Ministry of Science and Higher Education of the Republic of Kazakhstan

SDU University

SDU

1996

UNIVERSITY

Zhaniya Medeuova

Development of a Medical Decision Support System for Lung Disease Diagnosis Using Deep Learning

THESIS

Presented in Partial Fulfilment for the

Degree of Master of Technical Science in Computer Science

(degree code: 7M06102)

Department of Computer Science

Faculty of Engineering and Natural Sciences

Supervisor: **Batyrkhan Omarov, PhD**

Kaskelen, June 2025

SDU University
Faculty of Engineering and Natural Sciences
Department of Computer Science

Dean of Faculty of Engineering and Natural Sciences

Assistant Professor, PhD Akhmedov Ramis

« 13 » June 2025

Topic of the thesis:
Development of a Medical Decision Support System for Lung Disease Diagnosis Using Deep Learning

Thesis submitted as part of the requirements for the award of the MSc in "7M06102 - Computer Science", SDU University

Head of Department Zhanar Mukash

Academic Supervisor Batyrkhan Omarov

Master student Zhaniya Medeuova

Kaskelen, 2025

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Zhaniya Medeuova

June, 2025

Acknowledgements

I would like to express my sincere gratitude to everyone who supported me throughout this research and thesis writing. First and foremost, I am deeply thankful to my supervisor, Dr. Batyrkhan Omarov, for his invaluable guidance, encouragement, and constructive feedback during every stage of my work. I also thank the Faculty of Engineering and Natural Sciences at SDU University for providing a supportive academic environment and resources. Special thanks go to my family and friends for their unwavering support, patience, and motivation throughout this journey. Finally, I appreciate all those who contributed to my learning and growth, directly or indirectly, during this academic pursuit.

Dedication

This thesis is dedicated to those interested in medicine and AI.

Abstract

Lung diseases such as Chronic Obstructive Pulmonary Disease (COPD), asthma, and pneumonia remain a leading cause of morbidity and mortality worldwide, with developing countries like Kazakhstan experiencing a rising burden due to environmental and lifestyle factors. Traditional diagnostic methods, while effective, are often limited by cost, accessibility, and dependence on expert interpretation, especially in low-resource settings. To address these challenges, this study proposes a novel artificial intelligence DL based decision support system that classifies lung conditions into three categories like normal, COPD. Using multimodal data: chest X-rays and lung sound recordings. The system utilizes a dual-stream deep learning architecture with late fusion to extract and integrate features from each modality independently, removing the need for synthetically paired datasets. This approach not only improves diagnostic scalability and accuracy but also provides a practical, low-cost solution for early lung disease detection, particularly in underserved regions. The model's performance demonstrates the viability of using real-world, unpaired data for multimodal diagnostics, offering a significant step toward accessible, AI-driven respiratory healthcare.

Аңдатпа

Созылмалы обструктивті өкпе ауруы (СОӨА), астма және пневмония сияқты өкпе аурулары бүкіл әлемде сырқаттанушылық пен өлім-жітімнің басты себептері болып қала береді, ал Қазақстан сияқты дамушы елдерде бұл жағдай қоршаған орта мен өмір салтына байланысты ушығып келеді. Дәстүрлі диагностикалық әдістер тиімді болғанымен, олардың құны жоғары, қолжетімділігі шектеулі және сарапшының түсіндірмесіне тәуелді, әсіресе ресурстары аз аймақтарда. Осы қиындықтарды шешу үшін бұл зерттеуде көкірек қуысының рентген кескіндері мен өкпе дыбыстары сияқты мультимодальды деректерді пайдалана отырып, өкпе жағдайларын жіктейтін жасанды интеллектке негізделген шешім қабылдау жүйесі ұсынылады. Жүйе әрбір модалдықтан ерекшеліктерді дербес шығарып, біріктіруге арналған екі арналы терең оқыту архитектурасын пайдаланады, бұл синтетикалық түрде жұпталған деректердің қажеттілігін жояды. Бұл тәсіл диагностика дәлдігін арттырып қана қоймай, сонымен қатар шалғай және ресурстары шектеулі аймақтар үшін арзан, практикалық шешім ұсынады. Модельдің нәтижелері нақты өмірден алынған деректер негізінде мультимодальды диагностикаға арналған сенімді шешім бола алатынын көрсетеді.

Аннотация

Заболевания лёгких, такие как хроническая обструктивная болезнь лёгких (ХОБЛ), астма и пневмония, остаются одной из основных причин заболеваемости и смертности во всём мире, причём развивающиеся страны, такие как Казахстан, испытывают всё большую нагрузку из-за экологических и поведенческих факторов. Хотя традиционные методы диагностики эффективны, они часто ограничены высокой стоимостью, труднодоступностью и зависимостью от экспертной интерпретации, особенно в условиях с ограниченными ресурсами. В данном исследовании предлагается новая интеллектуальная система поддержки принятия решений, основанная на методах глубокого обучения, которая классифицирует состояния лёгких как нормальные, ХОБЛ и другие заболевания, используя мультимодальные данные: рентгеновские снимки грудной клетки и записи звуков дыхания. Архитектура с двойным потоком и поздним объединением признаков позволяет использовать немаркированные реальные данные, что делает систему более масштабируемой и доступной для ранней диагностики заболеваний лёгких. Результаты демонстрируют эффективность подхода и его потенциал в обеспечении доступной диагностики в условиях ограниченных ресурсов.

Abbreviations

ABG	Arterial Blood Gas
AI	Artificial Intelligence
AST	Audio Spectrogram Transformer
AUC	Area Under the Curve
BOW	Bag of Words
CFTR	Cystic Fibrosis Transmembrane Conductance Regulator
CNN	Convolutional Neural Network
COPD	Chronic Obstructive Pulmonary Disease
CT	Computed Tomography
CXR	Chest X-ray
DBNs	Deep Belief Networks
DL	Deep Learning
FEV	Forced Expiratory Volume
FVC	Forced Vital Capacity
GPU	Graphics Processing Unit
MFCCs	Mel Frequency Cepstral Coefficients
ML	Machine Learning
MRI	Magnetic Resonance Imaging
PANNs	Pretrained Audio Neural Networks
ReLU	Rectified Linear Unit
RF	Random Forests
ROC	Receiver Operating Characteristic
SE Block	Squeeze-and-Excitation Block
STE	Short-Time Energy
SVM	Support Vector Machines
TB	Tuberculosis
WHO	World Health Organization

Table of Contents

Declaration i

Acknowledgements ii

Dedication iii

Abstract iv

Аңдатпа v

Аннотация vi

List of Abbreviations vii

1 Background and motivations 1
- 1.1 Introduction . . . 1
- 1.2 Background . . . 3
  - 1.2.1 Human respiratory system . . . 3
  - 1.2.2 Lung Diseases Overview . . . 4
  - 1.2.3 Chronic Obstructive Pulmonary Disease (COPD) . . . 5
  - 1.2.4 Traditional Diagnostic Methods . . . 6
  - 1.2.5 Artificial Intelligence in Medical Diagnosis . . . 8
  - 1.2.6 Deep Learning in Healthcare . . . 8
  - 1.2.7 Multimodal Learning Approaches . . . 10
- 1.3 Thesis Organization . . . 11

2 Problem Statement 12
- 2.1 Global Burden of Lung Diseases . . . 12
- 2.2 Diagnostic Challenges in Resource-Limited Settings . . . 13

3 Literature Review 15
- 3.1 Deep Learning in Medical Imaging for Lung Disease . . . 15
- 3.2 Deep Learning for Lung Sound Analysis in Lung Disease . . . 17
- 3.3 Multimodal Approaches for Lung Disease Detection . . . 20

4 Methodology 25
- 4.1 Dataset . . . 26
- 4.2 Data Preprocessing . . . 27

4.3 Model Design 28
4.3.1 Image-based model 28
4.3.2 Audio-based model 29
4.3.3 Fusion strategy 29
4.4 Model Training 29
4.5 Evaluation 31
4.6 Experimental Environment 32

5 Experiments and Results **33**
5.1 Model Performance 33
5.2 Comparative Analysis 36
5.2.1 Singlmodal vs. Multimodal 36
5.2.2 Comparison with exicting work 37

6 Conclusions and future work **39**
6.1 Conclusions 39
6.2 Future work 39

Chapter 1

Background and motivations

1.1 Introduction

The respiratory system is an essential part of the body and involves the airways, lungs, and windpipe for respiration. It helps in the exchange of essential oxygen and carbon dioxide gases for cell respiration and removes waste products while ensuring that body tissues receive the oxygen they need, supporting all metabolic activities [1, 2]. This system, despite its robust nature, is still susceptible to numerous diseases that can affect its work and the health of an individual, reducing quality of life and increasing the risk of mortality. Lung diseases encompass various conditions affecting the lung and airways and can, in some cases, threaten the well-being of an individual [3]. Conditions that fall into this category are pneumonia, tuberculosis, lung cancer, chronic obstructive pulmonary disease, among others [4, 5]. Lung diseases, particularly chronic respiratory conditions, continue to pose major public health challenges worldwide. These include asthma, pneumonia, tuberculosis and, in particular, COPD a progressive illness characterized by chronic inflammation and obstruction of airflow in the lungs.
Globally, respiratory diseases account for millions of deaths each year, with COPD ranking as one of the leading causes. Recent statistics on the global burden of benign lung diseases were obtained from reliable sources. COPD is the leading cause of death compared to the other leading causes, causing millions of deaths per year and affecting the overall quality of life of human beings with this condition [6]. In 2019, there were 3.23 million cases of COPD worldwide, while the number of deaths from chronic respiratory diseases was 4.0 million [7, 8]. In addition, asthma burdens about 23.3 million adults and 6.6% of children in the United States alone [9]. According to recent estimates, the prevalence and impact of COPD are growing, particularly in developing countries. A striking example is Kazakhstan, which in 2024 is third in the world in terms of prevalence of COPD [10]. This alarming statistic reflects a combination of factors that include widespread tobacco use, occupational exposure, and a significant environmental issue such as air pollution [11]. COPD is primarily caused by long-term exposure to harmful substances that damage lung tissue and airways. These include cigarette smoke, industrial dust, chemical fumes, biomass fuel smoke, and air pollution [12]. The disease develops slowly and is often diagnosed only in its later stages, when symptoms such as

chronic cough, shortness of breath, and wheezing become more severe and irreversible damage has already occurred [12].
Traditional methods used to detect lung diseases include spirometry, which measures lung function, imaging techniques such as chest CT scans, X-rays and physical examination methods such as auscultation [4]. Although clinically valuable, each of these methods has significant limitations. Spirometry requires active patient cooperation and may be inaccurate in elderly or severely ill patients. CT scans offer detailed imaging, but are expensive and expose patients to ionizing radiation. Auscultation relies heavily on clinician expertise and may not detect early-stage disease. These drawbacks make traditional diagnosis both difficult to scale and inaccessible in many low-resource settings [6, 13].
Early detection is considered paramount in the fight to reduce the burden of lung diseases and further improve outcomes and survival rates among patients. Therefore, timely diagnosis enables proper intervention, while effective prevention and management efforts can considerably reduce the burden of these chronic, noncommunicable diseases [13, 14, 15]. Combined with the shortage of pulmonology specialists and limited access to diagnostic tools in remote areas, these factors underscore the urgent need for scalable, affordable, and accurate detection methods.
In response to these challenges, artificial intelligence has emerged as a promising solution in medical diagnostics. Specifically, deep learning, a subfield of AI, has demonstrated strong capabilities for analyzing complex medical data, such as chest X-rays and cough sound recordings [3, 6]. Deep learning offers several advantages. This solution complements the work of medical professionals by offering fast, scalable, and non-invasive analysis, helping streamline the diagnostic process.
However, most of the research done so far relies on single-modality approaches that, in general, rely on one data source such as X-rays or CT scans and other types of data [16]. While these methods have produced promising results, they often do not address the complexities of lung diseases that require a multimodal approach [16]. In fact, a wide variety of recent studies illustrate that the integration of more types of data sources, for example, lung sounds and imaging data, provides better diagnosis performance [16]. Although that has potential, current studies of this nature are very few in number, focusing mainly on single-modal approaches to date [16]. Recent research has attempted to overcome these limitations by combining multiple data modalities for diagnosis. However, most existing multimodal models are based on paired datasets, that is, X-rays and cough recordings taken from the same patient. In practice, such data are rare, limiting the feasibility and scalability of these approaches. Furthermore, very few studies specifically focus on distinguishing COPD from other lung conditions and healthy cases, even though this distinction is crucial for treatment planning and prognosis.
To address this gap, the present study proposes this approach: the development of a multimodal deep learning system that classifies lung conditions into two categories: normal, COPD lung disease using unpaired chest X-rays and lung sound data. This system will use dual-stream neural networks with late fusion, allowing it to extract and combine meaningful features from each modality independently. Importantly, it eliminates the need for synthetically paired datasets and instead leverages independently collected data from the real world. The central research

question is Can a multimodal deep learning model combining chest X-rays and cough sound data improve the accuracy of COPD detection compared to single-modal approaches? The objective of this research is to build and validate such a system, with the aim of improving early detection and reducing misdiagnosis, particularly in countries like Kazakhstan, where COPD is a major public health problem and access to specialized care is limited. This research contributes to a dual stream model trained on X-ray and sound data, introducing a practical and ethical approach that aligns with clinical needs. Avoids the pitfalls of synthetic data generation, introduces a meaningful binary class prediction scheme, and demonstrates how low-cost tools can be used for scalable lung disease screening. By addressing both the technical and the real world healthcare limit, this study aims to advance the field of AI in medicine and offer a solution that is directly applicable to the healthcare challenges faced in Kazakhstan and similar regions.

1.2 Background

1.2.1 Human respiratory system

The human respiratory system is a complex network of organs and tissues that allows the body to breathe and maintain vital gas exchange [1]. Its primary function is to deliver oxygen to the bloodstream and remove carbon dioxide, a metabolic waste product. This system plays a crucial role in maintaining cellular respiration and general physiological balance [9].

As shown in Figure 1.1 anatomically, the respiratory system is divided into two main parts: the upper and lower respiratory tracts. The upper respiratory tract includes the nose, nasal cavity, sinuses, throat, and larynx (voice box). These structures are responsible for filtering, warming, and humidifying the air before it enters the lungs. The lower respiratory tract comprises the trachea, bronchi, bronchioles, and the lungs themselves, which contain millions of tiny air sacs called alveoli. These alveoli are the sites of gas exchange, where oxygen is diffused into the blood and carbon dioxide is diffused out [17].

The act of breathing is controlled by the diaphragm, a dome-shaped muscle beneath the lungs. When the diaphragm contracts, it creates a negative pressure that draws air into the lungs. During exhalation, the diaphragm relaxes, allowing air rich in carbon dioxide to be expelled.

In addition to gas exchange, the respiratory system performs several secondary functions. Regulates blood pH by controlling the levels of carbon dioxide. Facilitates vocalization via vibration of the vocal cords. It contributes to the sense of smell by directing air to the olfactory receptors. Provides protective mechanisms such as mucus secretion and the action of cilia to trap and expel foreign particles [18, 19].

