Nakul_Phd_Thesis.docx

Delhi Technological University

Document Details

Submission ID
trn:oid:::27535:122649592

Submission Date
Nov 24, 2025, 1:19 PM GMT+5:30

Download Date
Nov 24, 2025, 1:24 PM GMT+5:30

File Name
Nakul_Phd_Thesis.docx

File Size
9.4 MB

113 Pages

32,800 Words

210,158 Characters

7% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Cited Text
- Small Matches (less than 10 words)

Exclusions

- 3 Excluded Sources

Match Groups

13 Not Cited or Quoted 1%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

139 Missing Citation 6%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

4% Internet sources
4% Publications
4% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **13** Not Cited or Quoted 1%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **139** Missing Citation 6%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

4% Internet sources
4% Publications
4% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Publication	Khushi Yadav, Yasha Hasija. "Integrated analysis of gene expressions and targete... <1%
2	Submitted works	Delhi Technological University on 2024-05-23 <1%
3	Internet	www.ncbi.nlm.nih.gov <1%
4	Internet	dtu.ac.in <1%
5	Publication	Kritika Sharma, Neeru Saini, Yasha Hasija. "Identifying the mitochondrial metabo... <1%
6	Submitted works	Aberystwyth University on 2016-04-12 <1%
7	Submitted works	National School of Business Management NSBM, Sri Lanka on 2025-11-14 <1%
8	Submitted works	University of Essex on 2025-11-19 <1%
9	Internet	link.springer.com <1%
10	Internet	vdocuments.site <1%

11 Internet
www.koreascience.or.kr <1%

12 Publication
Paola Paci, Giulia Fiscon, Federica Conte, Valerio Licursi et al. "Integrated transcri... <1%

13 Publication
Tian Sheng Chen, Ruenn Chai Lai, May May Lee, Andre Boon Hwa Choo, Chuen Ne... <1%

14 Internet
www.frontiersin.org <1%

15 Internet
bio3.giga.ulg.ac.be <1%

16 Submitted works
University of Technology, Sydney on 2024-10-31 <1%

17 Publication
Guan, Yixin. "Human-Centred Time Series Modeling of Daily Surgery Volumes in a... <1%

18 Internet
www.progettolibra.it <1%

19 Internet
pmc.ncbi.nlm.nih.gov <1%

20 Publication
Jaishree Meena, Yasha Hasija. "Application of explainable artificial intelligence in ... <1%

21 Submitted works
Australian Catholic University on 2024-04-28 <1%

22 Submitted works
American Sentinel University on 2021-03-15 <1%

23 Submitted works
The Cooperative University of Kenya on 2025-07-08 <1%

24 Internet
bmcgenomics.biomedcentral.com <1%

25 Publication
Sheetal Rajpal, Ankit Rajpal, Manoj Agarwal, Virendra Kumar, Ajith Abraham, Div... <1%

26 Internet
www.mdpi.com <1%

27 Internet
ses.library.usyd.edu.au <1%

28 Internet
www.researchsquare.com <1%

29 Publication
Constantinos Stathopoulos, Nikoleta Giarimoglou, Adamantia Kouvela, Argyris Al... <1%

30 Submitted works
Modern High School International on 2023-12-07 <1%

31 Internet
experts.umn.edu <1%

32 Submitted works
CSU, Chico on 2024-05-15 <1%

33 Publication
Moura, Joao, Elisabet Børsheim, and Eugenia Carvalho. "The Role of MicroRNAs in... <1%

34 Publication
Alfardus, Asma. "Evaluating Machine Learning for Intrusion Detection in CAN Bus... <1%

35 Publication
Ashok, Capoor Ishan. "An Association of Risk Factors Contributing to Chronic Obs... <1%

36 Publication
Amirhossein Ahmadi, Sharif Moradi. "In silico analysis suggests the RNAi-enhanci... <1%

37 Submitted works
Anglia Ruskin University on 2018-04-27 <1%

38 Publication
Pushpa Choudhary, Sambit Satpathy, Arvind Dagur, Dhirendra Kumar Shukla. "Re... <1%

39 Internet

5dok.net <1%

40 Publication

McDowell, Shana Monique. "Machine Learning Applications in Sweetpotato Prod... <1%

41 Submitted works

Sheffield Hallam University on 2025-07-16 <1%

42 Submitted works

EARLY MAKERS Group SA on 2024-10-05 <1%

43 Publication

Martin Riecan, Veronika Domanska, Cristina Lupu, Maulin Patel et al. "Tissue-spe... <1%

44 Submitted works

University of Baltimore on 2011-08-06 <1%

45 Internet

cgspace.cgiar.org <1%

46 Internet

ijirt.org <1%

47 Internet

www.jcancer.org <1%

48 Publication

Dothang Truong. "Demystifying AI - Data Science and Machine Learning Using IB... <1%

49 Submitted works

Imperial College of Science, Technology and Medicine on 2025-06-06 <1%

50 Publication

Li Zhang, Shangshang Hu, Jiasheng Chen, Shasha Ma, Fanghong Liu, Chuanmiao ... <1%

51 Submitted works

Midlands State University on 2025-06-17 <1%

52 Submitted works

Monash University on 2024-06-01 <1%

53 Publication
R Raksha, Jyoti Metan, Raghavendra M. Ichangi, P Suresh, Shazia Sulthana, M N A... <1%

54 Submitted works
University of Dundee on 2021-03-26 <1%

55 Submitted works
University of Queensland on 2025-11-06 <1%

56 Submitted works
University of Surrey on 2025-05-14 <1%

57 Publication
Y. Liu. "miR2Disease: a manually curated database for microRNA deregulation in ... <1%

58 Internet
aging-us.com <1%

59 Internet
arxiv.org <1%

60 Internet
cdn.bookey.app <1%

61 Internet
researchspace.ukzn.ac.za <1%

62 Internet
respiratory-research.biomedcentral.com <1%

63 Internet
worldwidescience.org <1%

64 Internet
www.nature.com <1%

65 Internet
www.science.gov <1%

66 Publication
Bilsborough, Helena. "Exploring Barriers and Motivations for Peer Volunteering i... <1%

67 Publication
Caramori, G.. "Mechanisms involved in lung cancer development in COPD", Intern... <1%

68 Publication
Claudia Cantoni, Laura Ghezzi, Jasmine Choi, Anne H. Cross, Laura Piccio. "Targeti... <1%

69 Publication
Gaetano Caramori, Paolo Ruggeri, Sharon Mumby, Antonio Ieni et al. "Molecular l... <1%

70 Submitted works
Grand Canyon University on 2014-08-10 <1%

71 Publication
Ido Givon, Nati Bor, Ran Matot, Lior Friedrich, Daya Gross, Gili Konforty, Arriel Be... <1%

72 Submitted works
Imperial College of Science, Technology and Medicine on 2024-08-14 <1%

73 Publication
Poonam Nandal, Mamta Dahiya, Meeta Singh, Arvind Dagur, Brijesh Kumar. "Pro... <1%

74 Publication
R. Nithya, T. Kokilavani, T. Lucia Agnes Beena. "Chapter 40 Cerebral Stroke Classif... <1%

75 Submitted works
The University of Manchester on 2025-09-01 <1%

76 Submitted works
University of Canberra on 2024-04-05 <1%

77 Submitted works
University of Teesside on 2024-10-31 <1%

78 Publication
Vivek P. Chavda, Mahesh T. Chhabria, Divya M. Teli. "Recent Advancements in Tub... <1%

79 Publication
Wu, Chenyu. "The Application of Machine Learning Algorithms on CO2 EOR", The ... <1%

80 Internet
ai.jmir.org <1%

81 Internet

archiv.ub.uni-heidelberg.de <1%

82 Publication

de Almeida Neves, Roseane Carvalho. "Inflammation-Related Risk Factors for Pan... <1%

83 Internet

ir.lib.uwo.ca <1%

AN INTEGRATED APPROACH TOWARDS THE IDENTIFICATION OF NOVEL BIOMARKERS IN RESPIRATORY DISORDERS

A Thesis Submitted
in Partial Fulfillment of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

by

Nakul Tanwar
(2K22/PHD/BT/507)

Under the Supervision of
Prof. Yasha Hasija
Department of Biotechnology
Delhi Technological University

Department of Biotechnology

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042. India

November, 2025

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE'S DECLARATION

I **Nakul Tanwar, 2K22/PHD/BT/507** hereby certify that the work which is being presented in the thesis entitled '**AN INTEGRATED APPROACH TOWARDS THE IDENTIFICATION OF NOVEL BIOMARKERS IN RESPIRATORY DISORDERS'** in partial fulfillment of the requirements for the award of the Degree of Doctor of Philosophy, submitted in the Department of Biotechnology , Delhi Technological University is an authentic record of my own work carried out during the period from **12th January 2023** to _____ under the supervision of **Prof. Yasha Hasija**.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Candidate's Signature

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE BY THE SUPERVISOR

Certified that **Nakul Tanwar, 2K22/PHD/BT/507** has carried out their search work presented in this thesis entitled **"AN INTEGRATED APPROACH TOWARDS THE IDENTIFICATION OF NOVEL BIOMARKERS IN RESPIRATORY DISORDERS"** for the award of Doctor of Philosophy from Department of Biotechnology, Delhi Technological University, Delhi, under my supervision. The thesis embodies results of original work, and studies are carried out by the student himself, and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Prof. Yasha Hasija
Professor & Head (BT)
Department of Biotechnology
Delhi Technological University

Date:

Table of Contents

List of Tables

List of Figures

Abstract

Respiratory disorders such as COPD, ILD, CPFE, and lung cancer are primarily lung diseases, yet they do not operate within isolated physiological boundaries. These conditions share a deeply interconnected inflammatory landscape, where chronic immune activation, oxidative stress, epithelial injury, and aberrant tissue repair collectively drive both disease progression and coexistence. This interconnectedness is evident in clinical practice, where patients frequently present with overlapping respiratory conditions such as COPD coexisting with ILD or lung cancer because they are shaped by the same underlying molecular and inflammatory pathways. The presence of such overlap points to a broader biological principle that chronic inflammation exists along a continuum across the body rather than remaining confined to a single organ. As a result, it emerges as a systemic process capable of linking diseases that traditionally appear unrelated. This becomes clearer when considering how circulating inflammatory mediators, dysregulated immune cells, and miRNA-driven signaling can influence tissues beyond the lungs. Within this continuum, some immune-mediated conditions for example, Multiple Sclerosis (MS) further demonstrate how shared inflammatory and immune-regulatory disturbances can bridge organ systems, reinforcing the idea that complex diseases are often unified by common immunological mechanisms rather than separated by anatomical boundaries.

Traditional diagnostic tools including imaging, pulmonary function tests, and histopathology frequently detect disease only at advanced stages. In parallel, although omics technologies have generated large-scale genomic and transcriptomic datasets, their clinical translation is hindered by the complexity of multi-omics signals and by the "black-box'' nature of most machine learning approaches. The present research addresses these gaps by integrating multi-omics analysis, machine learning, explainable artificial intelligence (XAI), miRNA–mRNA regulatory network exploration, and large language model (LLM)-based interpretability to uncover shared biomarkers, elucidate mechanistic relationships across diseases, and develop an accessible, clinically interpretable decision-support system for lung disease indication.

The first component of the study investigates coexistence among COPD, ILD, and CPFE through integrative transcriptomic and regulatory network analyses. Using the GSE47460 microarray dataset (582 lung tissue samples), rigorous preprocessing, quantile normalization, and class-balancing with SMOTE were applied, followed by a Random Forest classifier to distinguish COPD, ILD, and control samples. Explainable AI using SHAP revealed 20 key genes including OCIAD2, IRS2, TRIM2, MUC20, and CCDC109B—that consistently contributed to model performance across all classes. Functional enrichment analysis demonstrated that these genes participate in oxidative stress, immune activation, epithelial repair, extracellular matrix remodeling, and calcium signaling pathways, all of which underpin the shared pathogenesis of COPD, ILD, and CPFE. Subsequent validation via heatmaps, gene co-expression networks, single-cell expression analysis, and miRNA–mRNA regulatory mapping confirmed the biological relevance of these markers and identified their involvement in fibroblast activation, inflammatory fibroblast signatures, and altered epithelial homeostasis. Collectively, these findings provide strong evidence of convergent mechanisms underlying respiratory disease coexistence and highlight candidate biomarkers with diagnostic and therapeutic utility.

The second component explores systemic inflammatory connections among COPD, lung cancer, and MS using the GSE61741 peripheral blood miRNA dataset (237 samples) along with an independent validation dataset. Machine learning models, supported by SMOTE-based class

balancing and 5-fold cross-validation, achieved high predictive accuracy for all four classes. SHAP interpretability revealed 20 core miRNAs including hsa-let-7c, hsa-miR-223, hsa-miR-92a, and hsa-miR-454 that serve as central regulators across these diseases. These miRNAs converged on six shared inflammatory genes (IL6, IL10, CCL2, CCL5, MYC, and ITGB3), forming a cross-disease regulatory axis linking neuroinflammation, chronic respiratory inflammation, fibrosis, and oncogenesis. Downstream enrichment analyses identified common signaling pathways such as NF-κB, JAK-STAT, PI3K-Akt, cytokine–cytokine receptor interactions, and immune cell activation cascades. Single-cell expression mapping further demonstrated that these genes and miRNAs are enriched in inflammatory fibroblasts, macrophages, T cells, and epithelial populations, suggesting a shared pathological microenvironment across lung and neurological diseases. This objective provides a unified molecular explanation for the epidemiologically observed association between MS and COPD and for the heightened risk of lung cancer in COPD patients. It also identifies cross-disease miRNA signatures that hold promise as non-invasive biomarkers for early detection, risk stratification, and therapeutic targeting.

The third component translates these findings into a practical, interactive clinical tool through the development of a SHAP–LLM powered chatbot for lung disease indication. Using a structured dataset of 5,000 individuals with 17 clinical and behavioral features, an XGBoost classifier with monotonic constraints was trained to ensure biologically consistent predictions. The model achieved high accuracy, cross-validation stability, and strong performance on independent validation sets. SHAP-based interpretations were integrated into a conversational interface powered by an LLM, enabling users to query risk predictions, feature contributions, and disease mechanisms in natural language. The system automatically contextualizes SHAP explanations, interprets biomarker relevance, and supports free-text clinical queries, thereby bridging the gap between computational prediction and clinician/patient comprehension. This represents a novel fusion of clinical feature–based risk prediction, XAI-driven transparency, and LLM-powered interpretability, enabling real-time, user-friendly insights from questionnaire and physiological data. By integrating SHAP explanations with a conversational interface, the system transforms conventional tabular risk scores into intuitive, clinically meaningful guidance, with potential applications in telemedicine, early screening, patient counseling, and front-line clinical decision support.

Taken together, this thesis advances three major contributions: (i) the identification of shared multi-omics biomarkers and regulatory programs underlying the coexistence of COPD, ILD, CPFE, and related conditions; (ii) the discovery of cross-disease miRNA signatures and inflammatory axes connecting respiratory and neuroinflammatory disorders; and (iii) the development of an interpretable, LLM-augmented clinical decision-support system based on questionnaire-derived features rather than molecular biomarkers.. The findings offer a foundation for integrated biomarker panels for early diagnosis, unified therapeutic strategies targeting shared pathways, and AI-driven decision-support tools capable of enhancing clinical workflows. Future research can expand these models to include proteomics, metabolomics, longitudinal patient monitoring, and real-time wearable sensor integration. Further refinement of the chatbot into a clinically validated decision-support system may facilitate adoption in primary care and personalized respiratory healthcare. Ultimately, the study demonstrates how multi-omics analytics, explainable machine learning, and advanced language models can be combined to address long-standing challenges in understanding and managing complex respiratory disorders.

Acknowledgement

I am deeply grateful to all those who have supported and guided me throughout my Ph.D. journey. First and foremost, I would like to extend my sincere gratitude to Prof. Prateek Sharma, Vice Chancellor of DTU, Delhi, and Prof. Jai Prakash Saini, Former Vice Chancellor of DTU, for providing outstanding facilities and infrastructure that were instrumental in enabling my research.

I owe immense gratitude to my supervisor, Prof. Yasha Hasija, Head and DRC Chairperson, for her invaluable guidance, unwavering encouragement, and profound expertise. Her support, patience, and exceptional mentorship have been the cornerstone of my academic and research growth.

I am also deeply thankful to Prof. Pravir Kumar, Former Head and DRC Chairman, Department of Biotechnology, DTU, Delhi, for his insightful suggestions and valuable advice, which significantly enriched the quality of my work. My heartfelt thanks to the members of the DRC and SRC, for their continuous encouragement, constructive feedback, and invaluable guidance throughout this journey.

I am grateful to the technical staff members of the Department of Biotechnology, whose consistent support has been indispensable.

I deeply appreciate the camaraderie, collaboration, and motivation provided by my fellow Ph.D. scholars. Their engaging discussions, teamwork, and constant support, especially during challenging deadlines have made the past three years both memorable and fulfilling.

Finally, I would like to express my heartfelt gratitude to my family, whose unwavering support and love have been my greatest source of strength. I am profoundly thankful to my father, Mr. Tilak Raj, my mother, Mrs. Anita Rani, and my brother, Dr. Yash Tanwar (MBBS) for their constant encouragement, sacrifices, and belief in me. This achievement would not have been possible without them.

Nakul Tanwar

83

Chapter 1

1.1 Introduction

1.1.1 Burden of Respiratory Disorders

9

Respiratory disorders are among the leading global causes of morbidity and mortality, imposing a major health, social, and economic burden. Conditions such as Chronic Obstructive Pulmonary Disease (COPD), interstitial lung disease (ILD), combined pulmonary fibrosis and emphysema (CPFE), and lung cancer collectively account for millions of deaths annually [1]. According to the World Health Organization, chronic respiratory diseases are projected to cause over five million deaths per year by 2060. The progressive and often irreversible nature of these conditions, coupled with frequent comorbidities and late diagnoses, underscores the urgent need for improved

diagnostic tools and therapeutic strategies.

Chronic respiratory diseases (CRDs) such as COPD and lung cancer are among the most prevalent non-communicable diseases, collectively ranking as the third leading cause of death worldwide. Beyond mortality, these conditions lead to years of disability, diminished quality of life, and a growing financial strain on health systems. Monitoring the global burden of respiratory disorders is therefore central to designing targeted prevention and management strategies, and to achieving international goals such as the Sustainable Development Goal of reducing premature mortality from non-communicable diseases by one-third by 2030[2–4].

Among the various risk factors, tobacco use is one of the most prominent drivers of respiratory morbidity and mortality. Evidence from the Global Burden of Disease (GBD) (fig.1.1) study demonstrates that tobacco-related respiratory disease deaths and disability-adjusted life years (DALYs) have risen substantially over the past three decades, even as age-standardized mortality and DALY rates show a gradual decline. This apparent paradox reflects population growth and aging, which amplify the absolute burden despite relative improvements in disease control. Men and elderly individuals remain disproportionately affected, with the highest burdens concentrated in South Asia, East Asia, and Oceania. Such findings point to significant regional and socioeconomic disparities in the impact of respiratory disorders[4].

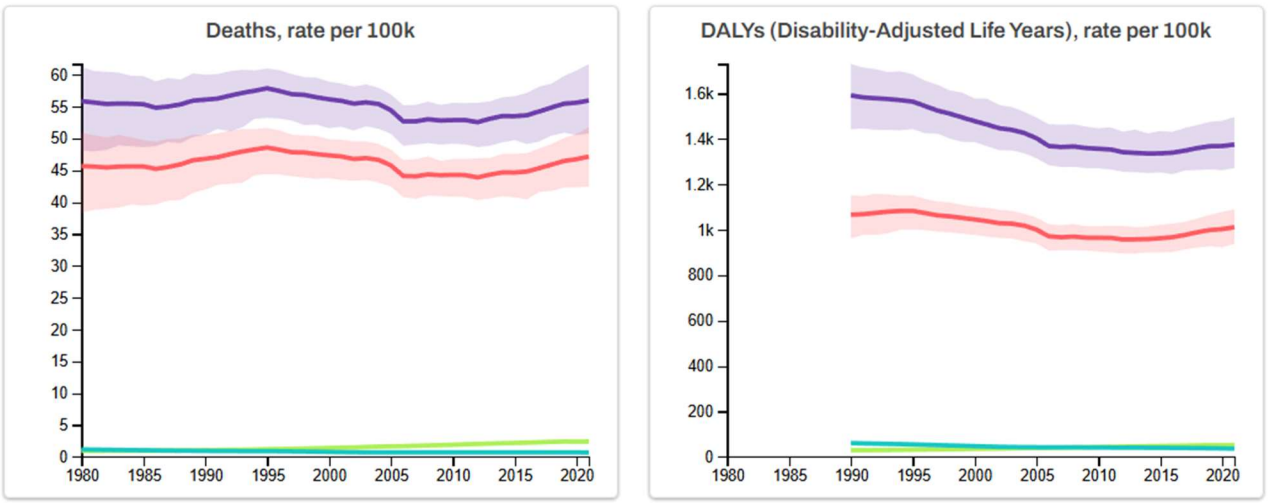


Figure 0.1.1 Global burden of chronic respiratory diseases (CRDs): age-standardized mortality rates (left) and disability-adjusted life years [DALYs] (right) per 100,000 population

.

Looking ahead, forecasts suggest a continued decline in age-standardized rates of mortality and disability from tobacco-related respiratory disorders up to 2036[2], indicating the positive effects of tobacco control policies and improved clinical management. However, the absolute number of cases and deaths will likely remain high, particularly in countries with medium levels of socioeconomic development where healthcare access is limited and tobacco use is widespread. These trends highlight the urgent need for robust, country-specific interventions that combine

prevention, early diagnosis, and effective treatment. By prioritizing high-risk populations and strengthening public health infrastructure, it is possible to reduce the overall burden of respiratory disorders and move closer to global targets for non-communicable disease control.

1.2 Research Rationale and Problem Identification

Respiratory diseases such as COPD, ILD, CPFE, and lung cancer contribute substantially to global morbidity and mortality, yet their diagnosis and prognosis remain challenging. Although imaging, pulmonary function tests, and histopathology are standard diagnostic modalities, they often detect disease only at advanced stages, limiting therapeutic options and survival outcomes. Hence, there is an urgent need for novel biomarkers and integrative frameworks that can provide early, accurate, and clinically interpretable insights into disease onset, coexistence, and progression.

In recent years, molecular regulators, particularly non-coding RNAs such as microRNAs (miRNAs), have emerged as promising candidates for biomarker discovery. Dysregulated miRNA–mRNA networks play crucial roles in inflammation, fibrosis, apoptosis, and oncogenesis. Circulating miRNAs, owing to their stability in body fluids, present themselves as non-invasive diagnostic and prognostic tools. However, despite compelling evidence, their clinical translation remains incomplete, with limited integration into routine patient care.

Further complexity arises from the coexistence of respiratory diseases with systemic or overlapping conditions. A well-established example is the frequent concurrence of COPD with lung cancer or ILD, which reflects shared molecular pathways such as oxidative stress, telomere attrition, and mitochondrial dysfunction. Additionally, systemic autoimmune disorders like Multiple Sclerosis (MS) are increasingly recognized to have inflammatory and molecular crosstalk with respiratory diseases. Recent meta-analyses show that people with MS have a significantly higher prevalence of COPD compared to healthy controls, indicating that MS-driven systemic inflammation and immune dysregulation may predispose to chronic lung pathology. These overlapping disease patterns remain insufficiently characterized in terms of their molecular networks, shared biomarkers, and clinical implications.

The rapid growth of multi-omics datasets (genomics, transcriptomics, proteomics, clinical data) provides unprecedented opportunities to unravel these complex disease mechanisms. Traditional machine learning (ML) and deep learning (DL) models have demonstrated success in biomarker identification, disease classification, and outcome prediction. However, their “black-box” nature significantly hampers trust, adoption, and clinical translation. Clinicians demand not only accurate predictions but also transparent, biologically interpretable explanations of why specific predictions are made.

This necessitates the integration of Explainable Artificial Intelligence (XAI). Methods such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) allow researchers to quantify the contribution of individual molecular features (e.g., miRNAs, mutations, proteins) to model outputs. When coupled with Large Language Models (LLMs), these tools open new avenues for clinical translation by transforming complex results into human-readable, interactive insights.

In this context, a SHAP-LLM powered chatbot for lung disease indication offers a transformative solution. Such a system would integrate miRNA biomarkers, multi-omics datasets, and XAI-driven outputs into an LLM interface, enabling clinicians, researchers, and patients to query disease risks, mechanisms, and treatment implications in a conversational manner. This approach addresses both the scientific need for transparent biomarker discovery and the practical need for

accessible, interpretable decision-support tools in respiratory healthcare.

Based on these identified gaps and opportunities, the present study has been designed with the following objectives.

1.3 Objectives

1.3.1 To investigate the coexistence and shared molecular mechanisms among respiratory diseases through integrative multi-omics and regulatory network analyses.

1.3.2 To explore systemic inflammatory crosstalk between neuro-inflammatory disorders and respiratory diseases.

1.3.3 To develop a SHAP-LLM powered chatbot for lung disease indication, which leverages explainable machine learning models and large language models.

1.4 Literature Review

1.4.1 Chronic Obstructive Pulmonary Disease (COPD)

COPD, according to (Global Initiative for Chronic Obstructive Lung Disease) GOLD 2023, is a broad lung disorder that comprises severe respiratory problems (dyspnea, cough, expectoration) triggered by bronchitis, bronchiolitis, and the alveolar malformations (emphysema) that produce chronic and frequently progressive airflow obstruction [16]. This condition is an inflammatory condition involving the airways, lung parenchyma, and pulmonary vasculature. The process is thought to involve oxidative stress and protease-antiprotease imbalances. Emphysema describes one of the structural changes seen in COPD where there is destruction of the alveolar air sacs (gas-exchanging surfaces of the lungs) leading to obstructive physiology. In emphysema, an irritant (e.g., smoking) causes an inflammatory response. Neutrophils and macrophages are recruited and release multiple inflammatory mediators. Oxidants and excess proteases leading to the destruction of the air sacs. The protease-mediated destruction of elastin leads to a loss of elastic recoil and results in airway collapse during exhalation[17].

1.4.2 COPD and Lung Cancer

Lung cancer is a leading cause of death in people with COPD [18] and National Lung Screening Trial found that the probability of acquiring lung cancer was 2.15 times higher in those with COPD than in those without COPD [19,20]. Both lung cancer and COPD have smoking as a common risk factor, which is why they frequently occur together. Since lung cancer is so fatal, it is rational to assume that the percentage of COPD patients with lung cancer-related death is indicative of the prevalence of lung cancer among COPD patients[21].

Lung cancer and COPD share several risk factors, including genetic predisposition, telomere shortening, mitochondrial dysfunction, and accelerated aging, suggesting that they may be different manifestations of the same disease. Antioxidants such anti-proteases, DNA repair mechanisms etc., may be able to handle the damage caused by smoking in most smokers. However, these fail in both cancer and COPD, leading, respectively, to mutations and cancer or to cell and protein damage that is too extensive to repair[22].

1.4.3 Non-coding RNAs and miRNA–mRNA Regulatory Networks

MicroRNAs (miRNAs) noncoding RNA molecules with 18 to 25 nucleotides. By either inhibiting

the production of proteins or speeding up the breakdown of mRNAs, both of which affect gene expression [23,24]and hence take part in a cell division, differentiation, apoptosis, disease onset and progression, and many other biological processes. According to various studies, multiple disorders, including COPD, have been linked to miRNA dysregulation. miRNAs (particularly circulating miRNAs) may become an integral part of the COPD assessment programs during future preventive treatment because they can also be articulated persistently in body fluids[23]. Numerous research analyzing tissue-derived miRNA profiles have demonstrated that abnormal miRNA profiles are linked to a wide variety of human disease disorders. (e.g., lung and breast cancer[25,26], or glioblastoma (GBM)[27]). Due to their exceptional stability, these small nucleic acids are being considered as potential diagnostic markers. Serum miRNA patterns have been studied extensively for many diseases, (TB, breast, prostate, and ovarian cancer[28]).

The Dicer process produces a pair of RNA that is approximately 21 nucleotides in length. When the miRNA has reached maturity, one of its two strands is placed in RISC. The "star" thread (the other strand) frequently breaks. miR-145* is appended with an asterisk (*) to indicate that it is the star strand. However, for some miRNAs, both strands may reach RISC at roughly the same rate. The "5p" strand originates from the 5' end of the stem-loop, while the "3p" strand originates from the 3' end. RISC loading may in fact strongly support the incorporation of a single strand. However, recent research utilizing next-generation sequencing (NGS) has revealed that for nearly all miRNA families, only a limited number of star strands are loaded. Even more complicated, the names of some miRNAs vary based on the cell type or biological state. Because of these factors, the use of random mature/star names is diminishing, and the use of 5p/3p name methods is increasing [29] .

Non-coding RNAs (ncRNAs) play crucial roles in regulating gene expression, among which microRNAs (miRNAs) are one of the most extensively studied classes. miRNAs are short non-coding RNAs (18–25 nucleotides) that regulate post-transcriptional gene expression by binding to complementary messenger RNAs (mRNAs), leading to mRNA degradation or translational repression. Through these interactions, miRNA–mRNA regulatory networks orchestrate essential biological processes including cell proliferation, differentiation, apoptosis, immune regulation, and inflammatory signaling.

Aberrant miRNA–mRNA interactions have been implicated in the pathogenesis of COPD, lung cancer, and other chronic diseases. For instance, integrative analyses have revealed dysregulated miRNA–mRNA networks associated with COPD phenotypes and disease progression[30].Similarly, combined transcriptomic and miRNA analyses in lung tissue have identified cross-disease regulatory signatures in COPD and ILD[31]. Beyond respiratory conditions, comparative studies have shown that COPD and non-small cell lung cancer share common miRNAs and target mRNAs in oncogenic and inflammatory pathways[32].
Recent multi-omics studies have even integrated miRNA–mRNA–protein networks, providing comprehensive insights into COPD pathogenesis [33]. These findings underscore the potential of miRNA–mRNA networks not only as mechanistic insights but also as clinically actionable biomarkers when coupled with explainable machine learning approaches.

1.4.4 Interstitial Lung Disease and CPFE

People with connective tissue disease associated with interstitial lung disease (CTD-ILD) were more likely to get lung cancer than people with ILD alone; the rates were 27% for men and 28%

for women. Lung cancer was 3.2 times more likely in men aged 40 to 49 who had both CTD and ILD than in men aged 40 to 49 who only had ILD[34]. It is known that people with combined pulmonary fibrosis and emphysema (CPFE) have a higher risk of death than people with COPD alone, but it is not clear what the risk is for people with idiopathic pulmonary fibrosis (IPF). Exacerbation, lung cancer, and pulmonary hypertension are all signs that a person with CPFE is not likely to get better. People with COPD often have interstitial lung abnormalities, which could be an early or mild form of ILD. These abnormalities are a sign of a bad outlook. Different ideas have been put forward about how CPFE affects the body. Biomarker studies suggest that this process may be more closely linked to the development of IPF than to COPD or emphysema. People who have CPFE should be told to stop smoking and get regular lung function tests and pulmonary therapy may also help. CPFE patients may benefit from several medications and surgery methods, but more research is needed[35].

1.4.5 Multiple Sclerosis and Systemic Inflammatory Crosstalk

Although primarily a neurological disorder, Multiple Sclerosis (MS) provides an important perspective on systemic inflammation relevant to respiratory health. MS is a chronic autoimmune disease of the central nervous system (CNS), characterized by demyelination, axonal degeneration, and persistent neuroinflammation ([36];[37]). The disease is driven by aberrant immune responses, including infiltration of autoreactive T cells, activation of B cells, and inflammatory signaling cascades that damage myelin and neuronal structures ([38];[39]).

In recent years, miRNA dysregulation has been recognized as a central feature of MS pathology. Specific miRNAs regulate T-cell differentiation, Th17 responses, and signaling pathways such as NF-kB, JAK/STAT, and PI3K-Akt, which are also implicated in COPD and lung cancer [40]. For example:

- miR-155 and miR-326 are overexpressed in MS and promote pro-inflammatory T-cell responses[41].
- miR-146a acts as a negative regulator of inflammation but is often dysregulated across autoimmune and respiratory disorders[42].

Recent evidence suggests that MS is associated with an elevated risk of developing chronic lung diseases, particularly COPD. A 2024 systematic review and meta-analysis of 40 studies involving 287,702 people with MS (pwMS) found pooled prevalences of asthma and COPD at 5.97% and 3.03%, respectively. Importantly, while asthma did not show a significant association with MS (OR 1.14; 95% CI 0.76–1.71), COPD was significantly more common in pwMS, with an odds ratio of 1.28 (95% CI 1.11–1.47, $p < 0.01$) compared to healthy controls[43]. These findings highlight that pwMS are more susceptible to COPD, whereas asthma prevalence appears coincidental rather than causally linked. The coexistence of COPD and MS may therefore reflect both shared systemic inflammatory pathways and neurogenic contributions to respiratory compromise(Tabel.1).

Table1.0.1. Overlap between MS, COPD, Lung Cancer and ILD

Disease / Condition	Global Burden & Mortality	Key Risk Factors	Molecular / Pathophysiological Features	Diagnostic Challenges	Overlap / Crosstalk
COPD	3.2 million deaths	Tobacco smoking, air	Oxidative stress, protease–	Often detected late; spirometry	Shares risk and molecular

	annually; major contributor to CRDs [5]	pollution, occupational exposure[6]	antiprotease imbalance, alveolar destruction [7]	underused in low-resource settings[7]	pathways with lung cancer, ILD; linked with systemic inflammation[8]
Lung Cancer	Leading cause of cancer death worldwide; 2.1× risk in COPD patients [9]	Smoking, aging, genetic predisposition	Telomere shortening, mitochondrial dysfunction, DNA repair failure → mutations	Late-stage detection; poor survival outcomes despite screening [9]	Frequently co-occurs with COPD; shares inflammatory and molecular pathways
ILD (incl. CTD-ILD)	Variable prevalence; higher mortality when combined with lung cancer [10]	Autoimmunity, smoking, environmental exposures	Fibrosis pathways: TGF-β, ECM remodeling [11]	Often diagnosed late; overlap with COPD hampers clarity [12]	ILD patients have ↑ risk of lung cancer; overlaps with CPFE [12]
CPFE	High mortality; worse prognosis than COPD or IPF alone	Smoking, aging	Mixed features of fibrosis + emphysema; vascular remodeling [12]	Difficult to classify; limited targeted therapies	Biomarker signatures closer to IPF; high risk of lung cancer [12]
Multiple Sclerosis (MS)	2.8M people worldwide; COPD prevalence ~3% in pwMS [13]	Autoimmunity, systemic inflammation	Neuroinflammation, demyelination, dysregulated miRNAs (miR-155, miR-326, miR-146a) [13,14]	Primarily neurological focus → respiratory links underexplored[13]	Systemic inflammatory crosstalk predisposes pwMS to COPD[13]
Shared Pathways	Collectively >5M CRD deaths projected annually by 2060 [15]	Smoking, pollution, genetic susceptibility, aging	Dysregulated miRNA–mRNA networks; oxidative stress; apoptosis; fibrosis; immune dysregulation [12–14]	Conventional tests (PFTs, imaging, histopathology) lack early detection power	Common miRNA signatures across COPD, lung cancer, ILD, and MS; highlight biomarker discovery need [12–14]

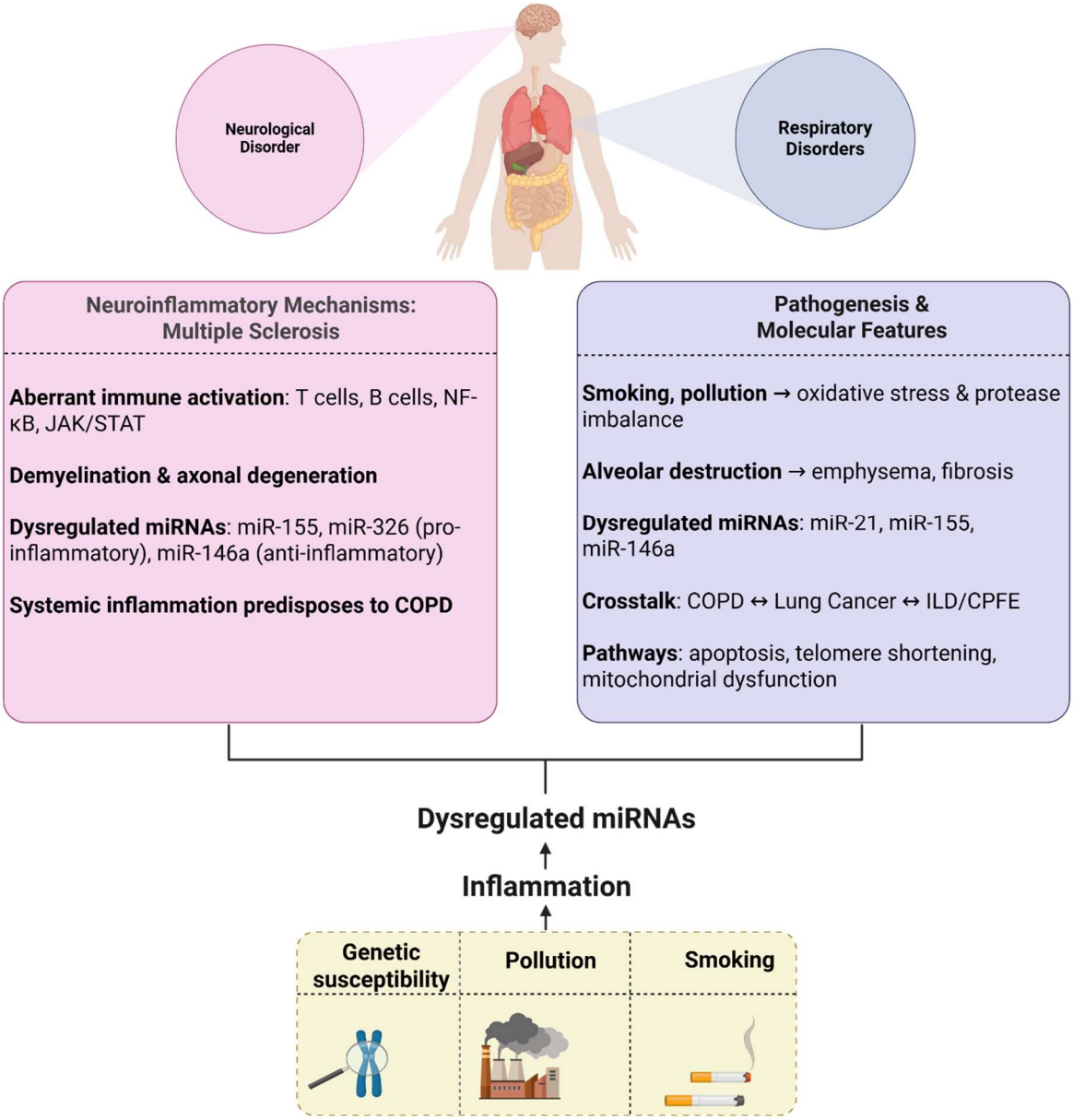


Figure 1.0.2 Integrated overview of shared neuroinflammatory and respiratory pathogenic mechanisms across Multiple Sclerosis, COPD, ILD, and related disorders, highlighting immune dysregulation, environmental triggers, and common miRNA-mediated inflammatory pathways.

