *Moraxella*, *Neisseria*, *Rothia*, *Veionella*, *Streptococcus*, *Leptotrichia* and *Actinomyces* are abundant in COPD samples.

Table 4.4. Taxonomy Classification on the basis of Feature Id's

Feature Id	Taxon	Confidence
000558507913a411464774ea3a394f3f	d__Bacteria; p__Bacteroidota; c__Bacteroidia; o__Bacteroidales; f__Prevotellaceae; g__Prevotella; s__Prevotella_denticola	0.9577189301849474
000babd2fd46d402a87ee8bebdaa07c2	d__Bacteria; p__Actinobacteriota; c__Actinobacteria; o__Micrococcales; f__Cellulomonadaceae; g__Tropheryma; s__Tropheryma_whipplei	0.9696374719813575
001cad1fad16168c4ae7002826de7efb	d__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Streptococcus	0.9999945001778408
002153336e8723d77d433cd4a25218f5	d__Bacteria; p__Bacteroidota; c__Bacteroidia; o__Flavobacteriales; f__Flavobacteriaceae; g__Capnocytophaga	0.999999620927301
003741c38ce0f69660c7c9e4d1d9d546	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pasteurellales; f__Pasteurellaceae; g__Haemophilus	0.9455821666772276
004ac61045280c32041b6052ca809b6b	d__Bacteria; p__Bacteroidota; c__Bacteroidia; o__Bacteroidales; f__Prevotellaceae; g__Prevotella	0.99988484507001
0066eeef3047a323b23ff2cddb9ab658	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Burkholderiales; f__Neisseriaceae; g__Neisseria; s__Neisseria_perflava	0.7878777440617253
006ad60c9804a581de76cb7d16b308bc	d__Bacteria; p__Firmicutes; c__Clostridia; o__Lachnospirales; f__Lachnospiraceae; g__Lachnospiraceae_NK3A20_group; s__uncultured_Lachnospiraceae	0.9871487963580632
006c98bce5e348c89a6722254f0aadab	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Burkholderiales; f__Neisseriaceae; g__Neisseria	0.999900266948179
00ac28004cee15b405952945d6bbf47	d__Bacteria; p__Firmicutes; c__Clostridia; o__Peptococcales; f__Peptococcaceae; g__Peptococcus; s__uncultured_bacterium	0.9750943690916912

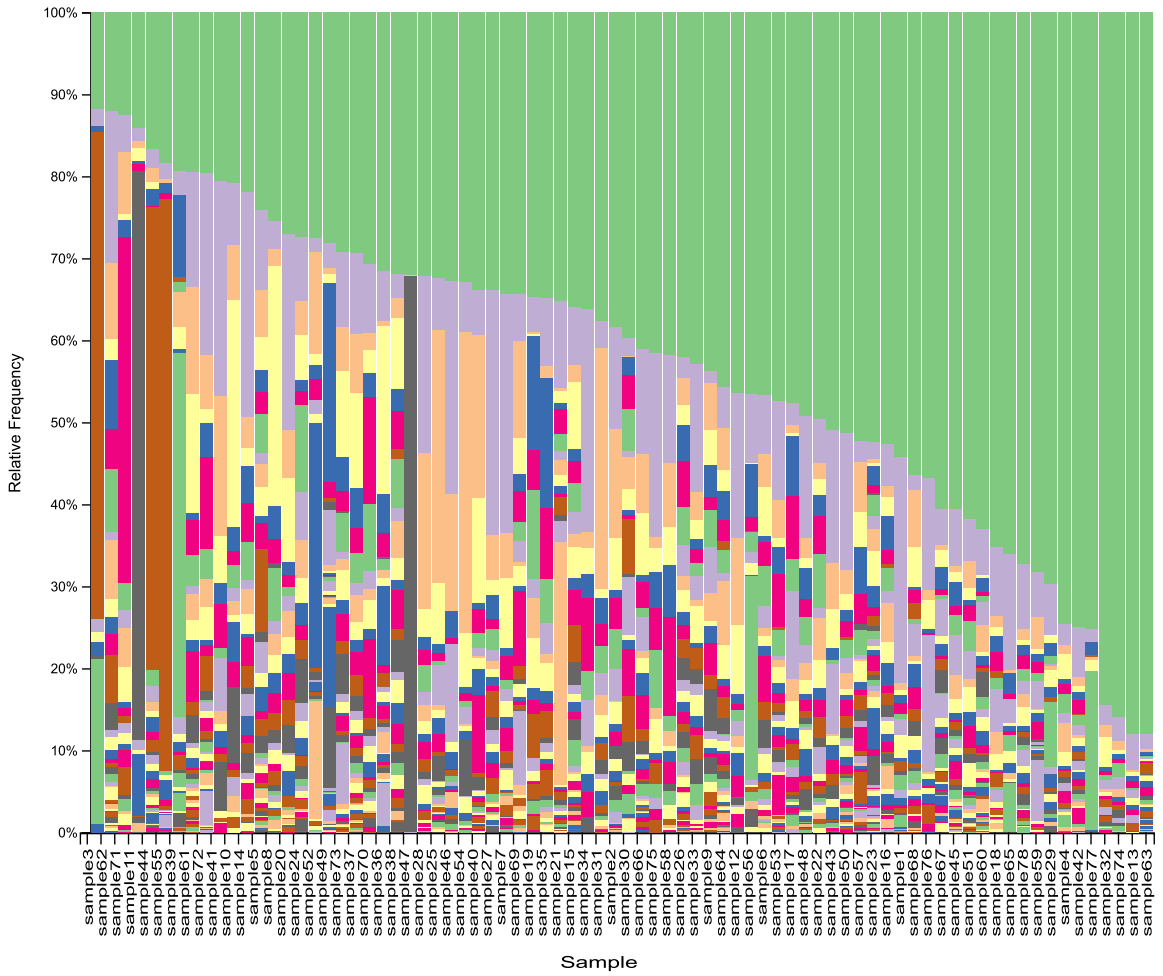


Figure 4.5: Bar Graph for Taxonomy Classification of COPD

4.1.2 COVID-19 Amplicon Results

The dataset used for COVID-19 analysis is PRJNA687143. It contains the 52 samples. The raw data files are downloaded from the ENA repository. As initial step is creation of path files and metadata files for both normal and diseased samples. All steps are performed separately for both control and infected patients samples. The path file and metadata file of COVID-positive samples are input to the QIIME2 environment to generate qza format. After that demultiplexed the reads, the summary of demultiplexing is shown in Table 4.5 and frequency histograms for forward and reverse reads are in Figure 4.6 and Figure 4.7. The Phred quality scores for both forward and reverse reads are shown in Figure 4.8 and Figure 4.9.

Table 4.5. Summary of Demultiplexed Sequences of COVID-19 Dataset

Statistics	Forward Reads	Reverse Reads
Minimum	1468	1468
Median	28532.0	28532.0
Mean	29340.166667	29340.166667
Maximum	86613	86613
Total	704164	704164

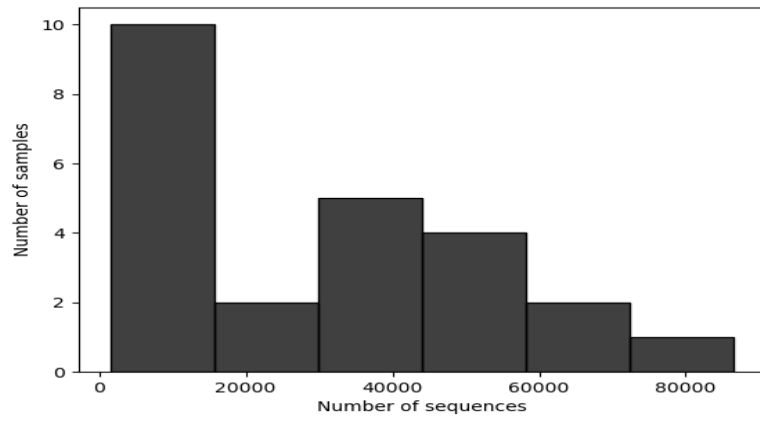


Figure 4.6: Forward Reads Frequency Diagram for COVID-19

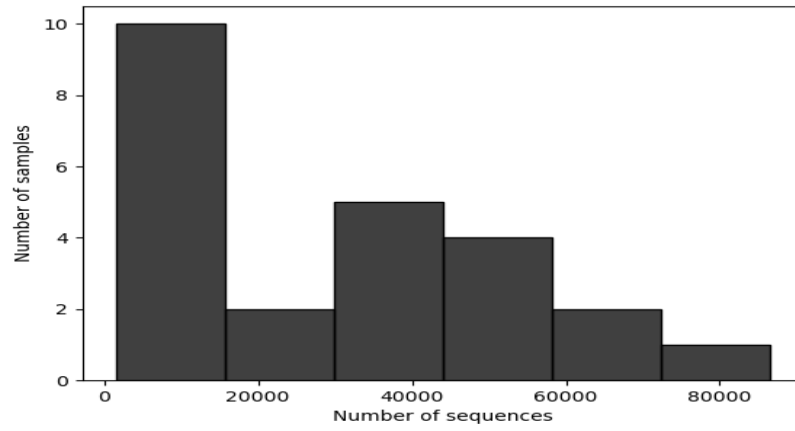


Figure 4.7: Reverse Reads Frequency Diagram for COVID-19

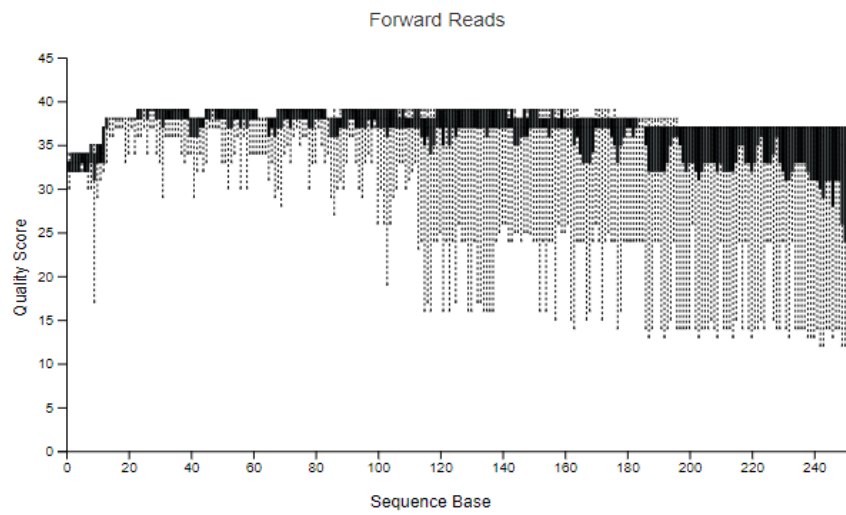


Figure 4.8: Quality Score Graph for Forward Reads of COVID-19

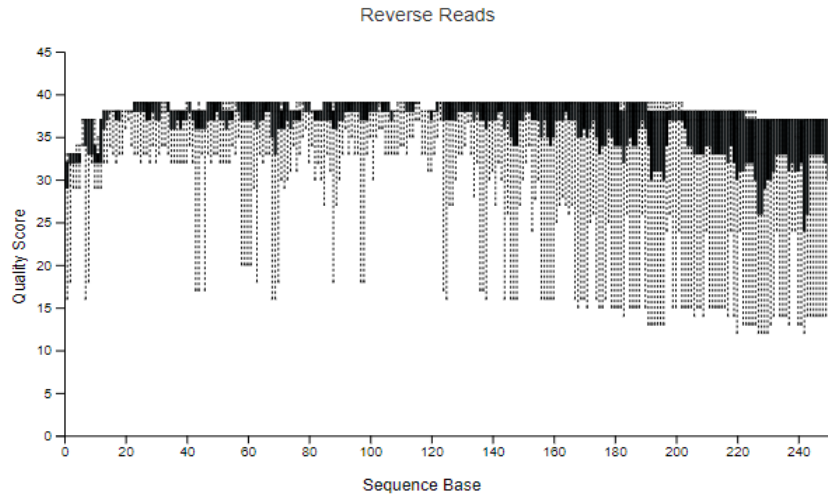


Figure 4.9: Quality Score Graph for Reverse Reads of COVID-19

The DADA2 tool is used to further improve the quality and low quality reads are truncated from 240 onwards for both forward and reverse reads. As the remaining reads are of good quality they are used for further downstream analysis. The tool outputs the statistical summary of reads, representative sequences and overall the complete feature table. Statistical summary of reads are shown in Table 4.6 and Table 4.7 represents the summary of feature table frequencies.