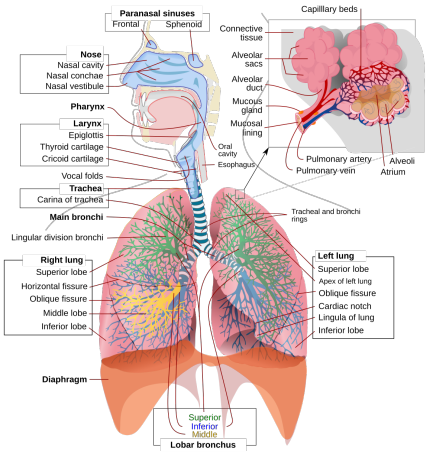

Figure 1.1 – Lung anatomy

1.2.2 Lung Diseases Overview

Lung diseases comprise a diverse group of disorders that alter the structure and function of the respiratory system, limiting the body's ability to maintain effective gas exchange. These diseases can be acute or chronic, infectious or non-infectious and can affect the airways, lung parenchyma, pleura, or the pulmonary vasculature [20]. Together, lung diseases are among the leading causes of morbidity and mortality worldwide, exerting a significant burden on healthcare systems, especially in low income and middle income countries [4, 6, 13].

Some of the most prevalent lung diseases include asthma, chronic obstructive pulmonary disease (COPD), pneumonia, lung cancer, tuberculosis (TB), bronchitis, and pneumothorax [4, 21]. Each of these conditions presents unique pathophysiological mechanisms and clinical manifestations.

Asthma is a chronic inflammatory disorder characterized by reversible airway obstruction, bronchial hyperresponsiveness, and variable respiratory symptoms such as wheezing, coughing, and dyspnea. In contrast, COPD is a progressive and largely irreversible disease that limits airflow due to chronic bronchitis or emphysema. Pneumonia, an acute infection of the lung parenchyma, often results from bacterial, viral, or fungal pathogens and is accompanied by symptoms such as fever, productive cough, and difficulty breathing [22].

Lung cancer, one of the deadliest cancers worldwide, involves the uncontrolled proliferation of abnormal cells within lung tissues. It is strongly linked to smoking and environmental pollutants [23]. Tuberculosis is a contagious bacterial infection affecting the lungs mainly, caused by Mycobacterium tuberculosis. Despite global public health efforts, TB remains endemic in many regions. Bronchitis refers to inflammation of the bronchial tubes and is usually associated with smoking, environmental exposure, or infection. Lastly, pneumothorax, or collapsed lung, occurs when air is leaking into the pleural space, compromising lung expansion and requiring immediate intervention [20, 22, 23].

The causes of lung diseases are multifactorial. Smoking remains the most prominent risk factor, contributing to conditions such as COPD, lung cancer, and bron-

chitis [22]. Air pollution, including exposure to particulates, asbestos, and industrial chemicals, also significantly increases the risk of developing chronic respiratory diseases [24]. Infectious agents such as bacteria, viruses, and fungi are common causes of pneumonia and TB [22]. Genetic factors, including mutations in genes such as CFTR in cystic fibrosis, also play a role. In addition, age, especially advanced age, is associated with a higher susceptibility to many chronic lung diseases [25].
The clinical manifestations of lung diseases vary by condition but often overlap. Common symptoms include shortness of breath, persistent cough, wheezing, chest tightness, fatigue, and, in infectious diseases, fever and chills [22]. Because these symptoms are nonspecific, an accurate and timely diagnosis is essential [4].

1.2.3 Chronic Obstructive Pulmonary Disease (COPD)

Chronic Obstructive Pulmonary Disease is a major global health concern and is one of the leading causes of death worldwide [13]. It is a progressive, incurable respiratory disorder characterized by persistent airflow limitation that is usually preventable and treatable. COPD encompasses two main pathological conditions: chronic bronchitis, which involves long-term inflammation and mucus production in the bronchial tubes, and emphysema, a condition that leads to destruction of the alveolar walls and reduced elastic recoil of the lungs [13].
The primary cause of COPD is long-term exposure to harmful particles and gases, most notably tobacco smoke. Other contributing factors include indoor air pollution, particularly from biomass fuel used for cooking and heating, occupational exposures to dust and chemicals, outdoor air pollution, and a history of respiratory infections during childhood [12]. Although traditionally seen as a disease of older adults and smokers, there is a growing recognition of non-smoking-related COPD, particularly in women and populations exposed to poor indoor air quality [12].
Pathophysiologically, COPD is marked by chronic inflammation in the lungs, leading to structural changes and narrowing of the small airways. The destruction of the alveolar walls results in a decreased surface area for gas exchange, causing hypoxia and hypercapnia in the advanced stages. This limitation of airflow is usually progressive and associated with an abnormal inflammatory response of the lungs to noxious particles or gases [13].
Clinically, COPD presents with symptoms such as chronic cough, sputum production, wheezing, and exertional dyspnea. As the disease progresses, patients may experience reduced tolerance to exercise, frequent respiratory infections, and episodes of acute exacerbations, which significantly worsen quality of life and increase the risk of hospitalization and death [6]. In severe cases, COPD can lead to respiratory failure, pulmonary hypertension, and right heart failure due to lung disease [26].
The diagnosis of COPD typically involves a combination of clinical history, physical examination, and objective tests. The gold standard for diagnosis is spirometry, which reveals a reduced ratio of forced expiratory volume in one second (FEV1) to forced vital capacity (FVC) ratio (FEV1/FVC < 0.70 post-bronchodilator) [12]. Imaging studies such as chest X-rays or CT may reveal hyperinflation or emphy-

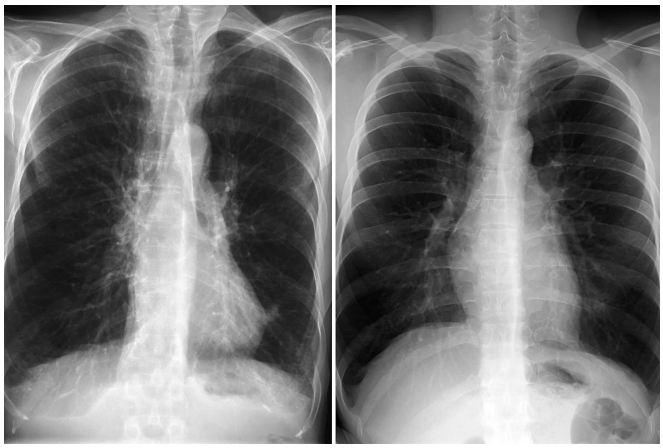

Figure 1.2 – X-rays showing COPD (left) versus normal lungs (right)

sematous changes, while arterial blood gas analysis is used in advanced cases to assess oxygenation and ventilation [4].
Management of COPD requires a multidisciplinary approach. Pharmacological treatments include bronchodilators, inhaled corticosteroids, and combination therapies that relieve symptoms and reduce the frequency of exacerbation. Non-pharmacological interventions, such as lung rehabilitation, nutritional support, and vaccination, are also crucial. In advanced stages, long-term oxygen therapy or surgical options such as lung volume reduction surgery or lung transplantation can be considered[27]. Lifestyle modifications, especially smoking cessation, are fundamental in slowing disease progression [12].
Despite advances in treatment, early detection of COPD remains a challenge. Accurate diagnosis often relies on the expertise of trained medical professionals and can be time-consuming, leading to delays in identification [16]. As a result, many people are diagnosed only in moderate to severe stages, when lung damage is already substantial. This underscores the pressing need for accessible, accurate and automated diagnostic tools to support timely detection and improve the management outcomes for COPD.

1.2.4 Traditional Diagnostic Methods

Traditional diagnostic methods have long served as the cornerstone of identifying and managing lung diseases, offering a combination of clinical judgment, radiographic imaging, functional testing, and laboratory analysis [4]. These approaches remain widely used due to their established clinical value, accessibility, and cost-effectiveness, especially in routine medical practice [6, 28]. The initial assessment typically begins with a thorough medical history and physical examination, which helps physicians gather essential information about symptom onset, severity, risk factors such as smoking or environmental exposures, and co-existing conditions. Although this step is non-invasive and offers quick preliminary information, its subjective nature and reliance on clinician expertise often limit diagnostic precision, especially in cases where respiratory symptoms overlap among different lung

diseases [28].
Imaging techniques, particularly chest X-rays (CXR), are among the most frequently employed tools due to their availability and speed. They are useful for detecting pneumonia, lung masses, or pleural effusion, but can miss subtle or early stage abnormalities and lack the resolution needed for detailed anatomical assessment [21]. In Figure 1.2 you can see healthy and non-healthy lung X-ray.
In contrast, computed tomography offers high-resolution, cross-sectional images that greatly improve diagnostic accuracy in conditions such as lung cancer, interstitial lung disease, or advanced emphysema [21]. However, CT imaging involves higher costs, increased radiation exposure, and limited availability in low-resource settings [28].
Another key diagnostic method is spirometry, which measures lung function and is particularly vital to the diagnosis and monitoring of chronic obstructive pulmonary disease (COPD). Spirometry provides objective metrics such as forced expiratory volume (FEV1) and the FEV1/FVC ratio, making it the gold standard for assessing airflow limitation [13]. However, it requires patient cooperation and is technique dependent, meaning that the results may be unreliable if not performed correctly [29].
In cases where infection is suspected, sputum analysis and microbiological testing play an essential role, especially for the identification of pathogens responsible for pneumonia or tuberculosis. These tests can inform targeted antibiotic therapy, but they are time consuming and susceptible to contamination or sampling errors, which can yield false negative results [13].
Despite their widespread use and historical importance, traditional diagnostic methods are increasingly constrained by their dependence on human expertise, availability of infrastructure, and logistic practicality. For example, the interpretation of imaging results, such as chest radiographs, often varies based on the experience of the radiologist, and the variability between observers remains a notable challenge to ensure consistent diagnostic outcomes. Furthermore, spirometry and other pulmonary function tests require not only calibrated equipment, but also well-trained operators and patient cooperation conditions that are not always guaranteed, particularly in pediatric, geriatric or low-literacy populations. These limitations can lead to delayed or missed diagnoses, misclassification of disease severity, and ultimately suboptimal clinical management.
In addition, as respiratory diseases become more complex due to emerging comorbidities, environmental risk factors, and pathogen variability, traditional approaches may lack the sensitivity and specificity needed for early stage detection or nuanced differentiation between disease types. In tuberculosis endemic regions, for example, overlapping radiographic features with pneumonia or lung cancer may mask the correct diagnosis without confirmatory lab testing, which itself may not be readily available. The need for more accurate, rapid, and scalable diagnostic tools is becoming increasingly urgent, especially in underserved areas where conventional diagnostics are insufficient to meet the population's healthcare demands. These shortcomings have created a critical gap that modern computational methods, particularly artificial intelligence and multimodal data integration, are uniquely positioned to address.

1.2.5 Artificial Intelligence in Medical Diagnosis

Artificial intelligence refers to the simulation of human intelligence processes by machines, especially computer systems. These processes include learning, reasoning, and self-correction. In the context of healthcare, AI aims to help in the analysis, interpretation, and understanding of complex medical data, ultimately supporting clinicians in decision making and improving patient care [30].

The application of AI in medical diagnostics dates back several decades. In the 1970s and 1980s, the first generation of AI systems in medicine emerged in the form of rule-based expert systems. MYCIN, developed at Stanford University, was designed to diagnose bacterial infections and recommend treatments using a structured set of rules. It demonstrated performance levels comparable to human experts [31]. Around the same time, the INTERNIST-I system was developed at the University of Pittsburgh to assist physicians with complex diagnostic cases in internal medicine [32]. Although these early systems showcased the promise of AI, they were heavily dependent on manually encoded knowledge and lacked the flexibility to adapt to new or nuanced cases, limiting their clinical utility.

With the rise of machine learning in the late 1990s and early 2000s, AI in medicine entered a new phase. Instead of relying on predefined rules, machine learning algorithms learned from data, allowing models to identify patterns and make predictions based on large and complex datasets [30]. In recent years, AI applications in diagnostics have expanded rapidly. For example, AI models are now widely used in radiology to detect abnormalities in X-rays and CT scans [30], in dermatology to classify skin lesions with accuracy comparable to that of dermatologists [33], and in pathology to analyze histopathological slides to detect cancerous changes [34]. These technologies have demonstrated the ability to reduce diagnostic errors, speed up workflow, and support earlier detection of diseases, ultimately contributing to better patient outcomes.

Despite these advancements, the integration of AI into clinical practice is not without challenges. Concerns such as the interpretability of AI decisions, the potential for algorithmic bias, and the need for large, diverse, and high quality datasets persist [35]. In addition, regulatory, ethical, and legal frameworks are still evolving to ensure that AI tools are used safely and effectively in real-world healthcare settings. It is also essential that AI complements, rather than replaces, the expertise of medical professionals, fostering collaboration between technology and clinical judgment.

One of the most transformative developments within AI is deep learning, a subfield that mimics the structure and function of the human brain through artificial neural networks. Deep learning models have shown an exceptional ability to process complex data types, such as medical images and sounds, without the need for manual feature extraction. These capabilities are particularly relevant to medical diagnostics, where data are often complex and high-dimensional.

1.2.6 Deep Learning in Healthcare

Machine learning, a subfield of artificial intelligence, refers to algorithms that allow computers to learn patterns from data and make predictions or decisions without

being explicitly programmed. In healthcare, ML has enabled data-driven tools for early diagnosis, personalized treatment, and outcome prediction by analyzing large-scale clinical data. A more advanced branch of ML, known as deep learning, utilizes multilayered artificial neural networks capable of automatically learning complex representations from data. Deep learning excels in processing unstructured data, such as medical images, sounds, or free text notes due to its hierarchical learning structure, which mimics aspects of human cognition [5, 36].

The conceptual roots of deep learning trace back to the 1950s and 1980s with early work on perceptrons and backpropagation, but deep learning did not flourish until the 2010s due to advances in computational power, big data availability, and algorithmic improvements. Convolutional Neural Networks (CNNs) have become a cornerstone of medical image analysis, effectively identifying features such as tumors, lesions, or fluid accumulations in modalities such as X-rays, magnetic resonance imaging, and CT scans [5, 37]. Likewise, DL models have been used to interpret biomedical audio signals such as lung and heart sounds by converting them into spectrograms and applying CNNs or other architectures [38].

Deep learning systems in healthcare are typically divided into single-modality and multimodal approaches. Single-modality models are trained using one type of data [16]. For example, a model might use only chest X-ray images to detect pneumonia or only lung sound recordings to identify respiratory abnormalities. The advantage of single-modality models lies in their relative simplicity, lower computational cost, and reduced data integration challenges. They are easier to train, interpret, and deploy in clinical environments where access to diverse data sources may be limited. However, their disadvantage is that they are entirely based on a single information stream, which can lack context or be ambiguous in certain cases, especially for complex diseases where visual, acoustic, and clinical symptoms all contribute to diagnosis.

In contrast, multimodal deep learning combines multiple data types, such as imaging, audio, and text, to enhance diagnostic performance. By integrating heterogeneous inputs, multimodal models can take advantage of complementary features, leading to more robust, accurate, and context-aware predictions [16]. This advantage becomes crucial in real-world scenarios, where diseases such as COPD may not present clear patterns in a single modality. Multimodal systems also offer greater flexibility and generalizability. However, their disadvantages include increased model complexity, higher training costs, and the need for carefully synchronized datasets. In addition, acquiring high-quality multimodal data from the same patient is often challenging in practice, especially in low-resource settings.