1.4.6 Artificial Intelligence and Explainable Models in Respiratory Research

The growing complexity of multi-omics data in COPD, ILD, lung cancer, and MS necessitates advanced computational approaches capable of managing high-dimensional datasets and uncovering hidden regulatory signatures. In recent years, bioinformatics integration with

healthcare has generated vast amounts of biological and clinical data. These large-scale datasets serve as the foundation for artificial intelligence (AI) approaches, where machine learning (ML) and deep learning (DL) algorithms can systematically extract patterns, identify biomarkers, and predict disease trajectories with greater accuracy and efficiency.ML has come to be as a valuable resource for detecting and forecasting chronic illnesses such as diabetes[44],hypertension[45], cholesterol[46], sleep disorders[47], CVDs[48], stroke[49], COVID-19[50] etc. In this research, COPD is of particular interest. Many investigations into this disease have been carried out with the use of supervised ML models[51]. In a study to determine which traits are crucial for case identification, the authors utilized Random Forest[52]. In [53] individuals admitted to the ED for asthma or COPD exacerbation were studied, and the authors assessed the accuracy of multiple ML methods for predicting two clinical outcomes. In a study[54] researchers systematically investigate the feasibility of using deep and non-deep ML models to forecast the readmission probability of COPD patients. In study [55], to improve the accuracy of COPD predictions, the authors examined three popular ML techniques (decision tree, naive Bayes, and Bayesian network) using the receiver operating characteristic (ROC) measure.

However, a critical limitation of conventional ML and deep learning approaches lies in their "black-box" nature, which restricts clinical adoption due to a lack of interpretability. Explainable Artificial Intelligence (XAI) methods, such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), have been developed to overcome this barrier. These methods not only improve trust in model predictions but also provide biologically meaningful insights by quantifying the contribution of individual features (e.g., genes, miRNAs, proteins) to outcomes[56].

More recently, Large Language Models (LLMs) have emerged as powerful tools capable of contextualizing complex biomedical knowledge and enabling natural language interactions with high-dimensional data. When combined with XAI techniques like SHAP, LLMs can translate computationally intensive results into clinically interpretable, conversational insights. Such integration holds the potential to revolutionize biomarker discovery, patient stratification, and clinical decision support in respiratory research. Thus, the convergence of multi-omics data, XAI, and LLM-driven interpretability lays the foundation for the present study.

Chapter 2

2.1 **Objective 1: To investigate the coexistence and Shared molecular mechanisms among major respiratory diseases through integrative multi-omics and regulatory network analyses.**

2.1.1 **Problem Statement and Rationale**:

Respiratory diseases rarely occur in isolation. Increasing evidence suggests that coexistence of

turnitin Page 31 of 122 - Integrity Submission Submission ID trn:oid:::27535:122649592

multiple respiratory conditions such as COPD, ILD, lung cancer, and CPFE is not merely coincidental but reflects shared risk factors, overlapping molecular pathways, and convergent pathological processes. Epidemiological studies have shown that patients with COPD have a significantly higher risk of developing lung cancer, while those with ILD, particularly idiopathic pulmonary fibrosis (IPF) or connective tissue disease-related ILD (CTD-ILD), demonstrate elevated incidences of emphysema and malignancy. Similarly, CPFE represents a distinct clinical entity that merges features of both emphysema and fibrosis, with worse outcomes compared to either disease alone.

The coexistence of these diseases poses unique diagnostic and therapeutic challenges. For instance, airflow obstruction in COPD may mask fibrotic changes of ILD on spirometry, while emphysema may complicate radiological interpretation in IPF or CPFE. Co-morbid lung cancer in COPD patients often leads to delayed detection, reduced treatment options, and poorer survival outcomes. From a mechanistic perspective, these diseases share common pathological hallmarks, including chronic inflammation, oxidative stress, epithelial injury, aberrant repair, extracellular matrix remodeling, and dysregulated immune responses. At the molecular level, evidence points to overlapping gene expression patterns and regulatory networks, suggesting that coexistent respiratory diseases may represent different manifestations of interconnected biological processes rather than entirely separate entities.

Understanding these interrelations is clinically significant. Patients with coexistent respiratory diseases generally have poorer prognoses, higher hospitalization rates, and reduced quality of life compared to those with single conditions. Moreover, the identification of shared biomarkers and pathways could facilitate earlier diagnosis, risk stratification, and the development of integrated therapeutic strategies targeting common mechanisms.

Therefore, the investigation of coexistence among respiratory diseases is not only relevant for advancing mechanistic insights but also holds direct translational value for personalized medicine. By adopting integrative approaches including transcriptomics, bioinformatics, and explainable machine learning frameworks this objective aims to systematically unravel the complex interplay between these diseases and to inform the design of more effective diagnostic and therapeutic interventions.

2.1.2 Background

COPD and ILD are clinically distinct conditions, yet they may coexist in some individuals[57]. This overlap is often attributed to common risk factors such as tobacco use. This specific coexistence is known as CPFE, characterized by the presence of emphysema, a form of COPD, primarily in the upper lung fields, along with the widespread presence of ILD[57]. In patients with COPD, interstitial lung abnormalities (ILA) are frequently observed, which might indicate an early or mild stage of ILD. ILAs are characterized by increased lung densities found in chest computed tomography (CT) scans of patients without a previous ILD diagnosis. Although the clinical features of these lung changes resemble those of ILD, they are generally less severe. This similarity suggests that ILAs could represent initial or mild forms of ILD. Approximately 13.5% of individuals with COPD exhibit these changes [58] .

The interplay between inflammation and the pathogenesis of chronic pulmonary diseases such as

CPFE, COPD, and ILD underscores a complex web of biological mechanisms. In CPFE, the uncertain pathogenesis is thought to involve various cytokines and signaling pathways, with overexpression of inflammatory mediators like PDGF, TNF-α, and TGF-β being linked to emphysema and fibrosis lesions [59]. Similarly, in COPD, epithelial cells play a crucial role in mediating inflammation triggered by inhaled toxins and microorganisms, leading to the production of cytokines, chemokines, and reactive oxygen species (ROS). Notably, smoking has been shown to induce CXCL14 expression by human epithelial cells, correlating with lung cancer development, as elaborated in another study[60]. ILD and lung fibrosis, characterized by varying fibrosis and inflammation degrees, highlight the importance of persistent low-grade inflammation. This persistent inflammation leads to cytokine, growth factor, and tissue proteinase imbalances, crucial for understanding these diseases' pathogenesis [61]. Collectively, these insights reveal the intricate role of inflammation in the progression of these pulmonary diseases and highlight potential areas for further investigation.

Discovering biomarkers for COPD and ILD can serve dual purposes: it aids in diagnosing these diseases and uncovers diagnostic markers for CPFE. This knowledge is crucial for developing innovative treatments that can improve patient survival. Microarray technology, known for its ability to simultaneously assess the expression levels of thousands of genes, has established itself as a valuable resource in the identification of biomarkers [62]

The complex interaction between ILD, COPD , and lung cancer poses considerable clinical difficulty due to their common risk factors and overlapping pathophysiologies. Smoking is a significant risk factor that not only makes individuals more likely to develop COPD and lung cancer, but also raises the risk of lung cancer in patients with ILD. The intricate pathways connecting fibrogenesis ILD to the development of lung cancer, in addition to the aggravating impact of COPD-induced oxidative stress and persistent inflammation [63,64], highlight a critical area of study. A study that investigated people with ILD, COPD, and metastatic cancer in the ICU gave us important information about their palliative care needs and results. The study showed that patients with COPD or ILD were less likely to get complete palliative care than those with metastatic cancer, even though they stayed in the ICU and hospital longer. This includes important things like figuring out how much pain someone is having, which shows a big difference in care and preparation for making end-of-life choices [65]. These results shed light on the bigger problem of how to care for people who have these two conditions at the same time, especially the need for better hospice care plans that consider how chronic lung diseases and cancer develop in different ways [65].

ML has come to be a valuable resource for detecting and forecasting chronic illnesses such as diabetes, hypertension, cholesterol, sleep disorders, CVDs, stroke, COVID-19 etc. In this research, COPD and ILD are of particular interest. Many investigations into this disease have been carried out with the use of supervised ML models. A study done gene expression biomarkers for COPD and ILD comprises mRMR (minimal Redundancy Maximal Relevance) and incremental feature selection (IFS) technique to highlight 38 gene biomarker that were fed to a SVM classifier [62]The fact that these techniques can be used to highlight gene in progression of condition, but these techniques lacks when it comes to explaining the ML classifiers.

To fill this gap, we analyzed expression dataset of 220 COPD, 254 ILD and 108 control with

accession GSE47460 and built an RF classier model for multiclass classification for classes Control, COPD and ILD. As the data is imbalanced, we applied SMOTE to avoid any bias towards majority class and achieved 88.1% accuracy. Following that, we implemented SHAP (python base library) to visualize the features (genes) that have impacted the model' performance. A total of 20 common genes were found to have contributed to models' performance and in progression of COPD and ILD. These genes can assist not only in understanding the molecular mechanisms of both the conditions but also can act as novel biomarkers for CPFE.

2.1.3 Methodology

2.1.3.1 Data collection:

The focus of this research is to investigate DEGs to support the coexistence of COPD and ILD and identify significant biomarkers through and machine learning approaches that can assist in treating the conditions and understanding their interrelated molecular mechanisms to get the insights of CPFE. To make our study more clinically significant, we have taken several measures: Considering the significant influence of inflammation on both disorders, inflammation was incorporated as a crucial aspect of our data mining step. Making sure that the COPD samples we used for our study have their relevant GOLD staging (*"0-At Risk", "1-Mild", "2-Moderate", "3-moderate" and "4-Very Severe COPD"*) details was a major focus of our dataset curation. To ensure that our findings are applicable in the clinic, gold staging must be included since it provides a commonly accepted way to evaluate the severity of COPD using spirometric measurements.
Microarray data has been acquired from the publicly available database Gene Expression Omnibus (GSE47460) [66] and with a pair of groups analyzed using distinct microarray platforms (GPL6480 and GPL14550). The dataset has been searched using the keywords 'COPD' and 'ILD'. The total RNA extracted from the whole lung homogenates of subjects undergoing thoracic surgery. These subjects were diagnosed as being controls or having ILD or COPD. The total sample consisted of 582 subjects, of whom 254 were ILD patients, 220 had COPD, and 108 were healthy controls.

2.1.3.2 Data preprocessing:

The raw data has been acquired from the GEO database, which needs to be normalized for ML to perform well. We decided to do quantile normalization for our dataset. We conducted a log2-transformation and quantile normalization on the expression data to generate boxplots and expression density plots for both the normalized and non-normalized datasets. The resulting dataset includes 15,261 genes. The dataset includes samples of COPD, ILD, and CONTROL, which were further used for the multi-class classification.

2.1.3.3 Machine learning:

Recent advances in computing capacity (such as massive parallel computing and GPUs) and data analytics and AI techniques (such as ML and DL) have created new opportunities in biomedicine and other sectors of data science. So, we chose to implement machine learning on our dataset. As our data is imbalanced, we chose to use the synthetic minority oversampling technique (SMOTE)[67] . Many machine learning algorithms have been shown to struggle with classification due to the class-imbalance problem. Utilizing this imbalance data can lead to biasness of model towards the majority class. To avoid that we have performed SMOTE technique. After that, we also performed cross validation using Stratified Fold (splits=5) step for model validation. Furthermore, the data is randomly split into a 90:10 ratio (90% training and 10% testing).

The random forest classifier (RF) combines the results of multiple classification trees into a single result. As the classifier can operate on classification as well as regression tasks, we chose a random forest classifier for our model, which was trained on a training dataset for multi-class classification between COPD, ILD, and CONTROL. The RF algorithms are available for numerous languages, and Python was the most used language for RF classifiers, due to which we chose Python to build the model. The performance of the model is judged on the bases of accuracy, confusion matrix, and classification report.

2.1.3.4 Explaining the trained model using SHAP:

To explain the trained RF classifier, we utilized the XAI algorithm SHAP, which is a library developed in Python. XAI provides a lens for the RF classifier to make the model more transparent and explain the model's performance. This is accomplished by emphasizing the features (genes) that have impacted the model's ability to classify between the classes COPD, ILD, and CONTROL. To provide a further explanation of the function that genes play in predicting all three classes, a SHAP bar plot was constructed. Following that, a SHAP-based local summary plot was developed to facilitate further interaction for each individual class, namely COPD, ILD, and CONTROL.

2.1.3.5 Relevance of identified genes:

To gain a more comprehensive understanding of the identified genes, we have employed a multifaceted approach. SRplot [68]is being incorporated for enrichment analysis of identified genes. After that the, focus was on evaluating the impact of these genes on model performance and their differential expression patterns. For this purpose, we employed IDEP 1.0 [69]for generating heatmaps of the top genes identified through SHAP analysis. These visual representations were crucial in elucidating the distinct expression profiles of these genes. Furthering our analysis, we reintroduced these top-performing genes into our model. This step was pivotal in assessing their influence on the model's accuracy and in generating a detailed confusion matrix.

To predict the functions of these genes, we turned to GeneMANIA[70]. This tool allowed us to explore the relationships and networks between the genes and their targets, facilitating the identification of gene clusters with similar characteristics or functions. Additionally, to delve into the role of these top genes at the single-cell level, we utilized the Single Cell Portal[71]. This analysis was complemented by comprehensive enrichment analyses, including gene ontology and pathway analysis. Lastly, we have incorporated miRNet 2.0 online tool [72] to create and assess the networks of interactions between miRNA and their targets(fig.2.1).

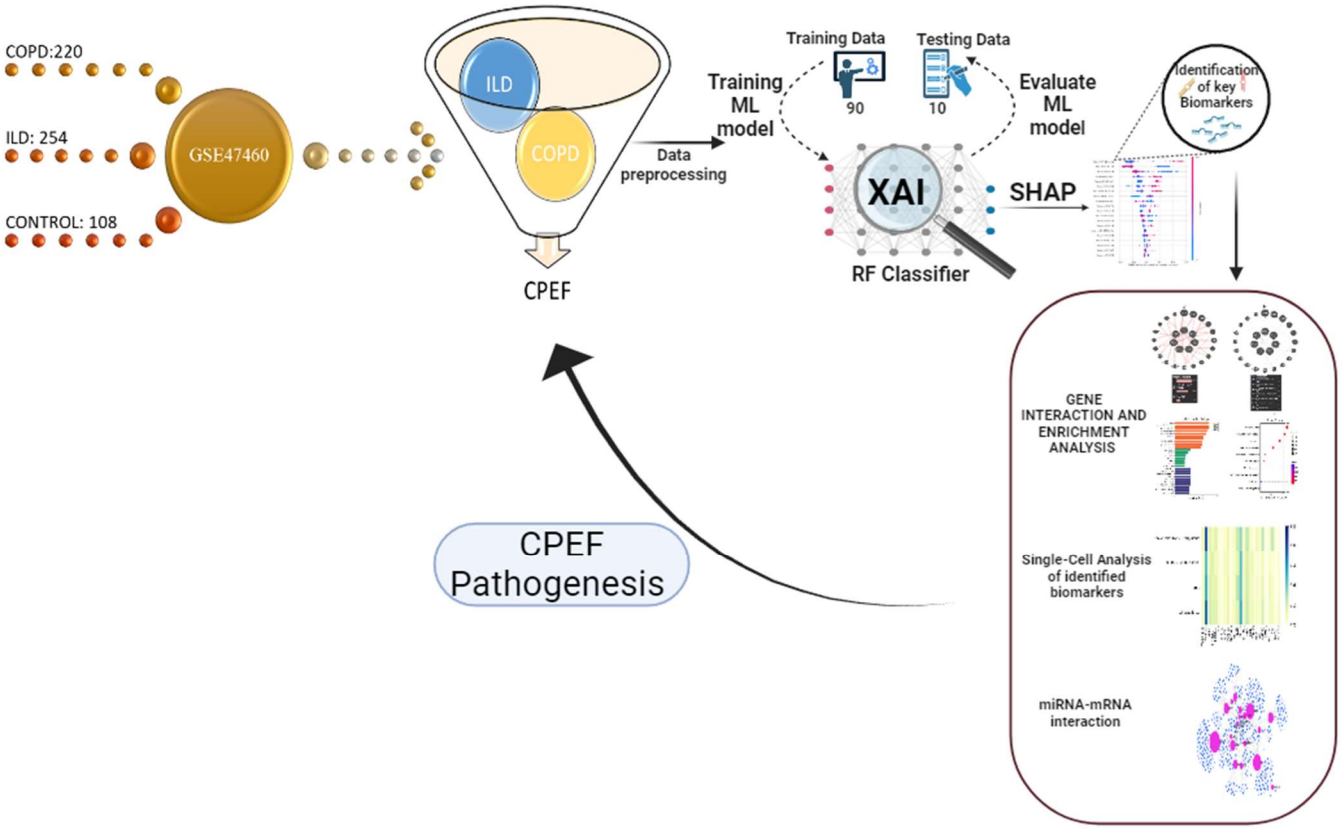


Figure 2.1Representation of proposed Model

2.2 Objective 2: To explore systemic inflammatory crosstalk between neuro-inflammatory disorders and respiratory diseases.

2.2.1 Problem Statement and Rationale

Although COPD, lung cancer, and MS are clinically distinct, there is growing evidence that they share common inflammatory and molecular pathways. Epidemiological studies highlight a higher prevalence of COPD among MS patients, and COPD itself is strongly associated with the development of lung cancer. Smoking, a common environmental trigger, further strengthens this triad of disease interactions. Despite this, the precise mechanisms of molecular convergence remain poorly characterized.

Current research has identified miRNAs as key regulators of inflammatory processes, capable of modulating immune cascades and maintaining homeostasis. Dysregulation of these miRNAs contributes to immune dysfunction, epithelial damage, fibrosis, and carcinogenesis. However, the role of miRNA-mediated regulation in linking neurological and respiratory disorders is still inadequately explored.
Clinically, the coexistence of these disorders poses challenges for early diagnosis, risk stratification, and therapeutic decision-making. Patients often present with overlapping symptoms and poorer prognoses, and the absence of shared biomarker panels limits precision medicine strategies.

Thus, the problem lies in the insufficient integration of omics data across neuro-inflammatory and respiratory diseases, and the lack of interpretable computational models to identify common biomarkers and pathways.

The rationale for this objective is that by applying machine learning and explainable AI to cross-disease miRNA expression datasets, it becomes possible to uncover regulatory networks that drive multiple diseases simultaneously. This integrative approach not only improves our understanding of inflammatory crosstalk but also opens avenues for diagnostic markers and therapeutic targets applicable across conditions like COPD, lung cancer, and MS.

2.2.2 Background

Lung cancer, Multiple sclerosis (MS), and chronic obstructive pulmonary disease (COPD) all have distinct clinical presentations, yet they are all related by persistent inflammation. Due to the weakening of respiratory muscles, neurogenic pulmonary edema, and decreased coughing efficiency, MS, a neuroinflammatory disorder, increases the risk and of aspiration lung infections, which raises morbidity and mortality [73]. Parallel to MS, COPD is characterized by persistent airway inflammation and has been associated with an increased risk of developing MS, especially in younger individuals diagnosed before to the age of 60 [74]. Studies have shown a dose-dependent association between smoking exposure and MS risk [75]. Smoking is a common environmental trigger for MS and also plays a significant role in the development of COPD and lung cancer. According to recent research, there may be a causal relationship between MS and lung cancer due to similar molecular mechanisms between the two ailments [76]. This suggests that these inflammatory pathways interact beyond the scope of individual diseases. This mutual dependence of diseases on immune system dysregulation points to shared pathways that may facilitate the development of these diseases.

Building on these common pathways, it becomes evident that immunological dysfunction and chronic inflammation are the governing factors for lung cancer, MS, and COPD, despite their targeting of distinct physiological systems. According to[77], the processes of demyelination, neuro-axonal damage, and gliosis in MS are driven by interactions among T cells, B cells, and

myeloid cells. This highlights the significance of therapies that modify immune responses from a pro-inflammatory to an anti-inflammatory state. Similarly, there is a pathological overlap between lung cancer and COPD, two diseases that are strongly associated with smoking, which further establishes their link. Not only is COPD common in lung cancer patients, it also acts as a risk factor for the illness. In both cases, there is a breakdown of the extracellular matrix, NF-κB activation, epithelial-to-mesenchymal transition, and immunological dysfunction. Smoking aggravates these processes further by promoting the development of the disease through angiogenesis, autophagy disturbances, and epigenetic alterations [78].

MicroRNAs, or miRNAs, regulate immune responses and modify inflammation at several levels, and this is at the core of these interrelated disorders. Through feedback loops, miRNAs adjust the inflammatory cascade, controlling the phases of inflammation that precede, intensify, and resolve. According to[79], the biogenesis of miRNAs is closely controlled during inflammatory processes, altering their synthesis, processing, and stabilization to maintain immunological homeostasis. This miRNA-mediated regulation offers a promising framework for identifying common biomarkers and formulating integrated therapeutic strategies that target the shared molecular mechanisms underlying MS, COPD, and lung cancer. Although these diseases impact distinct organ systems—neurological, respiratory, respectively they are all fundamentally shaped by chronic inflammation and immune dysregulation. We intentionally selected MS, COPD, and lung cancer to explore how systemic inflammatory processes manifest across different physiological contexts, with the hypothesis that a unified omics-based and XAI-driven analytical approach could reveal converging regulatory miRNA signatures.
By adopting this broader systems-level perspective, our goal was to move beyond disease-specific analyses and uncover shared molecular pathways that are often obscured in conventional studies. This cross-disease integration not only deepens our understanding of disease intersectionality but also expands the potential for biomarker discovery and the development of therapies that can target multiple inflammatory conditions simultaneously.

Building on this rationale, we recognize that the early diagnosis and effective treatment of chronic diseases such as MS, COPD, and lung cancer remain particularly challenging due to their complex and often overlapping molecular mechanisms. Although chronic inflammation is a well-established hallmark across these conditions, the precise role of miRNAs in modulating disease onset and progression remains only partially understood. Our study seeks to bridge this gap by systematically exploring miRNA-mediated regulatory networks across these diseases, with the goal of uncovering shared molecular signatures that may serve as clinically actionable biomarkers or therapeutic targets.
This need becomes even more urgent in light of growing evidence linking MS, COPD, and lung cancer to shared environmental triggers, immune dysregulation, and convergent inflammatory pathways. Identifying common biomarkers across these diseases could facilitate the development of integrated diagnostic tools and unified therapeutic strategies—an essential step toward personalized, cross-disease clinical management. To advance this goal, we turned to machine learning (ML) techniques, which are well-suited to uncovering complex, nonlinear patterns within high-dimensional biomedical datasets. To advance this goal, we turned to machine learning (ML) techniques, which are well-suited to uncovering complex, nonlinear patterns within high-dimensional biomedical datasets.
However, traditional ML models, though powerful in detecting complex patterns within high-

dimensional biomedical data, often lack transparency and biological interpretability. To overcome this limitation, we implemented explainable artificial intelligence (XAI) methodologies, enabling us to disentangle the model's decision-making process and extract biologically meaningful insights. This interpretability not only strengthens the clinical relevance of our findings but also enhances confidence in the translational potential of computational predictions.

To investigate shared molecular signatures across COPD, lung cancer, and MS, we utilized publicly available peripheral blood miRNA expression data from the Gene Expression Omnibus (GEO) database (GSE61741)[80] . This dataset includes 1,049 samples out of which 237 samples consisting 47 from COPD patients, 73 from individuals with lung cancer, 23 with MS, and 94 healthy controls.

The research identified 20 important miRNAs, such as hsa-let-7c and miR-223, that modulate relevant to inflammation, immunology, and disease development. These miRNAs revealed similar biological pathways and identified prospective biomarkers for diagnostic and therapeutic techniques in MS, COPD, and lung cancer. Furthermore, these miRNAs serve as key regulators for the six inflammatory genes: CCL2, CCL5, IL10, IL6, ITGB3, and MYC. Our study integrates ML with XAI techniques, specifically SHapley Additive exPlanations (SHAP), to bridge data-driven predictions and biological interpretability. This approach provides a novel framework for identifying cross-disease biomarkers, with particular emphasis on the immune dysregulation linking neurological and respiratory conditions. A comprehensive list of abbreviations used in this study is provided in Supplementary Table S1.

2.2.3 Methodology

2.2.3.1 Data Collection:

The miRNA profile array data has been acquired from NCBI's GEO database with the accession number GSE61741[80]. The data consists of peripheral blood profiles of individuals with a range of disorders, and those with healthy people. The overall design of the dataset comprises 1049 samples, of which 94 are controls, 15 are long-lived individuals, and 940 patients have been checked for the full set of miRNAs with respect to miRNAs V12–14. COPD, lung cancer, MS and healthy controls patients were chosen, and the rest were discarded, leaving 237 samples (Control = 94, COPD = 47, Lung Cancer = 73, MS= 23). Further, a independent dataset was chosen to validation purpose with GSE31568 [81] Table 2.1.

Tabel.2.1Summarized Data

Accession No.	Platform	Microarray Technology	Sample Breakdown	Total Samples	Normalization	Missing Values	Batch Effect Correction
GSE61741	GPL9040	*febit Homo sapiens miRBase 13.0*	MS: 23, COPD: 47, Control: 94, Lung Cancer: 74	238	Log2 + Quantile normalization	None (preprocessed, median of 7 replicates)	Not performed (same platform)
GSE31568 *(Independent Validation Dataset)*	GPL9040	*febit Homo sapiens miRBase 13.0*	MS: 24, COPD: 25, Control: 71, Lung	150	Log2 + Quantile normalization	None	Not performed (analyzed separately)

			Cancer: 33				

.

2.2.3.2 Data Preprocessing:

Normalizing the raw data acquired from the GEO database is essential for the efficient functioning of ML. We have opted to do quantile normalization for our dataset. We generated boxplots and expression density plots for both the normalized and non-normalized datasets by implementing a log2 transformation and quantile normalization on the expression data (fig.2.2). A total of 849 miRNAs were incorporated into the dataset. Further samples for the multi-class classification were extracted from the dataset, encompassing COPD, lung cancer, MS, and CONTROL. No additional filtering was applied to the 849 miRNAs, in order to preserve even low-abundance signals that may have functional relevance in disease-specific inflammatory processes. The dataset was obtained as a preprocessed matrix from GEO (GSE61741), where each miRNA was measured in at least seven replicates, and the median value was computed by the original authors. As a result, no missing values were present, and no imputation methods were required.

Furthermore, batch effect correction was not performed, since all samples originated from the same platform (GPL9040 – Agilent Human miRNA Microarray) and were processed under a consistent protocol. Importantly, the training and external validation datasets (GSE61741 and GSE31568) were analyzed independently to avoid any potential confounding due to dataset merging. This ensures the integrity of cross-dataset model evaluation without the need for cross-batch harmonization.

A) Non-normalized Expression vs Density plot B) Normalized Expression vs Density plot

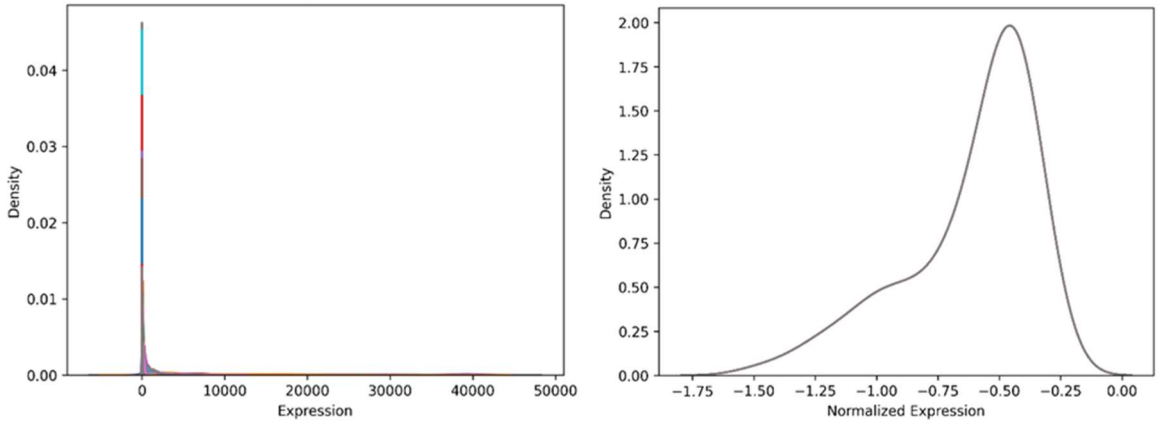


Figure 2.2Expression value for raw and normalized data with log2 transformation and quantile normalization.

2.2.3.3 Machine Learning:

Biomedicine and other fields of data science have recently benefited from new opportunities made possible by advances in computer power, data analytics, and artificial intelligence techniques, such as deep learning (DL) and ML. Considering these facts, we decided to incorporate ML. As our data is imbalanced, we chose to use the synthetic minority oversampling technique (SMOTE)[67]. To mitigate potential overfitting issues that may arise from SMOTE, we implemented several strategies. We optimized model parameters using hyperparameter tuning to avoid overfitting

owing to SMOTE and model complexity. We used Grid Search Cross-Validation to assess all critical hyperparameter combinations within a range. To find the optimal configuration, the mean cross-validation score was calculated for each set of parameters. For further validation, we then used Stratified Fold cross-validation (splits=5). 90% of the data are randomly assigned for training and 10% are assigned for testing in a 90:10 ratio. After that to make the model more robust and more reliable we conducted a model validation using independent validation dataset (GSE31568)[81]

Various classifiers including Random Forrest (RF), XGBoost (XGB), Linear Regression(LR) and Support vector machine (SVM) were compared on the bases of accuracy before further analysis. All ML, visualization, and interpretability tasks were performed using Python (v3.10) with the following libraries: scikit-learn (v1.2.2), imbalanced-learn (v0.11.0), shap (v0.42.1), matplotlib (v3.7.1), seaborn (v0.12.2), pandas (v1.5.3), and numpy (v1.24.2) (See Supplementary Table S2 for tool versions, links, and usage details).

71

2.2.3.4 A SHAP-based explanation of the trained model:

Shapley Additive Explanation can shed light on complicated models like DL and ML techniques, tree-based models, LR, and more. SHAP is based on Shapley's contribution to game theory , which outlines the mechanisms by which output can be modified through the application of different features and connects credit allotments to regional simplification. How much loss there is for all feature splits across all trees can help determine how essential features are in a tree ensemble model. This is inconsistent since it doesn't consider how essential feature order is in trees. This causes feature contributions to be unequal. Popular but unreliable approaches to gauge importance include the number of times each feature is split. Instead, SHAP values can attain consistency.

When 'x' is a vector containing all the model's feature values, ‘Øs’ can be computed using a linear corporative model. The following is the rationale given by SHAP:

$$ğ(ź) = Ø_0 + \sum_{i=1} Ø_i \acute{Z}_i$$

The explanation in the model is represented by ğ; in game theory, the coalition vector and its size are shown by ź = [⟦0,1⟧ ^M], ‘M’ respectively. The feature ‘i’ is denoted by Ø_i. Shapley's values are additive, efficient, symmetric, and fictitious. Since SHAP determines Shapley values, it satisfies both requirements. Several SHAP publications discuss inconsistencies between SHAP and Shapley traits. SHAP identified three desirable qualities: local accuracy, absence, and consistency[82] .

We employed the Python-based XAI technique SHAP to elucidate the foremost trained classifier. XAI provides the classifier with an enhanced understanding that elucidates the model's internal mechanisms and efficacy. The model's ability to differentiate among COPD, lung cancer, MS, and control is augmented by emphasizing the traits (miRNAs) that influence this knowledge. A SHAP bar plot was generated to provide a more detailed elucidation of the function miRNAs play in predicting all four classes. The subsequent phase involved generating a SHAP-based local summary plot for each class—COPD, lung cancer, MS, and control—to facilitate further interaction.

2.2.3.5 Profiling of identified miRNAs:

To visualize the impact of these SHAP identified miRNAs on model performance and their differential expression patterns, we decided to generate a heatmap of top identified miRNAs using iDEP[69]. The distinct expression characteristics of each miRNA were significantly elucidated

using these visual representations[12]. Following that, SHAP individual violin plots emphasize the most significant miRNAs across several classes. Common miRNAs for each category are selected from these summary plots. The prevalent miRNAs are inputted into MIENTURNET [83]to identify their target genes.

Further pathway enrichment and gene ontology were caried out using SRPlots[68]. At last single cell analysis was performed to visualize the expression of gene at single cell level using single cell portal[71]. This systematically offers insights into shared molecular mechanisms driving these conditions (Fig.2.3).

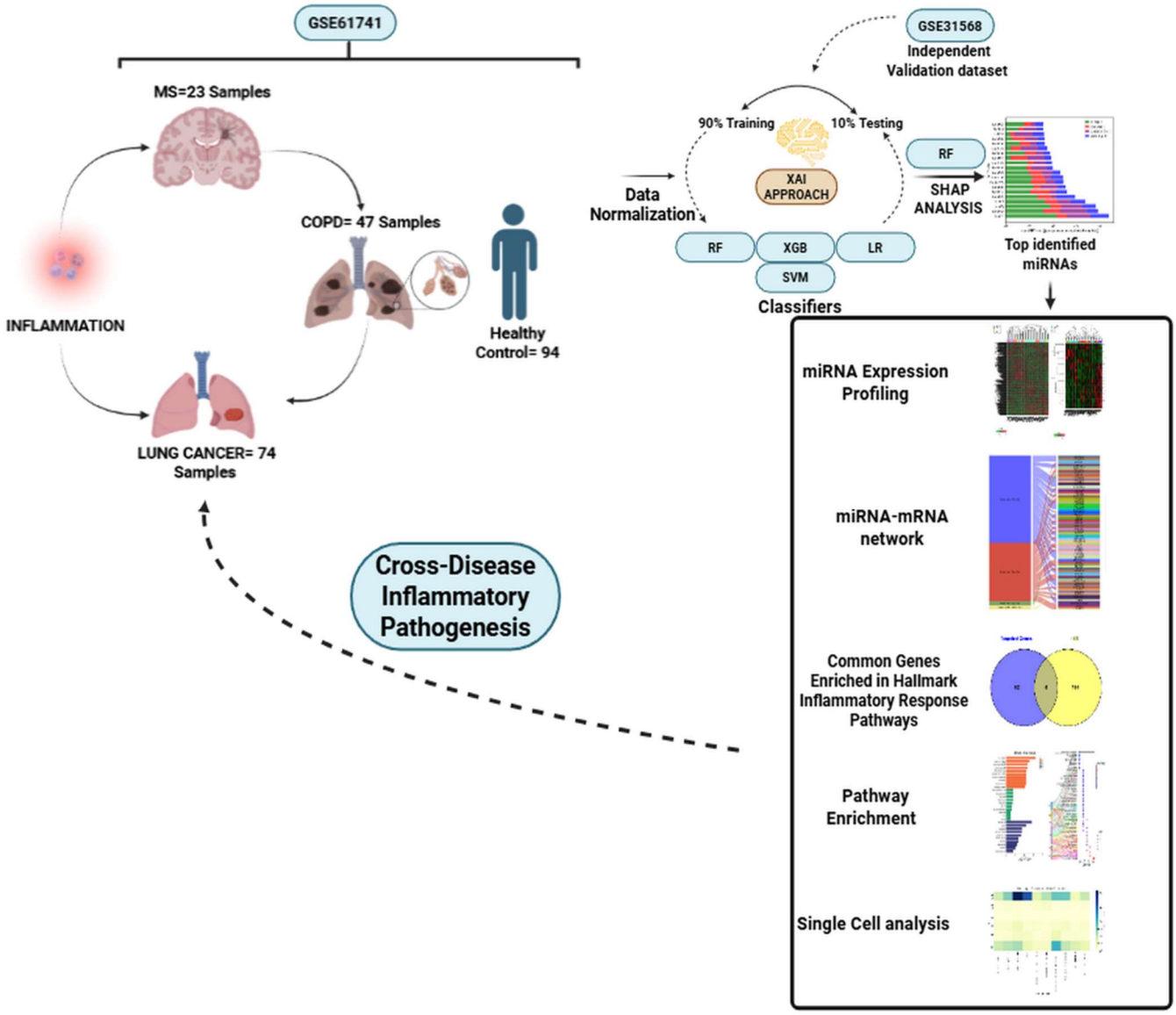


Figure 2.3Systematic representation of workflow

2.3 Objective 3: To develop a SHAP-LLM powered chatbot for lung disease indication, which leverages explainable machine learning models and large language models.

2.3.1 Problem Statement and Rationale

The rapid expansion of high-throughput omics technologies and clinical data has enabled the application of advanced computational models in respiratory research. While machine learning and explainable AI frameworks such as SHAP can highlight the contribution of molecular features to disease risk and progression, the communication of these outputs remains highly technical and inaccessible to non-computational experts. Clinicians, researchers, and patients require insights that are not only accurate but also interpretable, contextualized, and easily communicated. This lack of accessibility has limited the real-world integration of computational models into clinical decision-making.

Large Language Models (LLMs) provide a unique opportunity to address this challenge. Unlike static reporting systems, LLMs are capable of interpreting complex outputs, generating human-readable narratives, and facilitating dynamic, conversational interaction with biomedical data. When combined with SHAP, which provides the foundation for transparent feature attribution, LLMs can transform these outputs into clinically meaningful explanations that situate biomarkers within biological and pathological contexts. This integration allows predictions such as the contribution of a specific miRNA or gene to be explained in natural language, thereby bridging the gap between computational rigor and clinical usability.

The problem, therefore, lies not only in the interpretability of predictive models but in the absence of a platform that can communicate these insights in a form that directly supports real-time decision-making and patient engagement. The rationale for this objective is to design a SHAP-LLM powered chatbot that integrates biomarker discovery with conversational explainability, thereby creating an intelligent interface between complex datasets and their end-users. Such a system has the potential to enhance transparency, improve the accessibility of computational predictions, and promote the adoption of AI-driven insights in respiratory disease management.

2.3.2 Introduction

Pulmonary diseases such as chronic obstructive pulmonary disease (COPD), asthma, and lung cancer are among the leading causes of morbidity and mortality worldwide, posing a significant burden on healthcare systems and society at large[84][85][86]. Early and accurate identification of individuals at risk for these conditions is critical for improving patient outcomes, enabling timely intervention, and reducing long-term healthcare costs[84][87]. However, traditional diagnostic pathways often rely on subjective symptom reporting, clinician expertise, and resource-intensive procedures like spirometry or imaging, which can lead to variability in diagnosis and delayed treatment, particularly in resource-constrained settings[84] [87].

In recent years, artificial intelligence (AI) and machine learning (ML) have emerged as transformative tools in pulmonary medicine, offering the potential to enhance diagnostic accuracy, risk stratification, and personalized care[88][89]. AI-driven models have demonstrated superior performance compared to traditional clinical assessments in interpreting pulmonary function tests, analyzing imaging data, and predicting disease progression. For instance, studies have shown that AI-based software can outperform pulmonologists in the interpretation of pulmonary function tests and risk classification, while deep learning (DL) algorithms have achieved high accuracy in detecting undiagnosed cases of COPD from low-dose CT scans[84]. Furthermore, the integration of AI into telemedicine and remote monitoring platforms has enabled more dynamic, real-time management of chronic respiratory diseases, supporting both clinicians and patients in making informed decisions[89].

Despite these advances, the adoption of AI in clinical practice faces two major challenges: a lack of transparency in decision-making often referred to as the "black box" problem and limited accessibility of AI-generated insights for non-expert users. Explainable AI (XAI) frameworks, such as SHAP (SHapley Additive exPlanations), have been developed to address the first challenge by providing interpretable, instance-level explanations of model

predictions[56]. However, bridging the gap between technical outputs and actionable clinical or patient guidance remains a pressing need. .