Table 4.6. Statistical Aspects after denoising of COVID-19 Dataset

Sample-id	Input	Filtered	Passed Filter%	Denoised	Merged	Merged Filters%	Non-chimeric	Non-chimeric%
Sample1	210501	104581	49.68	104097	101588	48.26	16996	8.07
Sample10	171481	89081	51.95	87238	87238	50.87	10015	5.84
Sample11	176480	76604	43.41	75332	75332	42.69	17769	10.07
Sample12	215058	116606	54.22	114508	114508	53.25	14470	6.73
Sample13	127802	68706	53.76	67741	67741	53	17426	13.64
Sample14	176401	70257	39.83	67278	67278	38.14	15068	8.54
Sample15	144061	81368	56.48	78077	78077	54.2	17793	12.35
Sample16	104199	57923	55.59	55755	55755	53.51	14755	14.16
Sample17	396779	208566	52.56	1923787	192378	48.48	27610	6.96
Sample18	131099	71700	54.69	70223	70223	53.56	14571	11.11
Sample19	148286	66065	44.55	63480	63480	42.81	11092	7.84
Sample2	117782	54909	46.62	53499	53499	45.42	12336	10.47

Table 4.7. Summary of COPD Dataset Features

Metric	Sample
Number of Samples	24
Number of Features	898
Total Frequency	143,115

After denoising, we input the representative sequences and silva-138-99-nb-classifier that is used for the taxonomic assignment of reads. In the output we get a taxonomy file with feature id, taxon with their confidence scores. The format of taxonomy file is shared in Table 4.8. As we have the taxonomic file we represent the file in the form of bar charts. The bar chart for these particular diseased samples is displayed in the Figure 4.10. The same procedure is repeated with the COVID-negative samples. Proteobacteria, Firmicutes, Bacteroidota, Actinobacteriota, Campilobacterota and Spirochaetota are the abundant phylum in COVID positive patients. The most abundant families of bacteria are Pseudomonadaceae, Enterobacteriaceae, Pasteurellaceae, Staphylococcaceae, Streptococcaceae, Mycoplasmataceae and Flavobacteriaceae. Pseudomonas, Staphylococcus, Streptococcus, Escherichia-Shigella, Aggregatibacte are abundant at genus level. Some species e.g. *Streptococcus-constellatus*, *Schaalia-odontolytica*, *Capnocytophaga-sputigena*, *Neisseria-perflava* and *Capnocytophaga-gingivalis* etc are abundant in diseased samples. The COVID-negative samples has Firmicutes, Actinobacteriota and Pro bacteria phylum with Veillonellaceae, Prevotellaceae, Pasteurellaceae etc are abundant families.

Table 4.8. Taxonomy Classification on the basis of Feature Id's

Feature Id	Taxon	Confidence
[HTML]F5F5F5003a2f482726ce8e30005b3b3a979a64	d__Bacteria; p__Firmicutes; c__Clostridia; o__Peptostreptococcales-Tissierellales;	0.999892416086272
00a48663b44d647e7f1d8b467d623af3	f__Anaerovoracaceae; g__[Eubacterium]_nodatum_group	0.967432237041691
0160832fb548d29947148023a9f49c9a	d__Bacteria; p__Firmicutes; c__Bacilli; o__Staphylococcales;	
017083ec740b09a88c8ba26f27165231	f__Staphylococcaceae; g__Staphylococcus	
[HTML]F5F5F50194c42606eee82c7cd761b6584dac8c	d__Bacteria; p__Bacteroidota; c__Bacteroidia; o__Bacteroidales;	0.7117656406489767
01b3feea3b633a917a6ea64644ed157	f__Prevotellaceae; g__Prevotella; s__unidentified_eubacterium	0.8664528085764981
01c6c944f7f3bc2ef8127c1bce977565	d__Bacteria; p__Firmicutes; c__Bacilli; o__Erysipelotrichales;	
02a20a36215bb8e8be40e78e312bb5f	f__Erysipelotrichaceae; g__Solobacterium; s__uncultured_organism	
0330c8e1f7335450b134df6ad6525e60	d__Bacteria; p__Actinobacteriota; c__Actinobacteria;	0.9999825150006235
[HTML]F5F5F50333515843bfc579cd79e250682e6926	o__Actinomycetales; f__Actinomycetaceae; g__Actinomyces	0.8753038746220265
	d__Bacteria; p__Firmicutes; c__Bacilli; o__Mycoplasmatales;	
	f__Mycoplasmataceae; g__Mycoplasma; s__Metamycoplasma_faucium	
	d__Bacteria; p__Bacteroidota; c__Bacteroidia;	0.9999802749425311
	o__Bacteroidales; f__Prevotellaceae; g__Prevotella	
	d__Bacteria; p__Firmicutes; c__Clostridia; o__Lachnospirales;	0.7945943492614838
	f__Lachnospiraceae; g__Lachnospiraceae_UCG-004	
	d__Bacteria; p__Firmicutes; c__Clostridia;	0.9997963503878889
	o__Peptostreptococcales-Tissierellales; f__Peptostreptococcaceae	
	d__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales;	0.999993560565421
	f__Streptococcaceae; g__Streptococcus	

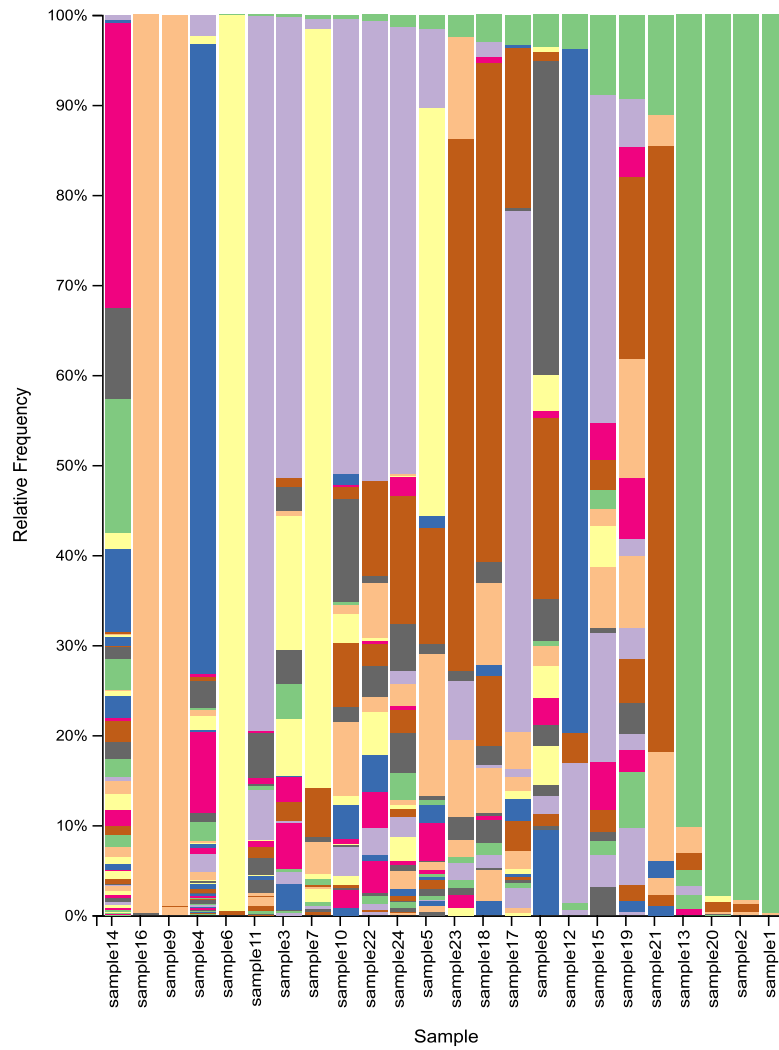


Figure 4.10: Bar Graph for Taxonomy Classification of COVID-19

4.2 Shotgun Analysis

Shotgun analysis is the non-targeted approach used to sequence the whole genome of all the micro-organisms present in a specified environment. The analysis consists of data retrieval, input, pre-processing or QC, Metaphlan analysis for taxonomical classification and finding their relative abundances.

4.2.1 In-Silico Shotgun Analysis of COPD

The dataset used for in-silico analysis for COPD is PRJEB9034. It consists of 18 samples; 8 are diseased and 10 are normal. The steps for in-silico analysis involve; input to the Galaxy, pre-processing, profiling, and relative abundances of microbial species. The Galaxy is a reliable platform for the analysis of high-throughput data. The raw reads are searched on the ENA database, and the reads are uploaded on the Galaxy platform using the Galaxy link on the ENA browser. After uploading the raw reads, the quality of the raw reads is checked using FastQC. The FastQC report has ten quality properties enlisting basic statistics, per base sequence quality, per tile sequence-quality, per sequence quality scores, per base GC content and N content, length distribution of reads, duplication level, over-represented sequences, and adaptor content.

All the diseased sequences have the adaptor content, per base sequence quality, and per sequence quality scores are also not good. To improve quality, we used the fastp tool. After utilizing fast, the quality of reads is improved and we used the reads for downstream analysis. The per base sequence quality is displayed in Figure 4.11 and the adaptor content is in Figure 4.12. After Fastp implementation, the quality score is improved and adaptor content is removed as shown in Figure 4.13 and Figure 4.14. The next step is utilizing Metaphlan4 primarily for taxonomic classification and their relative abundances. Additionally, before profiling microorganisms, Metaphlan4 does the aligning of reads. The abundance of each microbial species is output in a tab-separated value (CSV). Similarly, the analysis steps are repeated for the normal samples. The quality of control samples is also low to improve quality Fastp is used. The abundant phylum in diseased are Firmicutes, Proteobacteria and Actinobacteria. The Streptococcaceae, Veillonellaceae, Staphylococcaceae, Prevotellaceae, Actinomycetaceae, Moraxellaceae and Pasteurellaceae . At genus level *Streptococcus*, *Herbaspirillum*, *Actinomyces*, *Veillonella* etc and at specie level *Staphylococcus-aureus*, *Streptococcus-salivarius*, *Actinomyces-graevenitzii*, *Streptococcus-parasanguinis*, *Prevotella-histicola* and *Haemophilus-influenzae* etc are among the prominent species.