Despite these challenges, the integration of deep learning into healthcare continues to expand. Single-modality systems remain valuable in settings with limited data or technology infrastructure, while multimodal systems are increasingly being explored in academic research and specialized clinics due to their superior diagnostic capabilities. This progression underscores the need for intelligent decision support systems that can adapt to both the limitations and opportunities presented by diverse clinical environments.

1.2.7 Multimodal Learning Approaches

Multimodal learning refers to the integration of information from multiple data sources or modalities such as images, audio, text, or structured clinical data to improve the performance of machine learning models. In healthcare, this approach aims to mimic the way clinicians combine various forms of information during diagnosis, treatment planning, and patient monitoring. For example, a physician can simultaneously consider a patient's chest X-ray, auscultation sounds, symptoms, and medical history of a patient to reach a diagnosis. Similarly, multimodal deep learning systems are designed to capture and process this complex, complementary information to improve predictive accuracy, robustness, and generalization [16, 39].

The success of multimodal learning depends on the way information is combined from each modality. There are three primary strategies for data fusion in multimodal systems: early fusion, late fusion, and intermediate (or joint) fusion. Early fusion, also known as feature-level fusion, involves combining raw or preprocessed input features from each modality before feeding them into a unified model. This method allows for direct learning of cross-modal interactions, but requires careful alignment of inputs in time or space, which can be a challenge when the data types differ significantly. Late fusion, also called decision-level fusion, involves training separate models on each modality and then combining their output, often through averaging, voting, or a meta-classifier. Although this method offers greater flexibility and interpretability, it may miss subtle interactions between modalities that could enhance predictive power. Intermediate fusion, or joint representation learning, lies between these two extremes. Process each modality independently in the early layers, but merge them in deeper layers of the network, allowing specialized feature extraction and cross-modal learning [16].

Each fusion strategy comes with trade-offs. Early fusion benefits from tightly coupled data and is often more computationally efficient but can suffer from modality imbalance or missing data. Late fusion is robust to missing modalities and modular in design, but may underperform when strong intermodal relationships exist. Intermediate fusion seeks a balance, offering richer representations at the cost of increased model complexity and training difficulty. Choosing the optimal fusion strategy depends on the task, the characteristics of the dataset, and the computational resources. In the context of medical diagnosis, where multimodal data, such as chest X-rays and lung sounds, can be unpaired or asynchronous, late fusion is often preferred for its practicality and ability to handle heterogeneous inputs independently.

In general, multimodal learning represents a powerful paradigm shift in the design of medical AI systems. Using complementary signals from different data sources, these models can provide a more holistic and nuanced understanding of a patient's condition, paving the way for more accurate, timely, and context-sensitive diagnostics. This approach forms the methodological foundation of the system proposed in this thesis, which combines visual features from chest X-rays and acoustic features from lung sounds to detect and classify chronic respiratory diseases, particularly COPD.

1.3 Thesis Organization

The rest of this thesis is organized as follows. Section 2 defines the problem statement by highlighting the global burden of lung diseases and highlighting diagnostic challenges in resource-limited settings. Section 3 presents a comprehensive review of the literature, covering traditional diagnostic methods for lung diseases, deep learning applications in medical imaging and lung sound analysis, and recent multimodal approaches for lung disease detection. Section 4 describes the methodology, including the datasets used, preprocessing techniques, model design for both image and audio modalities, fusion strategy, training procedures, hyperparameter tuning, and evaluation metrics. Section 5 discusses the experiments and results, detailing the experimental environment, model performance, and comparative analysis between single-modality and multimodal models, as well as benchmarking against existing studies. Finally, Section 6 concludes the thesis by summarizing the key findings and suggesting directions for future research.

Chapter 2

Problem Statement

2.1 Global Burden of Lung Diseases

Lung diseases, particularly chronic respiratory conditions, represent a substantial global health burden, affecting hundreds of millions of people around the world [40]. Among these conditions, chronic obstructive pulmonary disease is among the most prevalent and deadly. According to the World Health Organization, COPD is currently the third leading cause of death worldwide, responsible for approximately 3.23 million deaths [6]. The disease is characterized by persistent respiratory symptoms and airflow limitation, often resulting from significant exposure to noxious particles or gases, such as tobacco smoke or air pollution [23]. Its progressive and irreversible nature contributes to long-term disability, reduced quality of life, and increased healthcare utilization.

The prevalence of COPD is particularly concerning in low and middle-income countries (LMICs), where more 90% of COPD-related deaths occur. Factors such as indoor air pollution from biomass fuel use, occupational dust exposure, and inadequate medical infrastructure exacerbate the incidence and impact of the disease. Although preventable and treatable, COPD is often undiagnosed until it reaches advanced stages. This underdiagnosis is due to the gradual onset of symptoms, limited awareness, and the lack of routine screening mechanisms in many regions [6, 13, 41].

In addition to COPD, other lung diseases such as asthma, lung cancer, pulmonary fibrosis, and tuberculosis also contribute significantly to global morbidity and mortality. The early detection and differentiation of these conditions is critical for effective treatment and disease management [6, 13, 41, 42]. However, reliance on specialized diagnostic tools such as spirometry, chest X-rays, and CT scans, combined with the need for expert interpretation, creates barriers to timely diagnosis, particularly in underserved areas [6]. As a result, the burden of lung diseases continues to increase, calling for innovative, accessible, and scalable diagnostic solutions.

Despite significant advances in global health policy, lung diseases remain underrepresented in research funding and public health initiatives compared to other noncommunicable diseases like cardiovascular conditions or diabetes. This underinvestment has contributed to a persistent gap in early diagnosis, public education,

and the development of innovative technologies for lung health. The stigma surrounding chronic lung diseases, particularly in socioeconomically disadvantaged communities, also deters individuals from seeking early diagnosis and treatment. Additionally, comorbidities such as cardiovascular disease, diabetes, and mental health disorders complicate clinical presentation and can divert attention from respiratory assessment, delaying intervention and leading to worse long-term outcomes.

Moreover, climate change and urbanization have begun to play a growing role in exacerbating lung disease prevalence worldwide. Rising levels of ambient air pollution, including fine particulate matter (PM2.5), nitrogen dioxide, and ozone, are linked to increased incidence of respiratory conditions, especially among children and the elderly. In many industrialized and rapidly urbanizing regions, exposure to environmental pollutants acts as a silent yet potent contributor to disease burden. This evolving landscape not only heightens the urgency for preventive strategies but also necessitates advanced diagnostic tools that can account for complex environmental and biological interactions in lung disease pathophysiology. Together, these emerging factors demand a rethinking of how lung diseases are monitored, detected, and managed in both clinical and public health contexts.

2.2 Diagnostic Challenges in Resource-Limited Settings

Accurate diagnosis of lung diseases such as COPD, pneumonia, and tuberculosis remains a major challenge in resource-limited settings, where access to advanced healthcare infrastructure and trained medical personnel is often scarce [43]. In many rural and underserved regions, diagnostic tools such as spirometry, which is essential to confirm airflow obstruction in COPD, are unavailable or underutilized due to cost, maintenance issues, and a lack of adequately trained personnel [13, 45]. Similarly, radiographic imaging, such as chest X-rays or CT scans, may not be easily accessible due to the high cost of equipment, the need for experienced radiologists, and the absence of digital health systems for image sharing and remote interpretation [44, 46, 47, 48].

These limitations result in delayed or inaccurate diagnosis, contributing to poor health outcomes, increased transmission of infectious respiratory diseases, and inefficient allocation of healthcare resources. Even when basic diagnostic tools are available, variability in clinical experience and subjective interpretation can lead to misdiagnosis [13, 23, 44].

The growing global emphasis on health equity highlights the need for low-cost, automated, and reliable diagnostic systems that can operate effectively without relying heavily on specialist input. Recent advances in artificial intelligence and deep learning offer promising avenues to address these gaps. Specifically, multimodal approaches that integrate audio signals and image-based data have the potential to provide accurate and scalable diagnostic support [16]. Using widely available technologies, such as mobile phones or portable devices, these systems can empower frontline healthcare workers and bridge the diagnostic divide in low-

resource environments.
In summary, the dual challenge of the global burden of lung diseases and diagnostic limitations in resource-constrained settings underscores the need for intelligent and accessible diagnostic tools. This thesis addresses this gap by proposing a multimodal deep learning-based medical decision support system for the early and accurate classification of lung conditions, particularly targeting COPD detection using audio and image data.

Chapter 3

Literature Review

3.1 Deep Learning in Medical Imaging for Lung Disease

In recent years, deep learning has made a big impact in medical diagnosis, especially for lung diseases. A comprehensive survey conducted by Kieu et al. (2020) meticulously examines 98 articles published between 2016 and 2020, shedding light on the utilization of various deep learning algorithms for the detection and classification of lung diseases from medical images [4]. The survey emphasizes that Convolutional Neural Networks are the most widely used algorithm in this field because of their exceptional capacity to automate feature extraction from several image modalities, including CT and X-rays [4, 23, 30, 48, 49]. In fact, CNN extracted features were used in 79% of the works surveyed, with an additional 13% employing a combination of CNN extracted features plus some other features. Moreover, transfer learning shows up as a powerful method that helps transfer knowledge from trained models to improve generalization and classification accuracy [4, 50].

Although alternative algorithms, such Deep Belief Networks and Bag of Words models, have been investigated, CNNs continue to be the preferred method because of their effectiveness and resilience in identifying complex patterns in medical images [4]. The survey also explores the ensemble of classifiers, emphasizing how combining several classifiers trained on various feature sets can increase detection accuracy [4]. The survey notes that although ensemble techniques have great potential, there are still issues that need to be investigated and improved upon, such as the high correlation of errors among base classifiers [4, 50].

Apart from the progress in algorithms, the survey offers understanding of the wide range of lung conditions that deep learning models aim to address. Deep learning algorithms demonstrate adaptability in diagnosing a range of pulmonary illnesses, including lung cancer, COVID-19, pneumonia, and tuberculosis. To address the particular difficulties connected to each disease kind, the survey emphasizes the need for more thorough databases and research initiatives. The survey indicates research gaps in managing data imbalance, big image sizes, and errors in ensemble techniques, despite advancements in algorithms and disease-specific investigations. Future research on developing more efficient medical decision support systems for

lung disease diagnosis will be motivated by addressing these obstacles[4].
A notable recent contribution to deep learning-based lung disease diagnosis is the LDDNet framework proposed by Podder et al., which extends the DenseNet201 architecture for the classification of infectious lung diseases such as COVID-19 and pneumonia using both chest CT scans and X-ray images [51]. The authors introduced an optimized architecture by integrating additional layers 2D global average pooling, dense and dropout layers, and batch normalization tailored to improve feature generalization and reduce overfitting. They further fine-tuned hyperparameters such as learning rate, batch size, and dropout rate across three diverse datasets, including a CT dataset of 1043 images and two CXR datasets comprising 5935 and 5002 images, respectively. Experimental results demonstrated that LDDNet outperformed state-of-the-art models like ResNet152V2 and XceptionNet, especially when trained using the Nadam optimizer. On CT scans, the model achieved a COVID-19 classification accuracy of 99.36% with an F1-score of 99%, while on the imbalanced X-ray dataset, it reached 99.55% accuracy, 100% precision, and 85% F1-score. These results highlight LDDNet's ability to maintain high precision even on imbalanced datasets, showcasing its potential for real-world deployment. The study underscores the importance of careful architectural tuning and optimizer selection in enhancing diagnostic performance for multiple image modalities in lung disease detection [51].
Another recent contribution to pulmonary disease detection using deep learning is PulmoNet, a deep convolutional neural network (DCNN) model proposed by Abdulahi et al. (2024), designed to identify COVID-19, bacterial pneumonia (BP), and viral pneumonia (VP) from radiographic images [52]. Developed in response to the urgent need for fast and accurate diagnosis particularly in low-resource settings the model was trained on a multiclass dataset consisting of 16,435 chest X-ray and CT images across four categories: healthy (10,325), COVID-19 (3,749), BP (883), and VP (1,478). PulmoNet incorporates image augmentation to improve generalizability and was benchmarked against traditional texture descriptor-based methods. The model achieved impressive detection accuracies across the four classes, including 95.4% for COVID-19, 99.4% for BP, and 98.3% for VP, while maintaining efficient training and inference times of approximately 60 and 50 seconds, respectively. These results suggest that PulmoNet offers a reliable and time-efficient diagnostic tool, especially for deployment in developing countries where computational resources and access to expert radiologists may be limited. The study highlights the continued evolution of CNN-based methods in handling multi-class lung disease classification tasks and the practical implications of deploying such models in real-world healthcare scenarios [52].
In 2020, Krit Shriporn et al. utilized three popular deep learning models: MobileNet, Densenet-121, and Resnet-50 to diagnose lung diseases from chest X-rays [5]. They achieved an impressive 98.88% accuracy with the Densenet-121 model using the Mish activation function and the Nadam optimizer. The study also examined the performance of other models such as MobileNet and Resnet-50, which achieved an accuracy of 93.28% and 97.59%, respectively. However, the study acknowledges hardware limitations, highlighting the need for future research to improve the performance of CNN models in medical image analysis [5].

In their 2022 publication, Yazan Al-Issa et al evaluated the performance of four deep learning models: VGG16, DenseNet201, DarkNet-19, and XceptionNet in differentiating between normal cases, pneumonia, COVID-19 and cases of lung opacities on chest radiographs [53]. Among these models, XceptionNet achieved the highest accuracy of 94.775%, and DarkNet-19 demonstrated efficient convergence and real-time detection capabilities. Ensemble features combining multiple models produced the highest accuracy of 97.79%. However, the study identified challenges such as forecasting time and resource requirements. Future studies could explore optimized models or ensemble methods to address these limitations and facilitate accurate and timely diagnosis, especially in settings with limited healthcare resources[53].
In their 2023 study titled 'Lung Diseases Detection Using Various Deep Learning Algorithms', researchers aimed to identify lung diseases like pneumonia, tuberculosis, and lung cancer using deep learning and X-ray images [3]. They experimented with different deep learning models: Sequential, Functional, and Transfer models, training them on existing datasets to assess their effectiveness. Their objective was to demonstrate that their models could outperform existing methods. Their findings revealed that the Sequential model performed the best, achieving high accuracy in identifying pneumonia and tuberculosis. Specifically, it achieved an F1 score of 98.55%, an accuracy of 98.43%, and a recall of 96.33% for pneumonia. For tuberculosis, it achieved an F1 score of 97.99%, an accuracy of 99.4%, and a recall of 98.88%. Regarding cancer detection, the Functional model emerged as the top performer, boasting an accuracy of 99.9% and a specificity of 99.89%. Notably, this model required fewer parameters and computational resources compared to pre-existing models. In summary, their research suggests that these new deep learning models show promise in accurately and efficiently diagnosing lung diseases. However, future work could explore optimization techniques, such as varying optimizers and learning rates, and implementing more extensive data augmentation to improve diagnostic outcomes for lung diseases[3].