Recent systematic reviews have underscored the untapped potential of AI-powered chatbots in chronic disease management, including respiratory conditions like COPD and asthma [90,91]. Despite the increasing prevalence of chronic diseases globally, only a limited number of studies have rigorously evaluated conversational agents tailored to these conditions, with most remaining at the prototype stage and lacking standardized development. These agents typically rely on natural language processing (NLP) and ML to provide functionalities such as symptom monitoring, patient education, medication reminders, and behavior modification [91,92]. Studies reviewed demonstrated promising outcomes such as improved treatment adherence, symptom recognition, and user engagement yet also highlighted persistent challenges related to transparency, clinical trustworthiness, and generalizability across diverse populations [92]. Our chatbot addresses these gaps by integrating XAI (via SHAP) with a LLM to generate both personalized risk predictions and conversational explanations, advancing the field from passive risk scoring tools toward dynamic, patient-facing decision support systems specifically built for pulmonary disease contexts.

We employed a publicly available Kaggle dataset due to its large sample size and diverse feature representation, which supports the development of robust machine learning models. Notably, recent peer-reviewed studies have also used Kaggle datasets for medical AI research, validating its suitability for scientific investigation[93]. By leveraging a dataset from Kaggle [94]of 5,000 clinical and behavioral records, the model employs domain-informed monotonic constraints to ensure biologically plausible outputs and uses SHAP for feature attribution, enhancing trust and interpretability. The dataset includes 17 features per patient, capturing demographic information, lifestyle habits, medical history, and symptomatology associated with lung cancer risk and pulmonary disease. The addition of LLM enables the translation of technical explanations into accessible narratives and supports natural language queries, making the tool both informative and user-friendly. To complement these capabilities with methodological rigor, we employed stratified data splitting into training, validation, and test sets, applied the Synthetic Minority Over-sampling Technique (SMOTE)[67] to mitigate class imbalance in the training data, and performed five-fold cross-validation to evaluate model generalizability. Additionally, SHAP-based feature attribution was used on the unseen test set to validate interpretability and ensure biologically consistent explanations for key predictors.

2.3.3 Methodology

2.3.3.1 Dataset and Preprocessing

The dataset used in this study was obtained from Kaggle and includes 5000 instances[94] (2,038 lung cancer cases & 2,964 normal cases) with 17 clinical and behavioral features relevant to pulmonary disease classification. Categorical string features were transformed using LabelEncoder[95], the target variable is binary, indicating the presence (1) or absence (0) of lung cancer, while continuous variables such as Oxygen Saturation were normalized using StandardScaler[96]. Additionally, the Energy Level feature was discretized into clinically meaningful categories: high (≥60), moderate (50–59), and low (<50), to reflect real-world fatigue levels relevant to pulmonary risk stratification.

To address class imbalance, the SMOTE[67] was applied only to the training set, ensuring improved sensitivity without introducing information leakage into validation or test sets.

2.3.3.2 Data Splitting and Cross-Validation

The dataset was stratified and divided into three parts: 60% for training, 20% for validation, and 20% for testing. To ensure robustness and prevent overfitting, 5-fold stratified cross-validation was conducted on the training set. Cross-validation folds were implemented using a pipeline that integrated SMOTE and classifier training to maintain balanced classes across splits.

2.3.3.3 Model Development

An XGBoost classifier [97]was selected due to its strong performance in tabular classification tasks and support for monotonic constraints. The model was configured with logloss as the objective function and trained using the training set, with hyperparameter tuning performed through cross-validation.

To improve clinical trust and consistency, monotonic constraints were explicitly applied to key features:

- **Positive monotonicity**: Increasing features like age, alcohol consumption, are expected to increase disease risk.
- **Negative monotonicity**: Increasing Oxygen Saturation and Energy Level are expected to reduce disease risk.

These constraints ensured that model predictions aligned with established medical knowledge and reduced the risk of biologically implausible outputs.

2.3.3.4 Model Interpretability

Model explainability was achieved using SHAP [82]. A SHAP explainer was initialized using a representative 100-sample subset of the training data. For each prediction, SHAP values were computed to estimate the marginal contribution of each feature to the final output. These values were visualized in ranked bar charts [12,98]within the user interface, providing transparent insight into the model's decision-making.

2.3.3.5 Web Application Architecture

A web-based application was developed using the Dash framework[99]. The platform allows clinicians and users to:

- Enter feature values via interactive widgets
- Receive real-time predictions with accompanying probability scores
- View SHAP-based visual explanations for transparency

The app is structured with tabs for prediction, automated explanation, and custom query support, enhancing usability in clinical or educational settings.

2.3.3.6 Natural Language Explanation Assistant

To facilitate interpretability for non-expert users, the platform integrates a conversational AI module using LangChain and the Ollama LLM (Mistral). For each prediction, an automated query template is filled with the SHAP summary and submitted to the LLM. This produces a clinically grounded, human-readable explanation. Additionally, users can pose free-text questions about their prediction or risk factors and receive interactive, context-aware responses. This enhances the tool's accessibility and supports patient education[100].

To enhance interpretability, SHAP values were computed using the SHAP package and visualized with Plotly. A local LLM powered by Ollama was integrated for automatic explanation and user query response. All components were run within a Conda-managed environment(fig.2.4). The full list of package versions is provided in Table 2.2.

Table 2.2Summary of packages utilized

Component	Version	Source / Notes
Python	3.11	Conda environment
Dash	2.16.1	Interactive UI framework
Plotly	5.22.0	For SHAP bar plot rendering
Pandas	2.2.2	Data handling
NumPy	1.26.4	Numerical operations
Scikit-learn	1.4.2	Preprocessing, train-test split
XGBoost	2.0.3	Classification with monotonic constraints
SHAP	0.45.1	Feature contribution explanation
Imbalanced-learn (SMOTE)	0.12.2	Synthetic oversampling
Joblib	1.4.2	Model persistence
LangChain	0.1.17	LLM pipeline construction
LangChain Community	0.0.34	Ollama interface wrapper
Ollama	0.1.29 (CLI)	Local LLM runner (https://ollama.com)
LLM Model	mistral	Automatically pulled via ollama run mistral

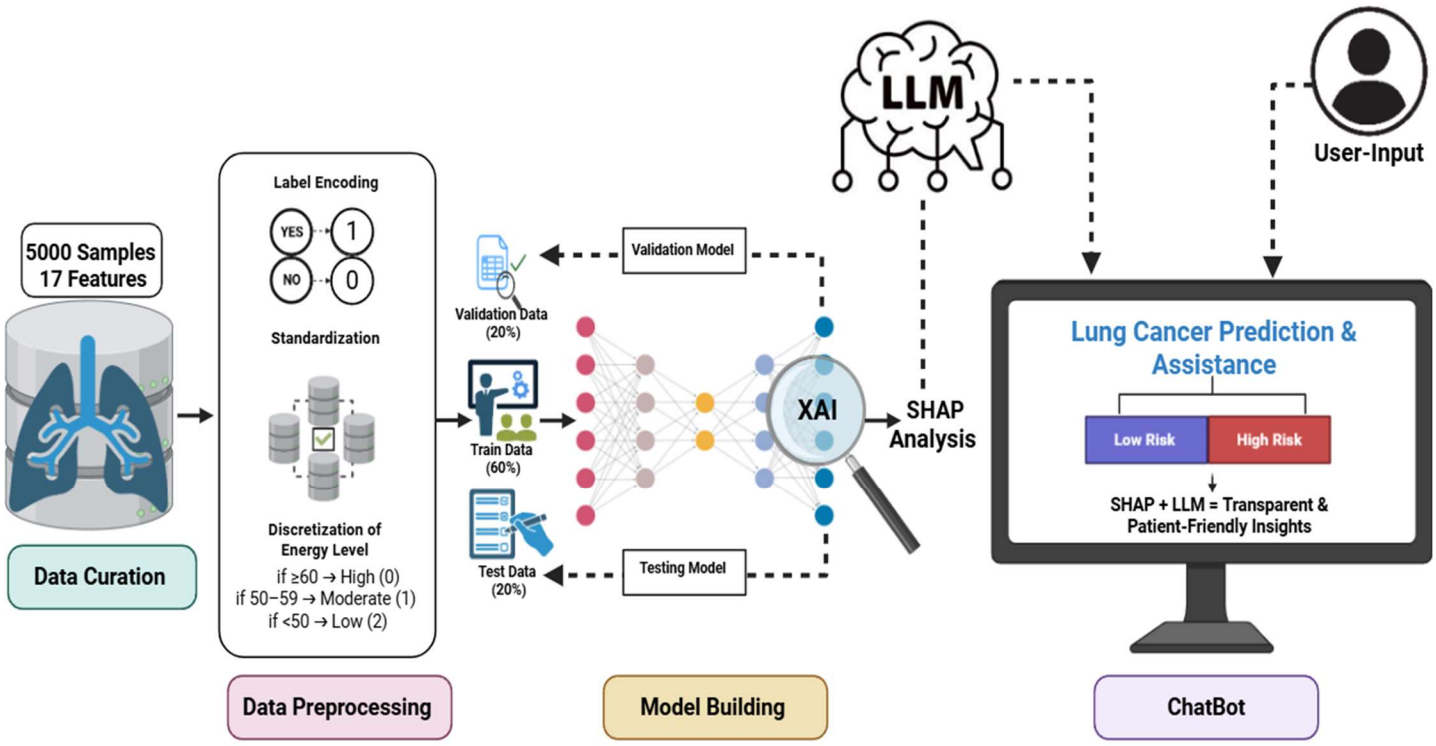


Figure 2.4Systematic Workflow

Chapter3

66

This chapter presents the findings of the study in accordance with the objectives defined in Chapter II. The results are organized objective-wise, beginning with the identification of candidate biomarkers , followed by the investigation of coexistence among respiratory diseases, and subsequently the exploration of systemic inflammatory crosstalk between neuro-inflammatory and respiratory disorders. The final section demonstrates the translational application of these findings through the development of a SHAP-LLM powered chatbot for lung disease indication.

3.1 Results for Objective 1: To investigate the coexistence and shared molecular mechanisms among major respiratory diseases through integrative multi-omics and regulatory network analyses.

The microarray data was acquired in its raw form from NCBI's Gene expression. Following that, Supplementary file 1 depicts the GOLD characterization of COPD samples for both the platforms i.e. GPL6480 (Table.1) and GPL14550(Table.3.1). Tabel.3.2 summarizes the data including GEO accession number, platform number ID and no. of sample.

Table3.1GOLD Characterization of COPD

GSE47460 - GPL6480

Accession	Title	Source name	Disease state	Characteristics	Gold stage
GSM1149949	LT001098RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 41.5 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 24 %predicted fvc (pre-bd): 73 %predicted fev1 (post-bd): 26 %predicted fvc (post-bd): 78 %predicted dlco: 34	4-Very Severe COPD
GSM1149953	LT007392RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 71.6 (PTX) smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 46 %predicted fvc (pre-bd): 76 %predicted fev1 (post-bd): 51 %predicted fvc (post-bd): 90 %predicted dlco: 45	2-Moderate COPD
GSM1149954	LT009099RL_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 0.121 smoker?: 3-Never %predicted fev1 (pre-bd): 88 %predicted fvc (pre-bd): 79 %predicted fev1 (post-bd): 91 %predicted fvc (post-bd): 79 %predicted dlco: 102	0-At Risk
GSM1149956	LT010491LL_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 3.767 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 73 %predicted fvc (pre-bd): 82 %predicted fev1 (post-bd): 78 %predicted fvc (post-bd): 83 %predicted dlco: 66	2-Moderate COPD
GSM1149960	LT024952RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 41.9 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 16 %predicted fvc (pre-bd): 47	4-Very Severe COPD

GSM1149961	LT026458RL_COPD	Flash frozen whole lung	COPD	%predicted fev1 (pre-bd): 54 %predicted fvc (pre-bd): 63 %predicted fev1 (post-bd): 67 %predicted fvc (post-bd): 77 %predicted dlco: 69	2-Moderate COPD
GSM1149964	LT032411RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 5.7 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 102 %predicted fvc (pre-bd): 111 %predicted fev1 (post-bd): 106 %predicted fvc (post-bd): 115 %predicted dlco: 76	1-Mild COPD
GSM1149965	LT032775RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 29.5 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 33 %predicted fvc (pre-bd): 62 %predicted fev1 (post-bd): 40 %predicted fvc (post-bd): 89 %predicted dlco: 46	3-Severe COPD
GSM1149967	LT038075LL_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 8.16 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 108 %predicted fvc (pre-bd): 108 %predicted fev1 (post-bd): 107 %predicted fvc (post-bd): 108 %predicted dlco: 103	0-At Risk
GSM1149968	LT038591LU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 44.8 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 22 %predicted fvc (pre-bd): 62	4-Very Severe COPD
GSM1149969	LT042988LU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 59.8 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 16 %predicted fvc (pre-bd): 23	4-Very Severe COPD
GSM1149972	LT046103LU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 10.5 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 73 %predicted fvc (pre-bd): 85 %predicted fev1 (post-bd): 80 %predicted fvc (post-bd): 86 %predicted dlco: 74	0-At Risk
GSM1149973	LT046180RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 14.2 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 25 %predicted fvc (pre-bd): 59 %predicted fev1 (post-bd): 28 %predicted fvc (post-bd): 66 %predicted dlco: 14	4-Very Severe COPD
GSM1149978	LT050246RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 41.7 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 18 %predicted fvc (pre-bd): 49 %predicted dlco: 27	4-Very Severe COPD
GSM1149979	LT051568RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 0.8 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 44 %predicted fvc (pre-bd): 55 %predicted fev1 (post-bd): 54 %predicted fvc (post-bd): 65 %predicted dlco: 103	2-Moderate COPD
GSM1149980	LT051993RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 27.7 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 28 %predicted fvc (pre-bd): 79 %predicted fev1 (post-bd): 32 %predicted fvc (post-bd): 86 %predicted dlco: 41	3-Severe COPD

44 GSM1149986	LT056627RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 2.6 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 25 %predicted fvc (pre-bd): 51 %predicted fev1 (post-bd): 47 %predicted fvc (post-bd): 89 %predicted dlco: 56	3-Severe COPD
GSM1149989	LT059224RL_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 0.9 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 113 %predicted fvc (pre-bd): 98 %predicted fev1 (post-bd): 123 %predicted fvc (post-bd): 105 %predicted dlco: 73	0-At Risk
GSM1149990	LT059865RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 30.1 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 35 %predicted fvc (pre-bd): 77 %predicted fev1 (post-bd): 37 %predicted fvc (post-bd): 84 %predicted dlco: 30	3-Severe COPD
GSM1149991	LT060406RL_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 11.7 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 60 %predicted fvc (pre-bd): 81 %predicted dlco: 62	2-Moderate COPD
GSM1149995	LT070021RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 2 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 49 %predicted fvc (pre-bd): 69 %predicted fev1 (post-bd): 63 %predicted fvc (post-bd): 81 %predicted dlco: 68	2-Moderate COPD
GSM1150000	LT075917RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 48.8 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 24 %predicted fvc (pre-bd): 37 %predicted fev1 (post-bd): 25 %predicted fvc (post-bd): 42 %predicted dlco: 31	4-Very Severe COPD
GSM1150002	LT079013RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 10.72 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 23 %predicted fvc (pre-bd): 43 %predicted dlco: 34	4-Very Severe COPD
GSM1150003	LT081077RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 3.6 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 67 %predicted fvc (pre-bd): 93 %predicted fev1 (post-bd): 74 %predicted fvc (post-bd): 100 %predicted dlco: 96	2-Moderate COPD
GSM1150005	LT084831RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 0.03 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 57 %predicted fvc (pre-bd): 69 %predicted fev1 (post-bd): 62 %predicted fvc (post-bd): 66 %predicted dlco: 49	
GSM1150006	LT085240RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 24.7 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 22 %predicted fvc (pre-bd): 48 %predicted dlco: 26	4-Very Severe COPD
GSM1150007	LT086542RL_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 0.3 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 70 %predicted fvc (pre-bd): 82 %predicted fev1 (post-bd): 81 %predicted fvc (post-bd): 90 %predicted dlco: 88	1-Mild COPD

GSM1150013	LT096554LU_COPD	Flash frozen whole lung	COPD	%predicted fev1 (pre-bd): 76 %predicted fvc (pre-bd): 91 %predicted fev1 (post-bd): 87 %predicted fvc (post-bd): 98	1-Mild COPD
GSM1150016	LT100984RL_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 0.1 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 34 %predicted fvc (pre-bd): 43 %predicted fev1 (post-bd): 34 %predicted fvc (post-bd): 44 %predicted dlco: 86	3-Severe COPD
GSM1150020	LT106557LU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 26.9 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 26 %predicted fvc (pre-bd): 68 %predicted dlco: 32	4-Very Severe COPD
GSM1150025	LT110408RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 44.4 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 21 %predicted fvc (pre-bd): 70 %predicted fev1 (post-bd): 22 %predicted fvc (post-bd): 75 %predicted dlco: 31	4-Very Severe COPD
GSM1150027	LT112339LI_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 3.5 smoker?: 1-Current %predicted fev1 (pre-bd): 74 %predicted fvc (pre-bd): 90 %predicted fev1 (post-bd): 78 %predicted fvc (post-bd): 91 %predicted dlco: 79	2-Moderate COPD
GSM1150029	LT113077LL_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 2.1 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 64 %predicted fvc (pre-bd): 74 %predicted fev1 (post-bd): 71 %predicted fvc (post-bd): 78 %predicted dlco: 68	2-Moderate COPD
GSM1150038	LT121841RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 8.8 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 76 %predicted fvc (pre-bd): 119 %predicted fev1 (post-bd): 85 %predicted fvc (post-bd): 120 %predicted dlco: 98	1-Mild COPD
GSM1150043	LT123457RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 1.4 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 93 %predicted fvc (pre-bd): 104 %predicted fev1 (post-bd): 101 %predicted fvc (post-bd): 108 %predicted dlco: 74	0-At Risk
GSM1150044	LT125775RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 0.4 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 73 %predicted fvc (pre-bd): 85 %predicted fev1 (post-bd): 74 %predicted fvc (post-bd): 84 %predicted dlco: 57	2-Moderate COPD
GSM1150045	LT126293RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 0.4 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 84 %predicted fvc (pre-bd): 100 %predicted fev1 (post-bd): 93 %predicted fvc (post-bd): 106 %predicted dlco: 72	1-Mild COPD
GSM1150046	LT126571RU_COPD	Flash frozen	COPD	%emphysema (f-950): 41.796 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 40 %predicted fvc (pre-bd): 13 %predicted fev1	4-Very Severe COPD

			whole lung		(post-bd): 18 %predicted fvc (post-bd): 36 %predicted dlco: 52
10 GSM1150047	LT127070RL_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 1.1 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 52 %predicted fvc (pre-bd): 64 %predicted fev1 (post-bd): 54 %predicted fvc (post-bd): 68 %predicted dlco: 67	2-Moderate COPD
GSM1150049	LT130350RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 43.768 smoker?: 3-Never %predicted fev1 (pre-bd): 20 %predicted fvc (pre-bd): 44 %predicted fev1 (post-bd): 25 %predicted fvc (post-bd): 51 %predicted dlco: 43	4-Very Severe COPD
GSM1150050	LT131432RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 7 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 70 %predicted fvc (pre-bd): 81 %predicted fev1 (post-bd): 74 %predicted fvc (post-bd): 96 %predicted dlco: 61	2-Moderate COPD
35 GSM1150052	LT132625RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 51.431 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 24 %predicted fvc (pre-bd): 64 %predicted fev1 (post-bd): 23 %predicted fvc (post-bd): 64 %predicted dlco: 17	4-Very Severe COPD
GSM1150059	LT144462RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 7.241 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 69 %predicted fvc (pre-bd): 87 %predicted fev1 (post-bd): 74 %predicted fvc (post-bd): 94 %predicted dlco: 64	2-Moderate COPD
GSM1150060	LT145086RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 43.3 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 21 %predicted fvc (pre-bd): 52	4-Very Severe COPD
GSM1150061	LT145162RM_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 7.3 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 55 %predicted fvc (pre-bd): 77 %predicted fev1 (post-bd): 66 %predicted fvc (post-bd): 102 %predicted dlco: 52	2-Moderate COPD
GSM1150062	LT147658RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 33.5 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 24 %predicted fvc (pre-bd): 61 %predicted fev1 (post-bd): 26 %predicted fvc (post-bd): 68 %predicted dlco: 36	4-Very Severe COPD
GSM1150063	LT150340LU_COPD	Flash frozen whole lung	COPD	smoker?: 1-Current %predicted fev1 (pre-bd): 51 %predicted fvc (pre-bd): 76 %predicted fev1 (post-bd): 60 %predicted fvc (post-bd): 87 %predicted dlco: 35	2-Moderate COPD
GSM1150064	LT150981RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 0.469 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 62 %predicted fvc (pre-bd): 70 %predicted fev1 (post-bd): 66 %predicted fvc (post-bd): 73 %predicted dlco: 69	

10 GSM1150065	LT152979LL_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 17.813 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 51 %predicted fvc (pre-bd): 98 %predicted fev1 (post-bd): 60 %predicted fvc (post-bd): 105 %predicted dlco: 47	2-Moderate COPD
GSM1150069	LT160089RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 1.82 smoker?: 3-Never %predicted fev1 (pre-bd): 85 %predicted fvc (pre-bd): 102 %predicted fev1 (post-bd): 93 %predicted fvc (post-bd): 110 %predicted dlco: 71	1-Mild COPD
GSM1150071	LT161707RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 21.1 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 26 %predicted fvc (pre-bd): 64 %predicted fev1 (post-bd): 29 %predicted fvc (post-bd): 68 %predicted dlco: 34	4-Very Severe COPD
GSM1150074	LT166565RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 36.6 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 19 %predicted fvc (pre-bd): 45 %predicted fev1 (post-bd): 50 %predicted fvc (post-bd): 53 %predicted dlco: 22	
GSM1150077	LT169564RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 0.5 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 92 %predicted fvc (pre-bd): 101 %predicted fev1 (post-bd): 91 %predicted fvc (post-bd): 99 %predicted dlco: 112	1-Mild COPD
GSM1150082	LT174536LU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 18 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 53 %predicted fvc (pre-bd): 75 %predicted fev1 (post-bd): 58 %predicted fvc (post-bd): 77 %predicted dlco: 64	2-Moderate COPD
GSM1150086	LT184347RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 6.5 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 57 %predicted fvc (pre-bd): 90 %predicted fev1 (post-bd): 64 %predicted fvc (post-bd): 92 %predicted dlco: 61	2-Moderate COPD
GSM1150093	LT193936RU_COPD	Flash frozen whole lung	COPD	%predicted fev1 (pre-bd): 40 %predicted fvc (pre-bd): 78 %predicted fev1 (post-bd): 61 %predicted fvc (post-bd): 94 %predicted dlco: 86	2-Moderate COPD
GSM1150099	LT204017RL_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 0.9 smoker?: 1-Current %predicted fev1 (pre-bd): 36 %predicted fvc (pre-bd): 48 %predicted fev1 (post-bd): 37 %predicted fvc (post-bd): 51 %predicted dlco: 45	3-Severe COPD
GSM1150102	LT206010LU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 0.7 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 63 %predicted fvc (pre-bd): 65 %predicted fev1 (post-bd): 66 %predicted fvc (post-bd): 73 %predicted dlco: 67	2-Moderate COPD
GSM1150105	LT211728RU_COPD	Flash frozen	COPD	%emphysema (f-950): 0.974 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 69	2-Moderate COPD

GSM1150112	LT230267RU_COPD	whole lung Flash frozen whole lung	COPD	%predicted fvc (pre-bd): 83 %predicted fev1 (post-bd): 79 %predicted fvc (post-bd): 88 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 13 %predicted fvc (pre-bd): 53 %predicted fev1 (post-bd): 14 %predicted fvc (post-bd): 56 %predicted dlco: 25	4-Very Severe COPD
GSM1150116	LT235584RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 61.7 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 19 %predicted fvc (pre-bd): 63 %predicted fev1 (post-bd): 21 %predicted fvc (post-bd): 82 %predicted dlco: 35	4-Very Severe COPD
GSM1150117	LT242420LU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 58.6 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 14 %predicted fvc (pre-bd): 59 %predicted fev1 (post-bd): 14 %predicted fvc (post-bd): 62 %predicted dlco: 24	4-Very Severe COPD
GSM1150120	LT262496LU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 45.3 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 16 %predicted fvc (pre-bd): 35	4-Very Severe COPD
GSM1150123	LT265199LU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 4.2 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 44 %predicted fvc (pre-bd): 71 %predicted fev1 (post-bd): 56 %predicted fvc (post-bd): 83 %predicted dlco: 75	2-Moderate COPD
GSM1150126	LT271679LU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 3.197 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 35 %predicted fvc (pre-bd): 39 %predicted fev1 (post-bd): 41 %predicted fvc (post-bd): 46 %predicted dlco: 44	3-Severe COPD
GSM1150127	LT271851LU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 9.3 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 68 %predicted fvc (pre-bd): 106 %predicted fev1 (post-bd): 76 %predicted fvc (post-bd): 106 %predicted dlco: 58	2-Moderate COPD
GSM1150129	LT274443LU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 29.7 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 36 %predicted fvc (pre-bd): 71 %predicted fev1 (post-bd): 35 %predicted fvc (post-bd): 83 %predicted dlco: 27	3-Severe COPD
GSM1150131	LT277036RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 2.1 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 17 %predicted fvc (pre-bd): 51 %predicted fev1 (post-bd): 23 %predicted fvc (post-bd): 66 %predicted dlco: 57	4-Very Severe COPD
GSM1150133	LT282031RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 35.6 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 32 %predicted fvc (pre-bd): 80 %predicted fev1 (post-bd): 44 %predicted fvc (post-bd): 99 %predicted dlco: 42	3-Severe COPD

GSM1150134	LT282467LL_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 5.6 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 36 %predicted fvc (pre-bd): 71 %predicted fev1 (post-bd): 50 %predicted fvc (post-bd): 81 %predicted dlco: 78	3-Severe COPD
GSM1150136	LT284144LU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 27.5 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 58 %predicted fvc (pre-bd): 114 %predicted dlco: 13	2-Moderate COPD
GSM1150138	LT291578RU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 62 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 19 %predicted fvc (pre-bd): 78 %predicted dlco: 16	4-Very Severe COPD
GSM1150139	LT291673RM_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 1 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 42 %predicted fvc (pre-bd): 49 %predicted fev1 (post-bd): 55 %predicted fvc (post-bd): 63 %predicted dlco: 95	2-Moderate COPD
GSM1150140	LT294945LU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 4.7 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 15 %predicted fvc (pre-bd): 21 %predicted fev1 (post-bd): 19 %predicted fvc (post-bd): 26	4-Very Severe COPD
GSM1150141	LT295167LU_COPD	Flash frozen whole lung	COPD	%emphysema (f-950): 41.9 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 27 %predicted fvc (pre-bd): 77	4-Very Severe COPD

GSE47460 - GPL14550

Accession	Title	Disease state	Gold stage	Characteristics
GSM1150145	LT012861RU_COPD	COPD	4-Very Severe COPD	%emphysema (f-950): 6.3 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 22 %predicted fvc (pre-bd): 52 %predicted dlco: 33
GSM1150147	LT030041RU_COPD	COPD	1-Mild COPD	%emphysema (f-950): 0.1 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 77 %predicted fvc (pre-bd): 105 %predicted fev1 (post-bd): 80 %predicted fvc (post-bd): 90

				%predicted dlco: 77
GSM1150162	LT155982RU_COPD	COPD	4-Very Severe COPD	%emphysema (f-950): 46.559 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 20 %predicted fvc (pre-bd): 32 %predicted fev1 (post-bd): 23 %predicted fvc (post-bd): 36 %predicted dlco: 31
GSM1150164	LT159988RL_COPD	COPD	4-Very Severe COPD	%emphysema (f-950): 45 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 16 %predicted fvc (pre-bd): 58 %predicted dlco: 26
GSM1150166	LT168204LL_COPD	COPD	0-At Risk	smoker?: 3-Never %predicted fev1 (pre-bd): 118 %predicted fvc (pre-bd): 92 %predicted fev1 (post-bd): 121 %predicted fvc (post-bd): 92 %predicted dlco: 54
GSM1150170	LT194990RU_COPD	COPD	4-Very Severe COPD	%emphysema (f-950): 59.7 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 16 %predicted fvc (pre-bd): 50 %predicted fev1 (post-bd): 16 %predicted fvc (post-bd): 56
GSM1150178	LT002501RL_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 6.168 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 46 %predicted fvc (pre-bd): 74 %predicted fev1 (post-bd): 52 %predicted fvc (post-bd): 81 %predicted dlco: 36
GSM1150182	LT017275LL_COPD	COPD	3-Severe COPD	%emphysema (f-950): 15.9 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 31 %predicted fvc (pre-bd): 67 %predicted fev1 (post-bd): 32 %predicted fvc (post-bd): 66 %predicted dlco: 50
GSM1150184	LT020426LU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 0.4 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 46 %predicted fvc (pre-bd): 56 %predicted fev1 (post-bd): 55 %predicted fvc (post-bd): 64 %predicted dlco: 82
GSM1150185	LT024106RU_COPD	COPD		%emphysema (f-950): 0.4 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 67 %predicted fvc (pre-bd): 68 %predicted fev1 (post-bd): 69 %predicted fvc (post-bd): 72
GSM1150186	LT024460RU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 0.8 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 73 %predicted fvc (pre-bd): 88 %predicted fev1 (post-bd): 77 %predicted fvc (post-bd): 91 %predicted dlco: 77

GSM1150187	LT025997RU_COPD	COPD	1-Mild COPD	%emphysema (f-950): 40.046 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 76 %predicted fvc (pre-bd): 88 %predicted fev1 (post-bd): 80 %predicted fvc (post-bd): 90 %predicted dlco: 20
GSM1150188	LT028044RU_COPD	COPD	3-Severe COPD	%emphysema (f-950): 13.886 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 35 %predicted fvc (pre-bd): 73 %predicted fev1 (post-bd): 35 %predicted fvc (post-bd): 76
GSM1150189	LT028427LU_COPD	COPD	3-Severe COPD	%emphysema (f-950): 42.5 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 27 %predicted fvc (pre-bd): 73 %predicted fev1 (post-bd): 32 %predicted fvc (post-bd): 71 %predicted dlco: 31
GSM1150190	LT030151RU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 11.616 smoker?: 1-Current %predicted fev1 (pre-bd): 60 %predicted fvc (pre-bd): 98 %predicted fev1 (post-bd): 65 %predicted fvc (post-bd): 116 %predicted dlco: 48
GSM1150192	LT034070LU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 16.4 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 53 %predicted fvc (pre-bd): 86 %predicted fev1 (post-bd): 60 %predicted fvc (post-bd): 87 %predicted dlco: 48
GSM1150193	LT034821RU_COPD	COPD	3-Severe COPD	%emphysema (f-950): 8.555 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 33 %predicted fvc (pre-bd): 72 %predicted fev1 (post-bd): 44 %predicted fvc (post-bd): 91 %predicted dlco: 35
GSM1150199	LT042552RL_ILD	COPD	2-Moderate COPD	%emphysema (f-950): 16.3 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 60 %predicted fvc (pre-bd): 70 %predicted fev1 (post-bd): 60 %predicted fvc (post-bd): 80 %predicted dlco: 43
GSM1150200	LT043343LU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 1.216 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 75 %predicted fvc (pre-bd): 83 %predicted fev1 (post-bd): 79 %predicted fvc (post-bd): 85 %predicted dlco: 82
GSM1150201	LT043798LU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 6.392 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 61 %predicted fvc (pre-bd): 76 %predicted fev1 (post-bd): 66

GSM1150203	LT057972LU_COPD	COPD	2-Moderate COPD	%predicted fvc (post-bd): 77 %predicted dlco: 61 %emphysema (f-950): 2.3 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 59 %predicted fvc (pre-bd): 86 %predicted dlco: 38
GSM1150204	LT058691LU_COPD	COPD	4-Very Severe COPD	%emphysema (f-950): 23.6 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 22 %predicted fvc (pre-bd): 50 %predicted fev1 (post-bd): 25 %predicted fvc (post-bd): 54 %predicted dlco: 22
GSM1150206	LT059975LU_COPD	COPD	3-Severe COPD	%emphysema (f-950): 14.325 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 43 %predicted fvc (pre-bd): 78 %predicted fev1 (post-bd): 41 %predicted fvc (post-bd): 78
GSM1150210	LT070403LL_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 2.5 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 48 %predicted fvc (pre-bd): 79 %predicted fev1 (post-bd): 56 %predicted fvc (post-bd): 85 %predicted dlco: 66
GSM1150211	LT072387LU_COPD	COPD	4-Very Severe COPD	%emphysema (f-950): 46 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 25 %predicted fvc (pre-bd): 50 %predicted fev1 (post-bd): 25 %predicted fvc (post-bd): 49 %predicted dlco: 39
GSM1150212	LT072808RL_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 6.664 smoker?: 1-Current %predicted fev1 (pre-bd): 52 %predicted fvc (pre-bd): 65 %predicted fev1 (post-bd): 57 %predicted fvc (post-bd): 75 %predicted dlco: 48
GSM1150213	LT075462RU_COPD	COPD	3-Severe COPD	%emphysema (f-950): 20.1 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 37 %predicted fvc (pre-bd): 59 %predicted fev1 (post-bd): 46 %predicted fvc (post-bd): 72 %predicted dlco: 57
GSM1150214	LT076181LI_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 1.5 smoker?: 2-Ever (>100) %predicted fev1 (post-bd): 55 %predicted fvc (post-bd): 63 %predicted dlco: 86
GSM1150215	LT076617LL_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 0.308 smoker?: 3-Never %predicted fev1 (pre-bd): 78 %predicted fvc (pre-bd): 91 %predicted dlco: 108
GSM1150220	LT080176RU_COPD	COPD	2-Moderate	%emphysema (f-950): 1.704 smoker?: 3-Never %predicted fev1 (pre-bd): 55

			COPD	%predicted fvc (pre-bd): 76 %predicted fev1 (post-bd): 52 %predicted fvc (post-bd): 69 %predicted dlco: 58 %emphysema (f-950): 0.307 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 92 %predicted fvc (pre-bd): 90
GSM1150223	LT083706RL_COPD	COPD	0-At Risk	%predicted dlco: 77 %emphysema (f-950): 1.5 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 69 %predicted fvc (pre-bd): 100
GSM1150224	LT083759RL_COPD	COPD	2-Moderate COPD	%predicted fev1 (post-bd): 74 %predicted fvc (post-bd): 103 %predicted dlco: 53 %emphysema (f-950): 0.012 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 49 %predicted fvc (pre-bd): 61
GSM1150225	LT084406RU_COPD	COPD	2-Moderate COPD	%predicted fev1 (post-bd): 53 %predicted fvc (post-bd): 64 %predicted dlco: 71 %emphysema (f-950): 9.647 smoker?: 1-Current %predicted fev1 (pre-bd): 63 %predicted fvc (pre-bd): 94
GSM1150226	LT084808LU_COPD	COPD	2-Moderate COPD	%predicted fev1 (post-bd): 63 %predicted fvc (post-bd): 95 %predicted dlco: 54 %emphysema (f-950): 0.173 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 51 %predicted fvc (pre-bd): 64
GSM1150228	LT089723LL_COPD	COPD	2-Moderate COPD	%predicted dlco: 55 %emphysema (f-950): 0.634 smoker?: 3-Never %predicted fev1 (pre-bd): 66 %predicted fvc (pre-bd): 81
GSM1150234	LT109231RU_COPD	COPD	2-Moderate COPD	%predicted fev1 (post-bd): 72 %predicted fvc (post-bd): 91 %predicted dlco: 77 %emphysema (f-950): 8.162 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 34 %predicted fvc (pre-bd): 80
GSM1150235	LT111643RU_COPD	COPD	3-Severe COPD	%emphysema (f-950): 0.2 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 82 %predicted fvc (pre-bd): 95
GSM1150236	LT112563LL_COPD	COPD	1-Mild COPD	%predicted fev1 (post-bd): 85 %predicted fvc (post-bd): 96 %predicted dlco: 91 %emphysema (f-950): 11.4 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 62 %predicted fvc (pre-bd): 72
GSM1150237	LT112597RU_COPD	COPD	2-Moderate COPD	%predicted fev1 (post-bd): 73 %predicted fvc (post-bd): 85 %predicted dlco: 70

GSM1150238	LT113005RU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 14 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 66 %predicted fvc (pre-bd): 81 %predicted fev1 (post-bd): 67 %predicted fvc (post-bd): 80 %predicted dlco: 30
GSM1150239	LT115251RU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 0.1 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 60 %predicted fvc (pre-bd): 70 %predicted fev1 (post-bd): 67 %predicted fvc (post-bd): 80 %predicted dlco: 80
GSM1150240	LT118064RL_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 0.777 smoker?: 1-Current %predicted fev1 (pre-bd): 59 %predicted fvc (pre-bd): 78 %predicted fev1 (post-bd): 60 %predicted fvc (post-bd): 78 %predicted dlco: 55
GSM1150242	LT122336LU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 8.7 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 50 %predicted fvc (pre-bd): 69 %predicted fev1 (post-bd): 54 %predicted fvc (post-bd): 76 %predicted dlco: 77
GSM1150243	LT126327LU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 14.5 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 73 %predicted fvc (pre-bd): 121 %predicted fev1 (post-bd): 80 %predicted fvc (post-bd): 128 %predicted dlco: 91
GSM1150246	LT134719RU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 6.406 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 75 %predicted fvc (pre-bd): 89 %predicted fev1 (post-bd): 76 %predicted fvc (post-bd): 91 %predicted dlco: 45
GSM1150248	LT136415RU_COPD	COPD	4-Very Severe COPD	%emphysema (f-950): 31.7 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 16 %predicted fvc (pre-bd): 50 %predicted fev1 (post-bd): 26 %predicted fvc (post-bd): 74 %predicted dlco: 30
GSM1150249	LT137832LU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 9.018 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 51 %predicted fvc (pre-bd): 79 %predicted fev1 (post-bd): 52 %predicted fvc (post-bd): 82 %predicted dlco: 64
GSM1150250	LT138418LU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 0.031 smoker?: 3-Never %predicted fev1 (pre-bd): 70 %predicted fvc (pre-bd): 83

GSM1150251	LT139051LU_COPD	COPD	1-Mild COPD	%predicted fev1 (post-bd): 69 %predicted fvc (post-bd): 82 %predicted dlco: 104 %emphysema (f-950): 7.7 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 99 %predicted fvc (pre-bd): 111
GSM1150253	LT140046RU_COPD	COPD	3-Severe COPD	%predicted fev1 (post-bd): 97 %predicted fvc (post-bd): 109 %predicted dlco: 34 %emphysema (f-950): 14.493 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 34 %predicted fvc (pre-bd): 60 %predicted dlco: 47
GSM1150254	LT140471RU_COPD	COPD	1-Mild COPD	%emphysema (f-950): 3.7 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 83 %predicted fvc (pre-bd): 96 %predicted fev1 (post-bd): 85 %predicted fvc (post-bd): 98 %predicted dlco: 81
GSM1150255	LT148377LU_COPD	COPD	3-Severe COPD	%emphysema (f-950): 39.5 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 41 %predicted fvc (pre-bd): 67 %predicted dlco: 49
GSM1150258	LT151920RL_COPD	COPD	1-Mild COPD	%emphysema (f-950): 4.232 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 69 %predicted fvc (pre-bd): 89 %predicted fev1 (post-bd): 82 %predicted fvc (post-bd): 91 %predicted dlco: 62
GSM1150261	LT152653LU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 2.3 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 47 %predicted fvc (pre-bd): 66 %predicted fev1 (post-bd): 51 %predicted fvc (post-bd): 75 %predicted dlco: 70
GSM1150262	LT154785RU_COPD	COPD	4-Very Severe COPD	%emphysema (f-950): 28.9 smoker?: 2-Ever (>100) %predicted fev1 (post-bd): 16 %predicted fvc (post-bd): 57 %predicted dlco: 34
GSM1150263	LT156041LU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 6.021 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 67 %predicted fvc (pre-bd): 89 %predicted fev1 (post-bd): 74 %predicted fvc (post-bd): 95 %predicted dlco: 67
GSM1150265	LT157177RU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 0.478 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 60 %predicted fvc (pre-bd): 69 %predicted fev1 (post-bd): 62 %predicted fvc (post-bd): 68
GSM1150266	LT158647RU_COPD	COPD	4-Very	%emphysema (f-950): 56.335