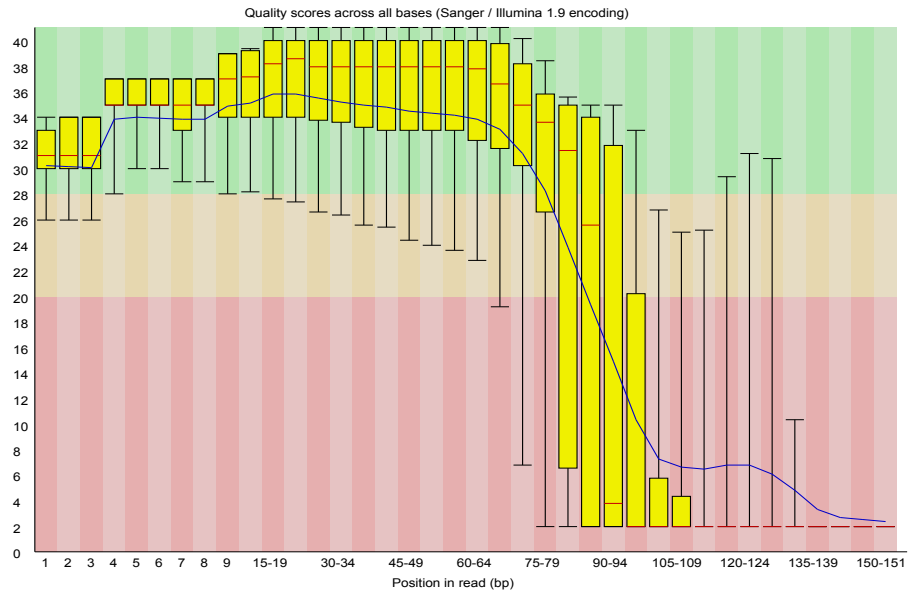


Figure 4.11: Graph representing the Phred Score per base for COPD Before Fastp

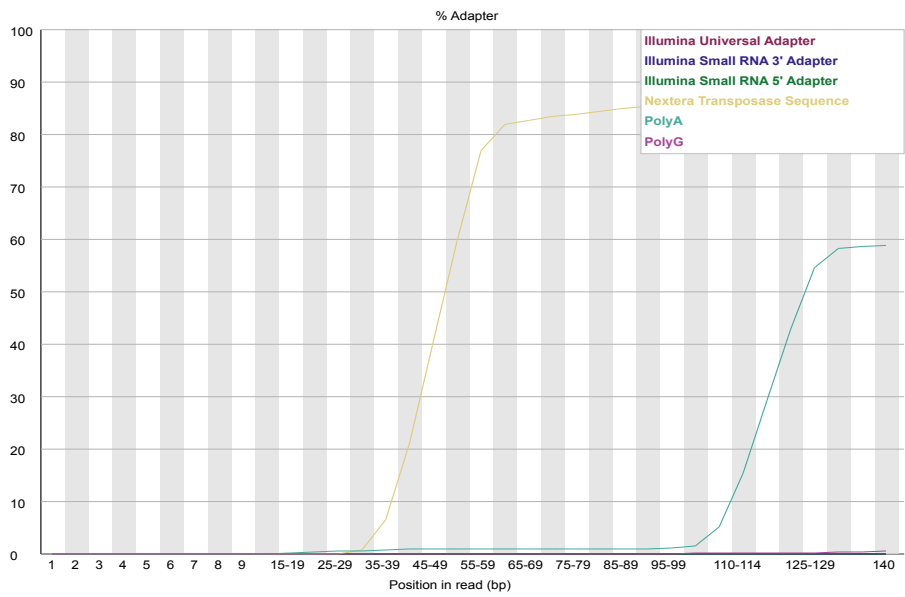


Figure 4.12: Graph representing the Adaptor Content in COPD Dataset Before Fastp

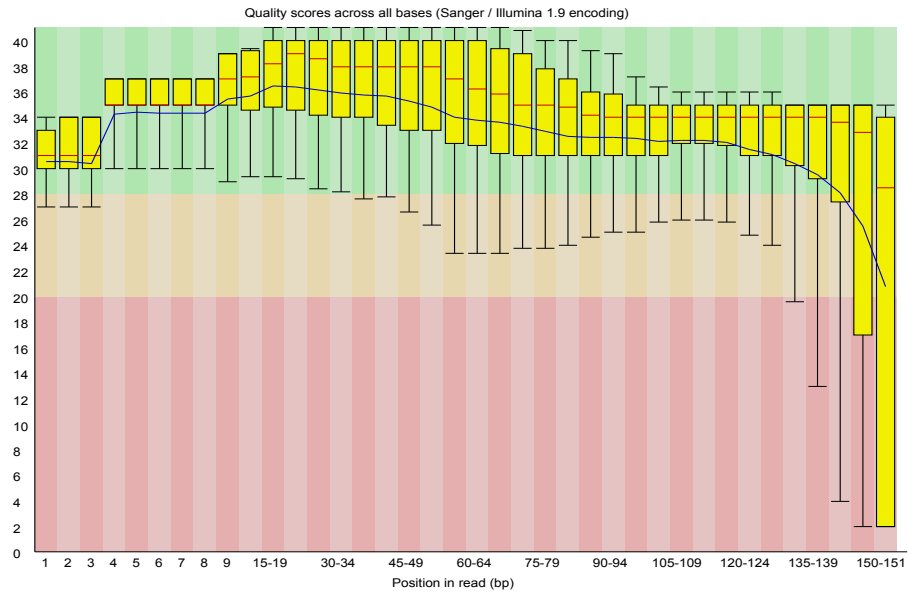


Figure 4.13: Graph representing the Phred Score per base for COPD After Fastp

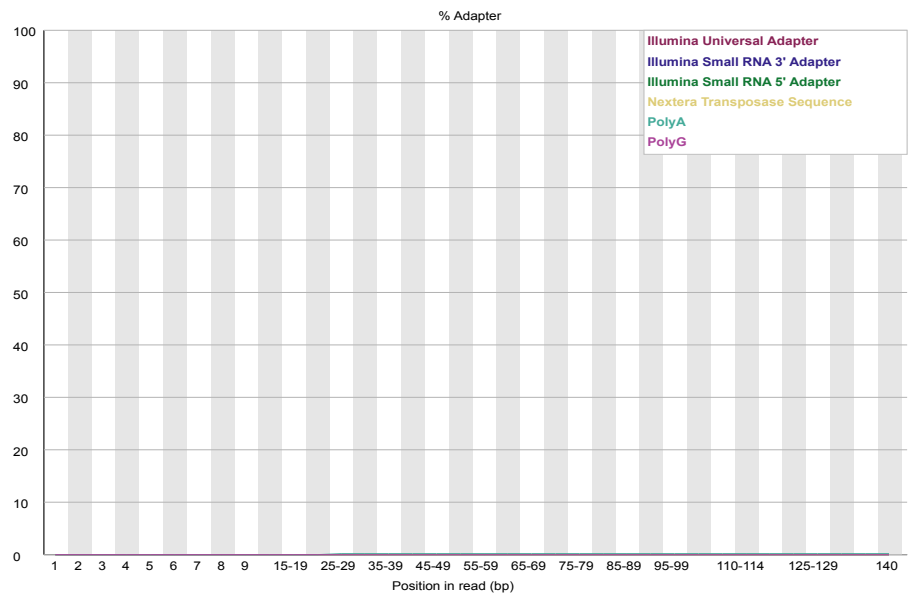


Figure 4.14: Graph representing the Adaptor Content in COPD Dataset After Fastp

4.2.2 In-Silico Shotgun Analysis of COVID-19

In-silico analysis of COVID-19 dataset was performed on Galaxy Platform. The ENA id for the dataset is PRJNA743981. It has 99 samples including 8 normal samples and

remaining samples are diseased. The fastq files of reads are uploaded on the Galaxy using the ENA browser. The quality of samples are checked using FastQC. The FastQC report is generated with the ten parameters including the basic statistical report. The quality of nearly all samples are good excluding eight diseased samples. Low-quality reads with scores below 20 are enhanced using the Fastp tool which uses trimming and other parameters for improvement of sequences. Taxonomic profiling is performed using Metaphlan4 which performs aligning of reads before profiling.

COVID-19 patients has also same phylums enlisting Firmicutes, Actinobacteria and Proteobacteria. The distinguishing families of bacteria present are Propionibacteriaceae, Malasseziaceae, Carnobacteriaceae, Staphylococcaceae, Corynebacteriaceae, Propionibacteriaceae, Moraxellaceae, Flavobacteriaceae, Streptococcaceae and Yersiniaceae. *Corynebacteriu-propinquum*, *Cutibacterium-acnes*, *Streptococcus-pneumoniae*, *Malassezia-restricta* and *Dolosigranulum-pigrum* are the prominent species present in diseased samples. All these families, genus and species are crucial in the infection development.

4.2.3 In-Silico Shotgun Analysis of Lung Cancer

The dataset used for analysis is PRJNA714488. It contains forty-seven samples. Among them 32 samples are diseased and 15 are non-cancerous. Broncho Alveolar Lavage (BAL) samples are used for the analysis. As the analysis initiated with the input of raw reads on the Galaxy platform using the ENA browser links of fastq files. After uploading the raw data files, an important step is their quality check. FastQC report shows that all the parameters and overall the quality of sequences are good. Hence we don't need Fastp or any other tool for quality improvement. As the quality is good we performed Metaphlan4 analysis for micro-organisms species profiling and their relative abundances are measured. The quality of base is measured through Phred Scores. Below 20 are the low quality reads and reads with high scores than 20 are considered as good quality reads. The Bacteroidetes, Proteobacteria, Firmicutes, Actinobacteria and Fusobacteria are abundant phylums. Streptococcaceae, Prevotellaceae, Selenomonadaceae, Porphyromonadaceae, Neisseriaceae, Halomonadaceae, Vibrionaceae and Thermoactinomycetaceae are families found with high percentage. The species *Halomonas-sp-LBP4*, *Vibrio-alginolyticus*, *Selenomonas*, *Campylobacter-jejuni*, *Streptococcus-pneumoniae*, *Prevotella-melaninogenica*, *Haemophilus-influenzae*, and *Escherichia-coli* are among the prominent ones and are critical in the disease.

4.2.4 In-Silico Shotgun Analysis of TB

The dataset we used for TB is PRJNA655567. It has sixty-one samples comprising both diseased and controls. The analysis starts with the basic step of uploading the reads, thirty-four diseased samples and twenty-seven are the control. The sequence files are in

fastq files and from ENA they are imported to the Galaxy website. The basic step is to analyze the quality of sequences, and for that, FastQC is used that generate a detailed report for the quality of reads. The report gives us information on various parameters e.g. quality score of reads per base, GC content, adaptor content, reads length distribution and duplication levels, etc. Quality of the reads in the datasets is high quality and can be used for downstream analysis. A further step is taxonomical profiling of species which is done using Metaphlan4 the more recent version of Metaphlan. Metaphlan4 does the reference-based assignment of microbes. The fastq files are input into Metaphlan4. Generally, it outputs the relative abundances table, BIOM file, Bowtie2, and SAM files. Bowtie2 and SAM files are the alignment files as Metaphlan4 has a plugin option for the alignment of sequences and then species profiling takes place. Bacteroidetes, Proteobacteria, Firmicutes and Candidatus-Saccharibacteria are the dominant phylums. Burkholderiaceae, Neisseriaceae, Peptostreptococcaceae, Streptococcaceae, Prevotellaceae and Porphyromonadaceae are abundant are the among the dominant families. *Neisseria-subflava*, *Prevotella-melaninogenica* , *Streptococcus-mitis*, *Ralstonia-pickettii* and *Porphyromonas-pasteri* are the species present in TB patients.

4.3 Machine Learning

The ML is performed on relative abundance tables resulting from Metaphlan4. Two main classifiers RF and SVM are used for classifying the diseased and normal bacterial species. After Metaphlan4 we have relative abundances, and then we merge all the tables of all diseased datasets COPD, COVID-19, lung cancer, and TB. The details of data used for ML is summarized in Table 4.9.

Table 4.9. Datasets Used for ML

Accession No.	Disease	Classes	No. of Samples
PRJEB9034	COPD	COPD, n	8, 10
PRJNA743981	COVID-19	Cov-2, n	78, 8
PRJNA714488	Lung Cancer	LC, n	32, 15
PRJNA655567	TB	TB, n	34, 27

4.3.1 Random Forest

After merging, the tables are pre-processed and transformed. There are four hundred and ninety features (bacterial species) and two hundred and eleven samples including all diseased and normal. The RF model is employed on all datasets and then on each dataset individually.

4.3.1.1 Chronic Obstructive Pulmonary Disease Dataset

The RF model is build and trained on the bacterial species of COPD dataset. The data is split into 80 and 20 %. The features are filtered based on chi-square statistical technique. The lasso is used for more significant feature selection and then model is trained on these features. The features selected using lasso are represented in the Figure 4.15. After training of model, the accuracy is 0.5 % and F1 score is also 0.45. The confusion matrix is used to evaluate the performance of model. The matrix displayed for COPD is displayed in Figure 4.16.