3.2 Deep Learning for Lung Sound Analysis in Lung Disease

Lung sound analysis has emerged as a promising non-invasive tool for respiratory disease diagnosis, particularly for conditions such as chronic obstructive pulmonary disease (COPD), asthma, bronchitis, pneumonia, and interstitial lung diseases. Traditional auscultation methods, although widely used, are highly dependent on clinician expertise and often suffer from subjectivity and variability [16, 28, 48]. In response, recent advances in deep learning (DL) have enabled the automatic and accurate interpretation of respiratory sounds by learning complex patterns from large datasets [39, 54]. This section reviews key deep learning approaches for lung sound classification and their contributions to lung disease diagnosis.
Early studies focused primarily on handcrafted features such as Mel Frequency Cepstral Coefficients (MFCCs), Zero Crossing Rate (ZCR), and Short-Time Energy (STE), which were then classified using shallow classifiers like Support Vector

Machines (SVM) and Random Forests (RF) [13, 45]. Although these approaches provided a baseline, they were limited in capturing hierarchical representations and contextual acoustic patterns. The advent of deep learning, especially Convolutional Neural Networks (CNNs), has transformed the field by enabling end-to-end learning directly from spectrogram representations of lung sounds [6, 37, 56, 65]. One of the most widely adopted methods involves converting raw audio into 2D mel spectrograms or log-mel spectrograms, which are then fed into CNN architectures such as ResNet50, VGG19, or DenseNet201 for feature extraction and classification [21, 44]. For example, a 2021 study used ResNet50 on mel spectrograms for binary COPD versus healthy classification and achieved an AUC of 0.996 and a F1 score of 0.98, showcasing the potential of pretrained image-based models when transferred to audio domains [55]. Similarly, VGG19 has been fine-tuned on lung sound spectrograms for multiclass disease classification, demonstrating superior sensitivity to subtle adventitious sounds such as wheezes and crackles [57].

In addition to image-based models, recent research has explored the diagnostic potential of respiratory sounds. Nadkarni et al. (2024) introduced AFEN (Audio Feature Ensemble Learning), a hybrid framework that combines Convolutional Neural Networks (CNNs) and XGBoost in an ensemble setting to classify respiratory diseases from audio recordings [58]. The model utilizes a carefully curated set of audio features, such as melfrequency cepstral coefficients and other spectral descriptors, which are separately fed into two classifiers: a multifeature CNN and an XGBoost model. The final prediction is generated through soft voting, enabling the model to benefit from both deep feature extraction and gradient boosted decision trees. Evaluated on a dataset of 920 respiratory sound samples and enhanced through data augmentation, AFEN demonstrated strong generalizability and reduced training time by 60% compared to traditional CNN only architectures. The model achieved state-of-the-art performance in terms of precision and recall, showcasing the potential of ensemble methods in capturing the temporal and frequency specific nuances of respiratory pathologies. This work highlights the effectiveness of hybrid architectures in audio-based disease detection and opens up new avenues for scalable and non-invasive diagnostic tools [58].

In a recent advancement targeting pediatric populations, TaghiBeyglou et al. proposed TRespNET, a novel dual-route deep learning architecture designed to identify adventitious respiratory sounds in children using auscultation audio signals [59]. Unlike most prior work focused on adult cohorts, this study leverages the SPRSound dataset comprising recordings from 288 children aged 1 month to 18 years. The authors explored multiple architectures traditional CNNs, CNNs integrated with transformer encoders, and Vision Transformers before developing TRespNET, which effectively fuses spectrotemporal representations from Mel spectrograms and raw audio time series. Further performance improvements were achieved by incorporating handcrafted acoustic features alongside learned representations. The proposed model attained a specificity of 0.98, a sensitivity of 0.84, and a harmonic mean score of 0.90, outperforming existing approaches for pediatric respiratory event detection. This study not only underscores the feasibility of automated auscultation for early disease diagnosis in children but also highlights the benefit of multimodal fusion (raw signal and spectral data) in boosting model

accuracy and generalizability [59].
In response to the need for accessible, real time respiratory disease diagnosis, Abadade et al. introduced a novel application of Tiny Machine Learning (TinyML) for classifying lung diseases from auscultation sounds using low-power, edge deployable models [60]. The study addresses limitations of traditional stethoscopes and cloud-based deep learning systems such as latency, internet dependency, and data privacy by implementing lightweight models that can be integrated into digital stethoscopes. Using bandpass filtered lung sounds and MFCC based feature extraction, the authors trained and quantized three models: a custom CNN, a custom LSTM, and an Edge Impulse CNN. The custom CNN demonstrated the best tradeoff between performance and computational efficiency, achieving 96% accuracy and 97% precision, recall, and F1-score, while maintaining low resource usage. This study demonstrates the viability of real time AI-assisted auscultation in low resource and remote settings, highlighting TinyML's potential in democratizing healthcare access and supporting early detection of respiratory illnesses without reliance on cloud infrastructure [60].
Other works have leveraged Transformer-based architectures like the Audio Spectrogram Transformer (AST) or PaSST, which incorporate self-attention mechanisms to model temporal dependencies in lung sounds [61]. These models have shown competitive performance in classifying lung sounds in noisy real-world datasets. Additionally, PANNs (Pretrained Audio Neural Networks) such as CNN14 have gained popularity for transfer learning on respiratory sound tasks due to their large-scale training on general-purpose audio datasets like AudioSet [62].
In terms of dataset usage, several studies used publicly available datasets such as the ICBHI 2017 Challenge or the COSWARA dataset [23]. The ICBHI dataset, which includes labeled recordings from 126 patients covering normal and abnormal lung sounds, has been instrumental in benchmarking deep learning models for disease-specific analysis. Some approaches have used segmentation to isolate individual respiratory cycles (inspiration or expiration) before feature extraction, while others applied data augmentation techniques such as noise injection, time shifting, and pitch alteration to enhance model robustness in low-resource settings [49].
Recent works have also begun exploring multichannel inputs, combining multiple respiratory cycles or sensor arrays to better represent spatial and temporal variability. For example, a 2023 study applied a dual-branch CNN that separately processed inspiratory and expiratory segments, leading to improved sensitivity in detecting fine-grained pathological features [63].
Despite promising results, challenges remain. A major issue is the limited availability of high-quality, labeled audio datasets, especially for underrepresented conditions like bronchiectasis or pulmonary fibrosis [30]. Furthermore, inter-patient variability, environmental noise, and differences in recording hardware can significantly affect generalization. Researchers have responded by incorporating denoising autoencoders, attention modules, and ensemble models to improve robustness and interpretability [55, 57].
In conclusion, deep learning methods have significantly advanced the field of lung sound analysis, enabling automated, scalable, and accurate detection of various

lung diseases. With the continued growth of open datasets and the refinement of multimodal fusion techniques, audio-based diagnostics are expected to play a pivotal role in early screening and remote monitoring, particularly in low-resource or rural healthcare settings where imaging tools are less accessible.

3.3 Multimodal Approaches for Lung Disease Detection

Recent advances in artificial intelligence and medical imaging have facilitated the emergence of multimodal deep learning frameworks for the detection of lung disease, leveraging the complementary strengths of various data types, such as images, audio signals, and textual records [37, 64, 65]. Multimodal models integrate heterogeneous modalities to enhance diagnostic accuracy, robustness, and generalization across diverse patient populations and disease types. This section reviews state of the art multimodal approaches that combine image (CT, X-ray, MRI), audio (cough or lung sounds), and text (clinical records or diagnostic reports) data for lung disease classification and prediction. And the recent studies wich used multimodal approach are summarized in Table 3.1

One of the earliest studies in this field explored COVID-19 detection using multiple imaging modalities. A 2024 study employed transfer learning with VGG16 on CT and X-ray data, achieving image-specific accuracies up to 98% for ultrasound, although no fusion was applied across modalities [14]. In contrast, subsequent work introduced intermediate fusion techniques to integrate chest X-rays with breathing sound recordings using an InceptionV3-MLP hybrid model, reaching 99.66% accuracy on X-ray inputs and 80% on audio [65]. These studies demonstrated the potential of combining visual and acoustic features for improving COVID-19 classification performance.

More recent work expanded multimodal inputs to include broader disease types. For instance, a 2023 study used CT, cough sounds, and X-ray images with a DenseNet201-based CNN ensemble and reported strong multi-class performance (Accuracy: 96.67%, AUC: 99.43%) across COVID-19, tuberculosis, pneumonia, lung cancer, and other conditions [44]. Similarly, another study that combined X-ray images with clinical blood test data used LSTM and CNNs with intermediate and late fusion techniques, demonstrating a 2.9% improvement in classification accuracy over unimodal counterparts [16]. This suggests that integrating temporal textual features with visual data enables richer clinical representation for diseases like pneumonia and chronic bronchitis.

Table 3.1 – Summary of multimodal deep learning approaches for lung disease diagnosis

Study	Modality	Datasets Used	Neural Network Architecture	Key Results
Kumar et al., 2023[16]	img + text	Manually collected (289 patients, future 65k records)	DenseNet121, ResNet50, LSTM, SVM fusion	Intermediate fusion improved accuracy by 2.9%
Malik et al., 2024[21]	img + audio	24 public datasets (CXR, Cough sound, RSNA, etc.)	CNN + BANL, RBAP, MWDG	Achieved SOTA performance across diseases
Kumar et al., 2024[42]	img + text	3,256 patient records (India)	CNN, Denoising Autoencoder, Cross-Modal Transformer	Addressed data imbalance, high accuracy for TB classification
Abhishek et al., 2024[45]	img + audio	1,979 respiratory sound recordings	Hybrid CNN-GRU model	High accuracy in common respiratory diseases, overfitting risk
Sangeetha et al., 2024[37]	img + text	TCIA, TCGA	MFDNN, CNN, DNN, Intermediate Fusion	92.5% accuracy in lung cancer classification
Varunkumar et al., 2024[64]	img + img	RIDER Lung CT, Kaggle X-ray	CNN with dilated convolutions, multimodal fusion	Limited dataset diversity, generalizability issues
Hamdi et al., 2021[66]	img + text	Public IPF dataset (33,026 CT + 1,549 records)	EfficientNet, DenseNet, LSTM, Attention Fusion	Multimodal integration improved prediction accuracy
Kumar et al., 2024[6]	img + audio + text	AIIMS, Raipur (CT, X-ray, cough, lung sounds)	EfficientNet, RNN, U-Net, OpenL3, RVFL neuro-fuzzy model	COPD prediction using multimodal fusion
Deng et al., 2024[47]	img + text	East China hospitals, Kaggle COVID-19 CT	CNN + Contrastive Learning + Early Fusion	Contrastive learning improved performance, Grad-CAM interpretation
Adeshina et al., 2022[28]	img + audio	COVIDx, SARS-CoV-2 CT-scan dataset	CNN, ResNet, DenseNet, XResNet, Self-Attention	91.07% accuracy, effective multimodal cascaded approach
Thandu et al., 2024[35]	img + audio	Chest X-ray (COVID-19 Radiography) + COUGHVID	DSPANN + Blockchain-based Privacy (ECHFA)	Data quality challenges, complex attention mechanisms
Liu et al., 2024[46]	img + text	4 hospitals (China), Chest CT, Clinical Features	DenseNet-201 + DNNs + Early Fusion	Outperformed junior radiologists, 11 key clinical features identified
Farhan et al., 2023[41]	img + img	CXRTD, PCXRA, CCSC, NIH Chest X-ray	CNN, LSTM, SVM, Decision Tree	Improved severity grading performance
Lay et al., 2022[50]	img + text	Shenzhen, Montgomery X-ray Dataset	EfficientNet, XGBoost, U-Net	AUC improved by 0.0213 over unimodal models
Mayya et al., 2021[54]	img + text	COVID-19 Chest X-ray, RSNA Pneumonia Dataset	ResNet18, NLP, Grad-CAM, Deep NN Ensemble	X-ray + diagnosis reports enhanced accuracy
Wu et al., 2021[39]	img + text	TCIA (422 NSCLC patients)	3D-ResNet, Clinical Embedding Layer, Fusion	Improved survival prediction using multimodal fusion

Thandu and Gera proposed a privacy-aware multimodal framework that integrates chest X-ray images and breathing sounds to detect COVID-19, emphasiz-

ing both diagnostic performance and data security [35]. Their system employs advanced preprocessing techniques, including tri-Gaussian filtering and Mel Frequency Cepstral Coefficients (MFCCs) for audio, and segmentation with feature extraction for images. A novel deformable fusion module, along with the DSPANN (Dual Sampling dilated Pre-activation residual Attention convolution Neural Network) architecture, enables effective integration of visual and acoustic features. To address data privacy, a blockchain-enabled encrypted federated learning mechanism is implemented, ensuring secure global model updates. The model achieved an accuracy of 98%, with a sensitivity of 98% and specificity of 97.02%, demonstrating that high-performing multimodal diagnostics can coexist with privacy-preserving architectures in real-world healthcare applications.

Among the most advanced efforts are those focusing on COPD, a particularly underrepresented disease in multimodal literature. The FuzzyGuard framework incorporated CT images, X-rays, and cough or lung sounds using a neuro-fuzzy architecture with EfficientNet, ResNet, LSTM, and VGGish modules [6]. The model achieved an accuracy of 98.65% and an F1-score of 98.34% for binary COPD detection, demonstrating the feasibility of high-precision multimodal COPD diagnosis. Another 2024 framework combined CT image texture features and MFCC-derived lung sound features with classifiers such as k-NN and SVM. Their ensemble achieved an overall accuracy of 97.5%, confirming the added value of harmonizing acoustic and imaging data. These works underscore how late fusion techniques and carefully engineered feature pipelines can support early detection and classification of complex chronic diseases.

In the broader landscape, multimodal systems have also targeted diseases such as asthma, tuberculosis, bronchiectasis, and idiopathic pulmonary fibrosis (IPF). For example, the MARL architecture integrated CT scans and patient history to predict disease progression through regression (R^2 score: 91%) and multi-class classification (accuracy: 92%) [66]. Additionally, the MFDNN model for lung cancer combined imaging, genomic, and clinical data to improve diagnostic precision using deep neural network layers for joint feature learning [37].

A notable recent contribution by Kumar et al. at 2023 introduced a novel multimodal framework for early diagnosis and classification of COPD by integrating CT scan images and lung sound recordings [13]. The framework addresses diagnostic challenges in underserved populations, such as tribal communities, by enabling remote screening and classification of COPD and related respiratory conditions including tuberculosis, bronchitis, and lower/upper respiratory tract infections. It extracts diverse features, such as texture and histogram intensity from CT scans and MFCCs from audio samples, and employs unsupervised feature selection with ensemble learning for robust classification. Trained on data from AIIMS Raipur and open source datasets, the framework achieved 97.5% accuracy using the multimodal fusion approach, outperforming single modality counterparts (98% for CT model and 95.3% for cough based model). This work not only demonstrates the practical feasibility of multimodal COPD diagnosis, but also emphasizes the value of integrating acoustic and imaging modalities in resource limited and remote healthcare contexts.