			Severe COPD	smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 23 %predicted fvc (pre-bd): 41 %predicted fev1 (post-bd): 21 %predicted fvc (post-bd): 42 %emphysema (f-950): 31.664
GSM1150268	LT162479RU_COPD	COPD	3-Severe COPD	smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 33 %predicted fvc (pre-bd): 51 %predicted fev1 (post-bd): 37 %predicted fvc (post-bd): 59 %predicted dlco: 22 %emphysema (f-950): 29.4 smoker?: 1-Current %predicted fev1 (pre-bd): 29 %predicted fvc (pre-bd): 52 %predicted fev1 (post-bd): 35
GSM1150269	LT163771RU_COPD	COPD	3-Severe COPD	%predicted fvc (post-bd): 66 %predicted dlco: 38 %emphysema (f-950): 0.173 smoker?: 1-Current %predicted fev1 (pre-bd): 61 %predicted fvc (pre-bd): 72 %predicted fev1 (post-bd): 67
GSM1150272	LT168128RU_COPD	COPD	2-Moderate COPD	%predicted fvc (post-bd): 76 %predicted dlco: 87 %emphysema (f-950): 6 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 79 %predicted fvc (pre-bd): 104 %predicted fev1 (post-bd): 92
GSM1150274	LT175949LL_COPD	COPD	1-Mild COPD	%predicted fvc (post-bd): 118 %predicted dlco: 91 %emphysema (f-950): 5.6 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 105 %predicted fvc (pre-bd): 112
GSM1150275	LT176562LL_COPD	COPD	1-Mild COPD	%predicted dlco: 76 %emphysema (f-950): 1.8 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 62 %predicted fvc (pre-bd): 69 %predicted fev1 (post-bd): 70
GSM1150276	LT177956LL_COPD	COPD	2-Moderate COPD	%predicted fvc (post-bd): 75 %predicted dlco: 68 %emphysema (f-950): 1 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 93 %predicted fvc (pre-bd): 95
GSM1150278	LT178929RL_COPD	COPD	1-Mild COPD	%predicted dlco: 72 %emphysema (f-950): 11.759 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 72 %predicted fvc (pre-bd): 87 %predicted fev1 (post-bd): 83
GSM1150279	LT178967RL_COPD	COPD	1-Mild COPD	%predicted fvc (post-bd): 98 %predicted dlco: 52 %emphysema (f-950): 0.4 smoker?: 1-Current %predicted fev1 (pre-bd): 78
GSM1150283	LT184772RL_COPD	COPD	2-Moderate COPD	%predicted fvc (pre-bd): 82

GSM1150284	LT185970RL_COPD	COPD	2-Moderate COPD	%predicted dlco: 60 %emphysema (f-950): 1.1 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 63 %predicted fvc (pre-bd): 92 %predicted fev1 (post-bd): 77 %predicted fvc (post-bd): 106 %predicted dlco: 96 %emphysema (f-950): 0.4 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 58 %predicted fvc (pre-bd): 68 %predicted fev1 (post-bd): 65 %predicted fvc (post-bd): 71 %predicted dlco: 39 %emphysema (f-950): 4.4 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 61 %predicted fvc (pre-bd): 85 %predicted fev1 (post-bd): 59 %predicted fvc (post-bd): 89 %predicted dlco: 43 %emphysema (f-950): 1.7 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 87 %predicted fvc (pre-bd): 100 %predicted dlco: 92 %emphysema (f-950): 1.3 smoker?: 3-Never %predicted fev1 (pre-bd): 37 %predicted fvc (pre-bd): 47 %predicted dlco: 72 %emphysema (f-950): 16.906 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 66 %predicted fvc (pre-bd): 90 %predicted fev1 (post-bd): 68 %predicted fvc (post-bd): 97 %predicted dlco: 42 %emphysema (f-950): 31.3 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 58 %predicted fvc (pre-bd): 100 %predicted fev1 (post-bd): 71 %predicted fvc (post-bd): 126 %predicted dlco: 28 %emphysema (f-950): 11.725 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 51 %predicted fvc (pre-bd): 79 %predicted fev1 (post-bd): 57 %predicted fvc (post-bd): 81 %predicted dlco: 55 %emphysema (f-950): 37.954 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 22 %predicted fvc (pre-bd): 67 %predicted fev1 (post-bd): 23 %predicted fvc (post-bd): 72 %predicted dlco: 47
GSM1150285	LT186521RU_COPD	COPD	2-Moderate COPD	
GSM1150288	LT190004RU_COPD	COPD	2-Moderate COPD	
GSM1150290	LT192758RU_COPD	COPD	1-Mild COPD	
GSM1150291	LT194473RU_COPD	COPD	3-Severe COPD	
GSM1150293	LT195871RU_COPD	COPD	2-Moderate COPD	
GSM1150294	LT197511LU_COPD	COPD	2-Moderate COPD	
GSM1150295	LT198062LL_COPD	COPD	2-Moderate COPD	
GSM1150296	LT198134LU_COPD	COPD	4-Very Severe COPD	

GSM1150297	LT198612RL_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 1.512 smoker?: 1-Current %predicted fev1 (pre-bd): 55 %predicted fvc (pre-bd): 77 %predicted fev1 (post-bd): 59 %predicted fvc (post-bd): 79 %predicted dlco: 66
GSM1150298	LT199384RL_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 1.523 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 54 %predicted fvc (pre-bd): 64 %predicted fev1 (post-bd): 57 %predicted fvc (post-bd): 69 %predicted dlco: 59
GSM1150300	LT203231RM_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 4.959 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 66 %predicted fvc (pre-bd): 88 %predicted fev1 (post-bd): 66 %predicted fvc (post-bd): 90 %predicted dlco: 48
GSM1150301	LT203541RU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 0.8 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 59 %predicted fvc (pre-bd): 91 %predicted fev1 (post-bd): 69 %predicted fvc (post-bd): 96 %predicted dlco: 70
GSM1150303	LT208505LU_COPD	COPD	1-Mild COPD	%emphysema (f-950): 0.5 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 84 %predicted fvc (pre-bd): 84 %predicted fev1 (post-bd): 92 %predicted fvc (post-bd): 97 %predicted dlco: 74
GSM1150304	LT208778RU_COPD	COPD	4-Very Severe COPD	smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 24 %predicted fvc (pre-bd): 78 %predicted fev1 (post-bd): 26 %predicted fvc (post-bd): 90
GSM1150305	LT210463LU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 0.8 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 65 %predicted fvc (pre-bd): 79 %predicted fev1 (post-bd): 74 %predicted fvc (post-bd): 84 %predicted dlco: 34
GSM1150307	LT212777RU_COPD	COPD	4-Very Severe COPD	%emphysema (f-950): 37.8 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 18 %predicted fvc (pre-bd): 58 %predicted fev1 (post-bd): 25 %predicted fvc (post-bd): 76 %predicted dlco: 30
GSM1150308	LT213352LU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 0.5 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 70 %predicted fvc (pre-bd): 75 %predicted dlco: 101
GSM1150309	LT213735RU_COPD	COPD	2-	%emphysema (f-950): 10.7 smoker?:

			Moderate COPD	2-Ever (>100) %predicted fev1 (pre-bd): 63 %predicted fvc (pre-bd): 92 %predicted fev1 (post-bd): 69 %predicted fvc (post-bd): 100 %predicted dlco: 51 %emphysema (f-950): 0.1 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 59 %predicted fvc (pre-bd): 63 %predicted fev1 (post-bd): 56 %predicted fvc (post-bd): 72 %predicted dlco: 48 %emphysema (f-950): 1 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 65 %predicted fvc (pre-bd): 84 %predicted fev1 (post-bd): 74 %predicted fvc (post-bd): 87 %predicted dlco: 94 %emphysema (f-950): 20.4 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 81 %predicted fvc (pre-bd): 100 %predicted fev1 (post-bd): 83 %predicted fvc (post-bd): 98 %predicted dlco: 42 %emphysema (f-950): 48.9 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 45 %predicted fvc (pre-bd): 91 %predicted fev1 (post-bd): 47 %predicted fvc (post-bd): 92 %predicted dlco: 39 %emphysema (f-950): 26.7 smoker?: 1-Current %predicted fev1 (pre-bd): 79 %predicted fvc (pre-bd): 96 %predicted fev1 (post-bd): 78 %predicted fvc (post-bd): 95 %predicted dlco: 66 %emphysema (f-950): 0.004 smoker?: 3-Never %predicted fev1 (pre-bd): 87 %predicted fvc (pre-bd): 89 %predicted fev1 (post-bd): 91 %predicted fvc (post-bd): 90 %predicted dlco: 92 %emphysema (f-950): 11.563 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 59 %predicted fvc (pre-bd): 76 %predicted fev1 (post-bd): 59 %predicted fvc (post-bd): 88 %emphysema (f-950): 17 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 24 %predicted fvc (pre-bd): 42 %predicted fev1 (post-bd): 27 %predicted fvc (post-bd): 47
6	GSM1150311	LT216419RL_COPD	COPD	2-Moderate COPD
	GSM1150312	LT220968RU_COPD	COPD	2-Moderate COPD
	GSM1150315	LT230415RU_COPD	COPD	1-Mild COPD
	GSM1150316	LT233620RU_COPD	COPD	3-Severe COPD
	GSM1150318	LT238531RU_COPD	COPD	2-Moderate COPD
	GSM1150321	LT242119LU_COPD	COPD	0-At Risk
	GSM1150331	LT242530RU_COPD	COPD	2-Moderate COPD
	GSM1150332	LT243058RU_COPD	COPD	4-Very Severe COPD

GSM1150334	LT244480LU_COPD	COPD	4-Very Severe COPD	%predicted dlco: 33 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 28 %predicted fvc (pre-bd): 69 %predicted fev1 (post-bd): 29 %predicted fvc (post-bd): 70 %predicted dlco: 50
GSM1150336	LT245031LL_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 2.6 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 61 %predicted fvc (pre-bd): 85 %predicted fev1 (post-bd): 71 %predicted fvc (post-bd): 104 %predicted dlco: 56
GSM1150338	LT245084RU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 0.5 smoker?: 1-Current %predicted fev1 (pre-bd): 48 %predicted fvc (pre-bd): 79 %predicted fev1 (post-bd): 58 %predicted fvc (post-bd): 82 %predicted dlco: 45
GSM1150339	LT245840RL_COPD	COPD	1-Mild COPD	%emphysema (f-950): 0.126 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 74 %predicted fvc (pre-bd): 103 %predicted fev1 (post-bd): 90 %predicted fvc (post-bd): 108 %predicted dlco: 105
GSM1150341	LT246702RU_COPD	COPD	1-Mild COPD	%emphysema (f-950): 0.419 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 89 %predicted fvc (pre-bd): 100 %predicted dlco: 66
GSM1150343	LT249917LL_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 0.083 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 77 %predicted fvc (pre-bd): 85 %predicted dlco: 83
GSM1150344	LT255244RU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 0.2 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 62 %predicted fvc (pre-bd): 73 %predicted fev1 (post-bd): 66 %predicted fvc (post-bd): 73 %predicted dlco: 48
GSM1150346	LT255718RU_COPD	COPD	4-Very Severe COPD	smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 23 %predicted fvc (pre-bd): 49 %predicted dlco: 27
GSM1150349	LT257433RU_COPD	COPD	0-At Risk	%emphysema (f-950): 0.146 smoker?: 3-Never %predicted fev1 (pre-bd): 101 %predicted fvc (pre-bd): 96 %predicted dlco: 85
GSM1150351	LT261141RU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 21.6 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 46 %predicted fvc (pre-bd): 75 %predicted fev1 (post-bd): 58 %predicted fvc (post-bd): 94 %predicted dlco: 69

GSM1150352	LT262371RM_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 4.1 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 48 %predicted fvc (pre-bd): 84 %predicted fev1 (post-bd): 51 %predicted fvc (post-bd): 94 %predicted dlco: 58
GSM1150353	LT263636RU_COPD	COPD	3-Severe COPD	%emphysema (f-950): 14.9 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 26 %predicted fvc (pre-bd): 71 %predicted fev1 (post-bd): 32 %predicted fvc (post-bd): 81 %predicted dlco: 48
GSM1150357	LT270247RU_COPD	COPD	4-Very Severe COPD	%emphysema (f-950): 38.5 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 25 %predicted fvc (pre-bd): 63 %predicted fev1 (post-bd): 27 %predicted fvc (post-bd): 60 %predicted dlco: 44
GSM1150358	LT271100LU_COPD	COPD	1-Mild COPD	%emphysema (f-950): 13.525 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 86 %predicted fvc (pre-bd): 104 %predicted fev1 (post-bd): 91 %predicted fvc (post-bd): 106 %predicted dlco: 47
GSM1150359	LT273284LL_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 2.5 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 60 %predicted fvc (pre-bd): 90 %predicted fev1 (post-bd): 70 %predicted fvc (post-bd): 94 %predicted dlco: 71
GSM1150362	LT286056RU_COPD	COPD	4-Very Severe COPD	%emphysema (f-950): 25.5 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 19 %predicted fvc (pre-bd): 44 %predicted fev1 (post-bd): 20 %predicted fvc (post-bd): 56 %predicted dlco: 41
GSM1150364	LT298520RU_COPD	COPD	1-Mild COPD	%emphysema (f-950): 2.291 smoker?: 3-Never %predicted fev1 (pre-bd): 93 %predicted fvc (pre-bd): 106 %predicted fev1 (post-bd): 97 %predicted fvc (post-bd): 107 %predicted dlco: 99
GSM1150369	LT005419RU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 4.6 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 78 %predicted fvc (pre-bd): 101 %predicted fev1 (post-bd): 74 %predicted fvc (post-bd): 98
GSM1150383	LT024967LU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 0.5 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 64 %predicted fvc (pre-bd): 83 %predicted fev1 (post-bd): 72

GSM1150388	LT037710RU_COPD	COPD	3-Severe COPD	%predicted fvc (post-bd): 83 %predicted dlco: 82 %emphysema (f-950): 43.7 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 28 %predicted fvc (pre-bd): 80 %predicted fev1 (post-bd): 35 %predicted fvc (post-bd): 99 %predicted dlco: 35 %emphysema (f-950): 50.1 smoker?: 1-Current %predicted fev1 (pre-bd): 35 %predicted fvc (pre-bd): 42
GSM1150389	LT041389RL_COPD	COPD	3-Severe COPD	%predicted fev1 (post-bd): 40 %predicted fvc (post-bd): 48 %predicted dlco: 43 %emphysema (f-950): 28.2 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 28 %predicted fvc (pre-bd): 73
GSM1150393	LT053283RU_COPD	COPD	3-Severe COPD	%predicted fev1 (post-bd): 31 %predicted fvc (post-bd): 84 %predicted dlco: 41 %emphysema (f-950): 33 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 22 %predicted fvc (pre-bd): 62
GSM1150396	LT060717LU_COPD	COPD	4-Very Severe COPD	%predicted fev1 (post-bd): 26 %predicted fvc (post-bd): 59 %predicted dlco: 26 %emphysema (f-950): 25.8 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 43 %predicted fvc (pre-bd): 75
GSM1150406	LT081498RL_COPD	COPD	3-Severe COPD	%predicted fev1 (post-bd): 44 %predicted fvc (post-bd): 84 %predicted dlco: 25 %emphysema (f-950): 31.9 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 22 %predicted fvc (pre-bd): 51
GSM1150408	LT083950RU_COPD	COPD	4-Very Severe COPD	%predicted fev1 (post-bd): 30 %predicted fvc (post-bd): 71 %predicted dlco: 18 %emphysema (f-950): 1 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 59 %predicted fvc (pre-bd): 81
GSM1150409	LT084038RM_COPD	COPD	2-Moderate COPD	%predicted fev1 (post-bd): 59 %predicted fvc (post-bd): 82 %predicted dlco: 67 %emphysema (f-950): 0.3 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 42 %predicted fvc (pre-bd): 62
GSM1150414	LT095342LU_COPD	COPD	3-Severe COPD	%predicted fev1 (post-bd): 45 %predicted fvc (post-bd): 64
GSM1150419	LT108067RU_COPD	COPD	4-Very Severe	%emphysema (f-950): 30.3 smoker?: 2-Ever (>100) %predicted fev1 (pre-

			COPD	bd): 17 %predicted fvc (pre-bd): 58 %predicted fev1 (post-bd): 17 %predicted fvc (post-bd): 56 %predicted dlco: 31 %emphysema (f-950): 26.3 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 28 %predicted fvc (pre-bd): 73
GSM1150430	LT134121RU_COPD	COPD	4-Very Severe COPD	%predicted fev1 (post-bd): 27 %predicted fvc (post-bd): 69 %predicted dlco: 36 %emphysema (f-950): 53.4 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 17 %predicted fvc (pre-bd): 58
GSM1150434	LT139691RU_COPD	COPD	4-Very Severe COPD	%predicted fev1 (post-bd): 19 %predicted fvc (post-bd): 65 %emphysema (f-950): 0.3 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 72 %predicted fvc (pre-bd): 85
GSM1150437	LT144769RL_COPD	COPD	2-Moderate COPD	%predicted fev1 (post-bd): 75 %predicted fvc (post-bd): 85 %predicted dlco: 115 %emphysema (f-950): 1.9 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 24 %predicted fvc (pre-bd): 46
GSM1150450	LT165114RU_COPD	COPD	3-Severe COPD	%predicted fev1 (post-bd): 32 %predicted fvc (post-bd): 55 %predicted dlco: 48 %emphysema (f-950): 2.2 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 50 %predicted fvc (pre-bd): 77
GSM1150458	LT172093RL_COPD	COPD	2-Moderate COPD	%predicted fev1 (post-bd): 54 %predicted fvc (post-bd): 87 %predicted dlco: 49 %emphysema (f-950): 4.3 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 74 %predicted fvc (pre-bd): 90
GSM1150462	LT176510LU_COPD	COPD	2-Moderate COPD	%predicted dlco: 64 %emphysema (f-950): 2.2 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 59 %predicted fvc (pre-bd): 76
GSM1150464	LT184423LL_COPD	COPD	2-Moderate COPD	%predicted fev1 (post-bd): 67 %predicted fvc (post-bd): 97 %predicted dlco: 87 %emphysema (f-950): 26.1 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 22 %predicted fvc (pre-bd): 59
GSM1150465	LT184901RL_COPD	COPD	4-Very Severe COPD	%predicted fev1 (post-bd): 21 %predicted fvc (post-bd): 61 %predicted dlco: 70
GSM1150476	LT196677RU_COPD	COPD	4-Very Severe	%emphysema (f-950): 25.4 smoker?: 2-Ever (>100) %predicted fev1 (pre-

			COPD	bd): 16 %predicted fvc (pre-bd): 38 %predicted fev1 (post-bd): 24 %predicted fvc (post-bd): 55 %predicted dlco: 32
GSM1150479	LT200930RL_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 20.3 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 52 %predicted fvc (pre-bd): 75
GSM1150486	LT215341RU_COPD	COPD	4-Very Severe COPD	%emphysema (f-950): 23.3 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 11 %predicted fvc (pre-bd): 33
GSM1150492	LT223106RU_COPD	COPD	4-Very Severe COPD	%emphysema (f-950): 48.5 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 14 %predicted fvc (pre-bd): 49 %predicted fev1 (post-bd): 15 %predicted fvc (post-bd): 55 %predicted dlco: 37
GSM1150496	LT229669RU_COPD	COPD	4-Very Severe COPD	%emphysema (f-950): 23.7 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 22 %predicted fvc (pre-bd): 62 %predicted dlco: 21
GSM1150499	LT233821RL_COPD	COPD	3-Severe COPD	%emphysema (f-950): 38 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 40 %predicted fvc (pre-bd): 70 %predicted fev1 (post-bd): 48 %predicted fvc (post-bd): 82 %predicted dlco: 41
GSM1150502	LT234774LU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 71.6 (PTX) smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 46 %predicted fvc (pre-bd): 76 %predicted fev1 (post-bd): 51 %predicted fvc (post-bd): 90 %predicted dlco: 42
GSM1150507	LT237439RU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 35.1 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 54 %predicted fvc (pre-bd): 70 %predicted fev1 (post-bd): 58 %predicted fvc (post-bd): 78 %predicted dlco: 87
GSM1150508	LT238765RL_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 6.2 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 50 %predicted fvc (pre-bd): 85 %predicted fev1 (post-bd): 62 %predicted fvc (post-bd): 86 %predicted dlco: 48
GSM1150512	LT244399LU_COPD	COPD	2-Moderate COPD	%emphysema (f-950): 7.8 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 63 %predicted fvc (pre-bd): 88 %predicted fev1 (post-bd): 70 %predicted fvc (post-bd): 95 %predicted dlco: 59
GSM1150514	LT245983LU_COPD	COPD	4-Very	%emphysema (f-950): 46.3 smoker?:

GSM1150520	LT249811RU_COPD	COPD	Severe COPD	2-Ever (>100) %predicted fev1 (pre-bd): 23 %predicted fvc (pre-bd): 72 %predicted fev1 (post-bd): 29 %predicted fvc (post-bd): 85 %predicted dlco: 30 %emphysema (f-950): 4.6 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 53 %predicted fvc (pre-bd): 69
GSM1150523	LT253131RU_COPD	COPD	2-Moderate COPD	%predicted fev1 (post-bd): 55 %predicted fvc (post-bd): 78 %predicted dlco: 54 %emphysema (f-950): 7 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 44 %predicted fvc (pre-bd): 62
GSM1150539	LT285906LL_COPD	COPD	3-Severe COPD	%predicted fev1 (post-bd): 50 %predicted fvc (post-bd): 76 %predicted dlco: 46 %emphysema (f-950): 12.6 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 45 %predicted fvc (pre-bd): 76
GSM1150544	LT291449RL_COPD	COPD	2-Moderate COPD	%predicted fev1 (post-bd): 54 %predicted fvc (post-bd): 85 %predicted dlco: 54 %emphysema (f-950): 21.7 smoker?: 2-Ever (>100) %predicted fev1 (pre-bd): 15 %predicted fvc (pre-bd): 33
			4-Very Severe COPD	%predicted fev1 (post-bd): 17 %predicted fvc (post-bd): 37 %predicted dlco: 28

Tabel.3.2Summarizes the Data

Accession No.	Platform	No. Of Sample
GSE47460	GPL6480 GPL14550	ILD: 254 COPD: 220 CONTROL: 108 TOTAL= 582

3.1.1 Data preprocessing:

The raw data has been normalized using quantile normalization, both normalized and non-normalized expression data were log2-transformed and quantile-normalized before being used to create box plots and expression density plots (Fig.3.1).

A) Non-normalized Expression vs Density plot B) Non-normalized Expression vs Density plot

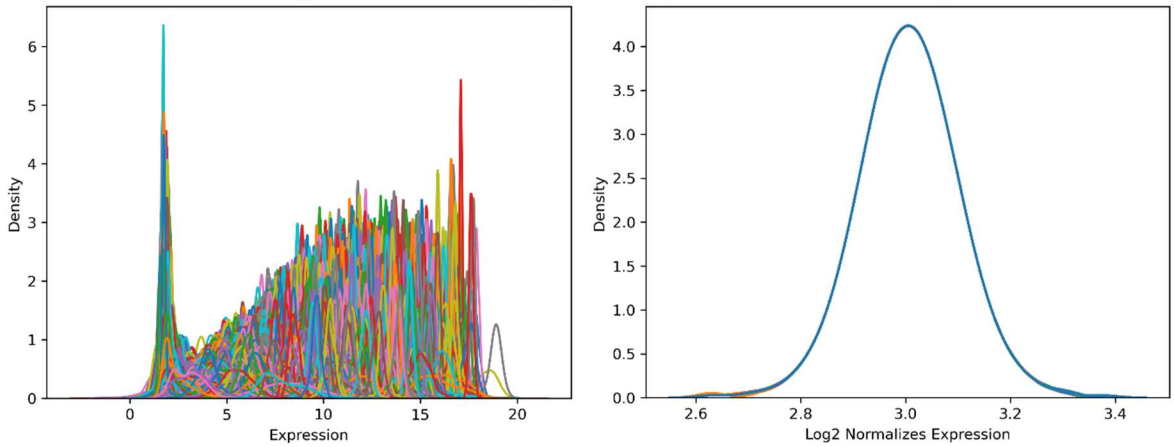


Figure 3.1Expression value for raw and normalized data with quantile normalization technique

3.1.2 Machine learning:

The dataset acquired from GEO (GSE47460) was imbalanced (Fig.3.2(A)). SMOTE technique was used to avoid the biasness caused by majority class. After that, the data is randomly split into 90:10 ratio (90% training and 10% testing). Once the classifier is trained on training data, it is evaluated based on test data and attains the accuracy of 88.1 %. Pre- and post-resampling accuracy of 79.6 Fig.3.2(B) and 88.1Fig.3.2(D) percent respectively have been attained by the model. After that, we also generated the classification report including Precision, Recall and F1 Score for each class (fig.3.2(C)(E)). The results demonstrate that SMOTE led to an increase in accuracy as well as precision recall, and F1-score for all classes. SMOTE's success in resolving the class imbalance problem indicates that the model has become better at classifying the classes namely COPD, ILD and CONTROL. Furthermore, we also validated our model with a StratifiedKFold (splits=5) and achieved the mean accuracy of 86.48 % with the mean deviation of 0.032 across the splits and the metrics are summarized in Fig3.2(F).

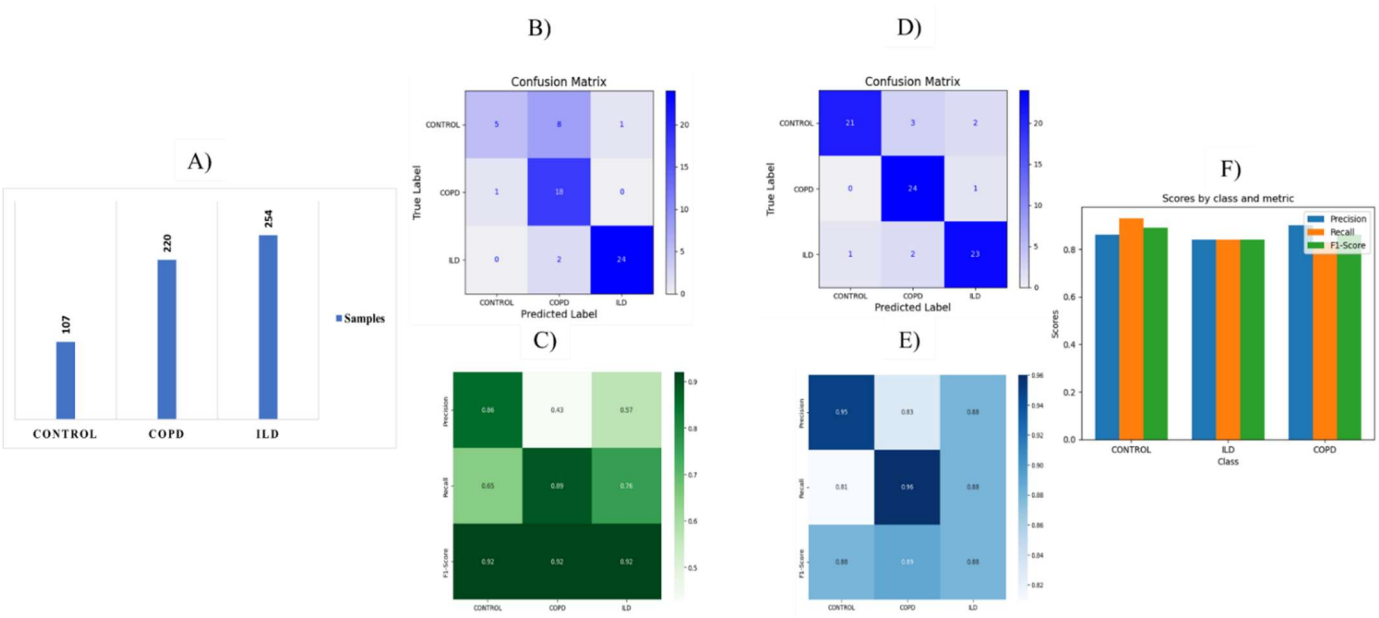


Figure 3.2In Fig.3.2(A) Sample Distribution across the Dataset. (B) and (D) Confusion matrix summarizing the impact of SMOTE resampling technique. In fig (3.2B) the model is biased towards the majority class i.e., COPD and ILD (79.6%), while in fig(3.2D) displays the effect of SMOTE on classification ability of model with an increase accuracy of 88.1%. Fig(3.2C) and 3.2(E) Heatmap summarization of effect of SMOTE on the classification report for the model. Furthermore, Fig.3.2(F) classification performance of classifier using five-split cross validation.

3.1.3 Integrating SHAP for explaining the model:

To evaluate the contribution of each gene in classifying the classes, namely: control, COPD, and ILD, we integrate the SHAP library to acquire the SHAP values. In classifying the conditions, the genes are arranged in descending order (measured by the average effects on model output magnitude) (Fig. 3.3). Each bar is split into three colors, blue, pink, and green, that represent the three classes to be classified ('1' = COPD, '0' = CONTROL, and '2 = ILD). Gene involvement in class prediction (as indicated by SHAP values) is depicted by the width of each color within a bar [98]. The contribution of the gene ADRB2 is found to have a more significant role in classifying ILD, followed by control and COPD. Similarly, the CDH3 gene contributes more to predicting COPD, ILD, and CONTROL. Furthermore, SHAP summary plots Figs. 3.3(a), 3.3(b), and 3.3(c) were generated to get the insight of positive and negative relation with the genes. The 20 most significant genes for ILD, COPD, and the control class, respectively are illustrated on y-axis. The genes are arranged in descending order according to their impact on the model. The red and blue color represents the level of impact (significant(red) or minimal (blue)) of each gene on the model's classification task. For instance, in Class=1 (COPD), gene ADRB2 has a positive impact on the model. The 'high' can be observed by 'red' color and the positive impact indicated on x-axis.

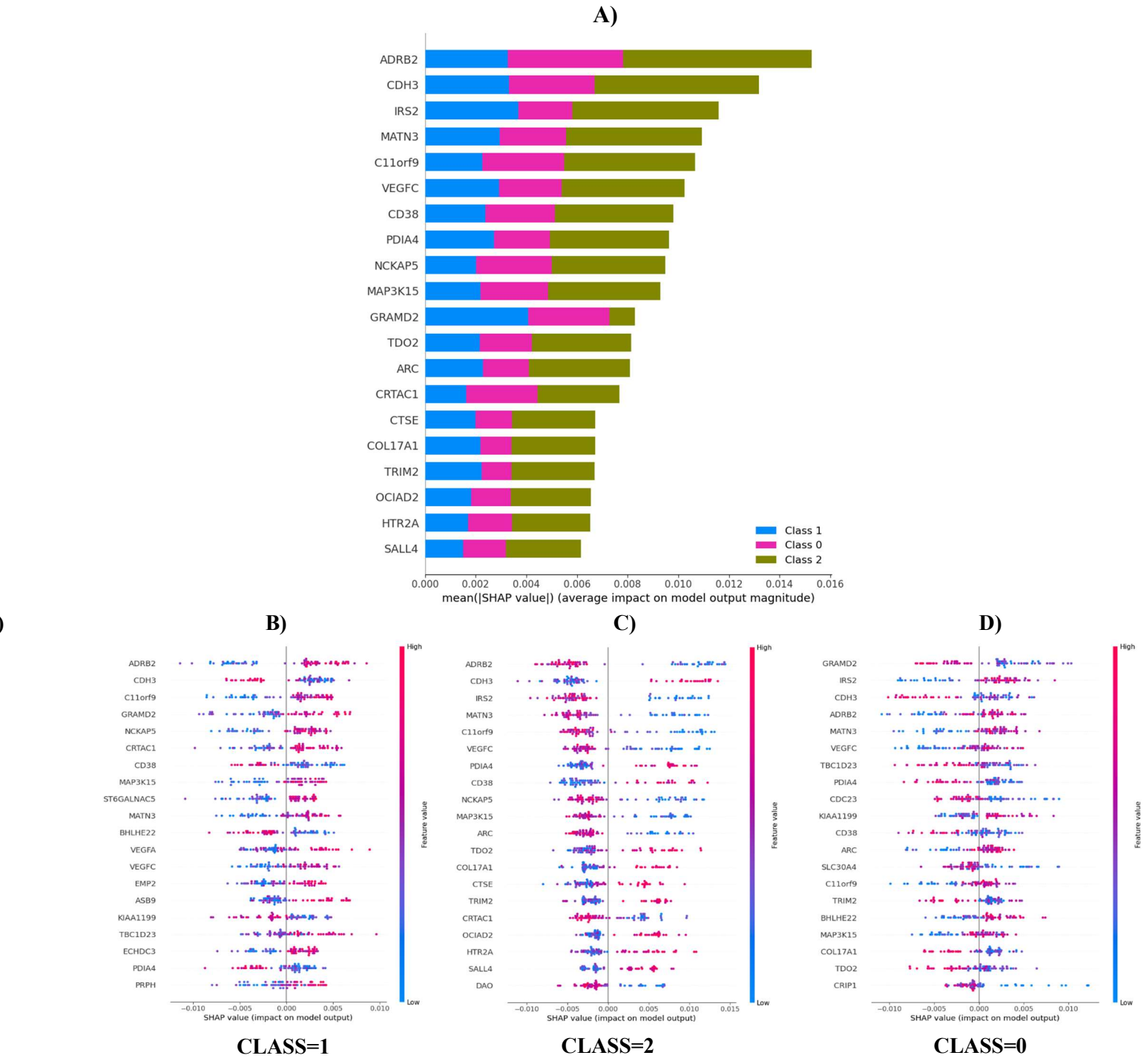


Figure 3.3In Fig.3.3(A)The SHAP bar plot displays the common genes in classifying the conditions COPD and ILD (Class = 1, Class =2) including control (Class=0). Fig 3.3(B,C,D). SHAP Summary plot that depicts the top 20 genes in predicting each class, where class1 is COPD, 2 is ILD and 0 for control.

In fig.3.3, the SHAP bar plot for multiple classes i.e., COPD, ILD and CONTROL, the gene ADRB2, is the most relevant gene in classifying all the three classes . Previously done studies, suggest that ADRB2 act as regulator for airway smooth muscle tone [101]and responsible for elevation of bronchodilation, ventricular function, and vasodilation [101]The high level of ADRB2 decreases lung function and is associated with COPD and asthma [101]. In another study reveals that ADRB2 signaling is crucial for regulating inflammation via the fast activation of IL-10[102]. The fact that ILD is characterized by inflammation that makes it difficult for the lungs to take in adequate oxygen [103]. Based on its known relationship with COPD and asthma, as well as its involvement in regulating inflammation and lung function, we speculate that ADRB2 not only involved in both ILD and COPD but also in CPFE. The second ranked gene CDH3 was found to be related to lung function and also act as a prognostic factor to evaluate the prognosis and survival status of IPF patients [104].The third ranked gene IRS2 was found to be up regulated in patients with COPD when compared to healthy patients [105].In another study the same gene IRS2 tends to contribute to anti-inflammatory and antiproliferative outcome during hypoxia by inhibiting the Akt-FOXO1 activation pathway in lungs[106]. The fourth gene MATN3 was found in a study on RNA-seq transcriptome analysis of lung tissues from 68 subjects revealed differential gene expression between younger and older groups. Genes such as MATN3, MAP3K15, CHRM2, GALNT13, COL17A1, and EDA2R exhibited age-related expression patterns. Multivariate analysis identified EDA2R as a significant risk factor for lung aging, providing insights into molecular mechanisms contributing to age-associated lung diseases [107].The Sixth ranked gene in SHAP bar plot is VEGF-C, one of the member of VEGF (vascular endothelial growth factor) family (also includes VEGF-A,B,C,D,E and F)which plays a significant role in physiological as well as pathological angiogenesis [108]. In a study, the authors observed the pattern of pulmonary VEGF-C and VEGFR-3 protein expression and the consistent presence of VEGF-C protein in tracheal aspirate fluid in human premature infants show that VEGF-C is important for the proper development of the lymphatic system in the lungs [109]. In other research, the finding suggests that the individuals with systemic sclerosis (SSc)-related ILD had lower levels of vascular endothelial growth factor A (VEGF-A) in bronchoalveolar lavage fluid (BALF) compared to both healthy controls and SSc patients without lung involvement. Blood levels of VEGF-A were directly correlated with the severity of ILD on both imaging and lung function tests [110].Similarly, a study suggested that VEGF and its corresponding receptors have a significant role in various pathophysiological processes associated with COPD, including bronchial wall remodeling, emphysema, and pulmonary hypertension [111]. The seventh ranked gene CD38 was involved in Enzymes involved in the metabolism of nicotinamide adenine dinucleotide (NAD) are increasingly being implicated in chronic airway diseases. One such enzyme, CD38, utilizes NAD to produce several metabolites, including cyclic ADP ribose (cADPR), which is involved in calcium signaling in airway smooth muscle (ASM) [112]. In another study, the eighth ranked gene namely PDIA4 was found to in the stress response within Alveolar Epithelial Type II cells (AECII) due to defects in surfactant processing, particularly in the context of lung diseases such as Idiopathic Pulmonary Fibrosis (IPF), a form ILD. The upregulation of PDIA4, alongside other chaperones, in response to the formation of β-amyloid structures by surfactant proteins SP-B and SP-C, highlights its potential role in mitigating protein misfolding and endoplasmic reticulum stress [113]. As PDIA4, including protein misfolding and cellular stress responses in lung epithelial cells, we can speculate that they are also relevant in the pathophysiology of COPD.

3.1.4 Analyzing gene sets and the pathways linked to identified key genes:

Functional enrichment analysis is a commonly employed approach for identifying commonalities in extensive biological datasets. In the realm of biomedicine, the analysis of gene expression data using functional enrichment is a prevalent method for unraveling disease mechanisms. Various methodologies have been developed to categorize regulated gene expression profiles into distinct functional groups. The choice of these functional categories, often derived from literature sources, typically reflects signaling or metabolic pathways. SRplot[68] platform is used for enrichment analysis on the genes obtained through SHAP values, aiming to comprehend their involvement in the disease. Dot plot fig.3.4(A) for pathway analysis and bar plot Gene ontologies including biological process (BP), cellular component (CC) and molecular function (MF) fig.3.4(B) are being acquired. In the pathway analysis the identified genes are enriched in calcium signaling, regulation of lipolysis in adipocytes, and salivary secretion. While in the case of GO BP the genes are enriched in smooth muscle contraction, negative regulation of synaptic transmission. The pathophysiology of asthma and COPD involves the contribution of airway smooth muscle to both contractility and inflammation. While in the case of GO CC the genes are significant in endoplasmic reticulum lumen. Exposure to environmental toxins and pathogens, common in respiratory diseases, can induce endoplasmic reticulum (ER) stress. The ER, crucial for protein synthesis and regulation, faces misfolded protein accumulation during stress, activating the unfolded protein response (UPR) [114]. Similarly, for MF the genes are enriched in G-coupled amine receptor activity which can enhance which either enhance bronchodilation or prevent bronchoconstriction [115].