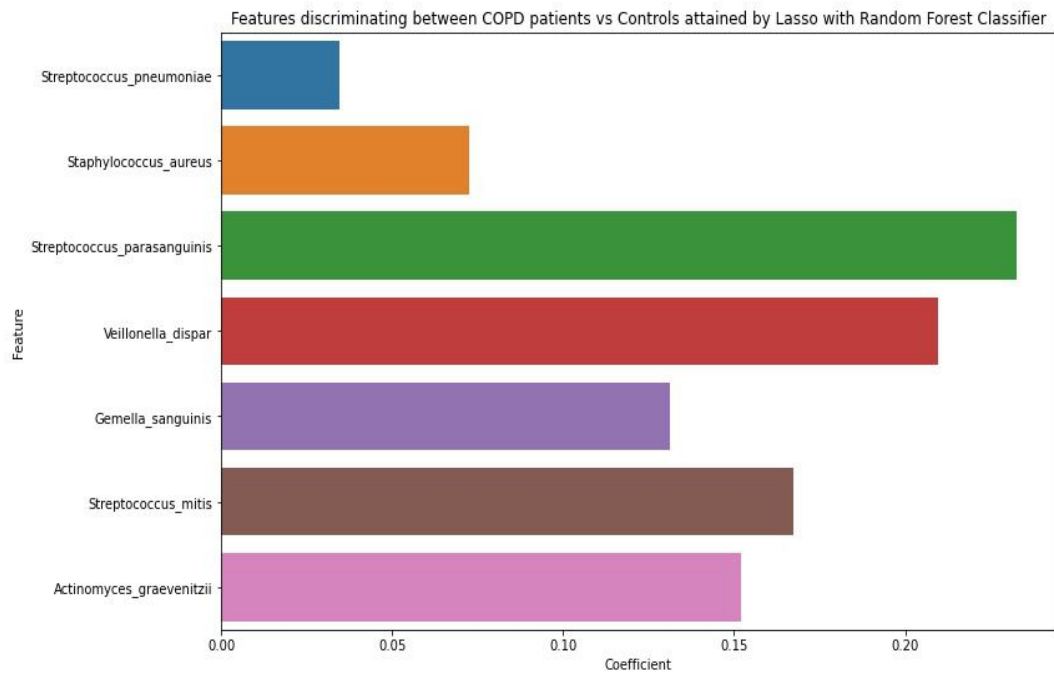


Figure 4.15: Feature Selected Using Lasso for COPD

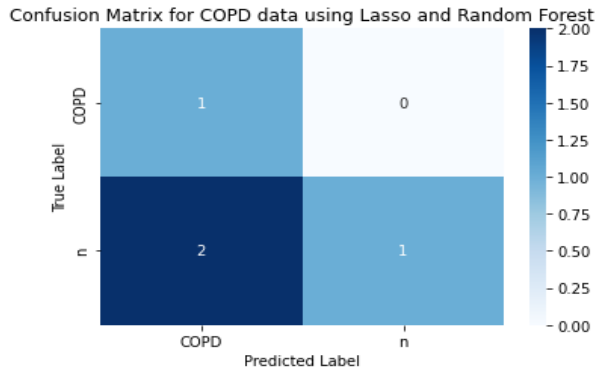


Figure 4.16: Confusion Matrix for COPD through RF

4.3.1.2 COVID-19 Dataset

The RF model is build and trained on the bacterial species of COVID-19 dataset. The features are filtered based on chi-square and lasso for more significant feature selection. After feature selection, training is performed. The features selected using lasso for COVID-19 are represented in the Figure 4.17. The model gives accuracy which is 0.94 % and F1 score is 0.936. Further, model is evaluated using the confusion matrix. The matrix displayed for COVID-19 is displayed in Figure 4.18.

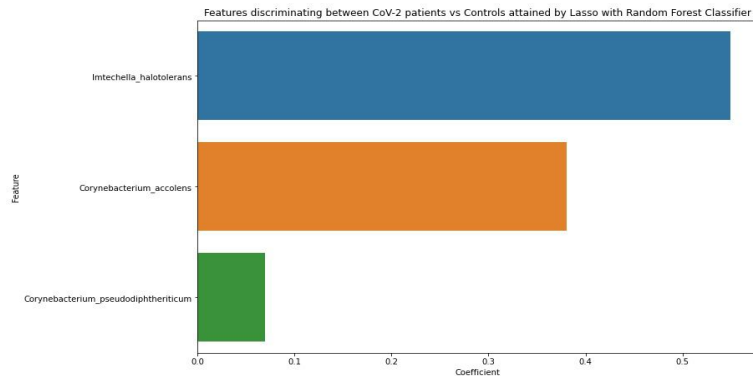


Figure 4.17: Feature Selected Using Lasso for COVID-19

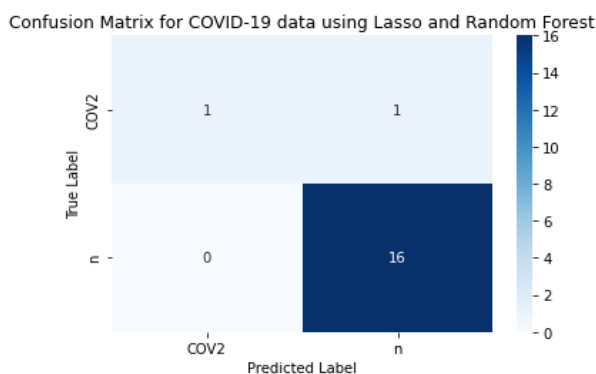


Figure 4.18: Confusion Matrix for COVID-19 through RF

4.3.1.3 Lung Cancer Dataset

The microbial species of lung cancer are used for the development and learning of RF model. Important features are chosen using chi-square technique for the learning of the algorithm. The Figure 4.19 is representing the remarkable features for lung cancer. After model training, the accuracy of model is 0.6 % and F1 score is 0.56. Confusion matrix for the lung cancer is shown in Figure 4.20.

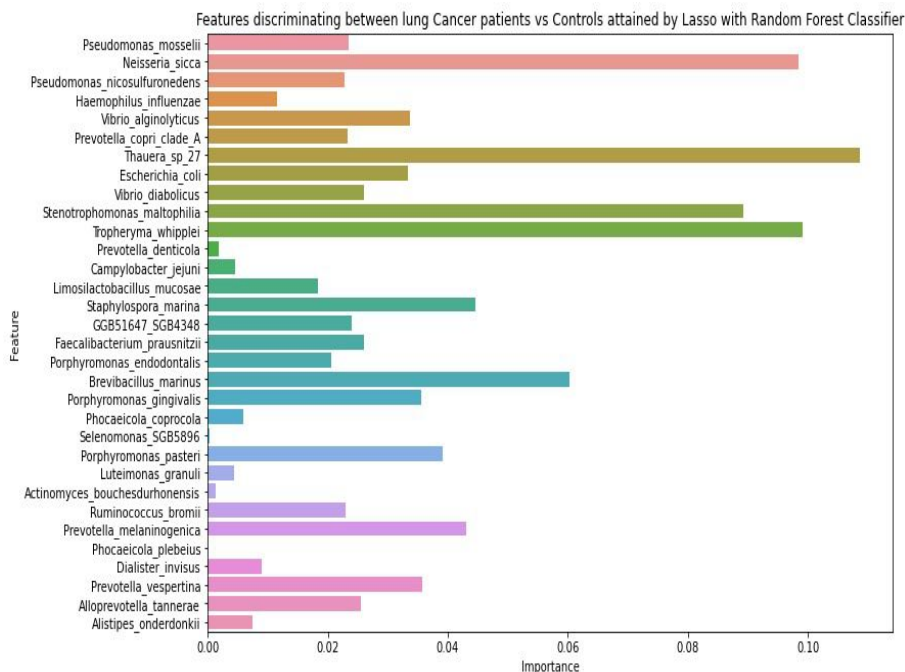


Figure 4.19: Feature Selected Using Lasso for Lung Cancer

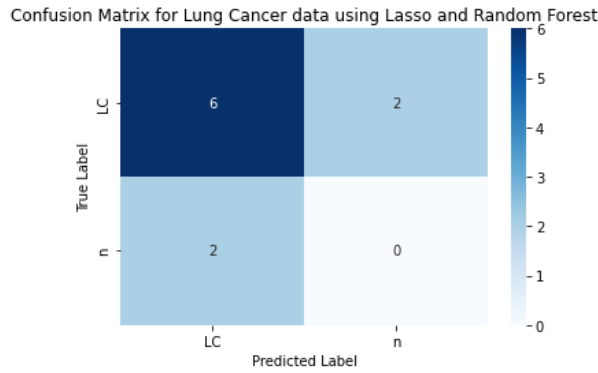


Figure 4.20: Confusion Matrix for Lung Cancer through RF

4.3.1.4 Tuberculosis Dataset

The fourth dataset contains information on microbial species of TB. After splitting of data, the algorithm is trained on the training data. Features that contributes to the development of disease are chosen using lasso and chi-square feature selection methods. Figure 4.21 is representing the discriminating features in disease and healthy conditions. The accuracy of model is 0.66 and F1 score is 0.62. Confusion matrix for TB dataset is shown in Figure 4.22.

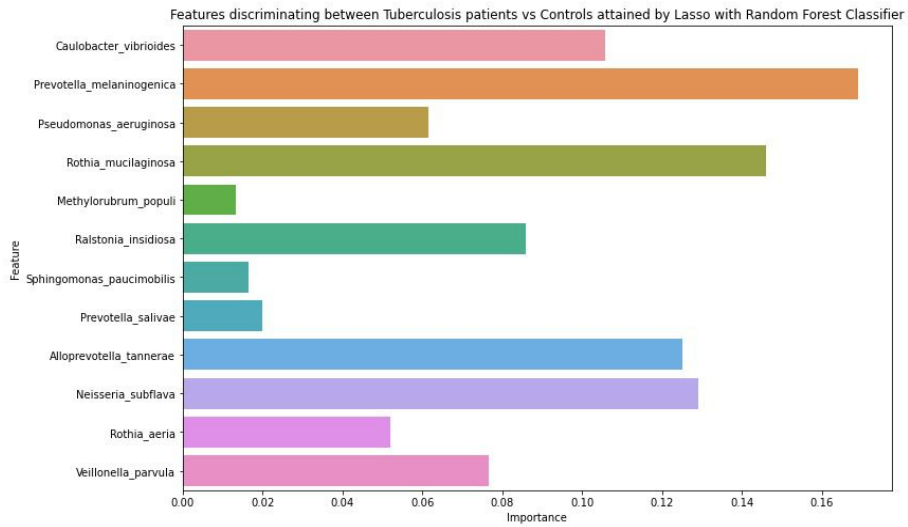


Figure 4.21: Feature Selected Using Lasso for TB

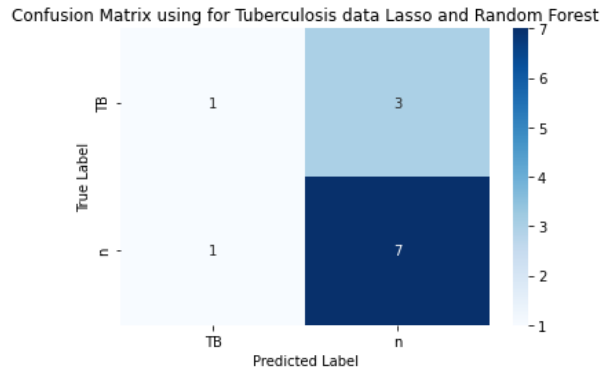


Figure 4.22: Confusion Matrix for TB through RF

4.3.1.5 All Datasets

A RF model is build using all four datasets together with healthy and diseased bacterial species. The features chosen for the model is shown in Figure 4.23 and confusion matrix is displayed in Figure 4.24. Accuracy of the model is 0.51 % and 0.52 is F1 score of the model. The F1 scores and accuracies for all models on all datasets are shown in Table 4.10.