Deng et al. introduced a novel training strategy called Diagnostic Report Su-

pervised Contrastive Learning (DRSCL) that leverages multimodal information specifically, chest CT scans and associated diagnostic reports to enhance the generalization and interpretability of COVID-19 detection models [47]. In this approach, diagnostic textual data is used during pretraining to supervise image feature learning, while only the pretrained image encoder is transferred for inference, thus decoupling inference from text input. To address the issue of repetitive report content often found in medical text, the authors refine the contrastive loss by merging identical image or text features, which stabilizes the pretraining process. Furthermore, they implement a hierarchical fine tuning strategy for performance evaluation. A custom image–text dataset was constructed using clinical data from hospitals in East China for pretraining, while fine-tuning was performed on a public SARS-CoV-2 dataset. The proposed DRSCL framework led to notable improvements in classification accuracy and model robustness across multiple image encoders. Moreover, Grad-CAM visualizations confirmed that DRSCL improves model interpretability by guiding attention toward clinically relevant regions, underscoring the potential of multimodal supervision in medical imaging tasks [47].
Liu et al. introduced MI-DenseCFNet, a robust multimodal diagnostic framework designed to differentiate between Staphylococcus aureus pneumonia (SAP) and Aspergillus pneumonia (ASP) by integrating high resolution chest CT images with structured clinical features [46]. The model extends the DenseNet architecture with a clinical feature fusion pipeline, leveraging a random forest-based dichotomous classifier to identify and incorporate eleven discriminative clinical variables. Evaluated on data from 60 patients across four tertiary hospitals, MI-DenseCFNet achieved impressive diagnostic performance, with an AUC of 0.92 on internal validation and 0.83 on external datasets outperforming junior and mid-level radiologists in diagnostic accuracy (78% vs. 40–75%). Notably, the system delivered diagnostic predictions in just 10.24 seconds for a batch of 20 cases, underscoring its practical utility in real-time clinical workflows. Beyond its classification efficacy, the model offers interpretability through its selection of key clinical markers and demonstrates the potential of multimodal fusion to bridge expertise gaps in resource limited or non-specialist healthcare settings. This work exemplifies the strength of combining imaging data with curated clinical inputs to enhance diagnostic precision in complex pulmonary infections [46].
Kumar and Sharma proposed an improved deep learning framework that leverages a transformer-based architecture for the multimodal diagnosis of lung diseases, focusing on tuberculosis detection [42]. Recognizing that accurate clinical decision making often relies on a combination of imaging and non imaging data, the authors developed a cross-attention transformer module to effectively integrate chest X-ray images with structured clinical information. This architecture enables the model to learn rich inter modal interactions by selectively attending to relevant features across the two data types. Their approach was validated on a newly constructed multimodal medical dataset for tuberculosis diagnosis, demonstrating a superior classification accuracy of 95%, significantly outperforming conventional feature-level and decision level fusion baselines. The results highlight the advantage of attention based mechanisms in capturing complex, complementary information from heterogeneous modalities. This study underscores the growing potential of trans-

former models in medical multimodal learning and affirms the feasibility of unified architectures for enhancing diagnostic reliability in clinical practice especially for diseases like tuberculosis that demand nuanced interpretation across modalities [42].

These studies highlight multiple trends. First, intermediate and late fusion techniques dominate the multimodal landscape, often outperforming early fusion due to their capacity to preserve modality specific feature hierarchies before integration. Second, disease-specific frameworks (COPD or COVID-19 models) tend to outperform generic multi-disease models, especially when fused representations are optimized for domain-specific pathophysiology. Third, data diversity that combines manually collected clinical data with public datasets remains a cornerstone in improving the generalizability of the model.

In conclusion, multimodal deep learning models are proving to be powerful tools to improve lung disease detection. Although substantial progress has been made in the detection of COVID-19 and pneumonia, multimodal COPD detection remains relatively less explored, indicating a need for more targeted research. The incorporation of acoustic features, clinical history, and imaging data when harmonized using appropriate fusion strategies offers a promising path toward robust, interpretable, and early-stage diagnosis in real-world healthcare settings.

Chapter 4

Methodology

This section outlines the methodological pipeline adopted for developing the proposed deep learning-based medical decision support system for the diagnosis of lung disease. The workflow comprises several key stages, including data acquisition, preprocessing of chest X-ray and lung sound modalities, model architecture design, training and evaluation of unimodal and multimodal models, and final classification. Figure 4.1 provides an overview of the complete process.

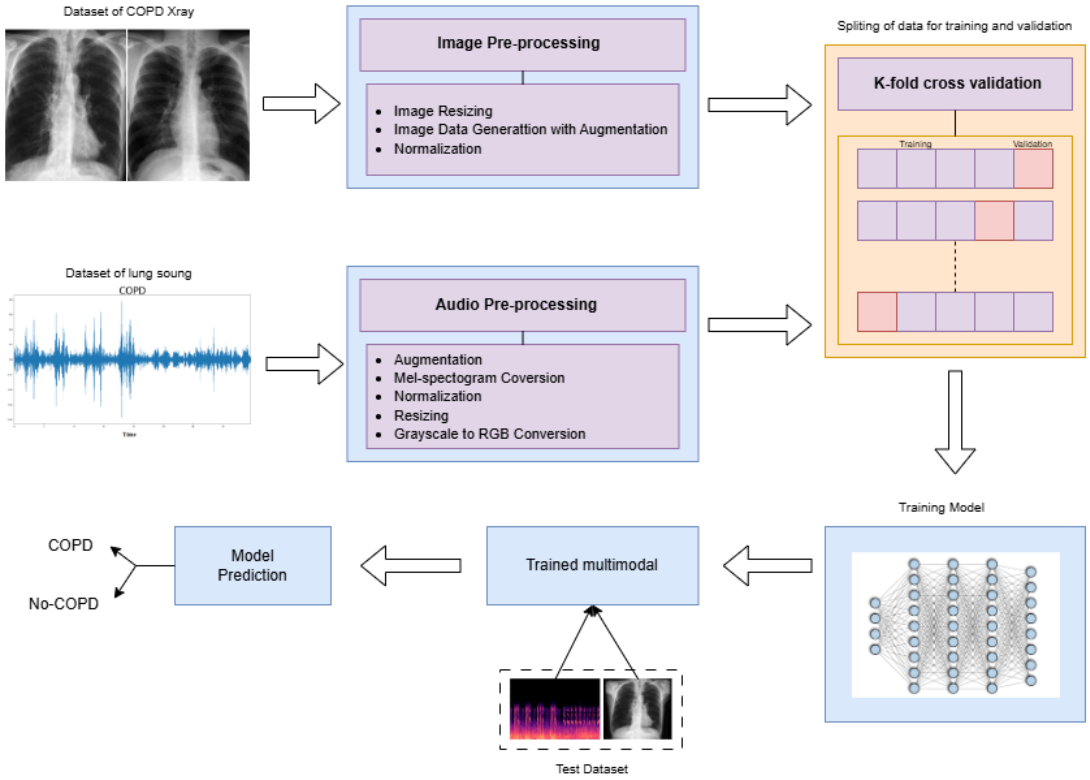

Figure 4.1 – Overview of the proposed methodology pipeline including preprocessing, model training, and classification.

4.1 Dataset

This study uses multiple publicly available datasets to construct a binary classification system for the detection of chronic obstructive pulmonary disease (COPD) from two complementary modalities: lung sounds and chest X-ray images. For lung sound modality, three datasets were used. First, the RespiratoryDatabase@TR introduced by Altan et al. [67]. provides 12-channel respiratory sound recordings for each patient, allowing multichannel analysis. It includes samples across five COPD severity levels (COPD0-COPD4), with each recording lasting at least 17 seconds, making it suitable for studying nuanced respiratory patterns. Second, the ICBHI 2017 Respiratory Sound Database, compiled for the International Conference on Biomedical and Health Informatics challenge, contains 920 annotated recordings from 126 patients, totaling 5.5 hours and 6,898 respiratory cycles. These were labeled by experts for adventitious sounds such as crackles and wheezes. Only samples labeled explicitly as "COPD" and "Healthy" were retained to match the binary classification goal, while ambiguous or multi-condition recordings were excluded. Third, the Lung Sound Dataset published on Mendeley includes recordings from Fortis Hospital, India, collected using an electronic stethoscope and annotated as normal or abnormal. The dataset comprises 10-second WAV files sampled at 44.1 kHz. Only the "normal" and "COPD" (abnormal respiratory disease) samples were retained, and recordings with other pulmonary diseases were discarded. For the chest X-ray modality, three major datasets were selected. The NIH ChestX-ray14 dataset, consisting of over 112,000 frontal-view images, was filtered to retain only those labeled with emphysema or chronic bronchitis, and “no finding” for the healthy class. The PadChest dataset, which includes over 160,000 annotated X-rays collected in Spain, was mined for frontal-view images labeled with COPD-related conditions (e.g., emphysema, chronic airway obstruction), while images labeled with unrelated or multi-label diagnoses were excluded. Lastly, from the Pneumonia Chest X-ray Dataset, only clean "normal" images were extracted to support the non-COPD class. Images showing signs of pneumonia, tuberculosis, fibrosis, or other lung pathologies were excluded across all datasets to ensure a clear distinction between the two target classes. All chest X-rays were reviewed to ensure consistency in image view or frontal and diagnostic relevance. An important consideration during dataset selection was ensuring diversity in recording environments and patient demographics to improve the model’s generalizability. The audio datasets span multiple countries and healthcare settings, including hospital and outpatient environments, which helped capture real-world variability in lung sound recordings. Likewise, the X-ray datasets included scans from both adult male and female patients of various age groups and clinical backgrounds, contributing to a representative sampling of COPD-related radiological features. Since the datasets were collected independently, most of the training data consisted of unpaired samples, which reflects the practical challenges in multimodal healthcare applications where simultaneous audio and imaging data are rarely available. However, the inclusion of a small but valuable paired test set from Hassaan et al. (2024) [21] enabled performance evaluation in a more realistic clinical scenario, where a model must integrate information from multiple sources for a single

patient. This blend of large-scale unpaired training data and paired test data supports robust model learning while preserving clinical relevance. To evaluate the final model in a real-world multimodal scenario, a test set compiled by Hassaan et al. (2024), which contains paired lung sounds and chest X-rays from the same patients with binary COPD labels, was used [21]. This combination of curated datasets ensures a clean, well-labeled input space for training and evaluating both unimodal and multimodal models in a binary classification setting.

4.2 Data Preprocessing

The preprocessing stage is a crucial step in preparing the raw lung sound and chest X-ray image data for effective training of deep learning models aimed at Chronic Obstructive Pulmonary Disease (COPD) detection. Given the multimodal nature of the dataset, distinct preprocessing pipelines were designed for the audio and image modalities, each tailored to optimize feature extraction and model performance.

Raw audio data consisted of lung sound recordings obtained from multiple publicly available datasets. These recordings were initially loaded using the librosa library, which enables efficient handling of audio data with variable sampling rates and formats, while preserving sound fidelity. Since deep learning models require fixed-size input representations, the continuous audio signals were converted into two-dimensional Mel spectrograms, a time-frequency representation emphasizing frequency bands most relevant to human auditory perception. This was achieved by computing Mel-scaled spectrograms with 128 Mel bands and converting them to the decibel scale for enhanced dynamic range visualization. To increase the robustness of the model and to simulate real-world variability in lung sounds, several augmentation techniques were applied to the audio signals. Random Gaussian noise was added to the original signal to mimic environmental or recording device noise. Signals were randomly shifted forward or backward in time to simulate variability in breath onset timing. The audio was speeded up or slowed down by factors of 1.2 and 0.8 respectively, to replicate differences in respiratory rate. Each augmented signal underwent the same Mel spectrogram transformation. Resulting spectrograms were normalized to a range between -1 and 1 to conform with the input expectations of the convolutional neural network (CNN) architecture, employed for feature extraction. Since ResNet models accept three-channel inputs akin to RGB images, the single-channel grayscale Mel spectrograms were duplicated across three channels to meet this requirement. Additionally, spectrogram images were resized to 224x224 pixels to align with the input dimensions of the pretrained ResNet model. The binary class labels were encoded as '1' for COPD and '0' for healthy cases. To handle class imbalance inherent in the datasets, class weights were computed and integrated into the training process, ensuring the model did not bias towards the majority class.

The chest X-ray images were subjected to a separate preprocessing pipeline tailored to exploit visual features indicative of lung pathology. Initially, images were loaded and resized to 224x224 pixels, matching the ResNet input size. The dataset was augmented using Keras' ImageDataGenerator to artificially enlarge the training

set and improve model generalization. Basic augmentation techniques included. These augmentations introduce variability that the model may encounter in real clinical images, such as patient positioning differences or varying imaging conditions. Each image was preprocessed using the ResNet-specific normalization function, which scales pixel values to the range and distribution used during the original ImageNet model training. This step is critical for leveraging transfer learning from the pretrained ResNet weights. To further enhance the representational power of the model, a squeeze-and-excitation (SE) block was introduced after the convolutional layers of the CNN architecture. The SE block adaptively recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels, thereby allowing the model to emphasize relevant features while suppressing less useful ones.
Both audio and image datasets were stratified and split into training and validation sets. This comprehensive preprocessing strategy ensured that both lung sounds and chest X-ray images were converted into standardized, normalized, and augmented inputs optimized for convolutional neural networks. By tailoring augmentation and normalization techniques to each modality, and incorporating attention mechanisms for the image data, the preprocessing pipeline laid a solid foundation for effective multimodal feature learning and improved COPD detection performance.

4.3 Model Design

The proposed model architecture is built upon a multimodal learning framework that independently processes audio and image data through modality-specific deep learning backbones and integrates their outputs using a late fusion strategy. This design enables the system to leverage complementary features from chest X-ray images and lung sound recordings for improved lung disease classification. The individual unimodal networks were initially trained and evaluated separately to identify the most effective model for each modality. ResNet50 was selected as the optimal architecture for the audio stream, while InceptionV3 demonstrated superior performance in the image modality. These backbones serve as feature extractors in the final multimodal pipeline. The fusion module combines the learned representations from both modalities before classification, allowing the model to make more informed predictions by jointly analyzing visual and acoustic cues related to pulmonary conditions.

4.3.1 Image-based model

For image-based lung disease classification, five pre-trained convolutional neural networks were considered: VGG19, ResNet50, DenseNet201, EfficientNetB0, and InceptionV3. These models were selected based on their proven effectiveness in medical image analysis tasks (Hassaan et al., 2024). Each network was fine-tuned on the chest X-ray dataset using transfer learning. InceptionV3 outperformed the other architectures in terms of validation accuracy, F1 score, and area under the curve (AUC), and was thus chosen as the image modality backbone. The final

image-based model replaces the top classification layers of DenseNet201 with a global average pooling layer followed by a fully connected layer and a sigmoid output unit for binary classification. The model efficiently captures complex radiological patterns associated with COPD and other pulmonary abnormalities.

4.3.2 Audio-based model

For the audio modality, mel spectrogram representations of lung sounds were used as input to deep convolutional networks. The same set of backbone models VGG19, ResNet50, DenseNet201, EfficientNetB0, and InceptionV3 were evaluated. Each model was fine-tuned on the spectrogram data, with class balancing techniques and callback functions (such as early stopping and model checkpointing) applied to prevent overfitting. Among these, ResNet50 achieved the highest classification performance, demonstrating its ability to extract discriminative temporal-frequency features from respiratory audio. The final model consists of the ResNet50 base with frozen convolutional layers and a custom classification head tailored for binary output.