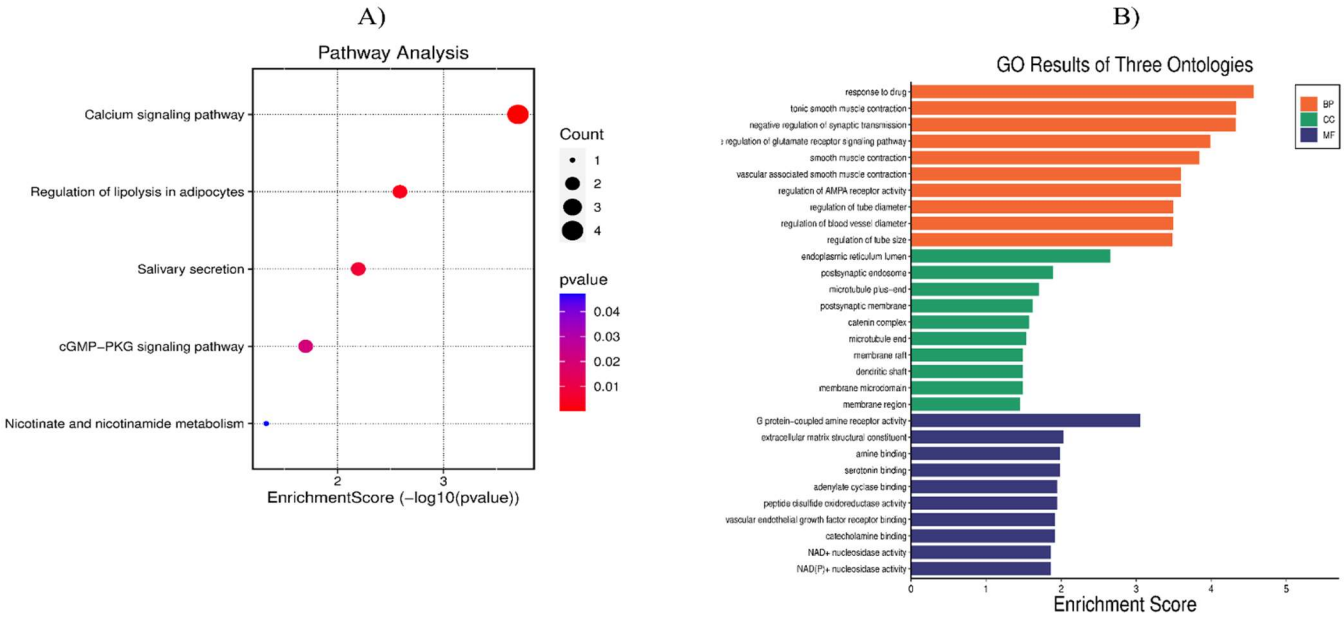


Figure 3.4(A) Pathway enrichment analysis of the identified key genes from SHAP values (B)Functional enrichment analysis of the identified key genes from SHAP values (GO ontology)

3.1.5 Gene Profiling and Validation:

In our study, we utilized the idep 1.0 tool to generate heatmaps, as illustrated in Figures 3.5(A) and 3.5(B), to showcase the differential expression of the top 20 genes identified from SHAP bar plots. These heatmaps effectively demonstrate the segregation of different classes based on biomarkers, with CONTROL being represented in red, COPD in green, and ILD in blue. This visual representation not only highlights the distinct gene expression profiles among these groups but also underscores the potential of these biomarkers in distinguishing between these conditions. After that**,** delving deeper into the implications of these findings, we reanalyzed the top genes, including ADRB2, CDH3, IRS2, MATN3, VEGFC, CD38, and PDIA4, by reintroducing them into our model. This subsequent analysis, the results of which are depicted in Figure 3.5(C), yielded an accuracy of 87.77%. This high accuracy rate ensures the robustness of our model and the pivotal role these genes play in its performance. Importantly, this approach not only validates the effectiveness of the identified genes in classifying CONTROL, COPD, and ILD cases but also opens avenues for further investigation into their biological relevance and potential therapeutic targets for CPFE also. We have also incorporated Matthews Correlation Coefficient (MCC) [116]for our model. Our model has achieved MCC of approximately 0.8282, demonstrates strong performance in multi-class classification tasks for COPD, ILD and CONTROL as depicted in fig 3.5 (D).’

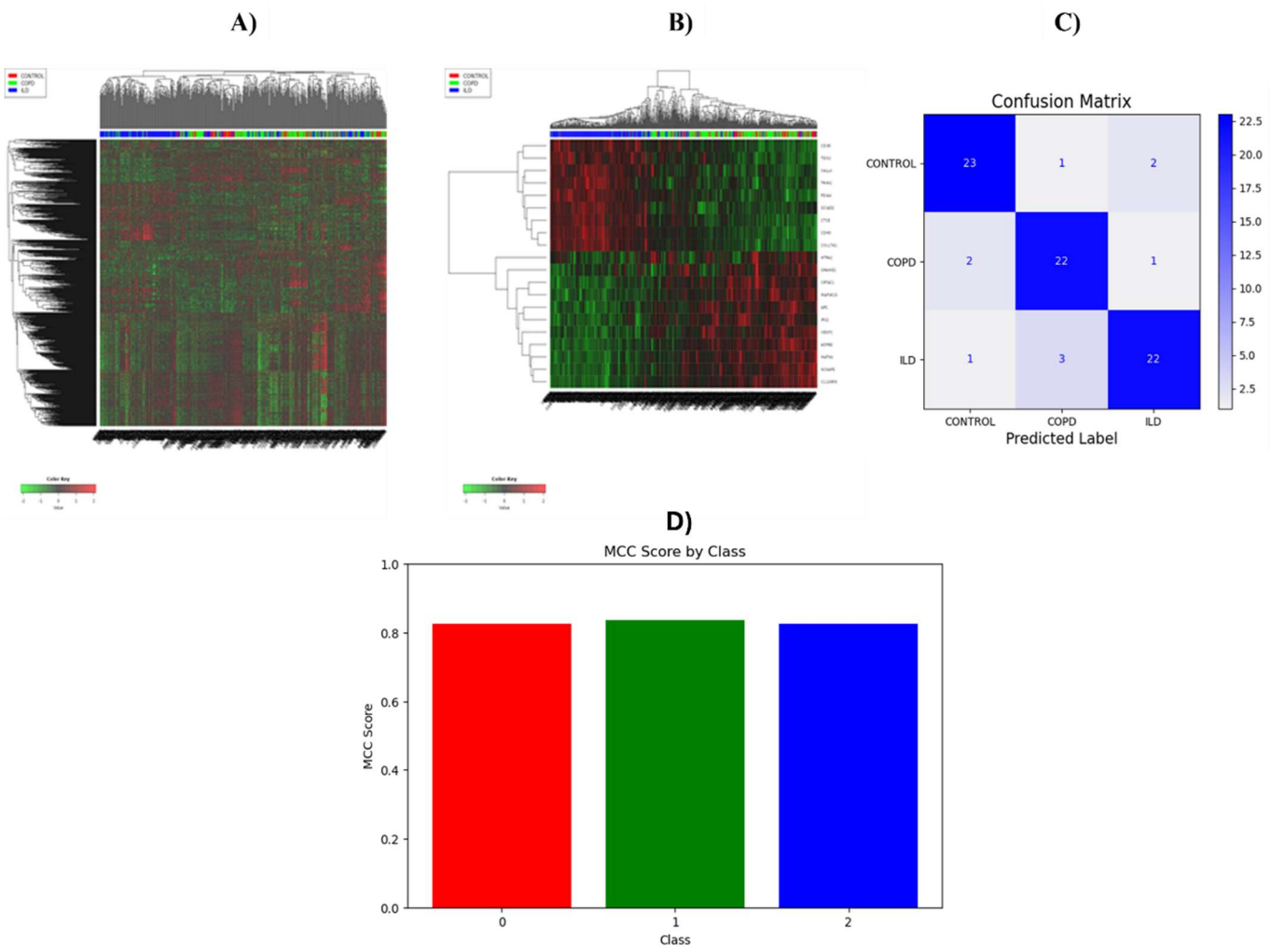


Figure 3.5Heatmap of the identified genes. A) Illustrates all the genes while B) shows top 20 genes. Genes with a higher expression than the mean are green and those with lower expression than the mean are in red. Genes without any significant expression are black Fig3.5(C) Confusion metrics with tops identified genes. Fig 3.5(D) MCC of 0.8282 with class=0 is control, class=1 COPD and class=2 is ILD.

Furthermore, we have incorporated GeneMania [70]to understand the interaction of top identified gene. Almost all genes were implicated in co-expression (77.7%) fig.3.6(A) and rest implicated with physical interaction (22.21%) fig.3.6(B). Next, we have incorporated the Single Cell portal to investigate the expression of identified genes at the single-cell level. Dot plots have been created for disease ontology and cell-type ontology, demonstrating gene expression patterns using scaled mean expression. The scaling is relative to each gene's expression across all cells within the selected annotation category. In fig.3.6 (C) the genes namely IRS2 and PDIA4 were the most significant in COPD, pulmonary systematic and pulmonary fibrosis when compared to normal and in fig.3.6 (D) the genes were enriched in various cell types including apoptosis fated cells, mitotic cell cycle and so on.

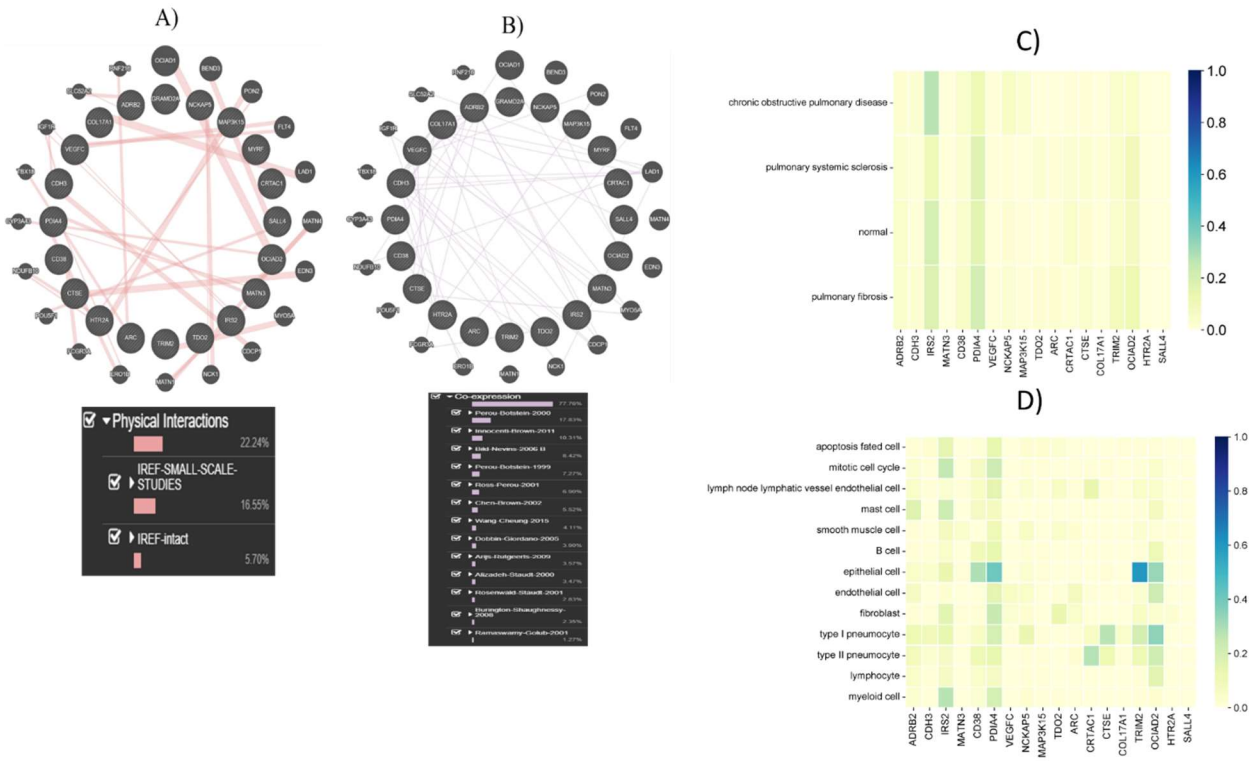


Figure 3.6GeneMania analysis reveals the interaction network of top identified genes. 3.6(A): Co-expression analysis illustrates the involvement of almost all genes (77.7%). 3.6(B): Physical interaction network highlights 22.21% of implicated genes. 3.6(C): Single -cell expression analysis using the Single Cell portal indicates significant association of genes with COPD, pulmonary systematic, and pulmonary fibrosis compared to normal. 3.6(D): Enrichment analysis across various cell types of identified genes.

Following this, we utilized the miRNet 2.0 database to construct a comprehensive miRNA-mRNA interaction network, essential for understanding the regulation of gene expression variations in these diseases. Our in-depth analysis involved mapping 20 crucial genes, identified as significant through SHAP analysis, against the human reference database in miRNet 2.0. This approach has culminated in the development of a robust network comprising 465 miRNA nodes and 18 gene nodes, interconnected through an extensive network of 696 edges (Fig 3.7).

This network revealed a complex array of miRNA-gene interactions. Prominent among the genes were OCIAD2, IRS2, and TRIM2, which demonstrated the highest degrees of connectivity (136, 97, and 93, respectively) and significant betweenness centrality (44626.78, 32257.86, and 31009.88, respectively), indicating their pivotal roles in the molecular landscape of COPD and ILD. Other genes such as VEGFC, MATN3, and SALL4 also emerged as key genes within this network.

Additionally, our analysis identified miRNAs with notable roles in this intricate network. For instance, hsa-mir-101-3p, hsa-mir-1343-3p, and hsa-mir-27a-3p were among those with significant interactions and network positions, suggesting their potential regulatory impact on the disease process fig3.8(A). The application of a degree cutoff of 5 to the miRNA nodes refined the network, highlighting key miRNAs like hsa-mir-101-3p, hsa-mir-1343-3p, and hsa-mir-27a-3p, each with significant degrees of interaction and betweenness fig3.8(B). This selective approach

enabled us to concentrate on the most influential miRNAs thus providing a focused view on the potential molecular biomarkers in COPD and ILD. The resultant network after the cutoff maintained a subset of the original nodes and edges, ensuring a targeted analysis while preserving the integrity and complexity of the biological interactions pertinent to these lung diseases.

This comprehensive miRNA-gene network not only sheds light on the complex regulatory mechanisms underlying COPD and ILD but also provides a valuable resource for further research in CPFE also. The identified genes and miRNAs could serve as potential biomarkers or therapeutic targets, offering new insights into the pathophysiology of these challenging lung diseases.

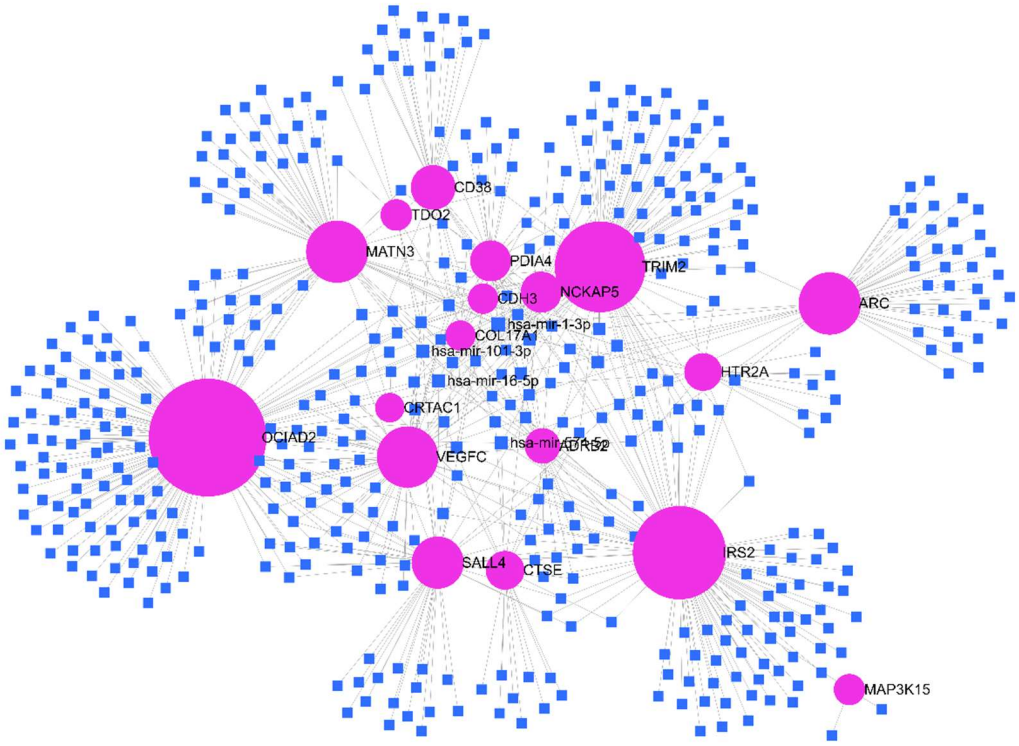


Figure 3.7miRNA-mRNA interaction of top 20 gene altered in COPD and ILD patients.

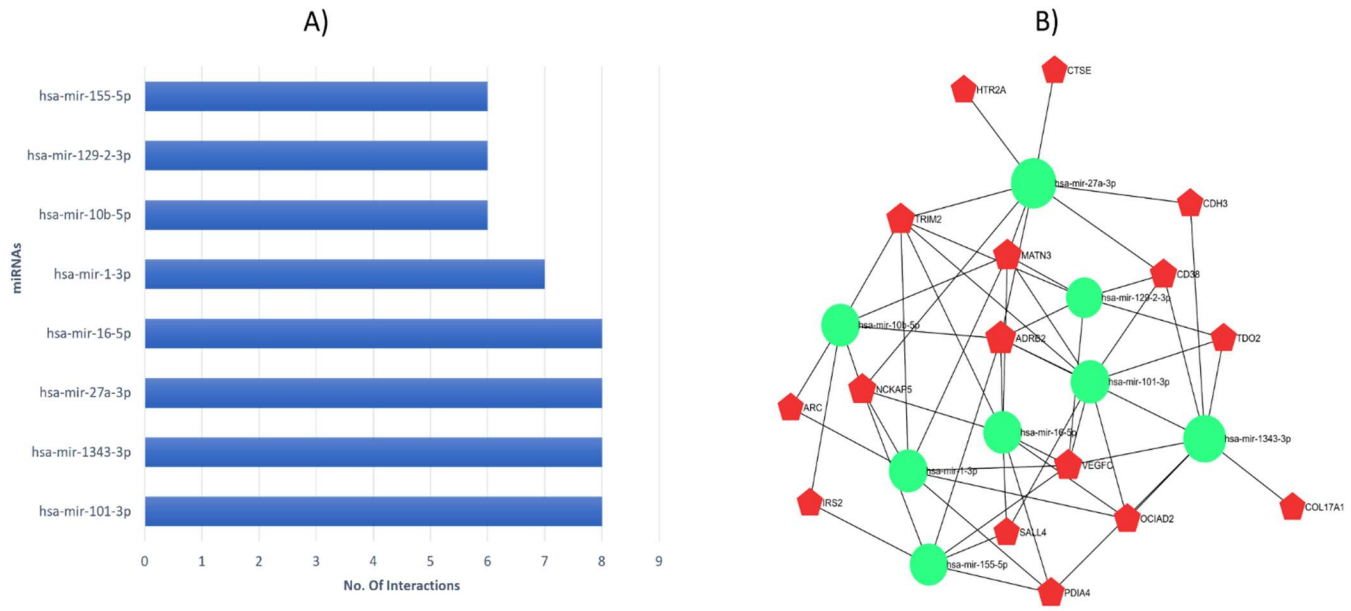


Figure 3.8Interactions between genes and miRNAs. (A) Significant miRNAs after the enrichment analysis are represented by a bar plot, displaying the number of its target genes (B) This network represents the mRNA-miRNA interaction based on degree cutoff of 5 is visualized using Cystoscope.

3.1.6 Key Findings

- Using the GSE47460 dataset (254 ILD, 220 COPD, 108 controls), a Random Forest model with SMOTE achieved 88.1% accuracy and an MCC of 0.8282, confirming robust classification of COPD, ILD, and controls.

- SHAP analysis highlighted 20 key genes, including ADRB2, CDH3, IRS2, MATN3, VEGFC, CD38, and PDIA4, linked to airway tone, fibrosis, angiogenesis, hypoxia, and ER stress. Functional enrichment confirmed involvement in cytokine signaling, ECM remodeling, calcium signaling, and smooth muscle contraction.

- Network and single-cell analyses underscored IRS2, PDIA4, VEGFC, and MATN3 as central drivers, while the miRNA–mRNA network identified hsa-miR-101-3p, hsa-miR-1343-3p, and hsa-miR-27a-3p as major regulators.

- Overall, the findings demonstrate that COPD–ILD coexistence and CPFE are driven by shared molecular signatures, with these genes and miRNAs emerging as promising biomarkers for early detection and patient stratification.

3.2 Results for objective 2 : To explore systemic inflammatory crosstalk between neuro-inflammatory disorders and respiratory diseases.

3.2.1 Incorporation of Machine Learning:

The dataset acquired from GEO (GSE61741) was imbalanced (Fig.3.9(A)). We used several ML classifiers to examine the unbalanced dataset and optimize model performance while reducing overfitting. We used Grid Search Cross-Validation hyperparameter adjustment to fine-tune the RF classifier after SMOTE addressed class imbalance. The best hyperparameters were max_depth = 10, min_samples_leaf = 2, min_samples_split = 10, and n_estimators = 200 (Supplementary Table S3). This included max_depth [5, 10, 15], min_samples_leaf [1, 2, 4], min_samples_split [2, 5, 10], and n_estimators [100, 150, 200, 300]. The optimal values were selected based on mean cross-validation accuracy.

As described in the Methods section, stratified K-fold cross-validation was used to ensure the dependability of our results. This method robustly assessed the model's generalizability to new data. The final model had 81.58% test accuracy and 97.62% AUC, indicating good class discrimination. Performance matrixes are summarized in Fig.3 and cross validation in fig.3.9(F). We also tested RF, XGB, LR, and SVM classifiers. In accuracy, the RF classifier performed best. The RF model has an accuracy of 81.58%, while other classifiers performed poorly, proving its efficacy. Comparison are summarized in fig. 3.9(G).

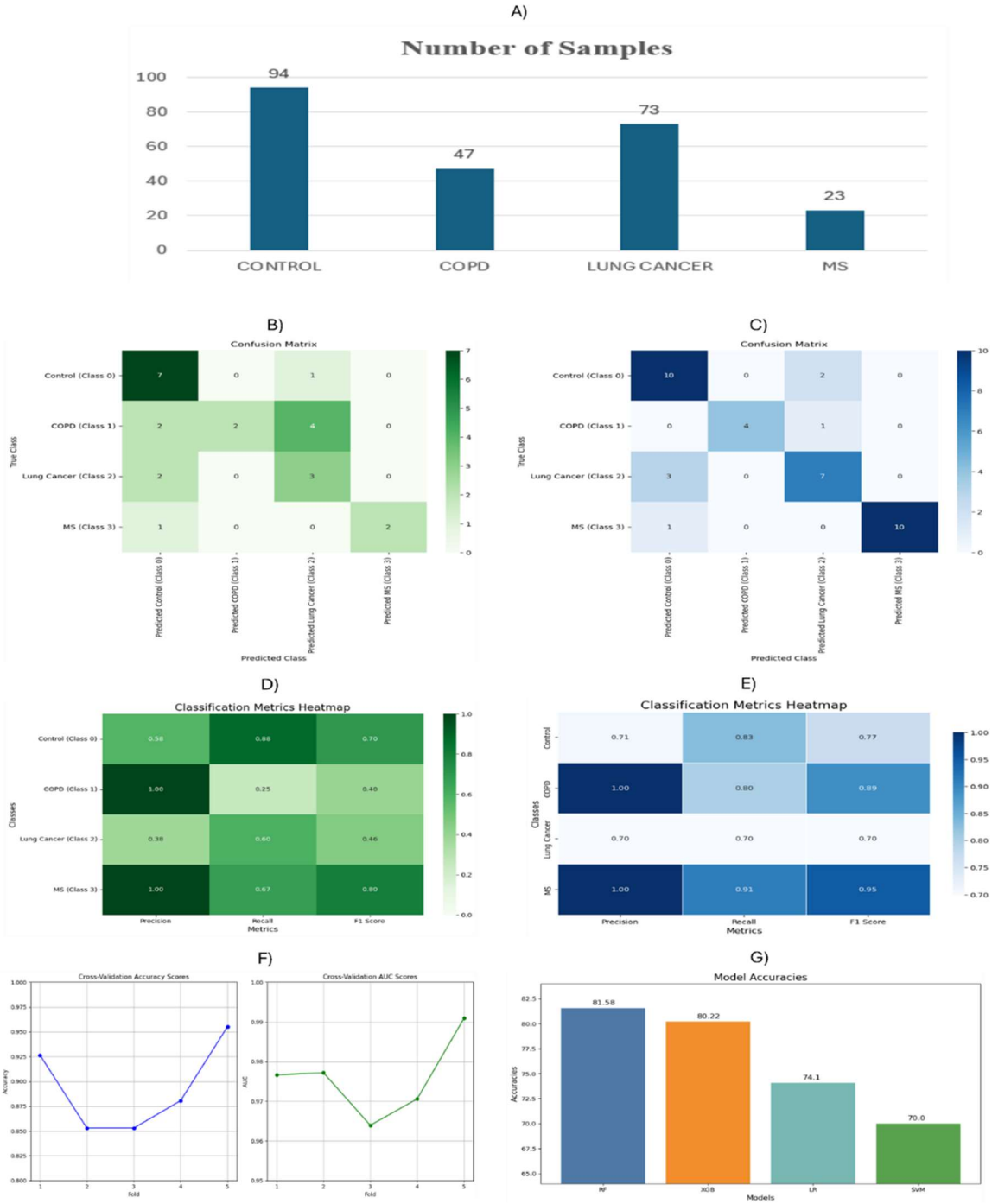


Figure 3.9 (A) Displays the distribution of samples across the four classes: Control, COPD, MS, and Lung Cancer, highlighting the imbalance present in the dataset. (B) and (C) compare model performance before and after applying SMOTE; while (B) shows poor classification with an accuracy of 58.33%, (C) reflects the enhanced accuracy of 81.58% after SMOTE. (D) and (E) present heatmaps of classification metrics pre- and post-SMOTE, illustrating significant improvement in class-wise precision and recall. (F) shows the stratified 5-Fold cross-validation results, confirming the robustness of the Random Forest model. (G) compares classifier performance across RF, XGBoost, SVM, and LR, establishing RF as the most accurate model.

Finally, to further validate the model we have done independent validation with dataset (GSE31568) and archived accuracy of 82.55%. Classification report, including class-wise precision, recall, and F1-scores alongside AUC values of 0.94 (Control), 0.99 (COPD), 0.93 (Lung Cancer), and 1.00 (MS), are summarized in fig.4(A&B). This consistency demonstrates that the model generalizes well to unseen data and performs reliably across datasets fig.3.10.

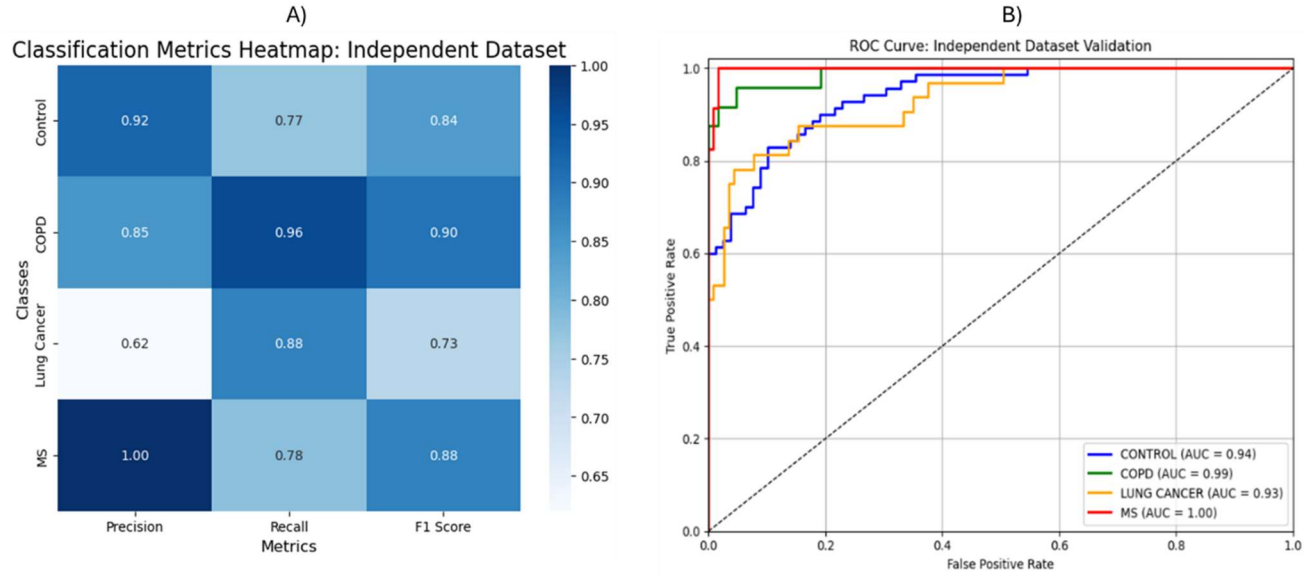


Figure 3.10(A) Presents the classification report of the Random Forest model on the independent validation dataset (GSE31568), including class-wise precision, recall, and F1-scores for Control, COPD, Lung Cancer, and MS. (B) illustrates the corresponding ROC curves for each class, showing high discriminatory power with AUC values of 0.94 (Control), 0.99 (COPD), 0.93 (Lung Cancer), and 1.00 (MS), confirming the model's generalizability and robustness.

3.2.2 Incorporation of SHAP for explaining the model:

We use the SHAP library to calculate SHAP values to assess each miRNA role in classifying control, COPD, MS, and lung cancer. The average effects on model output magnitude rank the miRNA in ascending order (measured by the average effects on model output magnitude) (Fig. 3.11). The four classifications that need to be identified are represented by the four colors and that each bar is divided into: green, red, purple, and blue ('1' = COPD, '0' = CONTROL, '2' = lung cancer, and '3' = MS respectively)[12,98]. The contribution of the miRNA in class prediction is measured by the width of the colors within a bar, which is calculated using SHAP values. For instance, in fig5 (A) the impact of hsa-let-7c miRNA is significant in classifying MS and lung cancer followed by COPD and control. Similarly, miRNA hsa-miR-454 has more impact in classifying MS and Lung cancer followed by COPD and control. Subsequently, miRNA hsa-92a has a significant role in classifying MS and COPD followed by lung cancer and control.

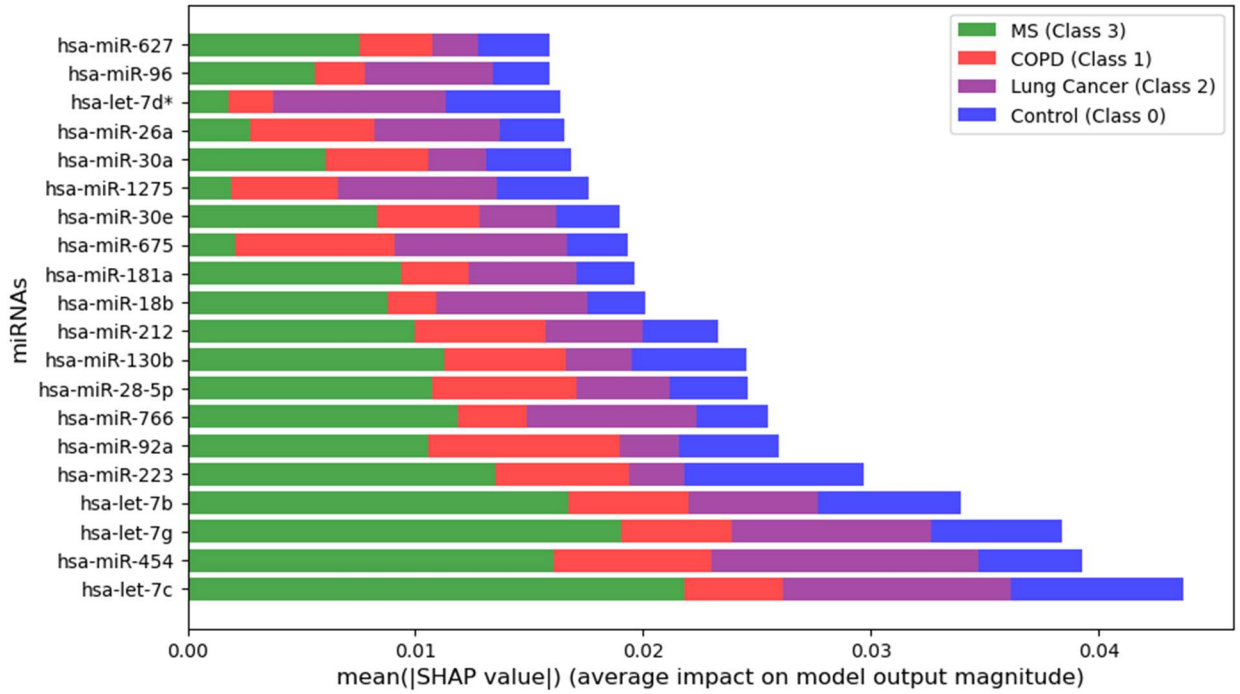


Figure 3.11Shows a SHAP summary bar plot ranking the top 20 miRNAs contributing to classification decisions across the four disease groups. The color segmentation within each bar (green for MS, red for COPD, purple for Lung Cancer, and blue for Control) represents the relative SHAP value contribution of each miRNA to class-specific predictions. The width of each color section indicates the strength of that miRNA's contribution to a given class, with hsa-let-7c, hsa-miR-454, hsa-miR-92a, and hsa-miR-223 emerging as key discriminators across multiple disease contexts.

In the SHAP bar plot, hsa-let-7c, let-7g, and let-7d ranked among the top four miRNAs contributing to the classification of all four classes—Control, COPD, Lung Cancer, and MS—highlighting their central role in shared pathophysiological mechanisms. Members of the let-7 family, including hsa-let-7c, let-7b, and let-7g, have been widely implicated in inflammation and immune regulation across MS, COPD, and lung cancer. Notably, hsa-let-7c-5p has been shown to reduce neuroinflammation and preserve neuronal function in MS by suppressing microglia and macrophage activation [117]. According to [118], who profiled miRNAs across CD4+ T cell subsets and found conserved expression patterns in MS patients, with let-7 family involvement. [119] further linked circulating miRNAs, including let-7 members, to MS stage and disability. [120] demonstrated that CSF levels of let-7b-5p—a close family member of let-7c—correlated with neuroinflammation and MS progression. In COPD, multiple studies reported downregulation of let-7c in patient lung tissue and animal models, with functional roles in inhibiting IL-6/STAT3 signaling, macrophage M2 polarization, and myofibroblast differentiation ([121], [122]). These findings reinforce its regulatory role in airway remodeling and emphysema progression. In lung cancer, hsa-let-7c acts as a tumor suppressor by inhibiting targets such as ITGB3 and MAP4K3, thereby reducing metastasis and epithelial invasion [123]. Low expression of let-7 family miRNAs is associated with advanced stage and poor prognosis in NSCLC. Collectively, these studies validate the biological importance of hsa-let-7c as a cross-disease miRNA biomarker, targeting

inflammatory, oncogenic, and immunoregulatory pathways such as TGF-β, PI3K-Akt, and MAPK, linking neurological and respiratory inflammation through shared molecular mechanisms. The second list in hsa-miR-454 which was found to be downregulated in peripheral blood of patients with all MS subtypes(primary progressive, secondary progressive and relapsing-remitting disease) [124]. Further, a bioinformatics and in vitro research have indicated that miR-454 is linked to the PI3K-Akt signaling pathway, which affects COPD pathology-related inflammatory and cellular processes. Overexpression of miR-454 in lung cancer worsens prognosis by increasing tumor proliferation, invasion, and apoptosis, making it a key target for drug treatment[125]. The fifth miRNA from the plot was hsa-miR-223 and the expression of miR-223 in myeloid-derived suppressor cells (MDSCs) from patients with multiple sclerosis (pwMS) is responsible for the diminished presence of MDSCs in these individuals, and inhibiting miR-223 could enhance MDSC-mediated regulation of T cell proliferation through STAT3 activation. The authors observed enhanced suppressive activities of MDSCs following GC therapy; however, this impact was independent of STAT3[126]. On the other hand, In COPD and lung cancer, miR-223 regulates inflammation and immunity. In COPD, miR-223 supports neutrophil recruitment and macrophage activation, regulating pro-inflammatory pathways such NF-κB and NLRP3 inflammasome. It modulates disease development by affecting airway remodeling and cellular proliferation. MiR-223 suppresses lung cancer by blocking pathways like TGF-β and p53 that promote cell proliferation, apoptosis, and invasion. Thus, miR-223 may be a diagnostic and therapeutic target in both disorders [127]. Further, hsa-mir-92a gets sixth position in bar plot and it was found that patients with multiple sclerosis had higher levels of miR-92a in their CD4+ T cells, and inhibiting this gene in these patients' T cells increased Treg formation but decreased Th17 differentiation [128]. Similarly, it was found that miR-92a-3p plays a role in COPD by influencing inflammation and immune response. Its elevated levels in blood affect gene regulation related to lung function and COPD severity, making it a potential biomarker for disease progression. Efforts aim to minimize its interference in detecting COPD-specific biomarkers[129].

3.2.3 miRNA profiling and validation:

In fig.3.12 (D, E, F, G) highlights the top 20 miRNAs for individual class prediction namely, Control, COPD, Lung cancer and MS. In these illustrations, each miRNA on the *y*-axis is accompanied by a horizontal bar that denotes the miRNA's impact on the model's prediction. The red and blue colors denote the degree of impact (major in red or minimal in blue) of each miRNA on the model's classification task. For example, in Class= '1' which is COPD, the miRNA hsa-miR-92a has a positive impact. The 'high' can be observed by 'red' color and the positive impact indicated on x-axis([12,98]).

To display the differential expression of the top 20 miRNAs found in SHAP bar plots, we used the idep 1.0 tool to create heatmaps, as shown in Figures 3.12(A) and 6(B). These heatmaps clearly show the split of many classes based on biomarkers, with CONTROL in red, COPD in green, MS in blue, and lung cancer in cyan. This picture depicts the distinct miRNA expression profiles of various groups and demonstrates the potential of these biomarkers in distinguishing between these disorders. To further validate, we decided to reintroduce top miRNAs including hsa-let7c, hsa-miR-454, hsa-let-7g, let-7b, hsa-miR-92a and miR-223 and achieved an accuracy of 89.47%[12,130]. This increase in accuracy rate ensures the resilience of our model and the critical role these miRNAs play in its performance. Furthermore, we found that hsa-let-7c, hsa-let-7g, hsa-let-7b, and hsa-miR-454 from the SHAP individual violin plots were common contributors across all classes (COPD, MS, Lung cancer and control). Selecting common miRNAs from SHAP plots can reveal shared biological pathways implicated in disease progression. This method helps us

identify miRNAs that are statistically significant in our model and biologically important to cross-disease interactions, enabling further research into their roles in disease interconnection(Fig.3.12.(H)).