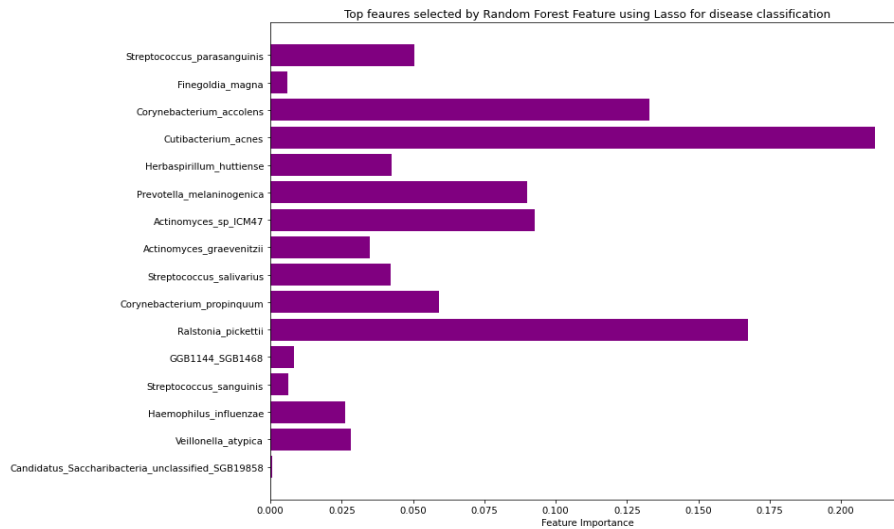


Figure 4.23: Feature Selected Using Lasso for All Datasets

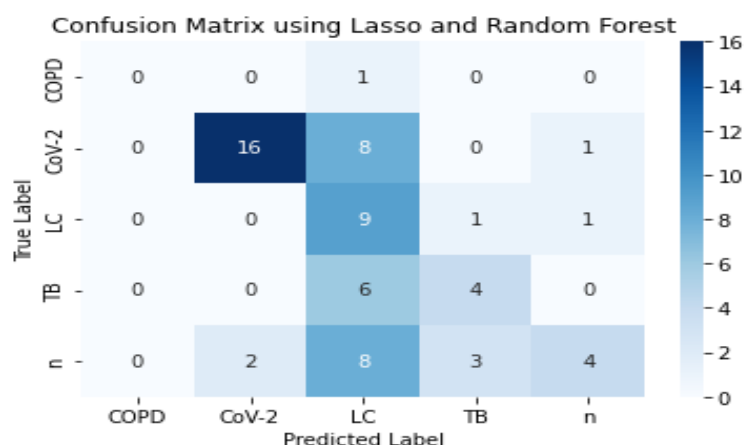


Figure 4.24: Confusion Matrix for All Datasets through RF

Table 4.10. Summary of RF Model Scores

Datasets	Accuracy	F1-Score
COPD	0.5	0.45
COVID-19	0.94	0.93
Lung Cancer	0.6	0.56
TB	0.66	0.62
All	0.51	0.52

4.3.2 Support Vector Machine

SVM is also an important classifier that can differentiate between the two or more classes in the data. It is also a supervised method. Here SVM is used on four datasets of different diseases and classify the bacterial species accordingly. Here four hundred and ninety features are used for the learning of the model.

4.3.2.1 Chronic Obstructive Pulmonary Disease Dataset

The relative abundances of bacteria are input into the ML model. The data is split into ratio of 20 and 80. The bacterial data is used for the training of SVM that can differentiate between the diseased and normal. The features which are bacterial species are used for the training of model and then testing is done to evaluate the performance of the model. The methods used for features selection are lasso and chi-square. The features selected for SVM are displayed in Figure 4.25. The confusion matrix for COPD are shown in Figure 4.26. The accuracy and F1 score for model is 0.45.

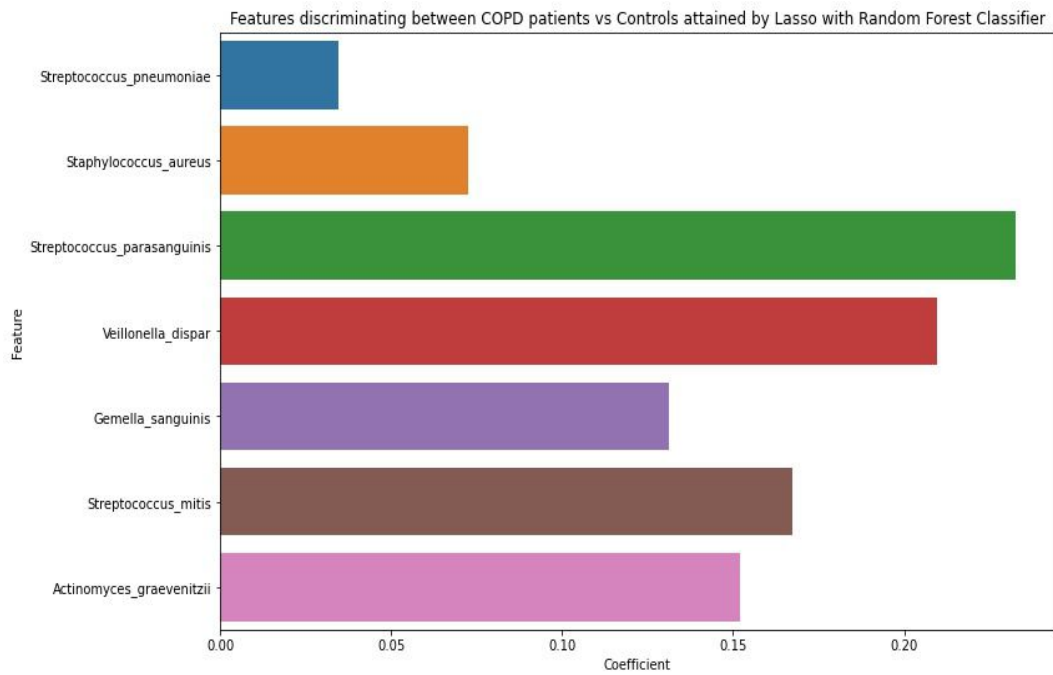


Figure 4.25: Feature Selected Using Lasso for COPD

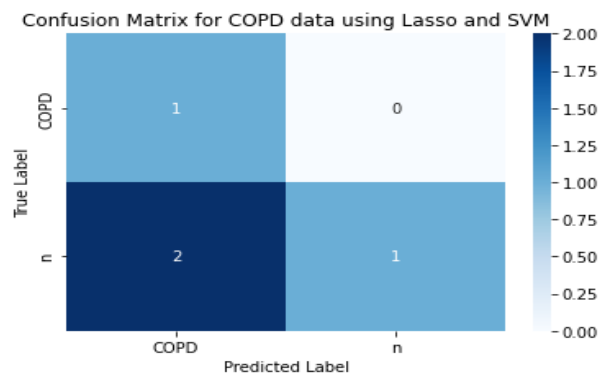


Figure 4.26: Confusion Matrix for COPD through SVM

4.3.2.2 COVID-19 Dataset

The model is develop only for COVID-19 disease and normal microbial species. The chosen microbial species are using lasso and shown in Figure 4.27. The confusion matrix is shown in Figure 4.28. The accuracy measured for the model is 0.88 % and calculated F1 score is 0.83.

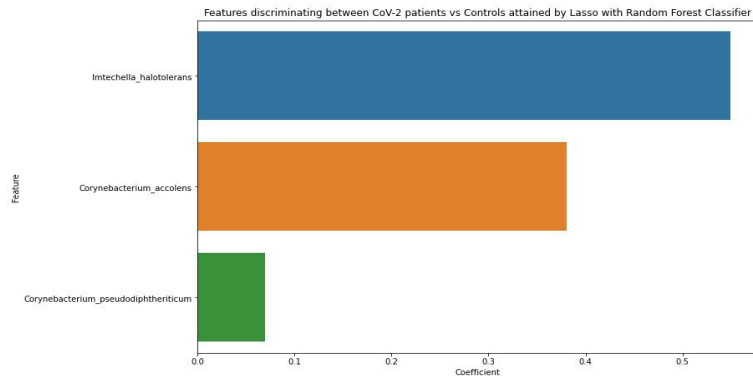


Figure 4.27: Feature Selected Using Lasso for COVID-19

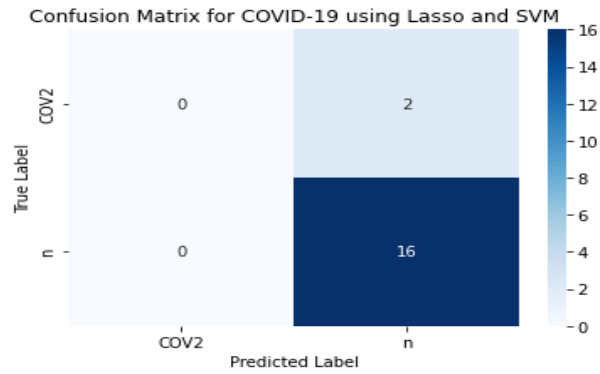


Figure 4.28: Confusion Matrix for COVID-19 through SVM

4.3.2.3 Lung Cancer Dataset

The SVM for lung cancer is build to classify the micro-organisms present in disease and control conditions. The chosen species are displayed in Figure 4.29 and confusion matrix is displayed in Figure 4.30. The model gives the accuracy 0.6 % and F1 score is 0.56.

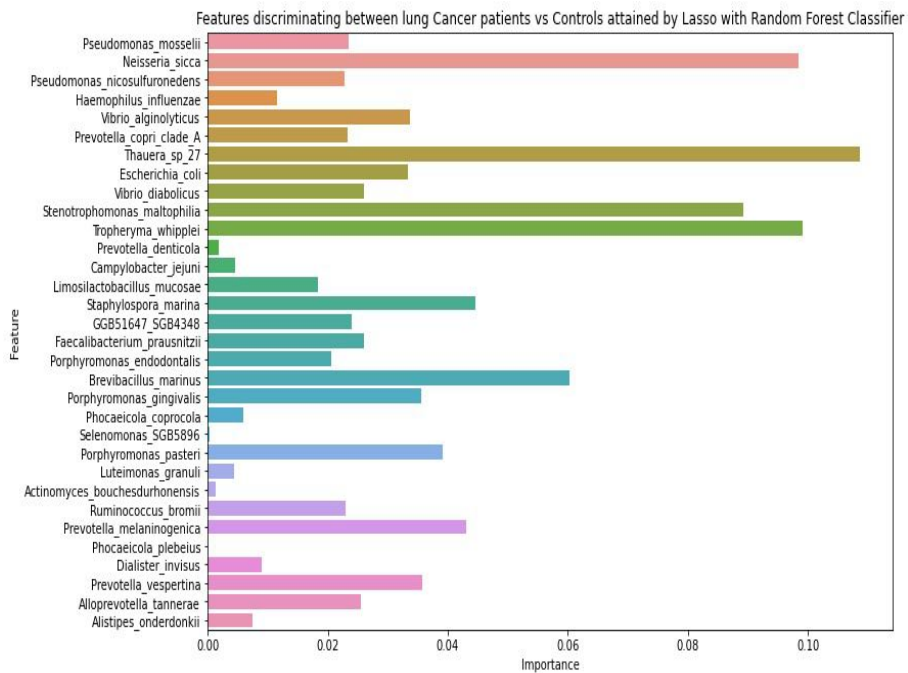


Figure 4.29: Feature Selected Using Lasso for Lung Cancer

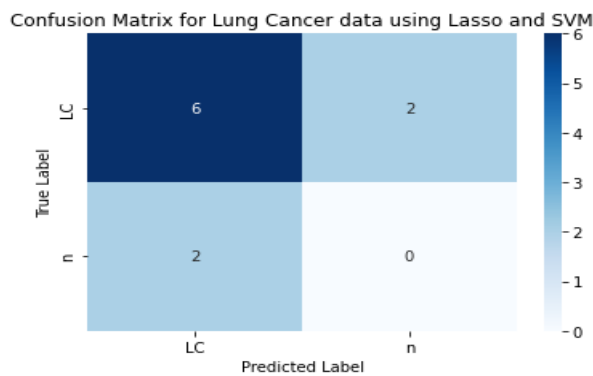


Figure 4.30: Confusion Matrix for Lung Cancer through RF

4.3.2.4 Tuberculosis Dataset

The SVM for TB is build on bacteria favoured by the chi-square and lasso feature selection techniques. Model is 0.75 % accurately classifying the infectious and control bacteria with 0.69 F1 score. The lasso discriminating features are displayed in Figure 4.31 and confusion matrix in Figure 4.32.