4.3.3 Fusion strategy

To effectively combine features from both modalities, a late fusion strategy was employed. In this setup, the feature vectors extracted from the penultimate layers of the ResNet50 (audio) and InceptionV3 (image) models are concatenated into a single joint representation. This fused feature vector is then passed through one or more fully connected layers, followed by a sigmoid activation function for final classification. The late fusion approach was chosen for its flexibility and robustness, allowing each modality to independently learn specialized features before integration. This design also facilitates modular experimentation, enabling future incorporation of additional modalities such as clinical text data without altering the modality-specific branches. The fusion model aims to capitalize on the complementary nature of visual and auditory data, improving the overall diagnostic accuracy for lung disease detection.

4.4 Model Training

The training phase of the multimodal deep learning framework focused on enabling the model to generalize well across both audio and image inputs, ensuring reliable detection of COPD in diverse patient data. After selecting the best-performing architectures for each modality ResNet50 for lung sound spectrograms and InceptionV3 for chest X-ray images the training process was carefully designed with attention to optimization strategy, data balance, and performance monitoring. The dataset was first partitioned into training, validation, and testing subsets using a stratified shuffle split, preserving the proportion of healthy and COPD cases across all subsets. This was essential to avoid training bias and ensure that model evaluation reflected realistic class distributions.
To address the class imbalance problem, which is common in medical datasets,

class weighting was applied during training. This technique adjusts the loss contribution of each class inversely proportional to its frequency, encouraging the model to treat minority class with greater importance, thus improving sensitivity and F1 score. The model was trained using the Adam optimizer, selected for its efficiency in handling sparse gradients and adaptive learning rates. The learning rate was initialized at a moderate value and adjusted experimentally to achieve optimal convergence. The binary cross-entropy loss function was used to measure classification error, as it is well-suited for binary classification tasks such as this one.

To prevent overfitting, two key techniques were incorporated into the training routine. Firstly in the model used early stopping function. By this function training was monitored using validation loss. If the validation loss did not improve over a defined number of epochs (patience threshold), training was halted automatically. This avoided unnecessary training cycles and reduced the risk of memorizing training data.

A checkpointing mechanism called Model Checkpointing was used to save the model weights corresponding to the best validation performance. This ensured that even if subsequent epochs led to overfitting or divergence, the best version of the model was retained. Batch size, number of epochs, and other training parameters were chosen based on hardware limitations and empirical tuning. Regularization methods such as dropout were also applied to the fully connected layers following fusion, further enhancing generalization capability. During training, performance metrics including accuracy, precision, recall, F1 score, and AUC (area under the ROC curve) were recorded for both the validation set and final test set. These metrics provided a comprehensive view of the model's behavior and helped determine the most effective training configuration.

An essential aspect of this training strategy was the integration of fine-tuning for the pre-trained base models. Fine-tuning involves unfreezing some of the upper layers of the pre-trained convolutional networks (ResNet50 and InceptionV3) so that they can be retrained on the target medical dataset. This enables the models to adapt high-level features specifically to COPD-related patterns while still leveraging their general visual and auditory representations learned from large-scale datasets like ImageNet. Additionally, several measures were taken to improve the model's ability to generalize. For instance, early stopping a form of regularization was used to monitor the validation loss during training. If no improvement was observed over a specified number of epochs (the "patience" threshold), training was stopped to avoid overfitting, a situation where the model memorizes training data instead of learning generalizable patterns. Another key regularization method was the use of class weights, which helped correct the bias introduced by an imbalanced dataset. By assigning higher weights to the minority class (COPD), the model was encouraged to improve recall and reduce false negatives, which is critical in medical diagnosis. These components collectively ensured that the trained model did not only perform well on training data but also maintained robust accuracy and sensitivity on unseen test samples.

Overall, this structured and carefully monitored training process contributed significantly to the robustness of the final multimodal model, allowing it to leverage

complementary strengths of lung sound and image features for accurate COPD classification.

4.5 Evaluation

To assess the effectiveness of both the unimodal and multimodal models for COPD classification, a comprehensive evaluation strategy was employed using multiple performance metrics. The objective was not only to achieve high predictive accuracy but also to ensure the robustness, generalizability, and clinical applicability of the proposed diagnostic framework. Given the medical significance of the problem, careful consideration was given to choosing evaluation metrics that reflect both model precision and the potential impact of misclassifications on patient outcomes.

The performance of the models was primarily evaluated using several standard metrics: accuracy, precision, recall, F1 score, and AUC-ROC (Area Under the Receiver Operating Characteristic Curve). Accuracy measures the overall correctness of the model by calculating the ratio of correctly predicted samples to the total number of predictions. However, in medical applications where class imbalance is common, accuracy alone may provide a misleading picture. Therefore, additional metrics were emphasized. Precision indicates the proportion of true positive predictions among all positive predictions, which reflects how reliable the model is when it identifies a patient as having COPD. Recall (or sensitivity) captures the model's ability to correctly identify all true COPD cases, a crucial aspect in medical diagnostics where missing a case can have serious consequences. The F1 score, as the harmonic mean of precision and recall, provides a balanced measure of the model's performance, especially when the cost of false positives and false negatives is unequal. Finally, the AUC-ROC evaluates the trade-off between sensitivity and specificity across different threshold settings, serving as a robust measure of the model's ability to distinguish between classes.

All performance metrics were computed using the held-out test set, which was stratified and kept entirely separate from training and validation phases to prevent data leakage and ensure an unbiased performance estimate. After training, predictions were generated using the final model checkpoints—those corresponding to the best validation F1 score or AUC. Validation loss and accuracy were continuously monitored during training to assess model convergence and detect signs of overfitting or underfitting. These values were used to guide early stopping and model checkpointing, ensuring that the selected model exhibited stable and optimal generalization on unseen data. Moreover, confusion matrices were constructed for each model to provide a detailed overview of classification performance, revealing the distribution of true positives, true negatives, false positives, and false negatives.

In addition to numerical results, visualizations such as ROC curves and metric progression plots (accuracy, loss, AUC, and F1 over epochs) were employed to analyze model convergence behavior and generalization. These plots provided critical insight into training dynamics and model stability across epochs. Furthermore, model comparisons were made not only between different unimodal architectures

(e.g., ResNet50 for lung sounds vs. EfficientNet for X-ray images) but also against the proposed multimodal fusion approach. This comparative evaluation enabled a holistic understanding of each model's strengths and limitations, highlighting the added value of integrating complementary modalities. The multimodal model's superior performance in terms of AUC and F1 confirmed the hypothesis that combining audio and image features can enhance diagnostic accuracy for complex respiratory conditions like COPD.

4.6 Experimental Environment

The entire pipeline is developed using Python, with the deep learning models implemented in either PyTorch or TensorFlow frameworks. Librosa is employed for audio processing tasks such as loading, feature extraction, and spectrogram generation. Scikit-learn is used for computing evaluation metrics and generating confusion matrices and ROC curves. The models are trained and tested on Google Colab or local servers.

All experiments and model training in this research were conducted on a high-performance desktop computer with the following specifications:

Operating System: Microsoft Windows 10 Enterprise (Version 10.0.19045, Build 19045). Processor: 12th Gen Intel® Core™ i9-12900K, 3.2 GHz, 16 cores, 24 threads. Memory (RAM): 128 GB DDR4. Motherboard: ASUS PRIME Z690-P D4. Storage: Multiple volumes, with a dedicated 19.0 GB page file on volume D. System Type: x64-based PC, UEFI boot mode.

The system was optimized for deep learning workloads and capable of handling memory-intensive tasks, including training deep neural networks with large datasets. GPU specifications are omitted here as the system's GPU details were not provided, but they may be added if relevant.

Chapter 5

Experiments and Results

This chapter presents the experimental setup, model performance, and comparative results obtained during the evaluation phase of the developed models for COPD detection. The study used multiple deep learning architectures across single modal and multimodal approaches. The performance was assessed using several key evaluation metrics, including Accuracy, Precision, Recall, F1-score, Area Under the ROC Curve (AUC), and Loss.

5.1 Model Performance

To establish baseline results, five convolutional neural network (CNN) architectures: ResNet-50, VGG-19, InceptionV3, EfficientNetB0, and DenseNet201, were trained separately on each data modality: lung sounds and chest X-ray images. Each model was fine-tuned on modality-specific inputs to determine its ability to distinguish between COPD and healthy cases.

For the lung sound modality, mel spectrograms were generated from raw audio recordings and served as the input. Among all models, DenseNet201 demonstrated the highest overall classification performance, achieving 100% training accuracy, 97.78% validation accuracy, and a 97% test accuracy, with a near-perfect F1-score of 0.9734 and AUC of 0.9953. These metrics suggest excellent learning capacity and generalization, with DenseNet201 capturing subtle acoustic patterns unique to COPD.

InceptionV3 and VGG-19 also performed remarkably well, each achieving 96% test accuracy, with InceptionV3 slightly outperforming VGG-19 in terms of F1-score (0.9651 vs. 0.9630) and test AUC (0.9894 vs. 0.9911). Interestingly, InceptionV3 achieved this with only 50 training epochs, suggesting that it quickly converged to an optimal solution. This indicates that InceptionV3 may offer computational efficiency without sacrificing performance, making it ideal for scenarios with limited training time or resources.

On the other hand, ResNet-50 achieved solid performance, with a test accuracy of 90% and a test AUC of 0.9697, indicating reliable performance despite being slightly outperformed by deeper models. Conversely, EfficientNetB0 lagged behind all other models with a test accuracy of 58%, test AUC of 0.6893, and F1-score of 0.7322, suggesting that its lighter architecture may not be sufficiently expressive

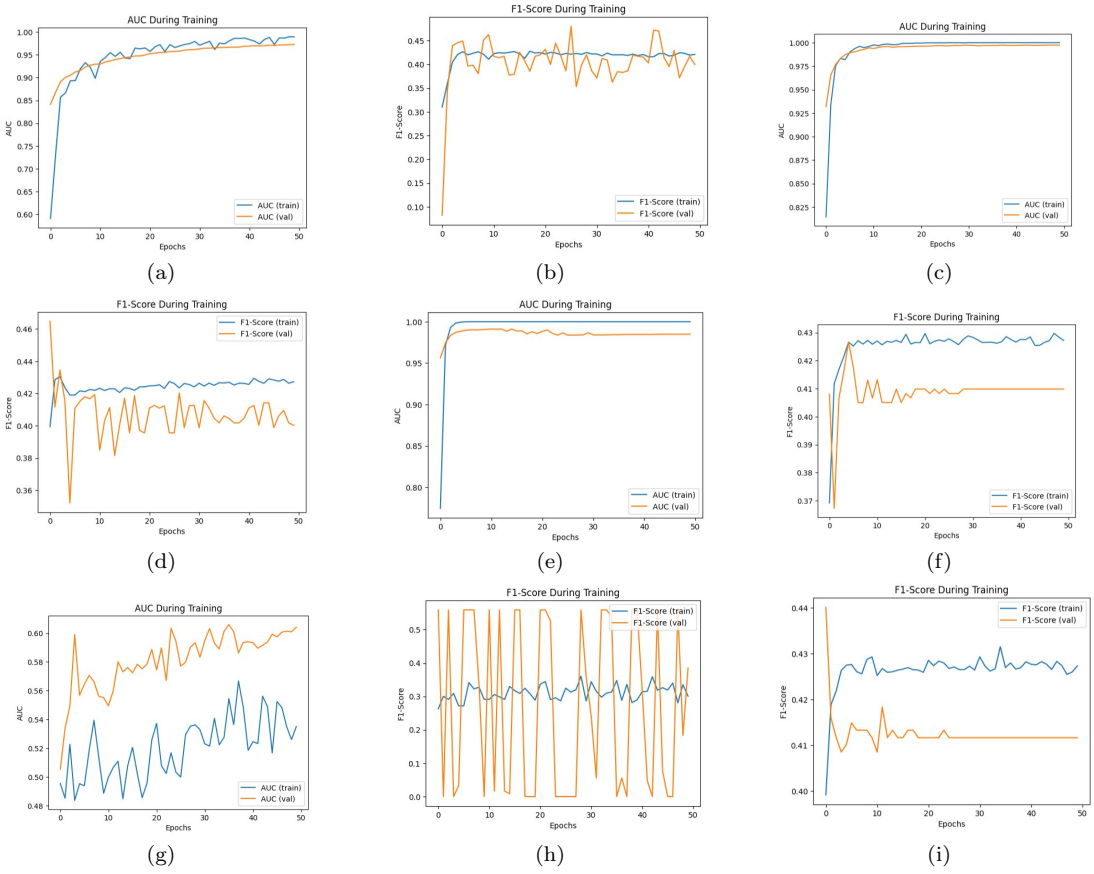

Figure 5.1 – Model performance plots: (a) ResNet50 AUC, (b) ResNet50 F1, (c) VGG19 AUC, (d) VGG19 F1, (e) InceptionV3 AUC, (f) InceptionV3 F1, (g) EfficientNetB0 AUC, (h) EfficientNetB0 F1, (i) DenseNet201 F1.

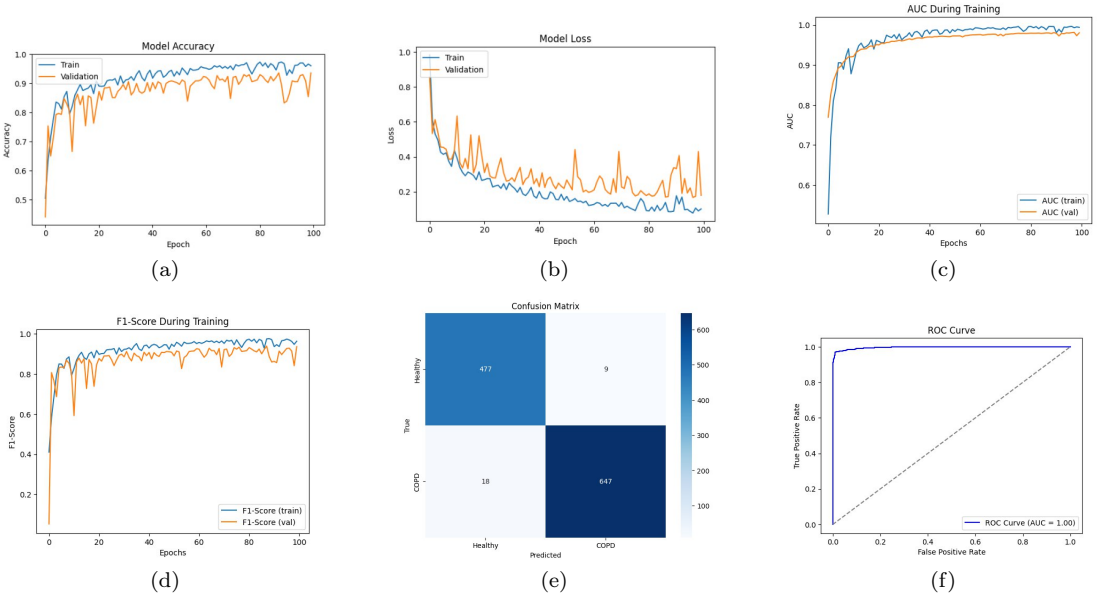

Figure 5.2 – Multimodal model performance. (a) Accuracy, (b) Loss, (c) AUC score, (d) F1-score, (e) Confusion Matrix, (f) ROC curve.

for capturing the complex acoustic signatures in lung sounds.
For the chest X-ray modality, the same set of CNN models was evaluated. Here, InceptionV3 emerged as the top performer again, achieving a test accuracy of 74%, F1-score of 0.7213, and the highest AUC of 0.8475 among image-based models. These results indicate InceptionV3's strength in identifying discriminative visual features associated with COPD-related structural changes in lung tissue.
DenseNet201 and ResNet-50 followed closely, with test AUCs of 0.8132 and 0.8241, respectively. DenseNet201 showed slightly better F1-score (0.7194) than ResNet-50 (0.6923), but ResNet had a more stable validation curve during training, making it more consistent overall. VGG-19 and EfficientNetB0 performed comparatively lower, with VGG-19 showing the weakest results (F1-score 0.6331 and AUC 0.7763), potentially due to its simpler architecture and deeper network overfitting.