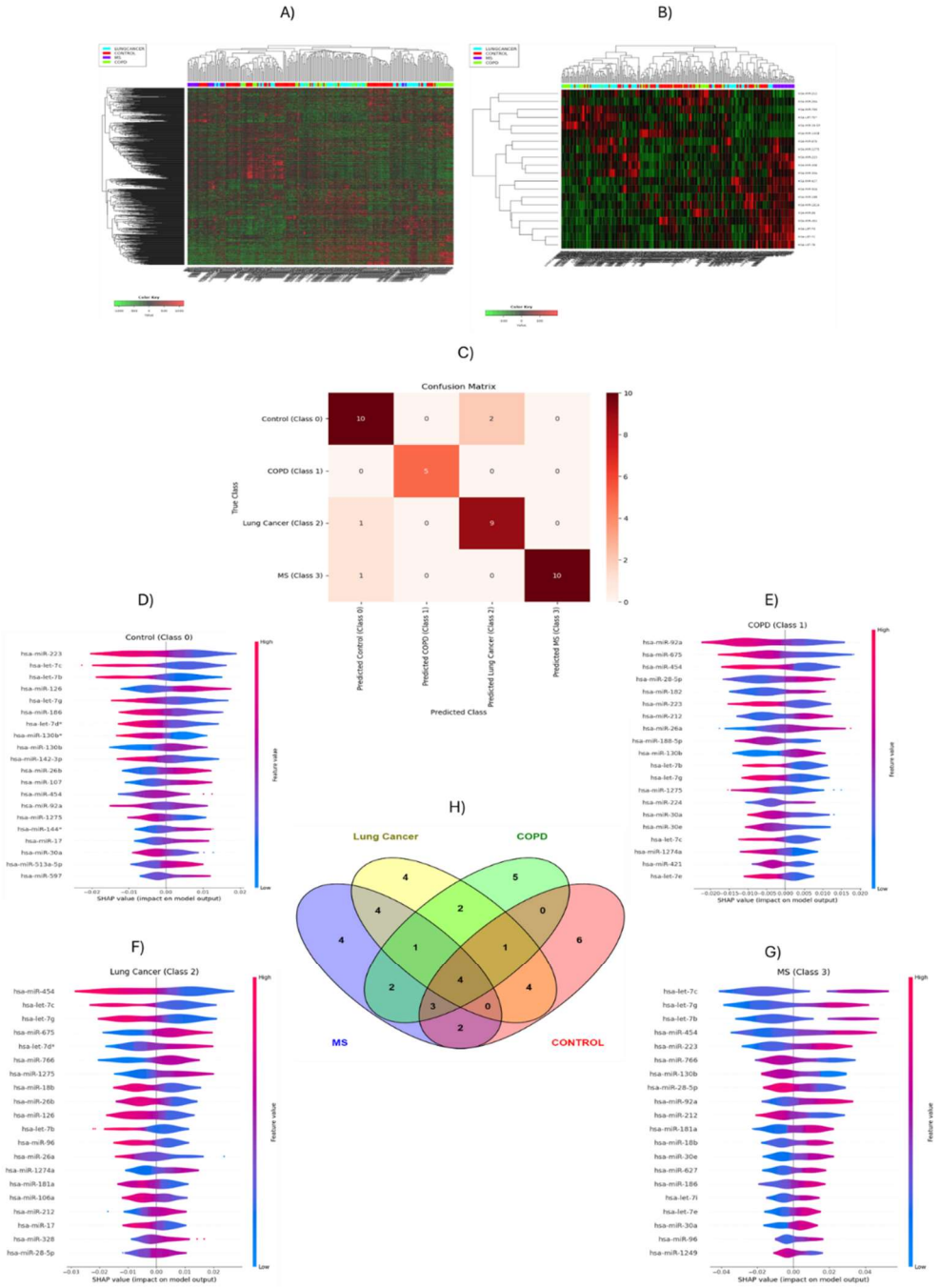


Figure 3.12(A)displays a heatmap of all 849 miRNAs across four disease groups, showing global expression trends. (B) focuses on the top 20 SHAP-identified miRNAs, emphasizing their ability to distinguish among the classes. (C) shows the confusion matrix for a refined model trained using only the top miRNAs, achieving an improved accuracy of 89.47%. (D), (E), (F), and (G) provide SHAP violin plots for each class—Control, COPD, Lung Cancer, and MS—highlighting the distribution and influence of each miRNA within specific predictions. (H) presents a Venn

diagram identifying four common miRNAs shared across all disease classes, reinforcing their biological and diagnostic relevance.

We have employed Mienturnet[83] to create a miRNA-mRNA interaction of four common miRNAs. The regulatory linkages between these miRNAs and their target mRNAs are shown by the resulting network, which is represented using a Sankey diagram (Fig3.13. (A)). In particular, hsa-let-7b-5p associated with 42 different mRNAs, making it the most active of the two, while hsa-let-7c-5p connected with 28 mRNAs. The interaction profiles of hsa-let-7g-3p and hsa-miR-454-3p were far more restricted, with just 2 mRNAs targeted by each (Fig3.13(B)).

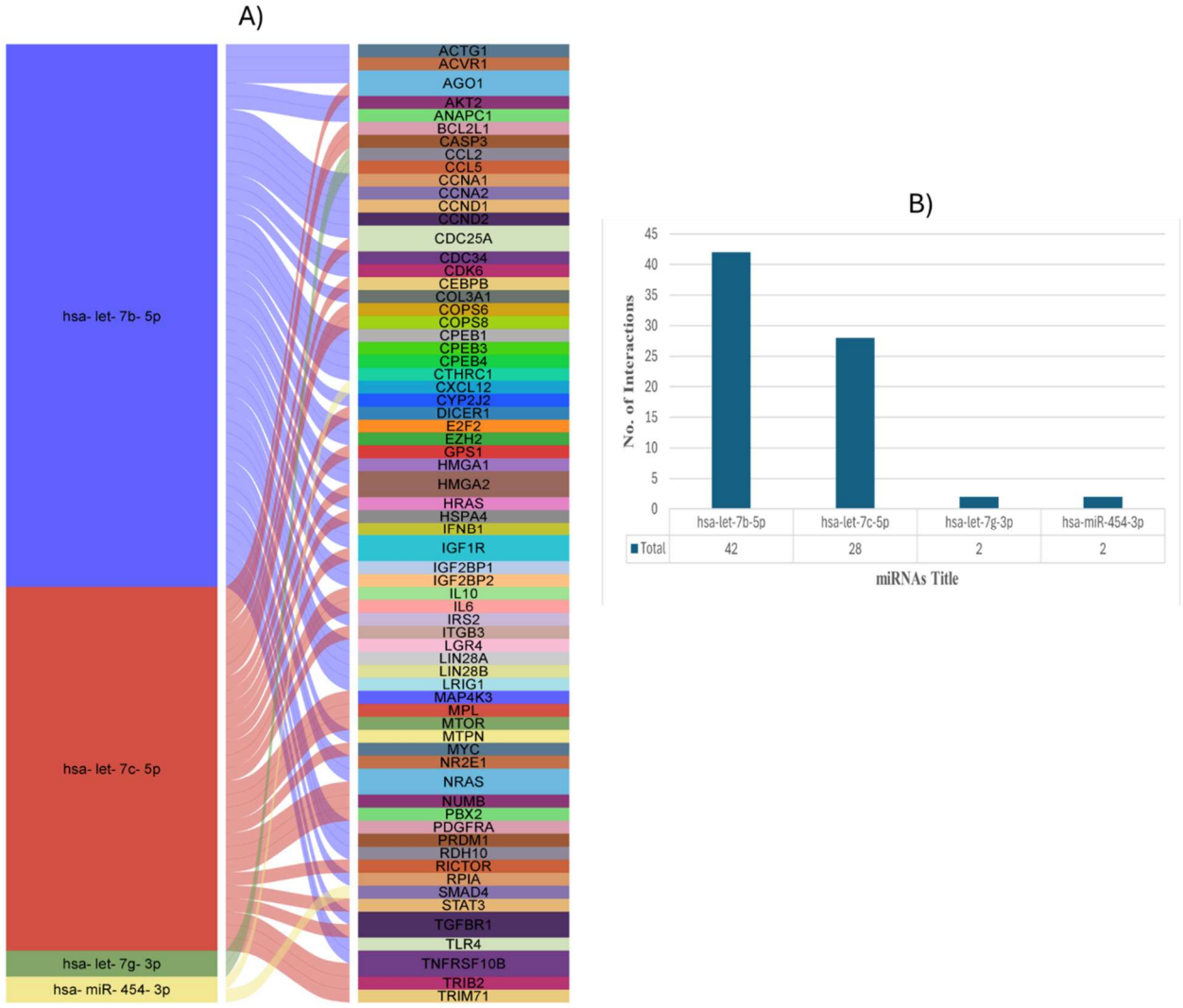


Figure 3.13(A) presents a Sankey diagram of miRNA–mRNA interactions for four common miRNAs (hsa-let-7c, hsa-let-7g, hsa-let-7b, and hsa-miR-454), visualizing the regulatory landscape. (B) quantifies the number of mRNA targets associated with each miRNA, with hsa-let-7b-5p showing the highest number of interactions.

13

The Molecular Signatures Database (MSigDB)[131], renowned for its extensive gene sets, features a particular "hallmark" collection aimed at reducing repetition and enhancing clarity regarding specific biological processes. These trademark sets offer more focused and straightforward enrichment analysis inputs, hence simplifying complex datasets.

Six genes—CCL2, CCL5, IL10, IL6, ITGB3, and MYC—were overlapping with the hallmark inflammatory gene (Supplementary Table.4) set which may considered as prominent within inflammatory pathways in case of COPD, MS and lung cancer. These genes are critical for the inflammatory response observed in conditions such as COPD, lung cancer, and MS. CCL2 and CCL5 play pivotal roles in lung cancer, COPD, and MS through their regulation of immune responses and inflammation. Elevated CCL2 levels in COPD are linked to enhanced macrophage recruitment and activation, with CCL2 deletion or CCR2 inhibition offering protection against alveolar injury and airway remodeling [132]. In lung cancer, CCL2 promotes tumor growth by recruiting immunosuppressive cells and enhancing cancer cell migration via interaction with CCR2 [133]. Similarly, CCL5 is implicated in COPD exacerbations, contributing to airway inflammation and mucus hypersecretion [134], while in lung cancer, it increases tumor invasiveness by recruiting macrophages [135]. In MS, elevated levels of CCL2 and CCL5 correlate with disease progression and may play a significant role in pathogenesis [136]. Similarly, IL-6 and IL-10 play critical but contrasting roles in MS, COPD, and lung cancer. In MS, IL-10-producing cells are protective, whereas IL-6 contributes to inflammation and disease progression [137]. In COPD, IL-10 levels are significantly elevated in patients with pulmonary infections, providing protective effects, while IL-6 is produced by both inflammatory and lung epithelial cells, contributing to disease exacerbations [138]. In lung cancer, reduced IL-6 activity results in decreased CD8 T cell activation and impaired tumor suppression, while IL-10 is important for activating CD8 T cells, aiding in tumor suppression through the induction of apoptosis in tumor cells [139].

ITGB3 plays a crucial role in lung cancer and MS by promoting tumor progression and facilitating key biological processes. In lung cancer, ITGB3 acts as a membrane receptor that promotes cancer through interactions with the tumor microenvironment and plays a pivotal role in tumor angiogenesis, which is essential for cancer growth and metastasis [140]. Additionally, ITGB3's activation of focal adhesion kinase (FAK) is critical for vesicle endocytosis, contributing to intracellular communication, which is vital for metastasis [141]. In MS, ITGB3 is associated with immune cell migration and activation, which can contribute to neuroinflammation and disease progression. While its specific role in COPD has not been well defined, ITGB3's involvement in cellular communication and inflammation suggests its potential participation in the underlying inflammatory mechanisms shared between COPD, lung cancer, and MS.

Further we also performed functional enrichment of these genes. Functional enrichment analysis is a popular tool for discovering commonalities across big biological datasets. In biomedicine, functional enrichment analysis of gene expression data is a standard method for understanding disease processes. Various approaches have been developed to categorize regulated gene expression profiles into functional groups. These functional categories are often selected based on literature sources and represent signaling or metabolic pathways[12].

A bar plot for Gene ontologies including biological process (BP), cellular component (CC) and molecular function (MF) fig.314(A) are being acquired. In this, the genes are enriched epithelial cell proliferation, positive regulation of epithelial cell in case of BP. In a study, In COPD and lung cancer, pulmonary epithelial cells regulate inflammation and repair. These diseases produce chronic inflammation and abnormal tissue remodeling by harming and killing epithelial cells,

which normally trigger immune responses and promote wound healing. DAMPs and epithelial signals exacerbate inflammation and fibrosis, progressing disease. Local immunological environment and tissue architecture dysregulation promotes lasting tissue damage, worsening COPD and lung cancer[142].In case of CC, the targeted genes were significant in plasma membrane signaling receptor complex and platelet alpha granule complex. In a study, Platelets are significant in MS beyond thrombosis. Prolonged MS activation enhances endothelial cell adhesion and neuroinflammation. Pro-inflammatory chemicals from platelet alpha granules cause CNS inflammation and damage, connecting neurons and astrocytes early. MS platelets begin neuroinflammatory and neurodegenerative processes before immune cell invasion. They are crucial disease progression players and therapeutic targets[143]. At last in case of MF, the genes were enriched in case of cytokine receptor binding and cytokine activity. Sankey Dot plot (fig.3.14(B)) for pathway analysis was done which illustrated the role of targeted genes are enriched, human cytomegalovirus infection and viral protein interaction with cytokine and cytokine receptor. The pathways emphasize their significance in COPD, lung cancer, and MS inflammation. In MS, these cytokines govern neuroinflammation; in lung cancer, they affect tumor growth and immune responses . Cytokines cause chronic airway inflammation and tissue damage in COPD[144]. This shared participation across all three diseases emphasizes their centrality in disease development and treatment potential

Subsequent to the analysis, we identified two miRNAs (hsa-let-7c-5p and hsa-let-7g-3p) within miRNA-mRNA network and made network using Cystoscope. These miRNAs are pivotal regulators of the six inflammatory genes: CCL2, CCL5, IL10, IL6, ITGB3, and MYC. These miRNAs were discovered to strongly interact with the identified genes involved in the inflammatory pathways associated with COPD, lung cancer, and MS. This concentrated investigation of miRNA-gene interactions elucidates the potential shared biological underpinnings of these disorders at the inflammatory level. By narrowing down these critical interactions, this analysis highlights potential points of interest for further therapeutic exploration, particularly in targeting shared pathways driving MS, COPD and Lung Cancer.

To further validate the findings, we utilized the Single Cell Portal[71] to generate heatmaps that provide single-cell level insights into the expression patterns of key inflammatory genes (CCL2, CCL5, IL10, IL6, ITGB3, and MYC) across various biological contexts. The first heatmap fig.3.14 (D) highlights their expression across distinct cell types within tissues, with a notable focus on inflammatory fibroblasts, which show elevated expression of genes like IL6 and CCL2. This underscores the role of fibroblasts not only in sustaining local inflammation but also in driving tissue remodeling and repair, linking their activity to both chronic respiratory and systemic inflammatory diseases. The second heatmap Fig.3.14(E) explores gene expression across multiple disease conditions, including idiopathic pulmonary fibrosis, interstitial lung disease, and Sjogren's syndrome. It highlights shared inflammatory pathways between respiratory and neurological diseases, such as elevated expression of CCL2 and CCL5, which are implicated in immune cell recruitment in both the central nervous system and lung tissue.

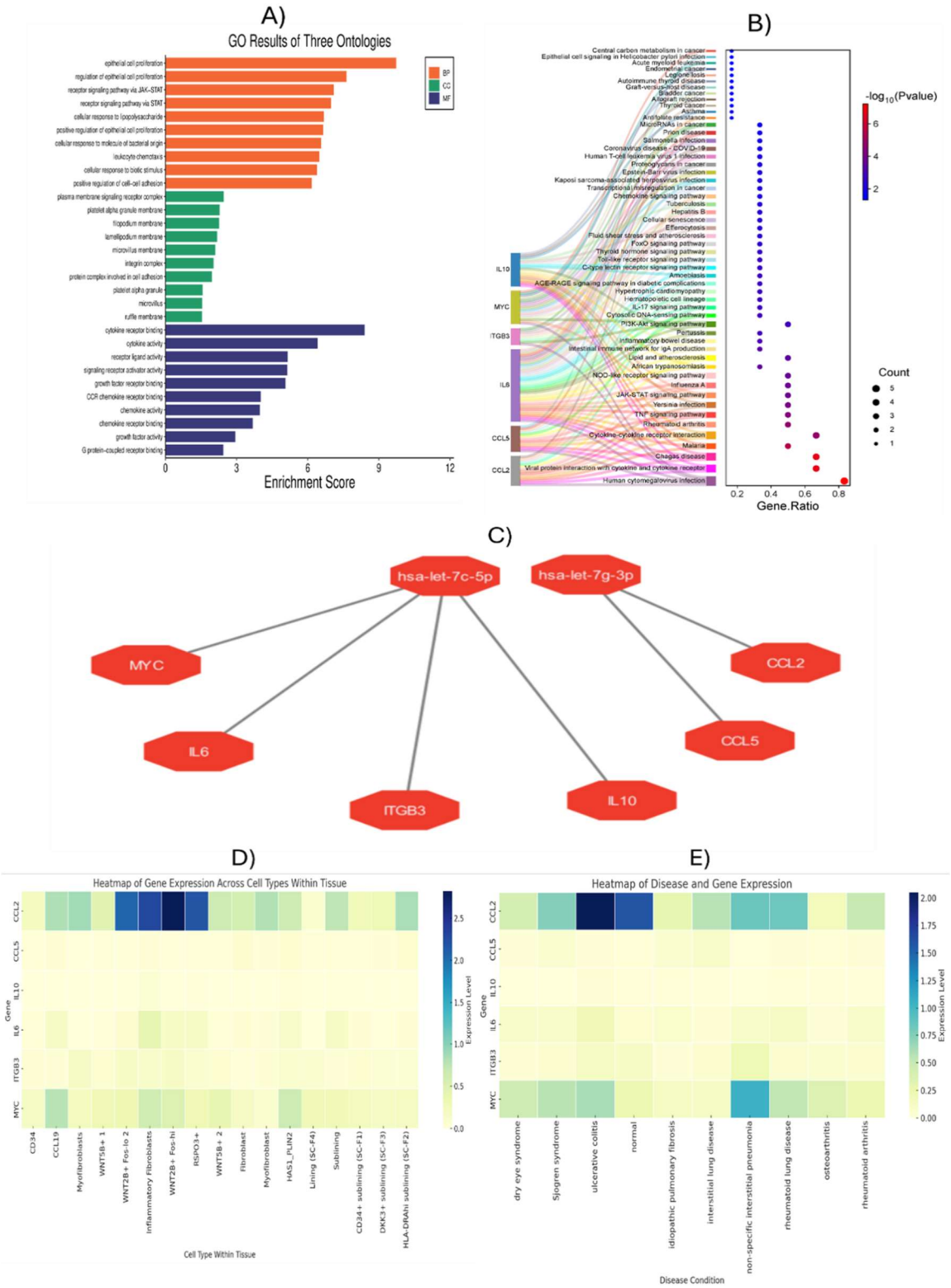


Figure 3.14(A)GO enrichment of six key inflammatory genes across biological processes, cellular components, and molecular functions. (B) Sankey dot plot showing enriched pathways, including cytokine signaling and viral interactions. (C) miRNA–mRNA network illustrating regulatory links between hsa-let-7c-5p, hsa-let-7g-3p, and target genes. (D) Single-cell heatmap

of gene expression across tissue-resident cell types, highlighting inflammatory fibroblasts. (E) Gene expression heatmap across disease conditions, showing shared inflammatory signatures.

To provide a consolidated view of our findings, Table.3.3 summarizes the key cytokines, miRNAs, and molecular signatures that are commonly dysregulated across COPD, MS, and lung cancer. In parallel, Supplementary Table 5 highlights experimentally validated roles of top SHAP-identified miRNAs—such as hsa-let-7c, miR-223, and miR-454—across cell and animal models, reinforcing the biological relevance of our computational discoveries.

Tabel.3.3Summary of shared cytokines, miRNAs, and molecular signatures across COPD, multiple sclerosis (MS)

Pathway/Marker	COPD	MS	Lung Cancer
IL6 (Interleukin-6)	↑IL6 linked to exacerbation and airway inflammation [138]	↑IL6 drives neuroinflammation and T-cell activation [137]	↑IL6 promotes tumor growth and immune suppression [139]
CCL2 (MCP-1)	Recruits macrophages, contributes to lung remodeling [132]	Elevated in CSF and lesions; facilitates immune cell trafficking [136]	Promotes tumor-associated macrophage infiltration [133]
hsa-let-7c	Regulates airway inflammation, linked to PI3K-Akt pathway [121]	Neuroprotective via inhibition of microglial activation *[117]*	Suppresses oncogenic targets (e.g., MYC); anti-tumor role *[123]*
hsa-miR-223	Controls neutrophil activation, NLRP3 inflammasome *[127]*	Regulates MDSCs and STAT3 in immune suppression *[126]*	Suppresses tumor progression via TGF-β and p53 pathways *[127,145]*
ITGB3	Involved in immune cell adhesion, potential ECM remodeling *(Fuentes et al., 2020)*	Facilitates immune cell migration into CNS [143]	Promotes angiogenesis and tumor metastasis *(C. Zhu et al., 2019; Fuentes et al., 2020)*

15

Shared regulatory networks involving miRNAs such as hsa-let-7c, hsa-miR-454, hsa-miR-92a, and hsa-miR-223 play key roles in modulating inflammatory mediators like CCL2, IL6, ITGB3, and MYC, which are central to cytokine signaling and immune regulation. These interactions converge on pivotal inflammatory pathways including PI3K/Akt, NF-κB, and cytokine–cytokine receptor interactions, as supported by our enrichment analyses. Recognizing the translational importance of these mechanisms, several therapeutic compounds—both natural agents and targeted inhibitors—have been identified that modulate these same pathways. For instance, fisetin, a dietary flavonoid, has demonstrated neuroprotective and anti-inflammatory properties via PI3K/Akt and NF-κB inhibition in preclinical models of inflammation and neurodegeneration [146]. Silymarin, validated in cigarette smoke-exposed models, attenuates airway inflammation and suppresses ERK/p38 MAPK signaling, suggesting potential utility in respiratory inflammatory diseases [147]. Additionally, repurposed kinase inhibitors such as ibrutinib and ponatinib have

demonstrated anti-inflammatory and anti-tumor activity in preclinical and clinical studies, with ibrutinib reducing NF-κB and inflammasome activation and ponatinib showing promise against aggressive, treatment-resistant cancers [148]. Finally, the PI3K inhibitor alpelisib, approved for PIK3CA-mutated cancers, targets the PI3K/AKT/mTOR pathway-a central axis in inflammation and cancer progression [149]. By synthesizing these molecular and pharmacological insights, the (Fig.3.15) underscores the translational potential of our findings and provides a visual roadmap for future therapeutic strategies targeting shared neuro-respiratory inflammatory mechanisms .

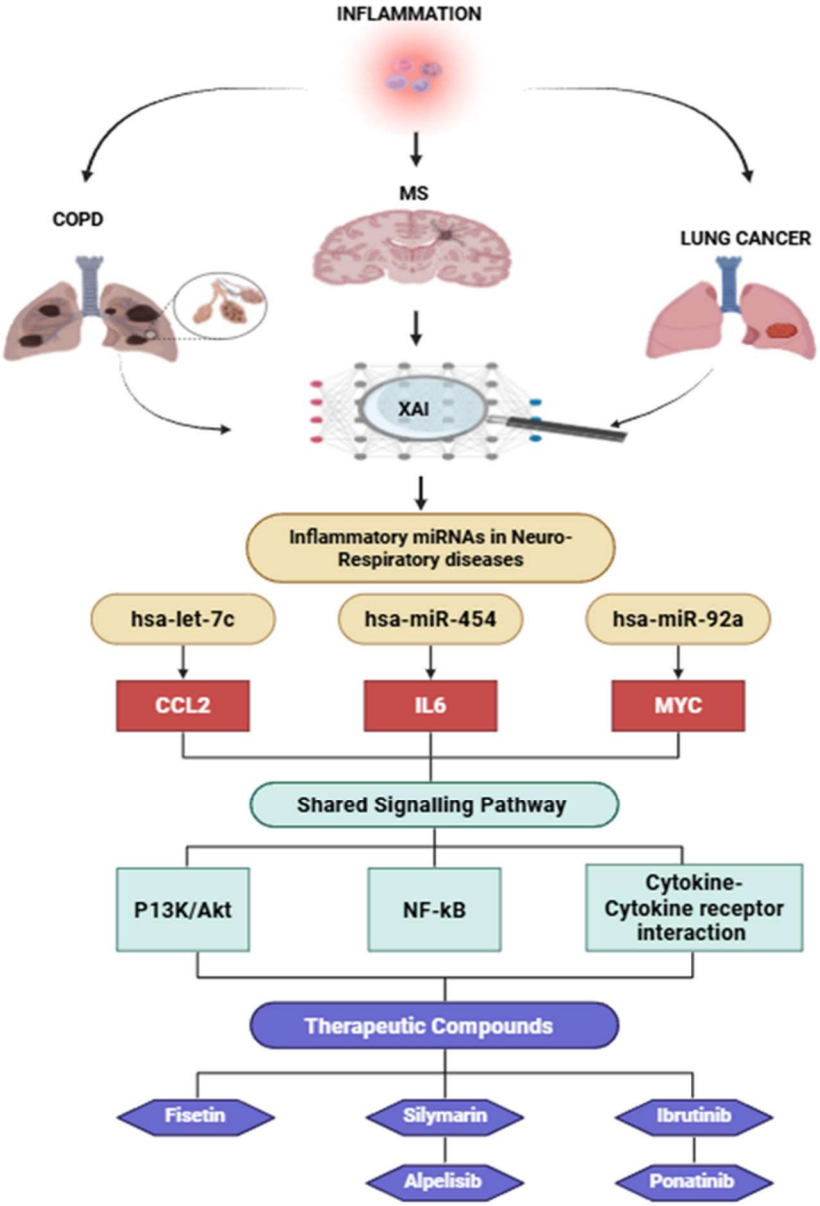


Figure 3.15Schematic overview linking shared inflammation in COPD, MS, and lung cancer with key miRNAs (hsa-let-7c, hsa-miR-454, hsa-miR-92a) and their targets (CCL2, IL6, MYC). These converge on common signaling pathways—PI3K/Akt, NF-kB, and cytokine–cytokine receptor interactions—highlighting potential therapeutic compounds such as fisetin, silymarin, alpelisib, ibrutinib, and ponatinib.

3.2.4 Key Findings

- Analysis of GSE61741 (237 samples: 94 controls, 47 COPD, 73 lung cancer, 23 MS) with Random Forest achieved 81.6% accuracy (AUC 0.97), validated on GSE31568 with 82.6% accuracy, confirming robust generalizability.

- SHAP explainability highlighted hsa-let-7c, hsa-let-7g, hsa-let-7b, hsa-miR-454, hsa-miR-92a, and hsa-miR-223 as central miRNAs across all classes. These miRNAs regulate key inflammatory mediators (CCL2, CCL5, IL6, IL10, ITGB3, MYC), converging on PI3K–Akt, NF-κB, cytokine–cytokine receptor signaling, and epithelial proliferation pathways.

- Reintroducing top miRNAs improved classification accuracy to 89.5%, validating their biomarker potential. Network and single-cell analyses further implicated fibroblasts and immune subsets as drivers of shared inflammation.

- Importantly, enrichment analyses identified translational opportunities, with compounds such as fisetin (PI3K–Akt/NF-κB inhibitor, neuroprotective), silymarin (anti-inflammatory, MAPK suppression), alpelisib (PI3K inhibitor), ibrutinib (NF-κB and inflammasome suppression), and ponatinib (anti-tumor, kinase inhibitor) showing potential for repurposing across COPD, MS, and lung cancer.

- Collectively, these findings confirm that neuro-respiratory inflammatory crosstalk is orchestrated by shared miRNA–mRNA networks, with both biomarker and therapeutic implications for integrated disease management.

3.3 Results for Objective 3 : To develop a SHAP-LLM powered chatbot for lung disease indication, which leverages explainable machine learning models and large language models.

3.3.1 Performance on the Test and Validation Sets

To address class imbalance and reduce model bias toward the majority class (i.e., Normal), SMOTE was applied during training. The impact of SMOTE on model performance is demonstrated through comparative analyses. Confusion matrices and classification reports before and after applying SMOTE are summarized in Figure.3.16. The predictive performance of the model was first evaluated on the held-out test set, comprising 20% of the total dataset. The confusion matrix (fig.3.16(A)) for the test set showed that the model correctly classified 509 normal cases and 352 lung cancer cases. Misclassifications included 84 false positives (normal cases predicted as lung cancer) and 55 false negatives (lung cancer cases misclassified as normal). The corresponding classification report (fig.3.16(B)) highlighted the model's strong generalization capability, with the lung cancer class achieving a precision of 0.81, recall of 0.86, and F1-score of 0.84. For the normal class, the model recorded a precision of 0.90, recall of 0.86, and F1-score of 0.88. These results reflect the model's effectiveness in distinguishing between diseased and healthy individuals, with particular emphasis on high recall for the lung cancer class, which is critical for minimizing missed diagnoses. The overall test accuracy was 0.86, indicating a balanced and clinically acceptable performance on previously unseen data.

Evaluation on the validation set, also comprising 20% of the data and used for internal tuning, showed similarly consistent results. The confusion matrix (Fig.3.16(C)) indicated 527 true negatives and 343 true positives, alongside 66 false positives and 64 false negatives. According to the classification report(fig.3.16(D)), the model achieved a precision, recall, and F1-score of 0.89 for the normal class. For the lung cancer class, all three metrics were 0.84, demonstrating that the model effectively captured disease-positive cases without significantly compromising precision. The close alignment between test and validation results reinforces the model's reliability, suggesting that it performs consistently across both internal validation and completely unseen samples (Table. 3.4-3.5).

Tabel.3.4Confusion Matrices with Accuracy

Set	True Positives (TP)	True Negatives (TN)	False Positives (FP)	False Negatives (FN)	Accuracy
Test Set	352	509	84	55	0.86
Validation Set	343	527	66	64	0.87

Tabel3.5Classification Report

Set	Class	Precision	Recall	F1-Score
Test Set	Normal	0.90	0.86	0.88
	Lung Cancer	0.81	0.86	0.84
Validation Set	Normal	0.89	0.89	0.89
	Lung Cancer	0.84	0.84	0.84

3.3.2 Cross-Validation Stability

To ensure the robustness and reliability of the model, stratified 5-fold cross-validation was performed on the training dataset using a pipeline that incorporated SMOTE for class balancing and XGBoost for classification. This approach allowed for an unbiased estimation of the model's stability across different training subsets. The cross-validation results, illustrated in the bar plot(fig.3.16(E)), indicate a mean classification accuracy of 0.880 with a low standard deviation of 0.004. Each of the five folds achieved accuracy in the range of 0.874 to 0.885, with no significant fluctuation observed between folds. The narrow confidence interval around the mean accuracy reinforces the model's reliability and its ability to perform consistently, irrespective of the specific data partition. These results confirm that the integration of SMOTE and monotonic constraints within the XGBoost framework leads to a model that is not only accurate but also generalizes well across varying subsets of data (Table.3.6).

Tabel.3.6 5-Fold Cross-Validation Results

Fold	**Accuracy**
Fold 1	0.874
Fold 2	0.885
Fold 3	0.881
Fold 4	0.879
Fold 5	0.881
Mean	**0.880**
Std Dev	**±0.004**

To evaluate further how well the model discriminates, a class-wise ROC analysis was performed using the test set predicted probabilities. Two-class ROC curves were created for Normal and Lung Cancer respectively, utilizing one-vs all scoring approaches. The model, as illustrated in Fig.3.16(F), both classed had AUCs of 0.90, which means that the classification performance for both Lung Cancer and Normal classes was good and showed balanced discrimination ability. There is a sharp inclination in both curves which prove that at low sensitivity settings the model achieves a high level of sensitivity and specificity, at a variety of decision thresholds. The equal AUC values in both classes suggest that the model is not biased towards one of the two classes, but rather performs consistently well in identifying true and false positives. Also, lack of intersection with the 50% chance line confirms that the model is not overfitting and, in contrast, generalizes well to unseen data.

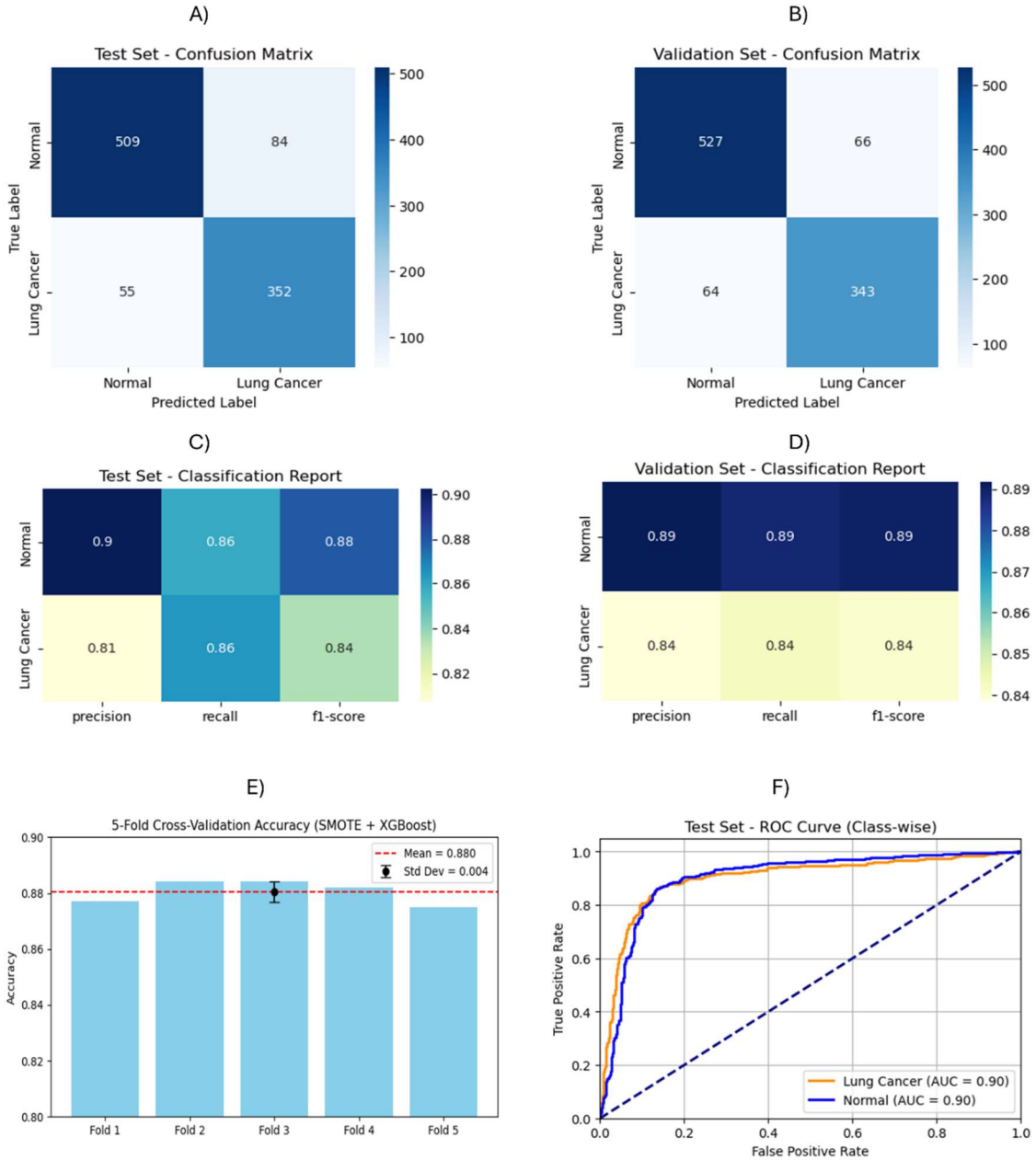
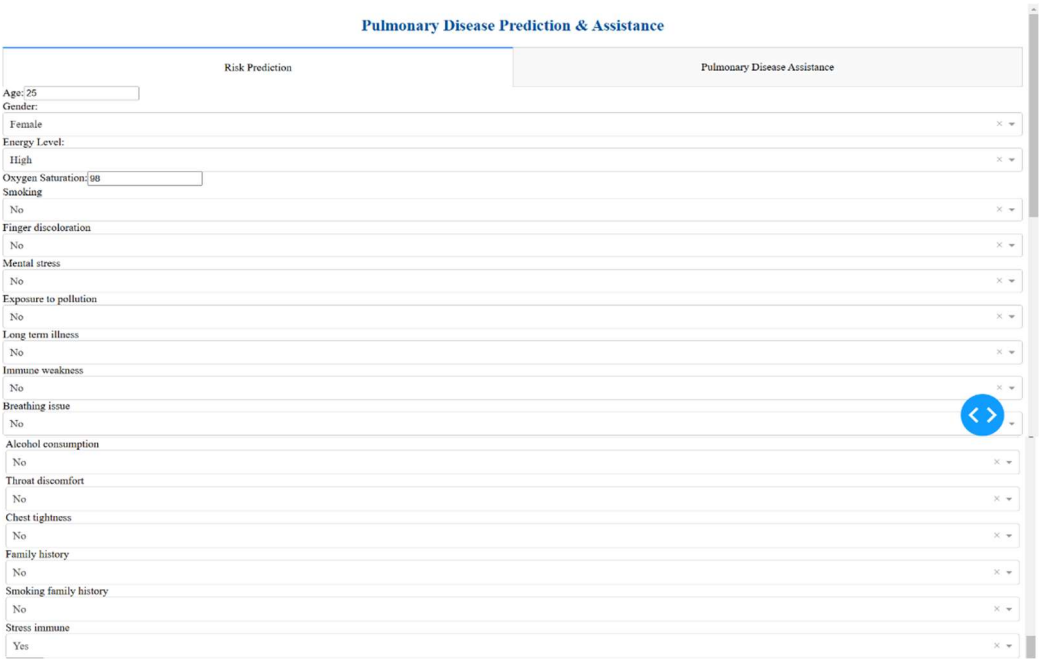


Figure 3.16This figure presents a comprehensive evaluation of the XGBoost model for pulmonary disease classification across test and validation sets. The test set confusion matrix (Fig. 3.16A) confirms the model's generalization capacity on unseen data, maintaining strong class-wise accuracy. The confusion matrix for the validation set (Fig. 3.16B) demonstrates the model's ability to correctly classify both normal and lung cancer cases, with minimal false positives and false negatives. Corresponding classification reports for the test sets (Fig. 3.16C) and validation (Fig. 3.16D) reveal balanced performance, with F1-scores exceeding 0.84 for both classes and consistently high recall values for the lung cancer class, which is critical in minimizing missed diagnoses. The model's robustness is further supported by the 5-fold cross-validation results (Fig. 3.16E), where a mean accuracy of 0.880 and a low standard deviation of 0.004 indicate stable performance across different data splits. ROC for both classes (Fig.3.16(F))

3.3.3 Chatbot Predictions, SHAP Interpretability, and Automated Assistance

To translate model predictions into a clinically interpretable and interactive format, a web-based chatbot interface was developed using Python Dash. This tool allows users to input individual-level health and lifestyle variables relevant to pulmonary disease risk and receive real-time predictions, feature attribution scores using SHAP, and automated guidance in plain language using LLM. The chatbot was evaluated using two representative cases—one with low predicted risk and another with high predicted risk.