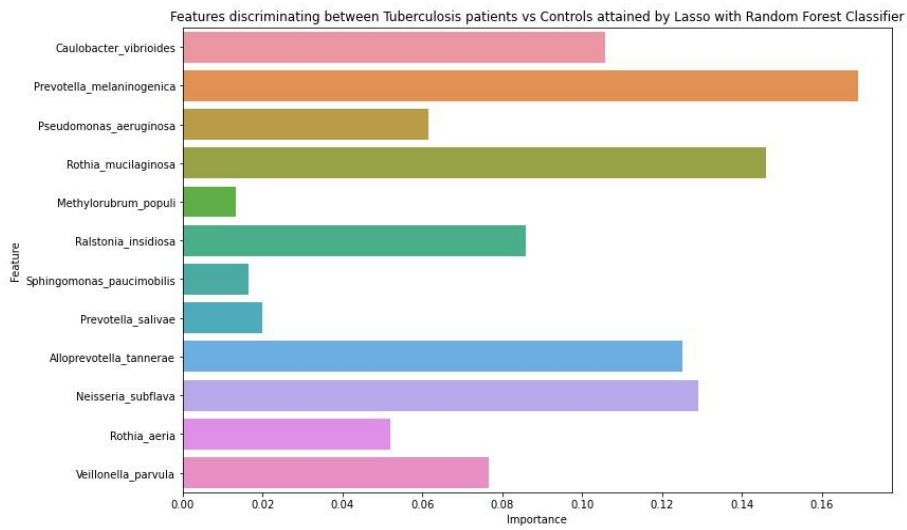


Figure 4.31: Feature Selected Using Lasso for TB

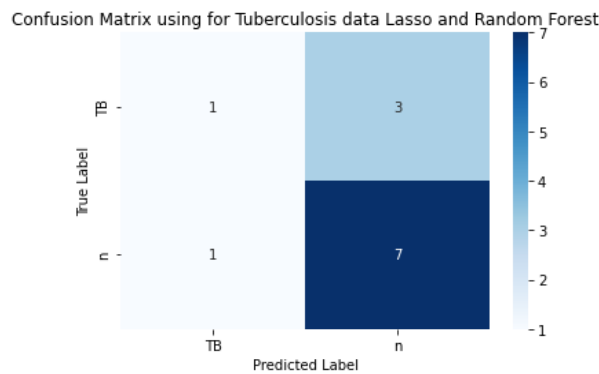


Figure 4.32: Confusion Matrix for TB through SVM

4.3.2.5 All Datasets

A SVM model with all combined datasets is developed that can differentiate between healthy and all diseased classes. To develop a model on the most significant data, chi-square and lasso models are used. Most notable micro-organisms used for model are shown in Figure 4.33. Model is tested with test data to check its performance. The algorithm is 0.54% accurately classifying the bacteria with F1 score of 0.46. Confusion matrix is used for the assessment of the SVM. Figure 4.34 is displaying the confusion matrix for all datasets generated through SVM. The summary of all models scores and accuracies are shown in Table 4.11.

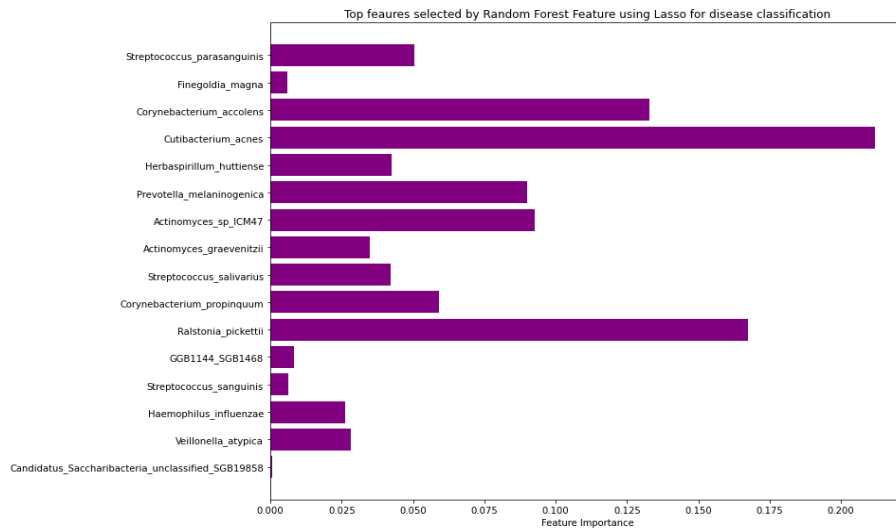


Figure 4.33: Feature Selected Using Lasso for All Datasets

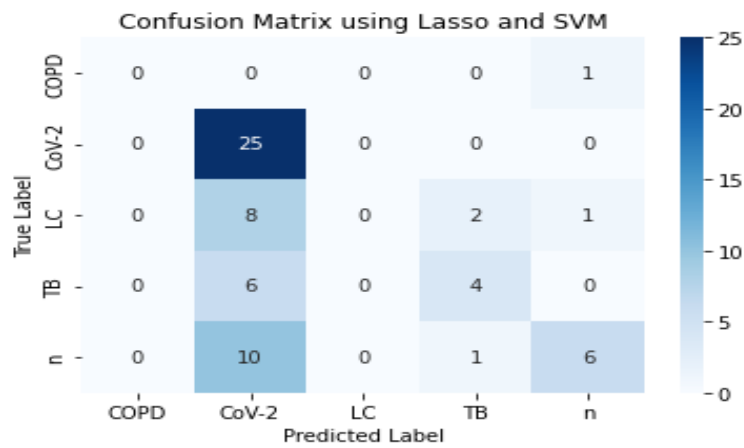


Figure 4.34: Confusion Matrix for All Datasets through SVM

Table 4.11. Summary of SVM Model Scores

Datasets	Accuracy	F1-Score
COPD	0.5	0.45
COVID-19	0.88	0.83
Lung Cancer	0.6	0.56
TB	0.75	0.69
All	0.54	0.46

Chapter 5

DISCUSSION

Respiratory diseases are among the top concerns due to their morbidity and mortality rate. Primarily they affect the lung functions and micro-organisms present in respiratory track and micro-organisms are among the major causes of development and progression of disease. Immediate attention is required for comprehensive understanding of micro-flora of pulmonary tract primarily the lungs. Novelty and enhancement is needed for the determination of pulmonary disorders to reduce the burdens of the chronic respiratory diseases.

The primary purpose of this study is to understand the microbiome of different parts of the pulmonary tract primarily lungs and determine their roles and relative abundances in the progression of different diseases using NGS technique. The diseases which are focused during this study are COPD, COVID-19, Lung Cancer and TB. Two different metagenomics techniques e.g. targeted (amplicon) and non-targeted (shotgun) are employed for deep understanding of the bacterial patterns in the diseases. Moreover, different bacterial species function in the respiratory disorders are evaluated using the comprehensive literature. Beside this, different ML model e.g. RF and SVM are build to differentiate between the bacterial species presence in the above mentioned diseases.

Six secondary datasets are used for metagenomics, two amplicon datasets and four shotgun datasets. In-silico analysis for taxonomical classification of amplicon was carried out using silva 138.1 database and shotgun analysis was performed using Metaphlan4. For detailed exploration and profiling at the maximum level ,shotgun analysis is preferred. Therefore, for specie and further profiling of microbiome, shotgun datasets were used. Originally, the 16S rRNA datasets provides insight on the phylum, families and some genera of bacteria. Furthermore, the shotgun provides valuable insight at the specie and some strains of bacteria.

Metagenomics sequencing is extensively used for the determination and understanding the functions of microbiome in different body sites of humans. Respiratory microflora also seek much attention due to their impact in different diseases states. The

abundant phylum in COPD in literature were Bacteroidetes, Proteobacteria and Firmicutes. *Selenomonas*, *Actinomyces*, *Corynebacterium* and *Actinobacillus* are among the most classified bacteria during profiling [37]. Similarly, metagenomics sequencing was also utilized to understand the composition of microflora of corona virus. The sequencing gives us the important phylum and families of bacteria that are Cyanobacteria, Firmicutes, Bacteroidetes and Actinobacteria while *Staphylococcus*, *Pseudomonas*, *Streptomyces* and many other genera are also determined [51]. Metagenomics was employed in lung cancer micro-organisms determination. There was over and under representation of micro-organisms in lung cancer patients. *Rothia*, *Moraxella*, *Haemophilus* and *Mycobacterium* are critical genera in diseased persons. Fusobacteria, Proteobacteria and Actinobacteria are effective phylum identified in the patients that shows their role and impact in the illness [63]. In case of TB, the microbes initiate their growth and has key effects on the disease. The genera in TB are a bit change from the other enlisted disorders. The Spirochaetes, Tenericutes, Fusobacterium, Proteobacteria, Campylobacter and Absconditabacteria in infected individuals [72].

The lung and overall pulmonary tract microbiota has a significant impact in development of disease. Our research results provide critical bacterial phyla, families, genus, species and some strains in aspect of respiratory diseases. The important bacteria for progression of selected disorders are Bacteroidota, Firmicutes and Proteobacteria that are present in all diseased datasets. Moreover Fusobacteria is discriminating in COPD amplicon. *Actinomyces*, *Streptococcus* and *Veionella* are same in amplicon and shotgun at genus level while Staphylococcaceae family is same in both amplicon and shotgun analysis in COVID-19. Actinobacteria and Fusobacteria are distinct phylum in lung cancer. Candidatus-Saccharibacteria is distinguish in TB samples.

Microbes play a significant role in pulmonary illness. The results show the presence of micro-organisms in different breathing disorders. Different phylum, families, genera and species of bacteria are classified with their relative abundances. Although, some phylum, families or genera are same in diseased and control sequences but their abundances or representation are altered that is a key modulator and progression of illness. As in our case in amplicon analysis of COVID-19, the Proteobacteria has 99.038 % relative abundance in infected samples but in control cases the maximum abundance observed is 88.098 % with different genera and species.

Additionally, ML models e.g. RF and SVM are developed and trained on the bacterial species. The data are bacterial species with their relative abundances that are input into scikit-learn. The tsv file is read through a function and data is transformed to bring the bacterial species into rows and samples relative abundances into columns. The tsv file contains bacterial species information of all diseases and non-diseased samples. The labels are added to differentiate between the datasets, with 80/20 ratio for splitting of data for training and testing of the models. The model can classify and differentiate between the bacterial species of all diseases. Lasso model and chi-square techniques are used for significant selection of features. The model is compiled in python, and

originally accuracy is used for measuring how accurately the model can classify between the given features. For model evaluation, test data is used which predicts the species with accurate labels for each dataset. The COVID-19 dataset shows the high accuracies among all with 94% in RF and 88% in SVM.

Chapter 6

CONCLUSIONS AND FUTURE RECOMMENDATIONS

In this research, several metagenomics datasets comprising amplicons (16S rRNA) and shotgun metagenomes of different respiratory disorders such as COPD, COVID-19, Lung cancer and TB are analyzed. Firmicutes, Proteobacteria and Bacteroidetes are the prominent phylums in all these disorders. Actinobacteria and Fusobacteria are distinct in lung cancer patients with Candidatus-Saccharibacteria is present in TB samples. In COPD sequences, Fusobacteria is discriminated in amplicon sequences while Staphylococcaceae family is same in both amplicon and shotgun analysis in COVID-19. These phylums have different classes and families of bacteria with few same and few discriminating in the infected sequences.

Shotgun provides more detailed perspective of taxonomical profiling by classifying the bacteria at the genus and specie level. The critical species identified in COPD are *Staphylococcus-aureus*, *Streptococcus-salivarius*, *Actinomyces-graevenitzii*, *Prevotella-histicola* and *Haemophilus-influenzae*. Similarly, impactful species in COVID-19 are *Corynebacteriu-propinquum*, *Cutibacterium-acnes*, *Streptococcus-pneumoniae*, *Malassezia-restricta* and *Dolosigranulum-pigrum*. In lung cancer sequences *Halomonas-sp-LBP4*, *Vibrio-alginolyticus*, *Selenomonas*, *Campylobacter-jejuni*, *Streptococcus-pneumoniae*, *Prevotella-melaninogenica*, *Haemophilus-influenzae* and *Escherichia-coli* plays their role in progression of disease. *Neisseria-subflava*, *Prevotella-melaninogenica*, *Streptococcus-mitis*, *Ralstonia-pickettii* and *Porphyromonas-pasteri* are significant species identified in TB samples. Furthermore, different ML models are employed to determine the microbial specie patterns in diseased and healthy samples. They can play their role in detection of different pulmonary disorders on the basis of bacterial species present in the samples.

To improve the authenticity of the research, datasets number implanted for analysis can be increased. For better understanding of microbial functions and their link with

proteins and pathways related to disorders functional analysis can be incorporated. Moreover, to increase the reliability of the results, it is recommended to incorporate the wet lab procedures for 16S rRNA and metagenome analysis. Bacteroidetes, Firmicutes and Proteobacteria are the significant taxonomic groups with further underlying families, genera and specie level information is identified in the analysis. For evaluation and confirmation of the research results, additional studies and more exploration is required.