Table 5.1 – Performance Metrics of Audio-Based Models (Lung Sounds)

Model	**Epoch**	**Train Acc**	**Val Acc**	**Test Acc**	**Test AUC**	**Test F1**
ResNet-50	50	0.9480	0.9053	0.90	0.9697	0.9140
VGG-19	48	1.0000	0.9735	0.96	0.9911	0.9630
InceptionV3	10	1.0000	0.9907	0.96	0.9894	0.9651
EfficientNetB0	48	0.4926	0.5758	0.58	0.6893	0.7322
DenseNet201	31	1.0000	0.9778	0.97	0.9953	0.9734

Table 5.2 – Performance metrics of fine-tuned image-based models using chest X-ray images.

Model	**Epoch**	**Train Acc**	**Val Acc**	**Test Acc**	**Test AUC**	**Test F1**
ResNet-50	34	0.7534	0.6801	0.70	0.8241	0.6923
VGG-19	50	0.7057	0.6335	0.65	0.7763	0.6331
InceptionV3	31	0.7887	0.7245	0.74	0.8475	0.7213
EfficientNetB0	48	0.7256	0.6543	0.67	0.8007	0.6823
DenseNet201	31	0.7456	0.6847	0.71	0.8132	0.7194

The detailed evaluation metrics for all audio and image-based models are summarized in Table 5.1 and Table 5.2, respectively. Additionally, Figure 5.1 visualizes the performance in terms of F1-score and AUC, allowing easy comparison across architectures and modalities.
Based on the above evaluations, the models selected for integration into the multimodal deep learning system were ResNet-50 for lung sounds and InceptionV3 for chest X-rays. This decision was not only based on absolute performance metrics but also took into account model stability, convergence speed, and potential for feature complementarity in a multimodal fusion setting.
The multimodal model, created using a late fusion strategy where features from both modalities were concatenated and passed through a shared classification head, demonstrated significantly improved performance. It achieved a test accuracy of 95%, an F1-score of 0.9696, and an AUC of 0.9765—surpassing the best single-modality models. This confirms the hypothesis that integrating heterogeneous features from complementary modalities provides richer information and enhances

diagnostic accuracy.
Figure 5.2 provides a graphical comparison of the multimodal model performance, further emphasizing the benefit of leveraging both structural (X-ray) and functional (lung sound) modalities for comprehensive COPD detection.
Despite the challenges of training deep learning models on relatively small datasets, especially for the image modality the results affirm the viability of constructing robust diagnostic models from limited but high-quality data. The chest X-ray dataset, though modest in size compared to public CT databases, still yielded meaningful results when paired with an appropriate architecture like InceptionV3. This underscores the importance of model selection and careful fine-tuning, particularly in domains where large-scale annotated data is scarce.
Crucially, the successful development of the multimodal model using only chest X-rays and lung sounds also demonstrates the potential of cost-effective and accessible data sources in clinical decision support systems. Unlike CT scans, which are expensive and less available in many parts of the world, especially in low- and middle-income countries, lung sounds and X-ray imaging represent tools already integrated into routine clinical practice. The effectiveness of combining these two modalities suggests a pathway to scalable, real-world diagnostic solutions that do not compromise performance for affordability. With this foundation established, the next section explores how this multimodal approach compares to single-modality systems and existing literature.

5.2 Comparative Analysis

5.2.1 Singlmodal vs. Multimodal

To illustrate the performance gain from combining modalities, a direct comparison was conducted between the best-performing single-modality models: ResNet50 for lung sounds and InceptionV3 for chest X-rays and the proposed multimodal architecture based on late fusion. The metrics considered for this comparison include Accuracy, F1-score, and Area Under the ROC Curve (AUC), providing a comprehensive understanding of each model's diagnostic reliability and class discrimination capabilities.
As shown in Table 5.3, the multimodal model clearly outperformed both unimodal counterparts across all evaluation metrics. The accuracy increased to 95%, compared to 90% for the lung sound-based model and 74% for the X-ray-based model. Similarly, the F1-score improved to 96.96%, reflecting a more balanced prediction across both COPD and non-COPD classes. Notably, this score represents an enhancement of more than 5 percentage points over the lung sound model and nearly 25 percentage points over the X-ray model, emphasizing the synergistic value of integrating modalities.
In terms of AUC, which reflects the model's overall ability to distinguish between classes regardless of threshold, the multimodal system achieved 97.65%, significantly surpassing the unimodal results (96.97% for lung sounds and 84.75% for X-rays). This consistent improvement across metrics indicates that the combined model not only makes more accurate predictions but is also better calibrated and

confident in its classification decisions.
The observed performance gain can be attributed to the complementary nature of the two modalities. Lung sounds provide functional information about airflow and breathing anomalies, while chest X-rays offer structural insights into lung morphology and tissue patterns. When used independently, each modality contributes partial evidence; however, when combined, they provide a holistic representation of the patient's pulmonary condition. This cross-modal integration allows the model to capture both latent and visible manifestations of COPD, some of which may remain undetected in unimodal analysis.
Moreover, late fusion enables each modality-specific network to learn features independently and at its own optimal capacity before merging. This flexible architecture prevents one modality from dominating the feature space, promoting balanced learning and robust generalization. The results reaffirm the hypothesis that leveraging multimodal data significantly enhances classification performance in complex clinical tasks like COPD detection.

Table 5.3 – Comparison of performance metrics between single-modality, multimodal models, and prior work.

Methods	**Accuracy**	**F1 Score**	**AUC**
Single X-ray modality (InceptionV3 base)	74%	72.13%	84.75%
Single Lung Sound modality (ResNet50 base)	90%	91.40%	96.97%
Multimodal with Late Fusion	95%	96.96%	97.65%
Malik et al., 2024	–	–	93%

5.2.2 Comparison with exicting work

In order to contextualize the achieved results, the performance of the proposed multimodal model was benchmarked against recent state of the art methods in the literature. One notable reference is the study by Malik et al. (2024) [21], which utilized a multimodal deep learning approach for multiclass classification of various respiratory conditions. Their model integrated CT scans, chest X-ray images, and lung sound recordings, resulting in a reported AUC of 93%. While Malik et al.'s approach demonstrated competitive performance, it relied on a richer and more resource-intensive input setup.
In contrast, the model proposed in this study focuses exclusively on binary classification (COPD vs. Healthy) using only two modalities: lung sounds and chest X-ray images. Despite the simpler and more cost-effective data sources, our multimodal model achieved an AUC of 97.65%, surpassing Malik et al.'s result by 4.65 percentage points. This is a significant outcome, as it highlights the potential of combining widely accessible and less expensive diagnostic tools to reach or even exceed the performance of systems that rely on high-resolution imaging such as CT scans, which may not be readily available in low-resource or rural settings.
Moreover, the training of chest X-ray models poses additional challenges due to the need for large, diverse datasets and greater computational resources. In our case, the dataset used for X-ray images was relatively limited, and yet the model

achieved a respectable test accuracy of 74% and AUC of 84.75% in the single-modality setup. This emphasizes the robustness of the multimodal fusion strategy, which compensates for the limitations of individual modalities by leveraging complementary information.

These findings collectively suggest that cost-effective and accessible diagnostic inputs, when properly combined through a well-designed deep learning architecture, can deliver high diagnostic performance that rivals or surpasses more complex systems. This contributes to the growing body of evidence supporting the viability of AI-powered screening tools in resource-constrained healthcare environments and underscores the importance of modality selection and efficient model design in real-world medical AI applications.

This comparison, illustrated in Table 5.3, highlights the competitive advantage of the proposed system. By effectively combining audio and image modalities through a late fusion mechanism, the model not only meets but exceeds existing performance benchmarks, suggesting a promising direction for future research in multimodal medical diagnostics.

Chapter 6

Conclusions and future work

6.1 Conclusions

This study addressed the critical challenge of improving COPD detection by designing a multimodal deep learning framework that integrates both lung sound recordings and chest X-ray images. Leveraging the complementary diagnostic strengths of audio and visual modalities, the proposed system aimed to enhance classification accuracy and reduce the reliance on specialist expertise, particularly in settings with limited medical infrastructure. Extensive experiments were conducted using state of the art convolutional neural network architectures, where five pretrained models ResNet50, DenseNet201, EfficientNetB0, VGG19, and InceptionV3 were fine-tuned and evaluated for their effectiveness in single-modality classification tasks.

Among the tested models, ResNet50 achieved the best performance for audio-based COPD detection, while InceptionV3 yielded the highest accuracy for image-based classification. The fusion of these two models in the multimodal framework demonstrated significant performance gains over their single-modality counterparts, as evidenced by improvements in F1 score, Area Under the Curve, and overall diagnostic accuracy. These findings underscore the value of integrating heterogeneous data sources to capture diverse and discriminative patterns associated with COPD pathology.

The study further highlighted the feasibility and potential of deploying such a system in real-world clinical environments, especially in resource-constrained regions where access to pulmonologists, spirometry, or radiologists is limited. By utilizing low-cost digital stethoscopes and portable X-ray devices, the proposed solution can facilitate early screening and triage, ultimately reducing disease progression and improving patient outcomes.

6.2 Future work

Despite the promising results achieved by the proposed multimodal model for COPD detection, several limitations should be noted. First, the dataset used was relatively small and lacked demographic diversity, which may limit the generalizability of the results to broader populations. Additionally, the lung sounds

and chest X-rays were unpaired collected from different individuals preventing the model from learning deeper cross-modal relationships. The model was also restricted to binary classification, whereas real-world clinical diagnosis involves distinguishing between multiple respiratory diseases. Furthermore, while the use of deep learning improved performance, the lack of model interpretability poses challenges for clinical trust and decision-making. While the results of this study are promising, several avenues remain open for further research and system enhancement. Future work could explore the integration of additional data types, such as text-based clinical symptoms. The incorporation of multimodal textual and temporal data may also enable the development of more context-aware and personalized diagnostic systems. The performance and reliability of deep learning models are heavily dependent on the diversity and quantity of training data. The current dataset, although sufficient for initial experimentation, should be expanded to include larger, more diverse, and clinically verified samples. Collaborations with hospitals, research institutions, and public health repositories will be pursued to build a more comprehensive dataset that includes patients from different age groups, ethnicities, and stages of disease progression. While this study focused primarily on binary classification, future work will aim to extend the model to multi-class settings that can distinguish between various pulmonary conditions such as asthma, pneumonia, tuberculosis, interstitial lung disease, and lung cancer. A more nuanced classification system could significantly enhance clinical decision-making and support differential diagnosis in complex respiratory cases. Although a late fusion strategy was employed in this study, future research may explore intermediate or hybrid fusion methods, including attention-based mechanisms, graph neural networks, or transformer-based architectures to better capture cross-modal dependencies and improve interpretability. A crucial next step involves the deployment of the trained model in real clinical environments for field testing. This includes developing a user-friendly mobile or web-based interface, integrating it with digital stethoscopes and imaging systems, and validating its performance on unseen patient populations in rural clinics or community health centers.

In conclusion, this research contributes a robust and scalable approach to medical decision support for lung disease diagnosis, demonstrating that multimodal deep learning can play a transformative role in the early detection and management of COPD.

Bibliography

[1] Y. Sugandi, I. Soesanti, and H. A. Nugroho, "A Systematic Literature Review of Convolutional Neural Network Architecture for Lung Disease Detection," in Proc. 2023 International Conference on Information and Communications Technology (ICOIACT), 2023, pp. 230-235. DOI: 10.1109/ICOIACT59844.2023.10455864.

[2] S. Seed, "Respiratory System", 2024.

[3] P. P. Jasmine, K. Kotecha, G. Rajini, K. Hariharan, K. Raj, K. Ram, V. Indragandhi, V. Subramaniyaswamy, and S. Pandya, "Lung Diseases Detection Using Various Deep Learning Algorithms," Journal of Healthcare Engineering, vol. 2023, pp. 1-13, 2023. DOI: 10.1155/2023/3563696.

[4] S. T. H. Kieu, A. Bade, M. H. A. Hijazi, and H. Kolivand, "A Survey of Deep Learning for Lung Disease Detection on Medical Images: State-of-the-Art, Taxonomy, Issues and Future Directions," Journal of Imaging, vol. 6, no. 12, p. 131, 2020. DOI: 10.3390/jimaging6120131.

[5] K. Sriporn, C.-F. Tsai, C.-E. Tsai, and P. Wang, "Analyzing Lung Disease Using Highly Effective Deep Learning Techniques," Healthcare, vol. 8, 2020.

[6] S. Kumar, A. V. Shvetsov, and S. H. Alsamhi, "FuzzyGuard: A Novel Multimodal Neuro-Fuzzy Framework for COPD Early Diagnosis," IEEE Internet of Things Journal, 2024. DOI: 10.1109/JIOT.2024.3467176.

[7] GBD 2019 Chronic Respiratory Diseases Collaborators, "Global Burden of Chronic Respiratory Diseases and Risk Factors, 1990-2019: An Update from the Global Burden of Disease Study 2019," EClinicalMedicine, vol. 59, p. 101936, 2023. DOI: 10.1016/j.eclinm.2023.101936.

[8] World Health Organization, "Chronic Obstructive Pulmonary Disease (COPD)," WHO, Nov. 6, 2024. Available: https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd).

[9] American Lung Association, "How Lungs Work," American Lung Association, Nov. 17, 2022. Available: https://www.lung.org/lung-health-and-diseases/how-lungs-work.

[10] A. Gafizkyzy, "Qazaqstan ökpe auruynan älem boiynşa üşınşı orynğa şyqqan",

Qazaqstan TV, Dec. 5, 2024. Available: https://qazaqstan.tv/news/203518/.

[11] R. Mutairov, "Qazaqstandyqtar 1-3 tamyz aralyğynda onkoortalyqtarda ökpe skriniñınen tegın öte alady", Kursiv Media, Aug. 1, 2024. Available: https://kz.kursiv.media/kk/2024-07-31/rnmtr-skpinig-tegin/.

[12] J.Błach, , M.Siedliński, W.Sydor, "Immunology in COPD and the use of combustible cigarettes and heated tobacco products", Eur J Med Res 28, 397 (2023). https://doi.org/10.1186/s40001-023-01374-2DOI: 10.1186/s40001-023-01374-2

[13] S. Kumar, V. Bhagat, P. Sahu, M. K. Chaube, A. K. Behera, M. Guizani, R. Gravina, M. Di Dio, G. Fortino, E. Curry, and S. H. Alsamhi, "A novel multimodal framework for early diagnosis and classification of COPD based on CT scan images and multivariate pulmonary respiratory diseases," *Computer Methods and Programs in Biomedicine*, vol. 243, p. 107911, 2024. DOI: 10.1016/j.cmpb.2023.107911.