For instance, in the low-risk case (Figure.3.17), a 25-year-old non-smoking female with no respiratory symptoms, no family history of lung disease, and normal physiological markers (e.g., oxygen saturation at 98%, high energy level) was assessed. The model predicted a lung disease probability of only 2% (confidence score: 0.02). The SHAP plot (Fig. 3.19(A)) revealed that the most influential factors reducing predicted risk were the absence of smoking, throat discomfort, and breathing issues, as well as younger age and normal oxygen saturation. Each of these features had strongly negative SHAP values, contributing to the overall low risk estimate. The automated explanation further translated these feature effects into clinically meaningful language, highlighting the protective role of non-smoking behavior, symptom absence, and overall physiological stability. This explanation was generated using an integrated LLM pipeline powered by Mistral via the LangChain framework, which was prompted with SHAP-based inputs and contextual feature values to deliver human-readable summaries tailored to each patient profile. This LLM is not merely a lifestyle advisor; it serves as a dynamic clinical companion capable of parsing SHAP-based feature importances, contextualizing individual predictions, explaining disease mechanisms, interpreting physiological variables, and guiding users through potential diagnostic and therapeutic pathways. Unlike static rule-based systems, this architecture allows the chatbot to adaptively respond to diverse and nuanced natural language inputs—including questions about symptoms, medications, or prognosis—while grounding its responses in the underlying model logic and medical best practices. This transforms the tool from a mere prediction interface into a continuously responsive, XAI assistant with genuine translational value in clinical and patient-facing environments.



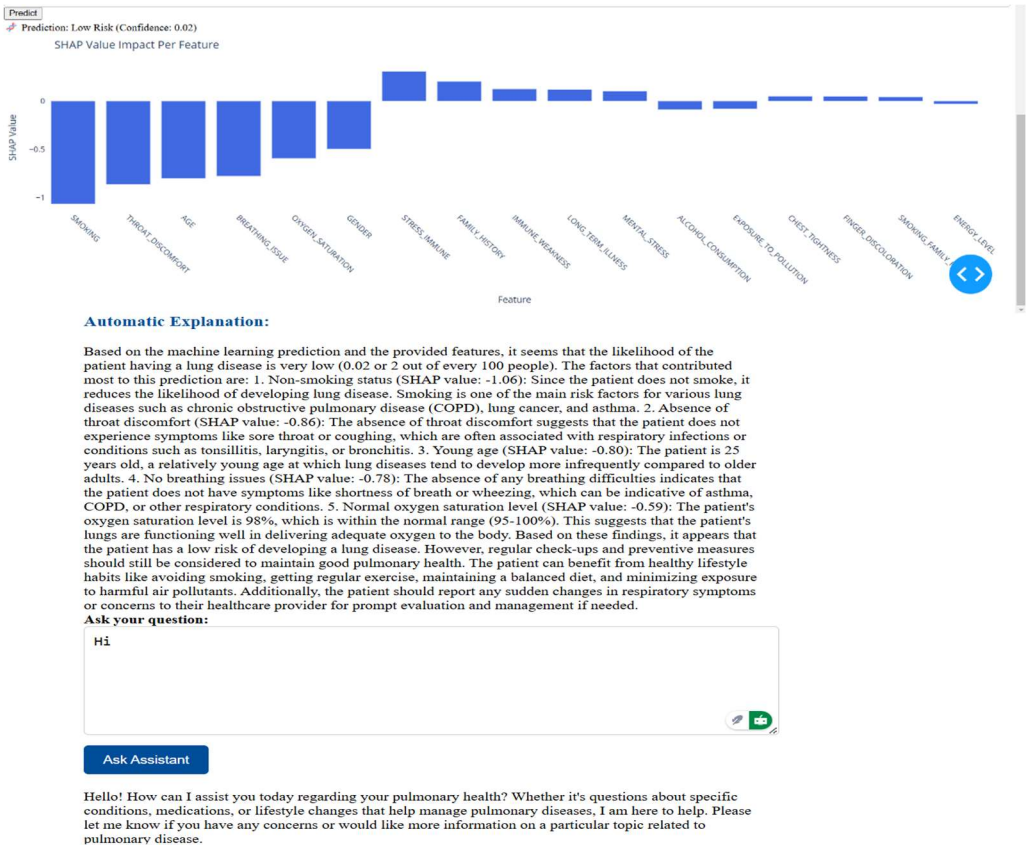


Figure 3.17 Low Risk Case

In contrast, the high-risk case involved a 40-year-old male smoker with moderate energy level, suboptimal oxygen saturation (94%), and reported symptoms such as throat discomfort, finger discoloration, and chest tightness. The model assigned a high lung disease probability of 98% (confidence score: 0.98), as shown in (Supplementary Figure.3.18). The SHAP explanation (Fig. 3.19(B)) identified smoking status, throat discomfort, family history, and the presence of breathing issues as the dominant contributors to the elevated risk. These features had strongly positive SHAP values, each pushing the prediction closer to the disease class. The generated explanation contextualized this output by suggesting possible diagnoses (e.g., COPD, asthma and lung cancer) and emphasizing the need for clinical follow-up, diagnostic imaging, and lifestyle interventions. The chatbot also successfully responded to queries regarding lifestyle changes and symptom interpretation, reinforcing its role as a supportive, patient-centric tool.


Pulmonary Disease Prediction & Assistance
Risk Prediction
Pulmonary Disease Assistance
Age: 40
Gender:
Male
Energy Level:
Moderate
Oxygen Saturation: 94
Smoking
Yes
Finger discoloration
Yes
Mental stress
Yes
Exposure to pollution
Yes
Long term illness
No
Immune weakness
No
Breathing issue
Yes
Alcohol consumption
Yes
Throat discomfort
Yes
Chest tightness
Yes
Family history
Yes
Smoking family history
Yes
Stress immune
No
Predict
Prediction: High Risk (Confidence: 0.98)
SHAP Value Impact Per Feature
SHAP Value
1.5
1
0.5
0
SMOKING
THROAT_DISCOMFORT
SMOKING_FAMILY_HISTORY
FAMILY_HISTORY
BREATHING_ISSUE
EXPOSURE_TO_POLLUTION
AGE
OXYGEN_SATURATION
GENDER
STRESS_IMMUNE
IMMUNE_WEAKNESS
CHEST_TIGHTNESS
LONG_TERM_ILLNESS
MENTAL_STRESS
FINGER_DISCOLORATION
ALCOHOL_CONSUMPTION
ENERGY_LEVEL
Feature

Automatic Explanation:

Based on the machine learning prediction, there is a high likelihood that the patient may have a lung disease. The factors that contributed most to this prediction are smoking, discomfort in the throat, a family history of smoking, a family history of any disease, and existing breathing issues. Smoking significantly increases the risk of developing various lung diseases such as chronic obstructive pulmonary disease (COPD), asthma, and lung cancer. Throat discomfort could be a symptom of conditions like gastroesophageal reflux disease (GERD) or postnasal drip, but it can also indicate inflammation or infection in the lungs. A positive family history increases the likelihood of genetic predispositions to lung diseases, and breathing issues such as shortness of breath, wheezing, or cough may be indicative of conditions like asthma or emphysema. Given these findings, it is strongly recommended that the patient undergo further diagnostic tests, such as chest X-rays, CT scans, spirometry, and possibly a bronchoscopy, to confirm or rule out the presence of a lung disease. The patient should be advised to quit smoking if they are currently smoking, and any other relevant lifestyle modifications should be suggested based on the specific diagnosis. Additionally, it would be prudent for the healthcare provider to discuss treatment options, management strategies, and prognosis with the patient once a definitive diagnosis has been made. Regular follow-ups and monitoring may also be necessary depending on the severity of the condition.

Ask your question:

What lifestyle changes you would suggest ?

Ask Assistant

1. Smoking Cessation: If the patient is a smoker, quitting smoking is crucial for managing pulmonary diseases and improving overall health. Various methods can be employed to help with this, such as nicotine replacement therapy, counseling, or medication. 2. Healthy Diet: A balanced diet rich in fruits, vegetables, lean proteins, and whole grains can help improve lung function and reduce inflammation. Foods high in antioxidants, vitamins C and E, beta-carotene, and omega-3 fatty acids are particularly beneficial for pulmonary health. 3. Regular Exercise: Light to moderate physical activity can help strengthen the lungs and improve breathing ability. Walking, swimming, and cycling are some examples of low-impact exercises that may be suitable for individuals with pulmonary conditions. 4. Maintain a Healthy Weight: Obesity can worsen symptoms of pulmonary diseases, so it's important to maintain a healthy weight through balanced nutrition and regular exercise. 5. Pulmonary Rehabilitation Programs: These programs combine education, exercise training, nutritional advice, and counseling to help patients improve their physical capacity, manage symptoms, and enhance their overall quality of life. 6. Avoid Respiratory Irritants: Exposure to air pollutants, dust, chemicals, or cigarette smoke should be minimized as much as possible to prevent further lung damage and aggravation of symptoms. 7. Stay Hydrated: Drinking plenty of fluids can help keep the airways moist and reduce the risk of developing respiratory infections. 8. Regular Medical Check-ups: Patients with pulmonary diseases should schedule regular check-ups with their doctors to monitor their condition, adjust treatment plans if necessary, and address any concerns that arise. 9. Flu and Pneumonia Vaccinations: These vaccines can help prevent respiratory infections, which are particularly dangerous for individuals with pre-existing lung conditions. 10. Good Air Quality at Home: Using air purifiers, maintaining cleanliness, and ventilating the home regularly can help improve indoor air quality and reduce exposure to allergens or irritants.

Figure 3.18 High Risk Case

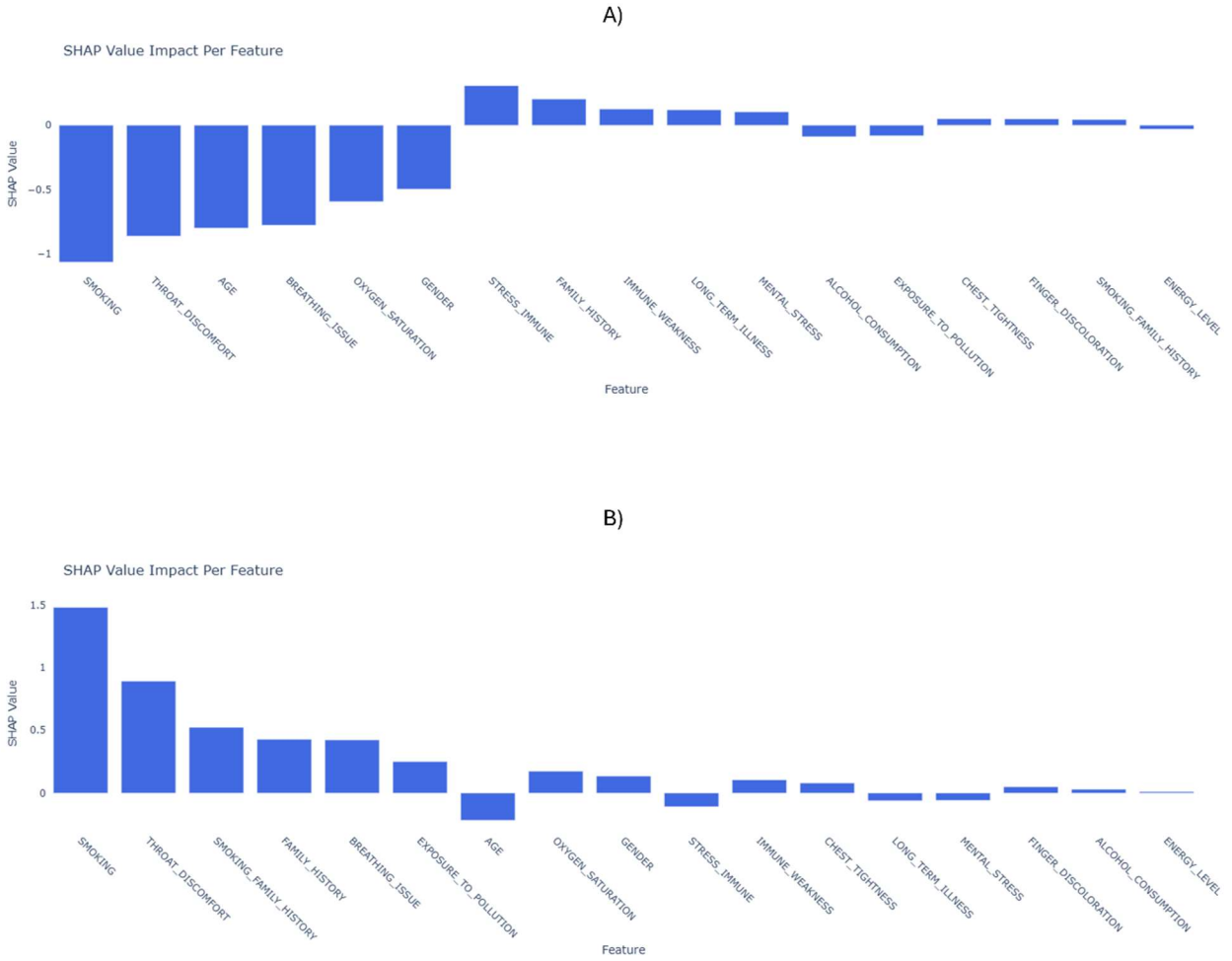


Figure 3.19 SHAP bar plot for low(A) and high risk(B)

Together, these results show that the chatbot can deliver accurate predictions, clear explanations, and useful support for decision-making. Each prediction is backed by SHAP values, which help explain why the model made that decision, making the system transparent and trustworthy. With the assistance of an LLM, users can pose follow-up questions to their concerns regarding risks, symptoms and next steps, and get answers in plain language for non-expert users. Most importantly, the system complies with clinical reasoning standards: high-risk cases are met with assertive, referral-driven support, while low-risk profiles are addressed with preventive health measures, devoid of false reassurance or oversimplification. This adaptable and nuanced approach demonstrates the chatbot's capability as a dependable, patient-facing assistant for the preliminary assessment of pulmonary disease risk.

3.3.4 Key Findings

- The XGBoost classifier, trained with SMOTE and optimized through stratified 5-fold cross-validation, achieved 86% accuracy on the test set and 87% on the validation set, with balanced performance across normal and lung cancer classes (precision/recall ~0.84–0.90). Independent ROC analysis confirmed robust discrimination with AUC = 0.90 for both classes, while cross-validation further supported model stability (mean accuracy = 0.880, ±0.004).

- Explainability via SHAP revealed class-specific feature contributions: non-smoking status, absence of respiratory symptoms, and high oxygen saturation strongly reduced predicted risk, whereas smoking, chest tightness, and family history markedly increased it. These feature attributions not only validated biological plausibility but also provided clinician-friendly insights into risk drivers.

- To enable real-world translation, predictions were embedded in a Python Dash–based LLM-integrated chatbot. The chatbot successfully contextualized SHAP outputs into natural language, explaining low-risk cases in terms of protective behaviors and physiological stability, and high-risk cases in terms of smoking, symptomatology, and hereditary burden. Importantly, it adapted explanations to user queries, suggesting potential diagnoses, lifestyle modifications, and referral needs, all while maintaining transparency and alignment with clinical reasoning.

- Collectively, this framework demonstrates that SHAP-guided, LLM-enabled chatbots can bridge AI predictions with human-interpretable insights, transforming black-box models into dynamic, patient-facing neuro-respiratory assistants that support preliminary risk stratification and clinical decision-making.

3.4 Discussion & Implementation

This study was designed with three interlinked objectives, each addressing a distinct but complementary aspect of chronic inflammatory diseases, and each progressively building toward translational implementation. Taken together, they illustrate how explainable artificial intelligence (XAI) can move from biomarker discovery to systemic integration and finally to patient-centric deployment.

The first objective focused on deciphering the molecular landscape of COPD and ILD, two respiratory diseases that frequently overlap to form the clinical entity Combined Pulmonary Fibrosis and Emphysema (CPFE). Our machine learning model, refined through SMOTE to counter dataset imbalance, reached an accuracy of 88.1%. More importantly, SHAP analysis revealed critical genes including ADRB2, CDH3, IRS2, MATN3, VEGFC, CD38, and PDIA4, which regulate pathways such as airway smooth muscle tone, angiogenesis, extracellular matrix remodeling, and ER stress. These findings are not only consistent with prior reports of ADRB2 involvement in airway reactivity and VEGFC in fibrotic remodeling, but also establish a mechanistic basis for why COPD and ILD converge clinically as CPFE. The discussion here marks the foundation of this thesis: gene-level biomarkers interpreted transparently through XAI.

Building on this respiratory foundation, the second objective expanded the scope to systemic crosstalk between respiratory and neurological disorders. Multiple sclerosis (MS), though classically neuroinflammatory, shares common immune dysregulation with COPD and lung cancer. By analyzing peripheral blood miRNA datasets, our model identified a panel of cross-disease regulators—hsa-let-7c, hsa-miR-454, hsa-miR-92a, and hsa-miR-223 along with key cytokines (IL6, IL10, CCL2, CCL5, and ITGB3). SHAP interpretability revealed that these miRNAs differentially influence disease classification across pulmonary and neurological contexts. Literature validation confirmed their roles: for instance, let-7c suppresses IL-6 signaling in COPD, preserves neuronal function in MS, and inhibits oncogenic ITGB3 in lung cancer, highlighting its position as a unifying molecular thread. This objective therefore elevated the thesis from organ-specific biomarker discovery to a systemic understanding of inflammation where neuro and respiratory pathways intersect at the level of immune signaling and miRNA regulation.

The third objective represented the translational leap embedding these insights into an interactive SHAP–LLM powered chatbot. Whereas Objectives 1 and 2 generated interpretable biomarkers and systemic networks, this final step demonstrated how such insights could be implemented in real-world decision support. The chatbot integrated SHAP explanations of model predictions with natural-language outputs generated by a large language model. Tested on low-risk and high-risk cases, it not only produced accurate predictions (86–89% accuracy across datasets) but also explained *why* those predictions were made, in clinically understandable terms. For example, in a low-risk individual, SHAP highlighted protective features (non-smoking, normal oxygen saturation), while the LLM translated these into plain-language reassurance and preventive advice. In a high-risk smoker with abnormal oxygen levels, SHAP emphasized pro-disease features, and the chatbot guided toward diagnostic follow-up. This combination of interpretability and communicability marks a crucial step beyond computational analysis, embedding XAI within patient- and clinician-facing workflows. Comparative analysis of proposed SHAP–LLM chatbot with other studies summarized in Table.3.7.

Tabel.3.7 Comparative analysis of proposed SHAP–LLM chatbot

Category	Proposed Work: LLM-Augmented XAI Chatbot	XGBoost (Pulmonary Inflammation Mortality Risk)	CNN (Chest X-ray Diagnosis)	CNN-XGBoost (Pneumonia Classification)	SVM, RF, KNN, DT (Lung Cancer Risk)	Deep Learning (InceptionV3, VGG16, ResNet-50)	ML Model + XAI (Multiple Disorders)
Data & Features	5,000 instances, 17 features (Kaggle)	2790 ICU patients	Imaging (Chest X-rays)	Chest X-ray images	Tabular (Kaggle and Data World datasets)	Chest X-ray images	PFT indices, multiple disorders
Methods /	XGBoost +	XGBoost	CNN, VGG16	CNN +	SVM, RF,	InceptionV3,	ML +

Algorithms	SHAP + LLM			XGBoost	KNN, DT	VGG16, etc.	Shapley values
Accuracy / F1 / AUC	Accuracy: 0.86 (test), F1: 0.84–0.88, AUC: 0.90	Accuracy: 0.889, F1: 0.891, AUC: 0.956	Accuracy: 95%, AUC: 0.92–0.97	Accuracy: 87%, Specificity: 89%, Sensitivity: 85%	Accuracy: 93–99% (varies)	AUC: 0.90–0.99 (class-dependent)	Accuracy: 74–82% (disease-dependent)
Explainability	SHAP + LLM-based case explanations	SHAP-based explanations	Not detailed	Not detailed	LIME, decision trees, boundary plots	Not detailed	Shapley values
Usability / Interface	Web-based chatbot, interactive	Interactive web page	Not specified	Not specified	Not specified	Not specified	Not specified
Key Advantages / Limitations	Combines high accuracy, XAI, clinical constraints, conversational LLM, real-time interaction	High accuracy, interpretable, but mortality-focused, lacks LLM	High imaging accuracy, low explainability	Pneumonia-specific, limited generalization	High accuracy, lacks patient-facing interactivity	High accuracy, poor transparency	Multi-disorder, but lower accuracy, no interactivity
References		[150]	[151]	[152]	[153]	[154]	[155]

As shown in Table.3.7, while deep learning approaches such as CNNs and transfer learning architectures (InceptionV3, VGG16, ResNet-50) achieve high accuracy, they are often opaque and lack clinician- or patient-facing interpretability. Classical ML models (SVM, RF, KNN, DT) provide transparency but remain limited in scope and generalizability. By contrast, our proposed SHAP–LLM chatbot strikes a balance between accuracy, interpretability, and usability, embedding AI insights into a conversational interface suitable for real-world deployment.

Taken together, these three objectives form a coherent hierarchical narrative that progressively advances from molecular discovery to clinical translation. Objective 1, focused on CPFE biomarkers, provided granular, gene-level evidence for overlapping respiratory disease mechanisms, demonstrating how explainable AI can illuminate the molecular underpinnings of COPD–ILD overlap. Building on this foundation, Objective 2 extended the scope to systemic neuro-respiratory crosstalk, uncovering shared inflammatory regulators and highlighting how common pathways connect distinct organ systems such as the lung and the central nervous system. Finally, Objective 3 transformed these discoveries into practice by developing a SHAP–LLM chatbot, an interactive tool capable of translating complex multi-omics and AI-driven insights into accessible explanations for risk assessment and early clinical guidance. Together, these objectives illustrate a continuum of discovery, integration, and implementation, unified under the framework of explainable AI.

This hierarchical progression from discovery to integration to implementation demonstrates how explainable AI can act as a bridge between data-driven research and real-world clinical needs. At every stage, the incorporation of XAI ensured that findings were not only statistically robust but also biologically interpretable and clinically meaningful, directly addressing the long-standing limitations of black-box AI in healthcare. This cumulative narrative naturally transitions into the implementation phase, where the validated biomarkers, systemic inflammatory signatures, and interpretable AI models are operationalized into interactive, patient- and clinician-facing tools, laying the foundation for precision diagnostics and translational impact.

Chapter 4

4.1 Conclusion, Future Scope and Social Impact

This thesis has traversed the continuum of modern translational research—from molecular discovery, through systemic integration, to digital implementation—anchored in the framework of explainable artificial intelligence. Beginning with Objective 1, we demonstrated how gene-level biomarkers illuminate the overlapping molecular signatures of COPD and ILD, especially within the CPFE phenotype. Objective 2 expanded this lens, bridging respiratory and neurological disorders to reveal shared inflammatory and immunoregulatory signatures across COPD, lung cancer, and MS. Objective 3 carried these findings forward into practice by developing a SHAP–LLM chatbot, translating complex multi-omics insights into accessible, real-time explanations for clinicians, patients, and researchers alike. Collectively, these contributions illustrate not just the value of integrative multi-omics and XAI, but also the possibility of reimagining respiratory medicine as a field where discovery and clinical utility coexist seamlessly.

The conclusion that emerges is twofold. First, respiratory diseases do not exist in isolation; they are connected to systemic and even neurological disorders via convergent inflammatory and regulatory pathways. Second, artificial intelligence—when explainable, interpretable, and integrated with domain knowledge—can act as a true bridge between molecular complexity and clinical decision-making. This thesis demonstrates that discovery (biomarkers), integration (neuro-respiratory signatures), and implementation (chatbot) are not sequentially isolated steps but part of a unifying continuum where each level strengthens the translational impact of the other.

Looking toward the future, several avenues unfold. The biomarker signatures identified here can be validated in prospective, multi-cohort studies, with an emphasis on stratifying patients not only by disease but by overlap syndromes such as CPFE or systemic crosstalk with neuroinflammatory disorders. Integration with digital twin technology could allow the construction of patient-specific models that simulate disease progression and therapeutic responses, driven by the same explainable AI frameworks validated here. The SHAP–LLM chatbot prototype can be expanded into a multi-modal platform, incorporating imaging, wearable biosensors, and EHR data, ultimately evolving into a clinician-augmented AI companion capable of supporting differential diagnosis, therapeutic planning, and patient education. Moreover, integration with federated learning could enable this system to learn from global datasets while preserving patient privacy, advancing both scalability and ethical responsibility.

The social impact of this work lies in its potential to democratize precision healthcare. Respiratory and neuro-respiratory diseases disproportionately affect populations in low-resource settings where access to early diagnostics, trained specialists, and advanced therapeutics is limited. By embedding interpretability at every stage, the XAI-driven framework proposed here can generate trust not only among clinicians but also among patients, empowering them with transparent explanations of their disease risk and therapeutic options. A chatbot that explains why a patient might be at higher risk, how biomarkers contribute to disease progression, and what preventive steps are actionable, transforms complex computational predictions into meaningful guidance at the community level. This democratization of knowledge where molecular discoveries no longer remain confined to specialized laboratories but reach the bedside and even the household represents the most profound societal contribution of this thesis.

In sum, this thesis advances a new paradigm where respiratory and systemic diseases are studied not in isolation but as interwoven processes; where artificial intelligence is not a black box but a transparent collaborator; and where the end point of research is not merely publication but implementation in tools that serve clinicians and patients alike. The integration of XAI and LLM-driven interfaces provides a blueprint for the future of precision medicine—one that is interpretable, equitable, and impactful. By addressing both the scientific complexities and the social imperatives of healthcare, this work positions itself at the intersection of discovery and responsibility, paving the way for research that not only explains but also transforms.

List of Publications:
International Refereed Journals

1. Tanwar, Nakul, and Yasha Hasija. *"Explicating miRNA-mediated regulation of inflammatory pathways in COPD, MS, and lung cancer using explainable artificial intelligence: insights from peripheral blood profiles."* **Integrative Biology**, 17 (2025): zyaf020.

Integrative Biology, 2025, 17, zyaf020
https://doi.org/10.1093/intbio/zyaf020
Original Article

Explicating miRNA-mediated regulation of inflammatory pathways in COPD, MS, and lung cancer using explainable artificial intelligence: insights from peripheral blood profiles

Nakul Tanwar and Yasha Hasija*

Department of Biotechnology, Delhi Technological University Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

*Corresponding author. Department of Biotechnology, Delhi Technological University Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India.
E-mail: yashahasija@dtu.ac.in

Abstract

Background: Chronic obstructive pulmonary disease (COPD), multiple sclerosis (MS), and lung cancer are linked by shared inflammatory pathways and immune dysregulation. miRNAs regulate these processes by influencing gene expression, yet their roles in the molecular mechanisms across neurological and respiratory systems are not fully understood.
Objective: This study aims to identify miRNAs and their target genes regulating inflammatory pathways, advancing the understanding of molecular genetics underlying COPD, MS, and lung cancer.
Methods: miRNA expression data (GSE61741) were analyzed using a Random Forest (RF) model optimized via Grid Search and validated with Stratified K-Fold cross-validation. Synthetic Minority Oversampling (SMOTE) addressed data imbalance, while SHapley Additive exPlanations (SHAP) identified key miRNAs. Functional enrichment and pathway analyses explored miRNA-gene interactions. Single-cell level analysis further validated the cell-specific roles of these genes. An independent dataset (GSE31568) was used for validation.
Results: Key miRNAs, including hsa-let-7c, hsa-miR-454, hsa-miR-92a, and hsa-miR-223, were identified as regulators of hallmark inflammatory genes such as CCL2, IL6, ITGB3, and MYC. These genes are critical for cytokine signaling, epithelial repair, and immune modulation. Single-cell analysis highlighted the role of inflammatory fibroblasts in localized inflammation and tissue remodeling. The RF model achieved an accuracy of 81.58%, validated at 82.55%. Pathway analysis emphasized cytokine-cytokine receptor interactions and shared pathways between neurological and respiratory diseases.
Conclusions: This study identifies miRNAs and their target genes as critical regulators of inflammation in COPD, MS, and lung cancer. Single-cell insights and pathway enrichment provide a comprehensive view of shared molecular mechanisms, contributing to biomarker discovery and therapeutic strategies for precision medicine in inflammatory diseases.

Received: February 18, 2025. **Revised:** May 18, 2025. **Accepted:** October 5, 2025

21

2. Tanwar, Nakul, and Yasha Hasija. *"Explicate molecular landscape of combined pulmonary fibrosis and emphysema through explainable artificial intelligence: a comprehensive analysis of ILD and COPD interactions using RNA from whole lung homogenates."* **Medical & Biological Engineering & Computing**, 62.8 (2024): 2557–2570.

Medical & Biological Engineering & Computing (2024) 62:2557–2570
https://doi.org/10.1007/s11517-024-03099-8

ORIGINAL ARTICLE

Explicate molecular landscape of combined pulmonary fibrosis and emphysema through explainable artificial intelligence: a comprehensive analysis of ILD and COPD interactions using RNA from whole lung homogenates

Nakul Tanwar[1] · Yasha Hasija[1]

Received: 9 February 2024 / Accepted: 14 April 2024 / Published online: 22 April 2024

Abstract
Combined pulmonary fibrosis and emphysema (CPFE) presents a unique challenge in respiratory disorders, merging features of interstitial lung disease (ILD) and chronic obstructive pulmonary disease (COPD). Using the random forest algorithm, our study thoroughly examines the molecular details of CPFE. Analyzing gene expression datasets from GSE47460 (ILD: 254, COPD: 220, control: 108), we identify key genes namely ADRB2, CDH3, IRS2, MATN3, CD38, PDIA4, VEGFC, and among twenty others, crucial in airway regulation, lung function, and apoptosis, shaping the complex pathogenesis of CPFE. Additionally, miRNAs (hsa-mir-101-3p, hsa-mir-1343-3p, hsa-mir-27a-3p, and miR-16-5p) showcase regulatory impacts on CPFE-related molecular pathways. Our machine learning model unveils these intricate interactions, offering a comprehensive insight into CPFE's molecular mechanisms. This research not only pinpoints potential therapeutic targets and biomarkers but also opens avenues for innovative approaches in managing CPFE, linking ILD and COPD within this complex respiratory condition.

Keywords CPFE · COPD · ILD · Explainable artificial intelligence · SHAP

1 Introduction

COPD and ILD are clinically distinct conditions, yet they may coexist in some individuals[1]. This overlap is often attributed to common risk factors such as tobacco use. This specific coexistence is known as CPFE, characterized by the presence of emphysema, a form of COPD, primarily in the upper lung fields, along with the widespread presence of ILD [1]. In patients with COPD, interstitial lung abnormalities (ILA) are frequently observed, which might indicate an early or mild stage of ILD. ILAs are characterized by increased lung densities found in chest computed tomography (CT) scans of patients without a previous ILD diagnosis. Although the clinical features of these lung changes resemble those of ILD, they are generally less severe. This similarity suggests that ILAs could represent initial or mild forms of ILD. Approximately 13.5% of individuals with COPD exhibit these changes [2].

The interplay between inflammation and the pathogenesis of chronic pulmonary diseases such as CPFE, COPD, and ILD underscores a complex web of biological mechanisms. In CPFE, the uncertain pathogenesis is thought to involve various cytokines and signaling pathways, with overexpression of inflammatory mediators like PDGF, TNF-α, and TGF-β being linked to emphysema and fibrosis lesions [3]. Similarly, in COPD, epithelial cells play a crucial role in mediating inflammation triggered by inhaled toxins and microorganisms, leading to the production of cytokines, chemokines, and reactive oxygen species (ROS). Notably, smoking has been shown to induce CXCL14 expression by human epithelial cells, correlating with lung cancer development, as elaborated in another study [4]. ILD and lung fibrosis, characterized by varying fibrosis and inflammation degrees, highlight the importance of persistent low-grade inflammation. This persistent inflammation leads to

✉ Yasha Hasija
yashahasija@dtu.ac.in

Nakul Tanwar
tbret41@gmail.com

[1] Department of Biotechnology, Delhi Technological University (DTU), Delhi 110042, India

Curriculum Vitae/Brief Profile

PROFILE

Computational Biologist and Ph.D. Researcher skilled in Explainable AI (XAI), Machine Learning, and Large Language Model (LLM) integration and multi-omics research. Authored multiple studies on respiratory and metabolic disorders, emphasizing interpretable AI, and contributed to a design patent advancing translational biotechnology innovation.

CONTACT

- **Phone**: 9810769156
- **Email**:nakultanwar_2k22phdbt607@dtu.ac.in/ thryt41@gmail.com
- **LinkedIn**: www.linkedin.com/in/nakul-tanwar-a714a4184

PUBLICATIONS:

Journal Articles:

i) **Tanwar, N.**, & Hasija, Y. (2025). Explicating miRNA-mediated regulation of inflammatory pathways in COPD, MS, and lung cancer using explainable artificial intelligence: insights from peripheral blood profiles. Integrative Biology, January 2025.

ii) **Tanwar, N.**, & Hasija, Y. (2024). Explicate molecular landscape of combined pulmonary fibrosis and emphysema through explainable artificial intelligence: a comprehensive analysis of ILD and COPD interactions using RNA from whole lung homogenates. Medical & Biological Engineering & Computing, April 2024.

iii) **Tanwar, N.**, & Hasija, Y.** (2025). An eXplainable AI and Conversational Decision Support System for Lung Cancer Diagnosis and Monitoring. (Manuscript under consideration https://zenodo.org/records/15570514)

Book Chapters:

i) Tanwar, N., & Hasija, Y. (2024). Overview of new trends on deep learning models for diabetes risk prediction. In Advances in Biomedical Engineering and Computing.

IEEE Conference Publications:

i) Yadav, K., **Tanwar, N.**, & Hasija, Y. (2024). Exploring the role of gut microbiota in hypertension: Insights from Machine Learning and eXplainable AI. 15th International Conference on Computing, Communication and Networking Technologies (ICCCNT), June 2024.

ii) **Tanwar, N.**, Meena, J., & Hasija, Y. (2022). Explicate Toxicity by eXplainable Artificial Intelligence. International Conference on Industry 4.0 Technology (I4Tech), September 2022.

iii) **Tanwar, N.**, & Hasija, Y. (2022). Explainable AI: Are We There Yet? IEEE Delhi Section Conference (DELCON), February 2022.

iv) **Tanwar, N.**, & Hasija, Y. (2022). Deep Learning: A Tool in Biomedical Science. International Conference for Advancement in Technology (ICONAT), January 2022.

INTELLECTUAL PROPERTY

Design Patent (Under Consideration)

i) **Portable Food Spoilage Detection Device**: Dual-sensor (gas + TDS/pH) system for real-time detection of food spoilage.
   - Application No.: **461910-001**
   - Status: FER Review (2025)

Nakul Tanwar

(26.09.1999)

EDUCATION

Ph.D. Scholar- *Department of Biotechnology, Delhi Technological university*
2022- Ongoing
Thesis Title*: An Integrated Approach Towards the Identification of Novel biomarkers in Respiratory Disorders.*
MSc Biotechnology- *Delhi Technological University*
2020-2022
Project Title: *Explaining Developmental Neurotoxicity By XAI*
B.Sc. (Honors) Biotechnology- Amity University, Gurugram
2017-2020

WORK EXPERIENCE

- ***Delhi Technological University / 2022–Present / Teaching Assistant***

Guided B.Tech and M.Tech students in bioinformatics, machine learning, and multi-omics projects, assisting in data analysis, model interpretation, and research documentation.

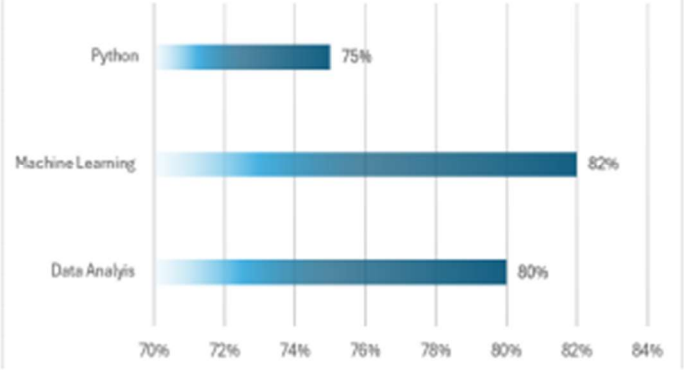
- ***Imperial life sciences/ MAY-JULY 2021/ Lab Intern***

Assisted in preparing and validating RT-PCR kits for COVID-19. Gained hands-on experience in molecular diagnostics, including process optimization, reagent prep, and quality control—contributing to public health efforts.

- ***Janakpuri Super Specialty Hospital/ May- July 2019/ Lab Intern***

Worked in hematology and pathology labs, performing CBC analysis, blood smears, and histopathology preparation. Built strong foundations in diagnostic techniques, reagent handling, and equipment maintenance.

SKILLS



References

[1] Zhu M, Ye M, Wang J, Ye L, Jin M. Construction of Potential miRNA–mRNA Regulatory Network in COPD Plasma by Bioinformatics Analysis. Int J Chron Obstruct Pulmon Dis 2020;15:2135. https://doi.org/10.2147/COPD.S255262.

[2] Feng H, Li Z, Zheng R. The global burden of chronic respiratory diseases attributable to tobacco from 1990 to 2021: a global burden of disease study 2021. BMC Public Health 2025;25:1–13. https://doi.org/10.1186/S12889-025-21680-0/FIGURES/5.

[3] Momtazmanesh S, Moghaddam SS, Ghamari SH, Rad EM, Rezaei N, Shobeiri P, et al. Global burden of chronic respiratory diseases and risk factors, 1990–2019: an update from the Global Burden of Disease Study 2019. EClinicalMedicine 2023;59. https://doi.org/10.1016/J.ECLINM.2023.101936.

[4] Abbafati C, Abbas KM, Abbasi-Kangevari M, Abd-Allah F, Abdelalim A, Abdollahi M, et al. Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. The Lancet 2020;396:1223–49. https://doi.org/10.1016/S0140-6736(20)30752-2.

[5] Safiri S, Carson-Chahhoud K, Noori M, Nejadghaderi SA, Sullman MJM, Ahmadian Heris J, et al. Burden of chronic obstructive pulmonary disease and its attributable risk factors in 204 countries and territories, 1990-2019: results from the Global Burden of Disease Study 2019. BMJ 2022;378. https://doi.org/10.1136/BMJ-2021-069679.

[6] Wang Z, Lin J, Liang L, Huang F, Yao X, Peng K, et al. Global, regional, and national burden of chronic obstructive pulmonary disease and its attributable risk factors from 1990 to 2021: an analysis for the Global Burden of Disease Study 2021. Respir Res 2025;26:1–11. https://doi.org/10.1186/S12931-024-03051-2/FIGURES/4.

[7] 2023 GOLD Report - Global Initiative for Chronic Obstructive Lung Disease - GOLD n.d. https://goldcopd.org/2023-gold-report-2/ (accessed September 25, 2025).