In this study, whole respiratory microbiota primarily lungs is explored with secondary data which is focused on the nasopharynx, BAL and BALF with different sites in respiratory tract. Hence, studies with primary data with one focused site in pulmonary tract can give more specific and detailed representation of the microbe's role and impact on the breathing disorders. In brief, the obtained results can be the base to the more valuable and thorough research related to respiratory microbes.

Bibliography

- [1] J. G. Natalini, S. Singh, and L. N. Segal, “The dynamic lung microbiome in health and disease,” *Nature Reviews Microbiology*, vol. 21, no. 4, pp. 222–235, 2023.
- [2] A. I. Ritchie and A. Singanayagam, “Metagenomic characterization of the respiratory microbiome. a piece de resistance,” 2020.
- [3] M.-M. Pust, L. Wiehlmann, C. Davenport, I. Rudolf, A.-M. Dittrich, and B. Tümmler, “The human respiratory tract microbial community structures in healthy and cystic fibrosis infants,” *npj Biofilms and Microbiomes*, vol. 6, no. 1, p. 61, 2020.
- [4] M. Bou Zerdan, J. Kassab, P. Meouchy, E. Haroun, R. Nehme, M. Bou Zerdan, G. Fahed, M. Petrosino, D. Dutta, and S. Graziano, “The lung microbiota and lung cancer: a growing relationship,” *Cancers*, vol. 14, no. 19, p. 4813, 2022.
- [5] R. Li, J. Li, and X. Zhou, “Lung microbiome: new insights into the pathogenesis of respiratory diseases,” *Signal Transduction and Targeted Therapy*, vol. 9, no. 1, p. 19, 2024.
- [6] C. Kumpitsch, K. Koskinen, V. Schöpf, and C. Moissl-Eichinger, “The microbiome of the upper respiratory tract in health and disease,” *BMC biology*, vol. 17, pp. 1–20, 2019.
- [7] N. L. Mankowski and B. Bordoni, “Anatomy, head and neck, nasopharynx,” 2020.
- [8] P. Zimmermann, “Exploring the microbial landscape of the nasopharynx in children: a systematic review of studies using next generation sequencing,” *Frontiers in Microbiomes*, vol. 2, p. 1231271, 2023.
- [9] M. Flynn and J. Dooley, “The microbiome of the nasopharynx,” *Journal of medical microbiology*, vol. 70, no. 6, p. 001368, 2021.
- [10] L. Qu, Q. Cheng, Y. Wang, H. Mu, and Y. Zhang, “COPD and gut–lung axis: how microbiota and host inflammasome influence COPD and related therapeutics,” *Frontiers in Microbiology*, vol. 13, p. 868086, 2022.

- [11] D. Yang, Y. Xing, X. Song, and Y. Qian, “The impact of lung microbiota dysbiosis on inflammation,” *Immunology*, vol. 159, no. 2, pp. 156–166, 2020.
- [12] J. Bouquet, D. E. Tabor, J. S. Silver, V. Nair, A. Tovchigrechko, M. P. Griffin, M. T. Esser, B. R. Sellman, and H. Jin, “Microbial burden and viral exacerbations in a longitudinal multicenter copd cohort,” *Respiratory research*, vol. 21, pp. 1–13, 2020.
- [13] R. J. Huijsmans, A. de Haan, N. N. ten Hacken, R. V. Straver, and A. J. van’t Hul, “The clinical utility of the gold classification of copd disease severity in pulmonary rehabilitation,” *Respiratory medicine*, vol. 102, no. 1, pp. 162–171, 2008.
- [14] M. Miravitlles and A. Ribera, “Understanding the impact of symptoms on the burden of copd,” *Respiratory research*, vol. 18, no. 1, p. 67, 2017.
- [15] M. A. Khan, “Monthly and seasonal prevalence of asthma and chronic obstructive pulmonary disease in the district dera ismail khan, khyber pakhtunkhwa, pakistan,” *The Egyptian Journal of Bronchology*, vol. 16, no. 1, p. 63, 2022.
- [16] A. Singh, S. Zaheer, N. Kumar, T. Singla, and S. Ranga, “Covid19, beyond just the lungs: A review of multisystemic involvement by covid19,” *Pathology-Research and Practice*, vol. 224, p. 153384, 2021.
- [17] G. Gabutti, E. d’Anchera, F. Sandri, M. Savio, and A. Stefanati, “Coronavirus: update related to the current outbreak of covid-19,” *Infectious diseases and therapy*, vol. 9, pp. 241–253, 2020.
- [18] S. Maryam, I. Ul Haq, G. Yahya, M. Ul Haq, A. M. Algammal, S. Saber, and S. Cavalu, “Covid-19 surveillance in wastewater: An epidemiological tool for the monitoring of sars-cov-2,” *Frontiers in cellular and infection microbiology*, vol. 12, p. 978643, 2023.
- [19] S. Khatiwada and A. Subedi, “Lung microbiome and coronavirus disease 2019 (covid-19): possible link and implications,” *Human microbiome journal*, vol. 17, p. 100073, 2020.
- [20] Y. Han, Z. Jia, J. Shi, W. Wang, and K. He, “The active lung microbiota landscape of covid-19 patients through the metatranscriptome data analysis,” *BioImpacts: BI*, vol. 12, no. 2, p. 139, 2022.
- [21] B.-Y. Wang, J.-Y. Huang, H.-C. Chen, C.-H. Lin, S.-H. Lin, W.-H. Hung, and Y.-F. Cheng, “The comparison between adenocarcinoma and squamous cell carcinoma in lung cancer patients,” *Journal of cancer research and clinical oncology*, vol. 146, pp. 43–52, 2020.

- [22] M. M. Saab, M. McCarthy, M. O’Driscoll, L. J. Sahm, P. Leahy-Warren, B. Noonan, S. FitzGerald, M. O’Malley, N. Lyons, H. E. Burns, *et al.*, “A systematic review of interventions to recognise, refer and diagnose patients with lung cancer symptoms,” *NPJ Primary Care Respiratory Medicine*, vol. 32, no. 1, p. 42, 2022.
- [23] A. Ali, M. F. Manzoor, N. Ahmad, R. M. Aadil, H. Qin, R. Siddique, S. Riaz, A. Ahmad, S. A. Korma, W. Khalid, *et al.*, “The burden of cancer, government strategic policies, and challenges in pakistan: A comprehensive review,” *Frontiers in nutrition*, vol. 9, p. 940514, 2022.
- [24] H. Guo, L. Zhao, J. Zhu, P. Chen, H. Wang, M. Jiang, X. Liu, H. Sun, W. Zhao, Z. Zheng, *et al.*, “Microbes in lung cancer initiation, treatment, and outcome: boon or bane?,” in *Seminars in Cancer Biology*, vol. 86, pp. 1190–1206, Elsevier, 2022.
- [25] T. Parbhoo, J. M. Mouton, and S. L. Sampson, “Phenotypic adaptation of mycobacterium tuberculosis to host-associated stressors that induce persister formation,” *Frontiers in Cellular and Infection Microbiology*, vol. 12, p. 956607, 2022.
- [26] L. Luies and I. Du Preez, “The echo of pulmonary tuberculosis: mechanisms of clinical symptoms and other disease-induced systemic complications,” *Clinical microbiology reviews*, vol. 33, no. 4, pp. 10–1128, 2020.
- [27] S. B. Obakiro, A. Kiprop, I. Kowino, E. Kigundu, M. P. Odero, T. Omara, and L. Bunalema, “Ethnobotany, ethnopharmacology, and phytochemistry of traditional medicinal plants used in the management of symptoms of tuberculosis in east africa: a systematic review,” *Tropical Medicine and Health*, vol. 48, pp. 1–21, 2020.
- [28] Y. Liu, J. Wang, and C. Wu, “Microbiota and tuberculosis: a potential role of probiotics, and postbiotics,” *Frontiers in Nutrition*, vol. 8, p. 626254, 2021.
- [29] C. Frioux, D. Singh, T. Korcsmaros, and F. Hildebrand, “From bag-of-genes to bag-of-genomes: metabolic modelling of communities in the era of metagenome-assembled genomes,” *Computational and Structural Biotechnology Journal*, vol. 18, pp. 1722–1734, 2020.
- [30] L. Zhang, F. Chen, Z. Zeng, M. Xu, F. Sun, L. Yang, X. Bi, Y. Lin, Y. Gao, H. Hao, *et al.*, “Advances in metagenomics and its application in environmental microorganisms,” *Frontiers in microbiology*, vol. 12, p. 766364, 2021.
- [31] W. Liebl, “Metagenomics,” in *Encyclopedia of geobiology*, Springer, 2010.

- [32] Y.-X. Liu, Y. Qin, T. Chen, M. Lu, X. Qian, X. Guo, and Y. Bai, “A practical guide to amplicon and metagenomic analysis of microbiome data,” *Protein & cell*, vol. 12, no. 5, pp. 315–330, 2021.
- [33] B. Mahesh, “Machine learning algorithms-a review,” *International Journal of Science and Research (IJSR).[Internet]*, vol. 9, no. 1, pp. 381–386, 2020.
- [34] A. Mathieu, M. Leclercq, M. Sanabria, O. Perin, and A. Droit, “Machine learning and deep learning applications in metagenomic taxonomy and functional annotation,” *Frontiers in Microbiology*, vol. 13, p. 811495, 2022.
- [35] T. Jiang, J. L. Gradus, and A. J. Rosellini, “Supervised machine learning: a brief primer,” *Behavior therapy*, vol. 51, no. 5, pp. 675–687, 2020.
- [36] S. Yu, H. Zhang, L. Wan, M. Xue, Y. Zhang, and X. Gao, “The association between the respiratory tract microbiome and clinical outcomes in patients with copd,” *Microbiological Research*, vol. 266, p. 127244, 2023.
- [37] C. Russo, V. Colaianni, G. Ielo, M. S. Valle, L. Spicuzza, and L. Malaguarnera, “Impact of lung microbiota on copd,” *Biomedicines*, vol. 10, no. 6, p. 1337, 2022.
- [38] H. Guo-Parke, D. Linden, S. Weldon, J. C. Kidney, and C. C. Taggart, “Mechanisms of virus-induced airway immunity dysfunction in the pathogenesis of copd disease, progression, and exacerbation,” *Frontiers in Immunology*, vol. 11, p. 516782, 2020.
- [39] M. Kaur, J. Chandel, J. Malik, and A. S. Naura, “Particulate matter in copd pathogenesis: an overview,” *Inflammation Research*, vol. 71, no. 7, pp. 797–815, 2022.
- [40] F. De Nuccio, P. Piscitelli, and D. M. Toraldo, “Gut–lung microbiota interactions in chronic obstructive pulmonary disease (copd): potential mechanisms driving progression to copd and epidemiological data,” *Lung*, vol. 200, no. 6, pp. 773–781, 2022.
- [41] L. Ren, R. Zhang, J. Rao, Y. Xiao, Z. Zhang, B. Yang, D. Cao, H. Zhong, P. Ning, Y. Shang, *et al.*, “Transcriptionally active lung microbiome and its association with bacterial biomass and host inflammatory status,” *MSystems*, vol. 3, no. 5, pp. 10–1128, 2018.
- [42] S. J. Cameron, K. E. Lewis, S. A. Huws, W. Lin, M. J. Hegarty, P. D. Lewis, L. A. Mur, and J. A. Pachebat, “Metagenomic sequencing of the chronic obstructive pulmonary disease upper bronchial tract microbiome reveals functional changes associated with disease severity,” *PLoS One*, vol. 11, no. 2, p. e0149095, 2016.