[14] M. Shimja and K. Kartheeban, "A Comparative Study of Lung Disease Classification Using Fine-Tuned CXR and Chest CT Images," Automatika, vol. 65, pp. 312–322, 2024. Available: https://doi.org/10.1080/00051144.2023.2293274.

[15] J. P. Allinson, N. Chaturvedi, A. Wong, I. Shah, G. C. Donaldson, J. A. Wedzicha, and R. Hardy, "Early Childhood Lower Respiratory Tract Infection and Premature Adult Death from Respiratory Disease in Great Britain: A National Birth Cohort Study," Lancet (London, England), vol. 401, no. 10383, pp. 1183–1193, 2023. DOI: 10.1016/S0140-6736(23)00131-9.

[16] S. Kumar, O. Ivanova, A. Melyokhin, and P. Tiwari, "Deep-learning-enabled multimodal data fusion for lung disease classification," Informatics in Medicine Unlocked, vol. 42, p. 101367, 2023. DOI: 10.1016/j.imu.2023.101367.

[17] National Cancer Institute (U.S.), "Introduction to the Respiratory System," NCI SEER Training Modules. Available: https://training.seer.cancer.gov/anatomy/respiratory/. Accessed: Apr. 2, 2024.

[18] M. G. Levitzky, "Pulmonary Physiology", 10th ed., New York, NY: McGraw Hill, 2022.

[19] National Heart, Lung, and Blood Institute (U.S.), "How the Lungs Work," NHLBI, Mar. 24, 2022. Available: https://www.nhlbi.nih.gov/health/lungs. Accessed: Apr. 2, 2024.

[20] B. Opitz, V. van Laak, J. Eitel, and N. Suttorp, "Innate immune recognition in infectious and noninfectious diseases of the lung," American Journal of Respiratory and Critical Care Medicine, vol. 181, no. 12, pp. 1294–1309, 2010. DOI: 10.1164/rccm.200909-1427SO.

[21] H. Malik and T. Anees, "Multi-modal deep learning methods for classification

of chest diseases using different medical imaging and cough sounds," PLoS One, vol. 19, no. 3, p. e0296352, 2024. DOI: 10.1371/journal.pone.0296352.

[22] S. P. Jadhav *et al.*, "Introduction to Lung Diseases," in *Targeting Cellular Signalling Pathways in Lung Diseases*, K. Dua, R. Löbenberg, Â. C. Malheiros Luzo, S. Shukla, and S. Satija, Eds. Singapore: Springer, 2021. DOI: 10.1007/978-981-33-6827-9_1.

[23] A. H. Sfayyih, N. Sulaiman, and A. H. Sabry, "A Review on Lung Disease Recognition by Acoustic Signal Analysis with Deep Learning Networks," Journal of Big Data, vol. 10, no. 1, p. 101, 2023. DOI: 10.1186/s40537-023-00762-z.

[24] K. A. Johannson, J. R. Balmes, and H. R. Collard, "Air Pollution Exposure: A Novel Environmental Risk Factor for Interstitial Lung Disease?," Chest, vol. 147, no. 4, pp. 1161–1167, 2015. DOI: 10.1378/chest.14-1299.

[25] W. Shi, S. Bellusci, and D. Warburton, "Lung Development and Adult Lung Diseases," Chest, vol. 132, no. 2, pp. 651–656, 2007. DOI: 10.1378/chest.06-2663.

[26] A. Zangiabadi, C. G. De Pasquale, and D. Sajkov, "Pulmonary hypertension and right heart dysfunction in chronic lung disease," *BioMed Research International*, vol. 2014, p. 739674, 2014. DOI: 10.1155/2014/739674.

[27] N. Patel, M. DeCamp, and G. J. Criner, "Lung transplantation and lung volume reduction surgery versus transplantation in chronic obstructive pulmonary disease," *Proceedings of the American Thoracic Society*, vol. 5, no. 4, pp. 447–453, May 2008. DOI: 10.1513/pats.200707-107ET.

[28] S. A. Adeshina and A. P. Adedigba, "Bag of Tricks for Improving Deep Learning Performance on Multimodal Image Classification," Bioengineering, vol. 9, no. 7, p. 312, 2022. DOI: 10.3390/bioengineering9070312.

[29] K. Lamb, D. Theodore, and B. S. Bhutta, "Spirometry," StatPearls, Treasure Island (FL): StatPearls Publishing; 2025 Jan–. [Updated 2023 Aug 17]. Available: https://www.ncbi.nlm.nih.gov/books/NBK560526/.

[30] T. Wanasinghe, S. Bandara, S. Madusanka, D. Meedeniya, M. Bandara, and I. De la Torre Díez, "Lung Sound Classification for Respiratory Disease Identification Using Deep Learning: A Survey," International Journal of Online and Biomedical Engineering (iJOE), vol. 20, pp. 1-15, 2024. DOI: 10.3991/ijoe.v20i10.49585.

[31] E. Shortliffe, "MYCIN: A Rule-Based Computer Program for Advising Physicians Regarding Antimicrobial Therapy Selection," pp. 739–739, 1976. DOI: 10.1145/1408800.1408906.

[32] R. A. Miller, M. McNeil, S. Challinor, F. Masarie, and J. Myers, "The INTERNIST-1/Quick Medical Reference project—status report", The Western Journal of Medicine, vol. 145, pp. 816–822, 1987.

[33] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and

S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, 2017. DOI: 10.1038/nature21056.

[34] I. Ahmad and F. Alqurashi, "Early cancer detection using deep learning and medical imaging: A survey" , Critical Reviews in Oncology/Hematology, vol. 204, p. 104528, 2024. DOI: 10.1016/j.critrevonc.2024.104528.

[35] A. L. Thandu and P. Gera, "Privacy-centric multi-class detection of COVID-19 through breathing sounds and chest X-ray images: Blockchain and optimized neural networks," IEEE Access, vol. 12, pp. 89968-89985, 2024. DOI: 10.1109/ACCESS.2024.3418202.

[36] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", Nature, vol. 521, pp. 436–444, 2015. DOI: 10.1038/nature14539.

[37] S. Skb, M. S. Kumar, P. Karthikeyan, H. Rajadurai, B. Shivahare, S. Mallik, and H. Qin, "An Enhanced Multimodal Fusion Deep Learning Neural Network for Lung Cancer Classification," Systems and Soft Computing, vol. 6, p. 200068, 2023. DOI: 10.1016/j.sasc.2023.200068.

[38] S. Purkovic, L. Jovanovic, M. Zivkovic, M. Antonijevic, E. Dolicanin, E. Tuba, M. Tuba, N. Bacanin, and P. Spalevic, "Audio analysis with convolutional neural networks and boosting algorithms tuned by metaheuristics for respiratory condition classification," Journal of King Saud University - Computer and Information Sciences, vol. 36, no. 10, p. 102261, 2024. DOI: 10.1016/j.jksuci.2024.102261.

[39] Y. Wu, J. Ma, X. Huang, S. Ling, and S. Su, "DeepMMSA: A Novel Multimodal Deep Learning Method for Non-small Cell Lung Cancer Survival Analysis," in Proc. IEEE SMC Conf., 2021, pp. 1468–1472. DOI: 10.1109/SMC52423.2021.9658891.

[40] A. D. Gunasinghe, A. C. Aponso, and H. Thirimanna, "Early Prediction of Lung Diseases," in *Proceedings of the 2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, Bombay, India, 2019, pp. 1–4. DOI: 10.1109/I2CT45611.2019.9033668.

[41] A. M. Q. Farhan, S. Yang, A. Q. S. Al-Malahi, and M. A. Al-antari, "MCLSG: Multi-modal classification of lung disease and severity grading framework using consolidated feature engineering mechanisms," Biomedical Signal Processing and Control, vol. 85, p. 104916, 2023. DOI: 10.1016/j.bspc.2023.104916.

[42] S. Kumar and S. Sharma, "An Improved Deep Learning Framework for Multimodal Medical Data Analysis," Big Data and Cognitive Computing, vol. 8, no. 10, p. 125, 2024. DOI: 10.3390/bdcc8100125.

[43] E. Seeram, "Computed Tomography: A Technical Review," Radiologic Technology, vol. 89, pp. 279CT-302CT, 2018.

[44] H. Malik, T. Anees, A. S. Al-Shamaylehs, S. Z. Alharthi, W. Khalil, and A. Akhunzada, "Deep Learning-Based Classification of Chest Diseases Using X-

rays, CT Scans, and Cough Sound Images," Diagnostics (Basel, Switzerland), vol. 13, no. 17, p. 2772, 2023. DOI: 10.3390/diagnostics13172772.

[45] S. Abhishek, A. Ananthapadmanabhan, T. Anjali, S. Remya, A. Perathur, and R. Bentov, "Multimodal Integration of Enhanced Novel Pulmonary Auscultation Real-Time Diagnostic System," IEEE MultiMedia, vol. PP, pp. 1–26, 2024. DOI: 10.1109/MMUL.2024.3422022.

[46] T. Liu, Z. Zhang, Q. Zhou, et al., "MI-DenseCFNet: Deep learning–based multimodal diagnosis models for Aureus and Aspergillus pneumonia," European Radiology*, vol. 34, pp. 5066–5076, 2024. DOI: 10.1007/s00330-023-10578-3.

[47] S. Deng, X. Zhang, and S. Jiang, "A diagnostic report supervised deep learning model training strategy for diagnosis of COVID-19," Pattern Recognition, vol. 149, p. 110232, 2024. DOI: 10.1016/j.patcog.2023.110232.

[48] E. Çallı, E. Sogancioglu, B. van Ginneken, K. G. van Leeuwen, and K. Murphy, "Deep Learning for Chest X-ray Analysis: A Survey," Medical Image Analysis, vol. 72, p. 102125, 2021. DOI: 10.1016/j.media.2021.102125.

[49] A. Ijaz, M. Nabeel, U. Masood, T. Mahmood, M. S. Hashmi, I. Posokhova, A. Rizwan, and A. Imran, "Towards Using Cough for Respiratory Disease Diagnosis by Leveraging Artificial Intelligence: A Survey," Informatics in Medicine Unlocked, vol. 29, p. 100832, 2022. DOI: 10.1016/j.imu.2021.100832.

[50] J. Lay and B. Pardamean, "Detection of pulmonary tuberculosis on chest X-ray images using multimodal ensemble," ResearchGate, 2022. DOI: 10.13140/RG.2.2.11678.61763.

[51] P. Podder, S. R. Das, M. R. H. Mondal, S. Bharati, A. Maliha, M. J. Hasan, and F. Piltan, "LDDNet: A Deep Learning Framework for the Diagnosis of Infectious Lung Diseases," *Sensors*, vol. 23, no. 1, p. 480, 2023. DOI: 10.3390/s23010480.

[52] A. T. Abdulahi, R. O. Ogundokun, A. R. Adenike, *et al.*, "PulmoNet: A Novel Deep Learning Based Pulmonary Diseases Detection Model," *BMC Medical Imaging*, vol. 24, no. 51, 2024. DOI: 10.1186/s12880-024-01227-2.

[53] Y. Al-Issa, A. Alqudah, H. Alquran, and A. Issa, "Pulmonary Diseases Decision Support System Using Deep Learning Approach," *Computers, Materials Continua*, vol. 73, 2022. DOI: 10.32604/cmc.2022.025750.

[54] V. Mayya, K. Karthik, S. S. Kamath, K. Karadka, and J. Jeganathan, "COVIDDX: AI-based clinical decision support system for learning COVID-19 disease representations from multimodal patient data," in Proc. International Conference on Health Informatics, 2021.

[55] A. Roy and U. Satija, "A Novel Melspectrogram Snippet Representation Learning Framework For Severity Detection of Chronic Obstructive Pulmonary Diseases," *TechRxiv*, Preprint, 2022. DOI: 10.36227/techrxiv.21758660.v1.

[56] Z. Tariq, S. K. Shah, and Y. Lee, "Multimodal lung disease classification using deep convolutional neural network," in Proc. 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2530–2537. DOI: 10.1109/BIBM49941.2020.9313208.

[57] A. Ashurov, Y. Zhou, L. Shi, Y. Zhao, and H. Liu, "Environmental Sound Classification Based on Transfer-Learning Techniques with Multiple Optimizers," *Electronics*, vol. 11, no. 15, p. 2279, 2022. DOI: 10.3390/electronics11152279.

[58] R. Nadkarni, E. Nikolakakis, and R. Marinescu, "AFEN: Respiratory Disease Classification using Ensemble Learning," *arXiv preprint arXiv:2405.05467*, 2024. DOI: 10.48550/arXiv.2405.05467.

[59] B. TaghiBeyglou, A. Assadi, A. Elwali, and A. Yadollahi, "TRespNET: A dual-route exploratory CNN model for pediatric adventitious respiratory sound identification," *Biomedical Signal Processing and Control*, vol. 93, 2024, Art. no. 106170. DOI: 10.1016/j.bspc.2024.106170.

[60] Y. Abadade, N. Benamar, M. Bagaa, and H. Chaoui, "Empowering Healthcare: TinyML for Precise Lung Disease Classification," *Future Internet*, vol. 16, no. 11, p. 391, 2024. DOI: 10.3390/fi16110391.

[61] L. Xiao, L. Fang, Y. Yang, and W. Tu, "LungAdapter: Efficient Adapting Audio Spectrogram Transformer for Lung Sound Classification," in *Proceedings of Interspeech 2024*, pp. 4738–4742, 2024. DOI: 10.21437/Interspeech.2024-106.

[62] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, and M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," *arXiv preprint* arXiv:1912.10211, 2019. DOI: 10.48550/arXiv.1912.10211.

[63] K. Venkatapathiah, D. Koppad, P. Kumar, N. Kantikar, and S. Ramesh, "Multi-task Learning for Lung Sound and Lung Disease Classification," *SN Computer Science*, vol. 6, 2024. DOI: 10.1007/s42979-024-03506-9.

[64] K. Varunkumar, M. Zymbler, and S. Kumar, "Multimodal Deep Dilated Convolutional Learning for Lung Disease Diagnosis," Brazilian Archives of Biology and Technology, vol. 67, 2024. DOI: 10.1590/1678-4324-2024231088.

[65] U. Sait, G. L. K. V, S. Shivakumar, T. Kumar, R. Bhaumik, S. Prajapati, K. Bhalla, and A. Chakrapani, "A deep-learning based multimodal system for Covid-19 diagnosis using breathing sounds and chest X-ray images," Applied Soft Computing, vol. 109, p. 107522, 2021. DOI: 10.1016/j.asoc.2021.107522.

[66] A. Hamdi, A. Aboeleneen, and K. Shaban, "MARL: Multimodal Attentional Representation Learning for Disease Prediction," in Proc. 3rd Int. Conf. Artif. Intell. Comput. Vis. (AICV 2021), Springer, 2021, pp. 14–27. DOI: 10.1007/978-3-030-87156-7_2.

[67] G. Altan and Y. Kutlu, "RespiratoryDatabase@TR (COPD Severity Analysis)," *Mendeley Data*, V1, 2020. DOI: 10.17632/p9z4h98s6j.1.