[8] de Oca MM, Perez-Padilla R, Celli B, Aaron SD, Wehrmeister FC, Amaral AFS, et al. The global burden of COPD: epidemiology and effect of prevention strategies. Lancet Respir Med 2025;13:709–24. https://doi.org/10.1016/S2213-2600(24)00339-4.

[9] DR A, AM A, CD B, WC B, JD C, RM F, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011;365:395–409. https://doi.org/10.1056/NEJMOA1102873.

[10] Watanabe S, Saeki K, Waseda Y, Murata A, Takato H, Ichikawa Y, et al. Lung cancer in connective tissue disease-associated interstitial lung disease: clinical features and impact on outcomes. J Thorac Dis 2018;10:799. https://doi.org/10.21037/JTD.2017.12.134.

[11] Verrecchia F, Mauviel A. Transforming growth factor-β and fibrosis. World Journal of Gastroenterology : WJG 2007;13:3056. https://doi.org/10.3748/WJG.V13.I22.3056.

[12] Tanwar N, Hasija Y. Explicate molecular landscape of combined pulmonary fibrosis and emphysema through explainable artificial intelligence: a comprehensive analysis of ILD and COPD interactions using RNA from whole lung homogenates. Med Biol Eng Comput (2024) 2024.

[13] Ghoshouni H, Rafiei N, Yazdan Panah M, Dehghani Firouzabadi D, Mahmoudi F, Aghariahmadabad M, et al. Asthma and chronic obstructive pulmonary disease (COPD) in people with multiple sclerosis: A systematic review and meta-analysis. Mult Scler Relat Disord 2024;85. https://doi.org/10.1016/j.msard.2024.105546.

[14] Tufekci KU, Oner MG, Genc S, Genc K. MicroRNAs and Multiple Sclerosis. Autoimmune Dis 2010;2011:807426. https://doi.org/10.4061/2011/807426.

[15] Chronic respiratory diseases: more than 80 million affected and many more undiagnosed, warns new WHO–ERS report - ERS - European Respiratory Society n.d. https://www.ersnet.org/news-and-features/news/chronic-respiratory-diseases-more-than-80-million-affected-and-many-more-undiagnosed-warns-new-who-ers-report/ (accessed September 25, 2025).
[16] Agustí A, Celli BR, Criner GJ, Halpin D, Anzueto A, Barnes P, et al. Global Initiative for Chronic Obstructive Lung Disease 2023 Report: GOLD Executive Summary. Am J Respir Crit Care Med 2023;207:819. https://doi.org/10.1164/RCCM.202301-0106PP.
[17] Agarwal AK, Raja A, Brown BD. Chronic Obstructive Pulmonary Disease. StatPearls 2023.
[18] Carr LL, Jacobson S, Lynch DA, Foreman MG, Flenaugh EL, Hersh CP, et al. Features of COPD as Predictors of Lung Cancer. Chest 2018;153:1326. https://doi.org/10.1016/J.CHEST.2018.01.049.
[19] Zhao G, Li X, Lei S, Zhao H, Zhang H, Li J. Prevalence of lung cancer in chronic obstructive pulmonary disease: A systematic review and meta-analysis. Front Oncol 2022;12. https://doi.org/10.3389/FONC.2022.947981/FULL.
[20] Young RP, Duan F, Chiles C, Hopkins RJ, Gamble GD, Greco EM, et al. Airflow limitation and histology shift in the National Lung Screening Trial: The NLST-ACRIN cohort substudy. Am J Respir Crit Care Med 2015;192:1060–7. https://doi.org/10.1164/RCCM.201505-0894OC/SUPPL_FILE/DISCLOSURES.PDF.
[21] Takiguchi Y, Sekine I, Iwasawa S, Kurimoto R, Tatsumi K. Chronic obstructive pulmonary disease as a risk factor for lung cancer. World J Clin Oncol 2014;5:660. https://doi.org/10.5306/WJCO.V5.I4.660.
[22] Durham AL, Adcock IM. The relationship between COPD and lung cancer. Lung Cancer 2015;90:121. https://doi.org/10.1016/J.LUNGCAN.2015.08.017.
[23] Zhu M, Ye M, Wang J, Ye L, Jin M. Construction of Potential miRNA–mRNA Regulatory Network in COPD Plasma by Bioinformatics Analysis. Int J Chron Obstruct Pulmon Dis 2020;15:2135. https://doi.org/10.2147/COPD.S255262.
[24] Bartel DP. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. Cell 2004;116:281–97. https://doi.org/10.1016/S0092-8674(04)00045-5.
[25] Yu SL, Chen HY, Chang GC, Chen CY, Chen HW, Singh S, et al. MicroRNA signature predicts survival and relapse in lung cancer. Cancer Cell 2008;13:48–57. https://doi.org/10.1016/J.CCR.2007.12.008.
[26] Ma L, Teruya-Feldstein J, Weinberg RA. Tumour invasion and metastasis initiated by microRNA-10b in breast cancer. Nature 2007;449:682–8. https://doi.org/10.1038/NATURE06174.
[27] Chan JA, Krichevsky AM, Kosik KS. MicroRNA-21 is an antiapoptotic factor in human glioblastoma cells. Cancer Res 2005;65:6029–33. https://doi.org/10.1158/0008-5472.CAN-05-0137.
[28] Keller A, Leidinger P, Vogel B, Backes C, ElSharawy A, Galata V, et al. miRNAs can be generally associated with human pathologies as exemplified for miR-144*. BMC Med 2014;12:1–12. https://doi.org/10.1186/S12916-014-0224-0.
[29] Scott M. Hammond. An overview of microRNAs. Adv Drug Deliv Rev 2015.
[30] Zhuang Y, Hobbs BD, Hersh CP, Kechris K. Identifying miRNA-mRNA Networks Associated With COPD Phenotypes. Front Genet 2021;12:748356. https://doi.org/10.3389/FGENE.2021.748356/BIBTEX.
[31] Pavel AB, Garrison C, Luo L, Liu G, Taub D, Xiao J, et al. Integrative genetic and genomic networks identify microRNA associated with COPD and ILD. Sci Rep 2023;13:1–14. https://doi.org/10.1038/S41598-023-39751-W;SUBJMETA.
[32] Fathinavid A, Ghobadi MZ, Najafi A, Masoudi-Nejad A. Identification of common microRNA between COPD and non-small cell lung cancer through pathway enrichment analysis. BMC Genom Data 2021;22:1–14. https://doi.org/10.1186/S12863-021-00986-Z/FIGURES/5.

[33] Li CX, Gao J, Sköld CM, Wheelock ÅM. miRNA–mRNA–protein dysregulated network in COPD in women. Front Genet 2022;13:1010048. https://doi.org/10.3389/FGENE.2022.1010048/BIBTEX.
[34] Choi WI, LDY, CHG et al. Lung Cancer development and mortality in interstitial lung disease with and without connective tissue diseases: a five-year Nationwide population-based study. Respir Res 2019.
[35] Choi JY, Song JW, Rhee CK. Chronic obstructive pulmonary disease combined with interstitial lung disease. Tuberc Respir Dis (Seoul) 2022. https://doi.org/10.4046/TRD.2021.0104.
[36] Filippi M, Bar-Or A, Piehl F, Preziosa P, Solari A, Vukusic S, et al. Multiple sclerosis. Nat Rev Dis Primers 2018;4:1–27. https://doi.org/10.1038/S41572-018-0041-4;SUBJMETA.
[37] Lubetzki C, Stankoff B. Demyelination in multiple sclerosis. Handb Clin Neurol 2014;122:89. https://doi.org/10.1016/B978-0-444-52001-2.00004-2.
[38] van Langelaar J, Rijvers L, Smolders J, van Luijn MM. B and T Cells Driving Multiple Sclerosis: Identity, Mechanisms and Potential Triggers. Front Immunol 2020;11:497059. https://doi.org/10.3389/FIMMU.2020.00760/FULL.
[39] Liu R, Du S, Zhao L, Jain S, Sahay K, Rizvanov A, et al. Autoreactive lymphocytes in multiple sclerosis: Pathogenesis and treatment target. Front Immunol 2022;13:996469. https://doi.org/10.3389/FIMMU.2022.996469/FULL.
[40] Davoli-Ferreira M, Thomson CA, McCoy KD. Microbiota and Microglia Interactions in ASD. Front Immunol 2021;12:676255. https://doi.org/10.3389/FIMMU.2021.676255/FULL.
[41] Wang L, Liang Y. MicroRNAs as T Lymphocyte Regulators in Multiple Sclerosis. Front Mol Neurosci 2022;15:865529. https://doi.org/10.3389/FNMOL.2022.865529.
[42] Fan W, Liang C, Ou M, Zou T, Sun F, Zhou H, et al. MicroRNA-146a Is a Wide-Reaching Neuroinflammatory Regulator and Potential Treatment Target in Neurological Diseases. Front Mol Neurosci 2020;13:90. https://doi.org/10.3389/FNMOL.2020.00090/FULL.
[43] Ghoshouni H, Rafiei N, Yazdan Panah M, Dehghani Firouzabadi D, Mahmoudi F, Aghariahmadabad M, et al. Asthma and chronic obstructive pulmonary disease (COPD) in people with multiple sclerosis: A systematic review and meta-analysis. Mult Scler Relat Disord 2024;85:105546. https://doi.org/10.1016/J.MSARD.2024.105546.
[44] Fazakis N, Kocsis O, Dritsas E, Alexiou S, Fakotakis N, Moustakas K. Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction. IEEE Access 2021;9:103737–57. https://doi.org/10.1109/ACCESS.2021.3098691.
[45] Dritsas E, Fazakis N, Kocsis O, Fakotakis N, Moustakas K. Long-Term Hypertension Risk Prediction with ML Techniques in ELSA Database. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 2021;12931 LNCS:113–20. https://doi.org/10.1007/978-3-030-92121-7_9.
[46] Fazakis N, Dritsas E, Kocsis O, Fakotakis N, Moustakas K. Long-term Cholesterol Risk Prediction using Machine Learning Techniques in ELSA Database. ICETE International Conference on E-Business and Telecommunication Networks (International Joint Conference on Computational Intelligence) 2021;2021-October:445–50. https://doi.org/10.5220/0010727200003063.
[47] Konstantoulas I, Kocsis O, Dritsas E, Fakotakis N, Moustakas K. Sleep Quality Monitoring with Human Assisted Corrections. ICETE International Conference on E-Business and Telecommunication Networks (International Joint Conference on Computational Intelligence) 2021;2021-October:435–44. https://doi.org/10.5220/0010727100003063.
[48] Dritsas E, Alexiou S, Moustakas K. Cardiovascular Disease Risk Prediction with Supervised Machine Learning Techniques. International Conference on Information and Communication Technologies for Ageing Well and E-Health, ICT4AWE - Proceedings 2022:315–21. https://doi.org/10.5220/0011088300003188.
[49] Heo JN, Yoon JG, Park H, Kim YD, Nam HS, Heo JH. Machine Learning-Based Model for

Prediction of Outcomes in Acute Stroke. Stroke 2019;50:1263–5. https://doi.org/10.1161/STROKEAHA.118.024293.

[50] Li L, Yang Z, Dang Z, Meng C, Huang J, Meng H, et al. Propagation analysis and prediction of the COVID-19. Infect Dis Model 2020;5:282–92. https://doi.org/10.1016/J.IDM.2020.03.002.

[51] Dritsas E, Alexiou S, Moustakas K. COPD Severity Prediction in Elderly with ML Techniques. ACM International Conference Proceeding Series 2022:185–9. https://doi.org/10.1145/3529190.3534748.

[52] Leidy NK, Malley KG, Steenrod AW, Mannino DM, Make BJ, Bowler RP, et al. Insight into Best Variables for COPD Case Identification: A Random Forests Analysis. Chronic Obstructive Pulmonary Diseases: Journal of the COPD Foundation 2016;3:406. https://doi.org/10.15326/JCOPDF.3.1.2015.0144.

[53] Goto T, Camargo CA, Faridi MK, Yun BJ, Hasegawa K. Machine learning approaches for predicting disposition of asthma and COPD exacerbations in the ED. Am J Emerg Med 2018;36:1650–4. https://doi.org/10.1016/J.AJEM.2018.06.062.

[54] Min X, Yu B, Wang F. Predictive Modeling of the Hospital Readmission Risk from Patients' Claims Data Using Machine Learning: A Case Study on COPD. Sci Rep 2019;9. https://doi.org/10.1038/S41598-019-39071-Y.

[55] Saleh L, Mcheick H, Ajami H, Mili H, Dargham J. Comparison of machine learning algorithms to increase prediction accuracy of COPD domain. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 2017;10461 LNCS:247–54. https://doi.org/10.1007/978-3-319-66188-9_22.

[56] Tanwar N, Hasija Y. Explainable AI; Are we there yet? 2022 IEEE Delhi Section Conference, DELCON 2022 2022. https://doi.org/10.1109/DELCON54057.2022.9752777.

[57] Choi JY, Song JW, Rhee CK. Chronic Obstructive Pulmonary Disease Combined with Interstitial Lung Disease. Tuberc Respir Dis (Seoul) 2022;85:122–36. https://doi.org/10.4046/TRD.2021.0141.

[58] Ohgiya M, Matsui H, Tamura A, Kato T, Akagawa S, Ohta K. The Evaluation of Interstitial Abnormalities in Group B of the 2011 Global Initiative for Chronic Obstructive Lung Disease (GOLD) Classification of Chronic Obstructive Pulmonary Disease (COPD). Intern Med 2017;56:2711–7. https://doi.org/10.2169/INTERNALMEDICINE.8406-16.

[59] Lin H, Jiang S. Combined pulmonary fibrosis and emphysema (CPFE): an entity different from emphysema or pulmonary fibrosis alone. J Thorac Dis 2015;7:767. https://doi.org/10.3978/J.ISSN.2072-1439.2015.04.17.

[60] King PT. Inflammation in chronic obstructive pulmonary disease and its role in cardiovascular disease and lung cancer. Clin Transl Med 2015;4:26. https://doi.org/10.1186/S40169-015-0068-Z.

[61] Hügle T. Immunology of fibrotic lung disease: managing infections whilst preventing autoimmunity? J Inflamm Res 2011;4:21. https://doi.org/10.2147/JIR.S10602.

[62] Yao Y, Gu Y, Yang M, Cao D, Wu F. The Gene Expression Biomarkers for Chronic Obstructive Pulmonary Disease and Interstitial Lung Disease. Front Genet 2019;10. https://doi.org/10.3389/FGENE.2019.01154/FULL.

[63] Durham AL, Adcock IM. The relationship between COPD and lung cancer. Lung Cancer 2015;90:121–7. https://doi.org/10.1016/J.LUNGCAN.2015.08.017.

[64] Naccache JM, Gibiot Q, Monnet I, Antoine M, Wislez M, Chouaid C, et al. Lung cancer and interstitial lung disease: a literature review. J Thorac Dis 2018;10:3829. https://doi.org/10.21037/JTD.2018.05.75.

[65] Brown CE, Engelberg RA, Nielsen EL, Curtis JR. Palliative care for patients dying in the intensive care unit with chronic lung disease compared with metastatic cancer. Ann Am Thorac Soc 2016;13:684–9. https://doi.org/10.1513/ANNALSATS.201510-667OC/SUPPL_FILE/DISCLOSURES.PDF.

[66] Tan J, Tedrow JR, Dutta JA, Juan-Guardela B, Nouraie M, Chu Y, et al. Expression of RXFP1 Is Decreased in Idiopathic Pulmonary Fibrosis. Implications for Relaxin-based Therapies. Am J Respir Crit Care Med 2016;194:1392–402. https://doi.org/10.1164/RCCM.201509-1865OC.
[67] Wang S, Dai Y, Shen J, Xuan J. Research on expansion and classification of imbalanced data based on SMOTE algorithm. Scientific Reports 2021 11:1 2021;11:1–11. https://doi.org/10.1038/s41598-021-03430-5.
[68] Tang D, Chen M, Huang X, Zhang G, Zeng L, Zhang G, et al. SRplot: A free online platform for data visualization and graphing. PLoS One 2023;18:e0294236. https://doi.org/10.1371/JOURNAL.PONE.0294236.
[69] Ge SX, Son EW, Yao R. iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data. BMC Bioinformatics 2018 19:1 2018;19:1–24. https://doi.org/10.1186/S12859-018-2486-6.
[70] Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res 2010;38:W214–20. https://doi.org/10.1093/NAR/GKQ537.
[71] Tarhan L, Bistline J, Chang J, Galloway B, Hanna E, Weitz E. Single Cell Portal: an interactive home for single-cell genomics data. BioRxiv 2023. https://doi.org/10.1101/2023.07.13.548886.
[72] Chang L, Zhou G, Soufan O, Xia J. miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology. Nucleic Acids Res 2020;48:W244–51. https://doi.org/10.1093/NAR/GKAA467.
[73] Tzelepis GE, McCool FD. Respiratory dysfunction in multiple sclerosis. Respir Med 2015;109:671–9. https://doi.org/10.1016/J.RMED.2015.01.018.
[74] Egesten A, Brandt L, Olsson T, Granath F, Inghammar M, Löfdahl CG, et al. Increased prevalence of multiple sclerosis among COPD patients and their first-degree relatives: A population-based study. Lung 2008;186:173–8. https://doi.org/10.1007/S00408-008-9081-Y/TABLES/3.
[75] Egesten A, Brandt L, Olsson T, Granath F, Inghammar M, Löfdahl CG, et al. Increased prevalence of multiple sclerosis among COPD patients and their first-degree relatives: a population-based study. Lung 2008;186:173–8. https://doi.org/10.1007/S00408-008-9081-Y.
[76] Ge F, Huo Z, Li C, Wang R, Wang R, Liu Y, et al. Lung cancer risk in patients with multiple sclerosis: a Mendelian randomization analysis. Mult Scler Relat Disord 2021;51. https://doi.org/10.1016/J.MSARD.2021.102927/ATTACHMENT/D8354462-28C2-4636-9E77-853C59670EDF/MMC7.XLSX.
[77] Haase S, Linker RA. Inflammation in multiple sclerosis. Ther Adv Neurol Disord 2021;14. https://doi.org/10.1177/17562864211007687/ASSET/IMAGES/LARGE/10.1177_17562864211007687-FIG1.JPEG.
[78] Yao H, Rahman I. Current concepts on the role of inflammation in COPD and lung cancer. Curr Opin Pharmacol 2009;9:375. https://doi.org/10.1016/J.COPH.2009.06.009.
[79] Das K, Rao LVM. The Role of microRNAs in Inflammation. Int J Mol Sci 2022;23:15479. https://doi.org/10.3390/IJMS232415479.
[80] Keller A, Leidinger P, Vogel B, Backes C, ElSharawy A, Galata V, et al. miRNAs can be generally associated with human pathologies as exemplified for miR-144. BMC Med 2014;12. https://doi.org/10.1186/S12916-014-0224-0.
[81] Keller A, Leidinger P, Bauer A, Elsharawy A, Haas J, Backes C, et al. Toward the blood-borne miRNome of human diseases. Nat Methods 2011;8:841–3. https://doi.org/10.1038/NMETH.1682.
[82] Meena J, Hasija Y. Application of explainable artificial intelligence in the identification of Squamous Cell Carcinoma biomarkers. Comput Biol Med 2022;146:105505. https://doi.org/10.1016/J.COMPBIOMED.2022.105505.

[83] Licursi V, Conte F, Fiscon G, Paci P. MIENTURNET: An interactive web tool for microRNA-target enrichment and network-based analysis. BMC Bioinformatics 2019;20:1–10. https://doi.org/10.1186/S12859-019-3105-X/TABLES/1.
[84] Choudhury S, Chohan A, Dadhwal R, Vakil AP, Franco R, Taweesedt PT. Applications of artificial intelligence in common pulmonary diseases. Http://WwwWjgnetCom/ 2022;3:1–7. https://doi.org/10.35711/AIMI.V3.I1.1.
[85] Bian H, Zhu S, Zhang Y, Fei Q, Peng X, Jin Z, et al. Artificial Intelligence in Chronic Obstructive Pulmonary Disease: Research Status, Trends, and Future Directions --A Bibliometric Analysis from 2009 to 2023. Int J Chron Obstruct Pulmon Dis 2024;19:1849. https://doi.org/10.2147/COPD.S474402.
[86] Gupta T, Shah A". Citation: The Role of Artificial Intelligence in Lung Disease Diagnosis n.d.
[87] Robertson NM, Centner CS, Siddharthan T. Integrating Artificial Intelligence in the Diagnosis of COPD Globally: A Way Forward. Chronic Obstructive Pulmonary Diseases 2024;11:114–20. https://doi.org/10.15326/JCOPDF.2023.0449.
[88] Khemasuwan D, Sorensen JS, Colt HG. Artificial intelligence in pulmonary medicine: computer vision, predictive model and COVID-19. European Respiratory Review 2020;29:200181. https://doi.org/10.1183/16000617.0181-2020.
[89] Kurniawan Y, Susilo B, Rahman SA. Revolutionizing COPD and Asthma Management with Artificial Intelligence. MedRxiv 2025:2025.03.18.25324219. https://doi.org/10.1101/2025.03.18.25324219.
[90] Easton K, Potter S, Bec R, Bennion M, Christensen H, Grindell C, et al. A virtual agent to support individuals living with physical and mental comorbidities: Co-design and acceptability testing. J Med Internet Res 2019;21:e12996. https://doi.org/10.2196/12996.
[91] Schachner T, Keller R, Wangenheim F v. Artificial intelligence-based conversational agents for chronic conditions: Systematic literature review. J Med Internet Res 2020;22:e20701. https://doi.org/10.2196/20701.
[92] Altom DS, Taha AIA, Hussein AAAM, Elshiekh MAI, Abdelmajed AHA, Ibrahim FIA, et al. Artificial Intelligence-Based Chatbots in Chronic Disease Management: A Systematic Review of Applications and Challenges. Cureus 2025;17. https://doi.org/10.7759/CUREUS.81001.
[93] Hassan N, Miah ASM, Suzuki K, Okuyama Y, Shin J. Stacked CNN-based multichannel attention networks for Alzheimer disease detection. Sci Rep 2025;15:5815. https://doi.org/10.1038/S41598-025-85703-X;SUBJMETA=117,258,639,705;KWRD=COMPUTER+SCIENCE,INFORMATION+TECHNOLOGY,MATHEMATICS+AND+COMPUTING.
[94] Irfan Ahmed. Lung Cancer Prediction Dataset . Kaggle 2025. https://doi.org/https://doi.org/10.34740/kaggle/dsv/10827884.
[95] Shah D, Xue ZY, Aamodt TM. Label Encoding for Regression Networks. ICLR 2022 - 10th International Conference on Learning Representations 2022.
[96] De Amorim LB V, Cavalcanti GDC, Cruz RMO. The choice of scaling technique matters for classification performance n.d.
[97] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016;13-17-August-2016:785–94. https://doi.org/10.1145/2939672.2939785/SUPPL_FILE/KDD2016_CHEN_BOOSTING_SYSTEM_01-ACM.MP4.
[98] Rajpal S, Rajpal A, Agarwal M, Kumar V, Abraham A, Khanna D, et al. XAI-CNVMarker: Explainable AI-based copy number variant biomarker discovery for breast cancer subtypes. Biomed Signal Process Control 2023;84:104979. https://doi.org/10.1016/J.BSPC.2023.104979.
[99] D. M. Dash: An Easy-to-Use Framework for Building Web Applications and Dashboards. Journal of Computing Sciences in Colleges 2023. https://doi.org/10.5555/3606388.3606401.

[100] (PDF) An Effective Query System Using LLMs and LangChain n.d. https://www.researchgate.net/publication/372529063_An_Effective_Query_System_Using_LLMs_and_LangChain (accessed May 31, 2025).
[101] HSP90AA1, ADRB2, TBL1XR1 and HSPB1 are chronic obstructive pulmonary disease-related genes that facilitate squamous cell lung cancer progression n.d. https://www.spandidos-publications.com/10.3892/ol.2020.11318 (accessed January 16, 2024).
[102] Ağaç D, Estrada LD, Maples R, Hooper L V., Farrar JD. The β2-adrenergic receptor controls inflammation by driving rapid IL-10 secretion. Brain Behav Immun 2018;74:176. https://doi.org/10.1016/J.BBI.2018.09.004.
[103] Cho SJ, Weiden MD, Lee CG. Chitotriosidase in the Pathogenesis of Inflammation, Interstitial Lung Diseases and COPD. Allergy Asthma Immunol Res 2014;7:14–21. https://doi.org/10.4168/AAIR.2015.7.1.14.
[104] Li Z, Wang S, Zhao H, Yan P, Yuan H, Zhao M, et al. Artificial neural network identified the significant genes to distinguish Idiopathic pulmonary fibrosis. Scientific Reports 2023 13:1 2023;13:1–15. https://doi.org/10.1038/s41598-023-28536-w.
[105] Lijing Wang 1 Hongjun Zhao, # 2, * Indu Raman, 3 Mei Yan, 3 Qiong Chen, 1 and Quan-Zhen. Peripheral Blood Mononuclear Cell Gene Expression in Chronic Obstructive Pulmonary Disease: miRNA and mRNA Regulation. J Inflamm Res 2022.
[106] Nakahara M, Ito H, Skinner JT, Lin Q, Tamosiuniene R, Nicolls MR, et al. The inflammatory role of dysregulated IRS2 in pulmonary vascular remodeling under hypoxic conditions. Am J Physiol Lung Cell Mol Physiol 2021;321:L416. https://doi.org/10.1152/AJPLUNG.00068.2020.
[107] Jeong I, Lim JH, Park JS, Oh YM. Aging-related changes in the gene expression profile of human lungs. Aging 2020;12:21391–403. https://doi.org/10.18632/AGING.103885.
[108] Masabumi Shibuya. Vascular Endothelial Growth Factor (VEGF) and Its Receptor (VEGFR) Signaling in Angiogenesis. GENE & CANCER 2011.
[109] Janér J, Lassus P, Haglund C, Paavonen K, Alitalo K, Andersson S. Pulmonary Vascular Endothelial Growth Factor-C in Development and Lung Injury in Preterm Infants. Https://DoiOrg/101164/Rccm200508-1291OC 2012;174:326–30. https://doi.org/10.1164/RCCM.200508-1291OC.
[110] Barratt SL, Flower VA, Pauling JD, Millar AB. VEGF (Vascular Endothelial Growth Factor) and Fibrotic Lung Disease. Int J Mol Sci 2018;19. https://doi.org/10.3390/IJMS19051269.
[111] Knox AJ, Stocks J, Sutcliffe A. Angiogenesis and vascular endothelial growth factor in COPD. Thorax 2005;60:88–9. https://doi.org/10.1136/THX.2004.030544.
[112] Gp Guedes A, Dileepan M, Jude JA, Deshpande DA, Walseth TF, Kannan MS. Role of CD38/cADPR signaling in obstructive pulmonary diseases. Curr Opin Pharmacol n.d.;2020:29–33. https://doi.org/10.1016/j.coph.2020.04.007.
[113] Korfei M, Mackenzie B, Meiners S. The ageing lung under stress. European Respiratory Review 2020;29:1–22. https://doi.org/10.1183/16000617.0126-2020.
[114] Bradley KL, Stokes CA, Marciniak SJ, Parker LC, Condliffe AM. Role of unfolded proteins in lung disease. Thorax 2021;76:92–9. https://doi.org/10.1136/THORAXJNL-2019-213738.
[115] Wendell SG, Fan H, Zhang C. G Protein–Coupled Receptors in Asthma Therapy: Pharmacology and Drug Action. Pharmacol Rev 2020;72:1. https://doi.org/10.1124/PR.118.016899.
[116] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 2020;21:1–13. https://doi.org/10.1186/S12864-019-6413-7/TABLES/5.
[117] Teuber-Hanselmann S, Meinl E, Junker A. MicroRNAs in gray and white matter multiple sclerosis lesions: impact on pathophysiology. J Pathol 2020;250:496–509. https://doi.org/10.1002/PATH.5399.

[118] Culina S, Commère PH, Turc E, Jouy A, Pellegrini S, Roux T, et al. MicroRNA signatures of CD4+ T cell subsets in healthy and multiple sclerosis subjects determined by small RNA-sequencing. J Neuroimmunol 2025;401:578531. https://doi.org/10.1016/J.JNEUROIM.2025.578531/ASSET/A6434A3D-ED75-4F67-80A3-DCDBDE5D2FE7/MAIN.ASSETS/GR2.JPG.
[119] Regev K, Healy BC, Paul A, Diaz-Cruz C, Mazzola MA, Raheja R, et al. Identification of MS-specific serum miRNAs in an international multicenter study. Neurol Neuroimmunol Neuroinflamm 2018;5. https://doi.org/10.1212/NXI.0000000000000491/ASSET/F593A038-703C-4A43-A3F1-63F65C294925/ASSETS/GRAPHIC/NXI.0000000000000491F1.JPEG.
[120] Mandolesi G, Rizzo FR, Balletta S, Bassi MS, Gilio L, Guadalupi L, et al. The microRNA let-7b-5p Is Negatively Associated with Inflammation and Disease Severity in Multiple Sclerosis. Cells 2021, Vol 10, Page 330 2021;10:330. https://doi.org/10.3390/CELLS10020330.
[121] Liu T, Zhang Z, Shen W, Wu Y, Bian T. MicroRNA Let-7 Induces M2 Macrophage Polarization in COPD Emphysema Through the IL-6/STAT3 Pathway. International Journal of COPD 2023;18:575–91. https://doi.org/10.2147/COPD.S404850,.
[122] Di T, Yang Y, Fu C, Zhang Z, Qin C, Sai X, et al. Let-7 mediated airway remodelling in chronic obstructive pulmonary disease via the regulation of IL-6. Eur J Clin Invest 2021;51:e13425. https://doi.org/10.1111/ECI.13425;REQUESTEDJOURNAL:JOURNAL:13652362;PAGE:STRING:ARTICLE/CHAPTER.
[123] Zhu X, Kudo M, Huang X, Sui H, Tian H, Croce CM, et al. Frontiers of MicroRNA Signature in Non-small Cell Lung Cancer. Front Cell Dev Biol 2021;9:643942. https://doi.org/10.3389/FCELL.2021.643942/XML/NLM.
[124] Ma X, Zhou J, Zhong Y, Jiang L, Mu P, Li Y, et al. Expression, Regulation and Function of MicroRNAs in Multiple Sclerosis. Int J Med Sci 2014;11:810. https://doi.org/10.7150/IJMS.8647.
[125] Zhang Y, Sheng Y, Gao Y, Lin Y, Cheng B, Li H, et al. Exploration of the Pathogenesis of Chronic Obstructive Pulmonary Disease Caused by Smoking—Based on Bioinformatics Analysis and In Vitro Experimental Evidence. Toxics 2023;11:995. https://doi.org/10.3390/TOXICS11120995/S1.
[126] Cantoni C, Ghezzi L, Choi J, Cross AH, Piccio L. Targeting miR-223 enhances myeloid-derived suppressor cell suppressive activities in multiple sclerosis patients. Mult Scler Relat Disord 2023;76:104839. https://doi.org/10.1016/J.MSARD.2023.104839.
[127] Roffel MP, Bracke KR, Heijink IH, Maes T. miR-223: A Key Regulator in the Innate Immune Response in Asthma and COPD. Front Med (Lausanne) 2020;7:196. https://doi.org/10.3389/FMED.2020.00196.
[128] Fujiwara M, Raheja R, Garo LP, Ajay AK, Kadowaki-Saga R, Karandikar SH, et al. microRNA-92a promotes CNS autoimmunity by modulating the regulatory and inflammatory T cell balance. J Clin Invest 2022;132. https://doi.org/10.1172/JCI155693.
[129] Zhuang Y, Hobbs BD, Hersh CP, Kechris K. Identifying miRNA-mRNA Networks Associated With COPD Phenotypes. Front Genet 2021;12:748356. https://doi.org/10.3389/FGENE.2021.748356/BIBTEX.
[130] Tanwar N, Meena J, Hasija Y. Explicate Toxicity By eXplainable Artificial Intelligence. 2022 International Conference on Industry 40 Technology, I4Tech 2022 2022. https://doi.org/10.1109/I4TECH55392.2022.9952865.
[131] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–50. https://doi.org/10.1073/PNAS.0506580102/SUPPL_FILE/06580FIG7.JPG.
[132] Dong Y, Dong Y, Zhu C, Yang L, Wang H, Li J, et al. Targeting CCL2-CCR2 signaling pathway alleviates macrophage dysfunction in COPD via PI3K-AKT axis. Cell

Communication and Signaling 2024;22:1–19. https://doi.org/10.1186/S12964-024-01746-Z/FIGURES/6.

[133] Xu M, Wang Y, Xia R, Wei Y, Wei X. Role of the CCL2-CCR2 signalling axis in cancer: Mechanisms and therapeutic targeting. Cell Prolif 2021;54:e13115. https://doi.org/10.1111/CPR.13115.

[134] Coperchini F, Chiovato L, Ricci G, Croce L, Magri F, Rotondi M. The cytokine storm in COVID-19: Further advances in our understanding the role of specific chemokines involved. Cytokine Growth Factor Rev 2021;58:82–91. https://doi.org/10.1016/j.cytogfr.2020.12.005.

[135] Zeng Z, Lan T, Wei Y, Wei X. CCL5/CCR5 axis in human diseases and related treatments. Genes Dis 2022;9:12–27. https://doi.org/10.1016/J.GENDIS.2021.08.004.

[136] Salim AA, Alsaimary IE, Alsudany AAK. The Role of Chemokines (CCL2, CCL5 and CXCL10) in the Neuroinflammation Among Patients with Multiple Sclerosis. Mult Scler Relat Disord 2023;80:105174. https://doi.org/10.1016/J.MSARD.2023.105174.

[137] Ireland SJ, Monson NL, Davis LS. Seeking Balance: Potentiation and Inhibition of Multiple Sclerosis Autoimmune Responses by IL-6 and IL-10. Cytokine 2015;73:236. https://doi.org/10.1016/J.CYTO.2015.01.009.

[138] Tang L, Zhang L, Mei X, Yu J, Jiang G. Pulmonary infection is associated with an increased IL-6 in acute exacerbation chronic obstructive pulmonary disease. Eur J Inflamm 2023;21. https://doi.org/10.1177/1721727X221149534/ASSET/IMAGES/LARGE/10.1177_1721727X221149534-FIG1.JPEG.

[139] Qu Z, Sun F, Zhou J, Li L, Shapiro SD, Xiao G. Interleukin-6 prevents the initiation but enhances the progression of lung cancer. Cancer Res 2015;75:3209–15. https://doi.org/10.1158/0008-5472.CAN-14-3042/651688/AM/INTERLEUKIN-6-PREVENTS-THE-INITIATION-BUT-ENHANCES.

[140] Zhu C, Kong Z, Wang B, Cheng W, Wu A, Meng X. ITGB3/CD61: a hub modulator and target in the tumor microenvironment. Am J Transl Res 2019;11:7195.

[141] Fuentes P, Sesé M, Guijarro PJ, Emperador M, Sánchez-Redondo S, Peinado H, et al. ITGB3-mediated uptake of small extracellular vesicles facilitates intercellular communication in breast cancer cells. Nature Communications 2020 11:1 2020;11:1–15. https://doi.org/10.1038/s41467-020-18081-9.

[142] Burgoyne RA, Fisher AJ, Borthwick LA. The Role of Epithelial Damage in the Pulmonary Immune Response. Cells 2021;10:2763. https://doi.org/10.3390/CELLS10102763.

[143] Orian JM, D'Souza CS, Kocovski P, Krippner G, Hale MW, Wang X, et al. Platelets in Multiple Sclerosis: Early and Central Mediators of Inflammation and Neurodegeneration and Attractive Targets for Molecular Imaging and Site-Directed Therapy. Front Immunol 2021;12:620963. https://doi.org/10.3389/FIMMU.2021.620963.

[144] Bernhagen J. A new cytokine target for chronic obstructive pulmonary disease? EBioMedicine 2021;69:103479. https://doi.org/10.1016/J.EBIOM.2021.103479.

[145] Roffel MP, Maes T, Brandsma CA, Van Den Berge M, Vanaudenaerde BM, Joos GF, et al. MiR-223 is increased in lungs of patients with COPD and modulates cigarette smoke-induced pulmonary inflammation. Am J Physiol Lung Cell Mol Physiol 2021;321:L1091–104. https://doi.org/10.1152/AJPLUNG.00252.2021,.

[146] Hassan SS ul, Samanta S, Dash R, Karpiński TM, Habibi E, Sadiq A, et al. The neuroprotective effects of fisetin, a natural flavonoid in neurodegenerative diseases: Focus on the role of oxidative stress. Front Pharmacol 2022;13:1015835. https://doi.org/10.3389/FPHAR.2022.1015835.

[147] Li D, Hu J, Wang T, Zhang X, Liu L, Wang H, et al. Silymarin attenuates cigarette smoke extract-induced inflammation via simultaneous inhibition of autophagy and ERK/p38 MAPK pathway in human bronchial epithelial cells. Sci Rep 2016;6:1–10. https://doi.org/10.1038/SREP37751;TECHMETA=80,82;SUBJMETA=1426,1785,308,692,699;KWRD=EXPERIMENTAL+MODELS+OF+DISEASE,RESPIRATORY+TRACT+DIS

EASES.
[148] Kusuma FK, Prabhu A, Tieo G, Ahmed SM, Dakle P, Yong WK, et al. Signalling inhibition by ponatinib disrupts productive alternative lengthening of telomeres (ALT). Nat Commun 2023;14. https://doi.org/10.1038/S41467-023-37633-3.
[149] Gremke N, Besong I, Stroh A, von Wichert L, Witt M, Elmshäuser S, et al. Targeting PI3K inhibitor resistance in breast cancer with metabolic drugs. Signal Transduction and Targeted Therapy 2025 10:1 2025;10:1–19. https://doi.org/10.1038/s41392-025-02180-4.
[150] Zhai Y, Lan D, Lv S, Mo L. Interpretability-based machine learning for predicting the risk of death from pulmonary inflammation in Chinese intensive care unit patients. Front Med (Lausanne) 2024;11:1399527. https://doi.org/10.3389/FMED.2024.1399527/BIBTEX.
[151] Krishna SU, Lakshman ANB, Archana T, Raja K, Ayyadurai M. Lung Cancer Prediction and Classification Using Decision Tree and VGG16 Convolutional Neural Networks. Open Biomed Eng J 2024;18. https://doi.org/10.2174/0118741207290271240322061032,.
[152] Hedhoud Y, Mekhaznia T, Amroune M. An improvement of the CNN-XGboost model for pneumonia disease classification. Pol J Radiol 2023;88:e483–93. https://doi.org/10.5114/PJR.2023.132533,.
[153] Pathan RK, Shorna IJ, Hossain MS, Khandaker MU, Almohammed HI, Hamd ZY. The efficacy of machine learning models in lung cancer risk prediction with explainability. PLoS One 2024;19:e0305035. https://doi.org/10.1371/JOURNAL.PONE.0305035.
[154] Zak M, Krzyżak A. Classification of Lung Diseases Using Deep Learning Models. Computational Science – ICCS 2020 2020;12139:621. https://doi.org/10.1007/978-3-030-50420-5_47.
[155] Das N, Happaerts S, Gyselinck I, Staes M, Derom E, Brusselle G, et al. Collaboration between explainable artificial intelligence and pulmonologists improves the accuracy of pulmonary function test interpretation. European Respiratory Journal 2023;61. https://doi.org/10.1183/13993003.01720-2022.

turnitin Page 122 of 122 - Integrity Submission Submission ID trn:oid:::27535:122649592