- [43] C.-Y. Yang, S.-W. Li, C.-Y. Chin, C.-W. Hsu, C.-C. Lee, Y.-M. Yeh, and K.-A. Wu, “Association of exacerbation phenotype with the sputum microbiome in chronic obstructive pulmonary disease patients during the clinically stable state,” *Journal of Translational Medicine*, vol. 19, pp. 1–14, 2021.
- [44] D. Yesudhas, A. Srivastava, and M. M. Gromiha, “Covid-19 outbreak: history, mechanism, transmission, structural studies and therapeutics,” *Infection*, vol. 49, pp. 199–213, 2021.
- [45] S. R. Paludan and T. H. Mogensen, “Innate immunological pathways in covid-19 pathogenesis,” *Science immunology*, vol. 7, no. 67, p. eabm5505, 2022.
- [46] E. Sharifipour, S. Shams, M. Esmkhani, J. Khodadadi, R. Fotouhi-Ardakani, A. Koohpaei, Z. Doosti, and S. Ej Golzari, “Evaluation of bacterial co-infections of the respiratory tract in covid-19 patients admitted to icu,” *BMC infectious diseases*, vol. 20, pp. 1–7, 2020.
- [47] R. Aquino-Martinez and S. Hernández-Vigueras, “Severe covid-19 lung infection in older people and periodontitis,” *Journal of Clinical Medicine*, vol. 10, no. 2, p. 279, 2021.
- [48] R. F. Kullberg, J. de Brabander, L. S. Boers, J. J. Biemond, E. J. Nossent, L. M. Heunks, A. P. Vlaar, P. I. Bonta, T. van Der Poll, J. Duitman, *et al.*, “Lung microbiota of critically ill patients with covid-19 are associated with nonresolving acute respiratory distress syndrome,” *American journal of respiratory and critical care medicine*, vol. 206, no. 7, pp. 846–856, 2022.
- [49] C. Merenstein, F. D. Bushman, and R. G. Collman, “Alterations in the respiratory tract microbiome in covid-19: current observations and potential significance,” *Microbiome*, vol. 10, no. 1, p. 165, 2022.
- [50] Q. Miao, Y. Ma, Y. Ling, W. Jin, Y. Su, Q. Wang, J. Pan, Y. Zhang, H. Chen, J. Yuan, *et al.*, “Evaluation of superinfection, antimicrobial usage, and airway microbiome with metagenomic sequencing in covid-19 patients: A cohort study in shanghai,” *Journal of Microbiology, Immunology and Infection*, vol. 54, no. 5, pp. 808–815, 2021.
- [51] M. N. Hoque, M. S. Rahman, R. Ahmed, M. S. Hossain, M. S. Islam, T. Islam, M. A. Hossain, and A. Z. Siddiki, “Diversity and genomic determinants of the microbiomes associated with covid-19 and non-covid respiratory diseases,” *Gene reports*, vol. 23, p. 101200, 2021.
- [52] N. Haiminen, F. Utro, E. Seabolt, and L. Parida, “Functional profiling of covid-19 respiratory tract microbiomes,” *Scientific Reports*, vol. 11, no. 1, p. 6433, 2021.

- [53] P. Gaibani, E. Viciani, M. Bartoletti, R. E. Lewis, T. Tonetti, D. Lombardo, A. Castagnetti, F. Bovo, C. S. Horna, M. Ranieri, *et al.*, “The lower respiratory tract microbiome of critically ill patients with covid-19,” *Scientific reports*, vol. 11, no. 1, p. 10103, 2021.
- [54] S. Ke, S. T. Weiss, and Y.-Y. Liu, “Dissecting the role of the human microbiome in covid-19 via metagenome-assembled genomes,” *Nature Communications*, vol. 13, no. 1, p. 5235, 2022.
- [55] L. Corrales, R. Rosell, A. F. Cardona, C. Martin, Z. L. Zatarain-Barron, and O. Arrieta, “Lung cancer in never smokers: The role of different risk factors other than tobacco smoking,” *Critical reviews in oncology/hematology*, vol. 148, p. 102895, 2020.
- [56] A. Acha-Sagredo, B. Uko, P. Pantazi, N. G. Bediaga, C. Moschandrea, L. Rainbow, M. W. Marcus, M. P. Davies, J. K. Field, and T. Liloglou, “Long non-coding rna dysregulation is a frequent event in non-small cell lung carcinoma pathogenesis,” *British journal of cancer*, vol. 122, no. 7, pp. 1050–1058, 2020.
- [57] Y. Zhao, Y. Liu, S. Li, Z. Peng, X. Liu, J. Chen, and X. Zheng, “Role of lung and gut microbiota on lung cancer pathogenesis,” *Journal of Cancer Research and Clinical Oncology*, vol. 147, no. 8, pp. 2177–2186, 2021.
- [58] A. Karvela, O.-Z. Veloudiou, A. Karachaliou, T. Kloukina, G. Gomatou, and E. Kotteas, “Lung microbiome: An emerging player in lung cancer pathogenesis and progression,” *Clinical and Translational Oncology*, vol. 25, no. 8, pp. 2365–2372, 2023.
- [59] F. Perrone, L. Belluomini, M. Mazzotta, M. Bianconi, V. Di Noia, F. Meacci, M. Montrone, D. Pignataro, A. Prelaj, S. Rinaldi, *et al.*, “Exploring the role of respiratory microbiome in lung cancer: A systematic review,” *Critical Reviews in Oncology/Hematology*, vol. 164, p. 103404, 2021.
- [60] Q. Dong, E. S. Chen, C. Zhao, and C. Jin, “Host-microbiome interaction in lung cancer,” *Frontiers in Immunology*, vol. 12, p. 679829, 2021.
- [61] O.-H. Kim, B.-Y. Choi, D. K. Kim, N. H. Kim, J. K. Rho, W. J. Sul, and S. W. Lee, “The microbiome of lung cancer tissue and its association with pathological and clinical parameters,” *American Journal of Cancer Research*, vol. 12, no. 5, p. 2350, 2022.
- [62] K. L. Greathouse, J. R. White, A. J. Vargas, V. V. Bliskovsky, J. A. Beck, N. von Muhlinen, E. C. Polley, E. D. Bowman, M. A. Khan, A. I. Robles, *et al.*, “Interaction between the microbiome and tp53 in human lung cancer,” *Genome biology*, vol. 19, pp. 1–16, 2018.

- [63] Q. Chen, K. Hou, M. Tang, S. Ying, X. Zhao, G. Li, J. Pan, X. He, H. Xia, Y. Li, *et al.*, “Screening of potential microbial markers for lung cancer using metagenomic sequencing,” *Cancer Medicine*, vol. 12, no. 6, pp. 7127–7139, 2023.
- [64] L. Zheng, R. Sun, Y. Zhu, Z. Li, X. She, X. Jian, F. Yu, X. Deng, B. Sai, L. Wang, *et al.*, “Lung microbiome alterations in nscl patients,” *Scientific reports*, vol. 11, no. 1, p. 11736, 2021.
- [65] G. Churchyard, P. Kim, N. S. Shah, R. Rustomjee, N. Gandhi, B. Mathema, D. Dowdy, A. Kasmar, and V. Cardenas, “What we know about tuberculosis transmission: an overview,” *The Journal of infectious diseases*, vol. 216, no. suppl_6, pp. S629–S635, 2017.
- [66] G. B. Migliori, C. W. Ong, L. Petrone, L. D’Ambrosio, R. Centis, and D. Goletti, “The definition of tuberculosis infection based on the spectrum of tuberculosis disease,” *Breathe*, vol. 17, no. 3, 2021.
- [67] S. S. Alsayed and H. Gunosewoyo, “Tuberculosis: pathogenesis, current treatment regimens and new drug targets,” *International journal of molecular sciences*, vol. 24, no. 6, p. 5202, 2023.
- [68] S. R. Hegde, “Computational identification of the proteins associated with quorum sensing and biofilm formation in mycobacterium tuberculosis,” *Frontiers in Microbiology*, vol. 10, p. 479640, 2020.
- [69] Y. Hu, Y. Kang, X. Liu, M. Cheng, J. Dong, L. Sun, Y. Zhu, X. Ren, Q. Yang, X. Chen, *et al.*, “Distinct lung microbial community states in patients with pulmonary tuberculosis,” *Science China Life Sciences*, vol. 63, pp. 1522–1533, 2020.
- [70] T. Shah, Z. Shah, Z. Baloch, and X. Cui, “The role of microbiota in respiratory health and diseases, particularly in tuberculosis,” *Biomedicine & Pharmacotherapy*, vol. 143, p. 112108, 2021.
- [71] G. Xiao, Z. Cai, Q. Guo, T. Ye, Y. Tang, P. Guan, J. Zhang, M. Ou, X. Fu, L. Ren, *et al.*, “Insights into the unique lung microbiota profile of pulmonary tuberculosis patients using metagenomic next-generation sequencing,” *Microbiology Spectrum*, vol. 10, no. 1, pp. e01901–21, 2022.
- [72] M. R. Ticlla, J. Hella, H. Hiza, M. Sasamalo, F. Mhimbira, L. K. Rutaihwa, S. Droz, S. Schaller, K. Reither, M. Hilty, *et al.*, “The sputum microbiome in pulmonary tuberculosis and its association with disease manifestations: a cross-sectional study,” *Frontiers in Microbiology*, vol. 12, p. 633396, 2021.

- [73] Y. Hu, M. Cheng, B. Liu, J. Dong, L. Sun, J. Yang, F. Yang, X. Chen, and Q. Jin, “Metagenomic analysis of the lung microbiome in pulmonary tuberculosis—a pilot study,” *Emerging microbes & infections*, vol. 9, no. 1, pp. 1444–1452, 2020.
- [74] J. Li, A. Xiong, J. Wang, X. Wu, L. Bai, L. Zhang, X. He, and G. Li, “Deciphering the microbial landscape of lower respiratory tract infections: insights from metagenomics and machine learning,” *Frontiers in Cellular and Infection Microbiology*, vol. 14, p. 1385562, 2024.
- [75] C. Y. Zhao, Y. Hao, Y. Wang, J. J. Varga, A. A. Stecenko, J. B. Goldberg, and S. P. Brown, “Microbiome data enhances predictive models of lung function in people with cystic fibrosis,” *The Journal of infectious diseases*, vol. 223, no. Supplement_3, pp. S246–S256, 2021.
- [76] A. McDowell, J. Kang, J. Yang, J. Jung, Y.-M. Oh, S.-M. Kym, T.-S. Shin, T.-B. Kim, Y.-K. Jee, and Y.-K. Kim, “Machine-learning algorithms for asthma, copd, and lung cancer risk assessment using circulating microbial extracellular vesicle data and their application to assess dietary effects,” *Experimental & Molecular Medicine*, vol. 54, no. 9, pp. 1586–1595, 2022.
- [77] K. Pienkowska, M.-M. Pust, M. Gessner, S. Gaedcke, A. Thavarasa, I. Rosenboom, P. Morán Losada, R. Minso, C. Arnold, S. Hedtfeld, *et al.*, “The cystic fibrosis upper and lower airway metagenome,” *Microbiology Spectrum*, vol. 11, no. 2, pp. e03633–22, 2023.
- [78] I. Abellan-Schneyder, M. Matchado, S. Reitmeier, A. Sommer, Z. Sewald, J. Baumbach, M. List, K. Neuhaus, and S. Tringe, “Primer, pipelines, parameters: issues in 16s rna gene sequencing. msphere 6: e01202-20,” 2021.
- [79] D. Straub, N. Blackwell, A. Langarica-Fuentes, A. Peltzer, S. Nahnsen, and S. Kleindienst, “Interpretations of environmental microbial community studies are biased by the selected 16s rna (gene) amplicon sequencing pipeline,” *Frontiers in Microbiology*, vol. 11, p. 550420, 2020.
- [80] K. Chappell, B. Francou, C. Habib, T. Huby, M. Leoni, A. Cottin, F. Nadal, E. Adnet, E. Paoli, C. Oliveira, *et al.*, “Galaxy is a suitable bioinformatics platform for the molecular diagnosis of human genetic disorders using high-throughput sequencing data analysis: five years of experience in a clinical laboratory,” *Clinical Chemistry*, vol. 68, no. 2, pp. 313–321, 2022.
- [81] S. Krakau, D. Straub, H. Gourel, G. Gabernet, and S. Nahnsen, “nf-core/mag: a best-practice pipeline for metagenome hybrid assembly and binning,” *NAR genomics and bioinformatics*, vol. 4, no. 1, p. lqac007, 2022.

- [82] A. Blanco-Míguez, F. Beghini, F. Cumbo, L. J. McIver, K. N. Thompson, M. Zolfo, P. Manghi, L. Dubois, K. D. Huang, A. M. Thomas, *et al.*, “Extending and improving metagenomic taxonomic profiling with uncharacterized species using metaphlan 4,” *Nature Biotechnology*, vol. 41, no. 11, pp. 1633–1644, 2023.
- [83] J. Cai, J. Luo, S. Wang, and S. Yang, “Feature selection in machine learning: A new perspective,” *Neurocomputing*, vol. 300, pp. 70–79, 2018.