# Pulmonary Infectious Disease Detection from Chest X-ray Images and Medical Records Using Deep Learning

Lidia Mekuanint Bantelay

A Thesis Submitted to the Department of Computer Science and Engineering

School of Electrical Engineering and Computing

Presented in Partial Fulfilment of the Requirement for the degree of Master's in

Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

September 2022

Adama, Ethiopia

# Pulmonary Infectious Disease Detection from Chest X-ray Images and Medical Records Using Deep Learning

Lidia Mekuanint Bantelay

Advisor: Dr. Mesfin Abebe

A Thesis Submitted to the Department of Computer Science and Engineering

School of Electrical Engineering and Computing

Presented in Partial Fulfilment of the Requirement for the degree of Master's in

Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

September 2022

Adama, Ethiopia

# Declaration

I hereby declare that this Master Thesis entitled "**Pulmonary Infectious Disease Detection from Chest X-ray Images and Medical Records using Deep Learning**" is my original work. That is, it has not been submitted for the award of any academic degree, diploma or certificate in any other university. All sources of materials that are used for this thesis have been duly acknowledged through citation.

| Lidia Mekuanint | _____ | _____ |
|---|---|---|
| Name of the Student | Signature | Date |

## Recommendation

I, the advisor of this thesis, hereby certify that I have read the revised version of the thesis entitled "**Pulmonary Infectious Disease Detection from Chest X-ray Images and Medical Records using Deep Learning**" prepared under my guidance by Lidia Mekuanint submitted in partial fulfilment of the requirements for the degree of Mater's of Science in Computer Science and Engineering.

Therefore, I recommend the submission of revised version of the thesis to the department following the applicable procedures.

| | | |
|---|---|---|
| Dr. Mesfin Abebe | _____ | _____ |
| Major Advisor | Signature | Date |

# Approval Sheet

I, the advisor of the thesis entitled "**Pulmonary Infectious Disease Detection from Chest X-ray Images and Medical Records using Deep Learning**" and developed by Lidia Mekuanint, hereby certify that the recommendation and suggestions made by the board of examiners are appropriately incorporated into the final version of the thesis.

| Dr. Mesfin Abebe | _____ | _____ |
|---|---|---|
| Major Advisor | Signature | Date |

We, the undersigned, members of the Board of Examiners of the thesis by Lidia Mekuanint have read and evaluated the thesis entitled "**Pulmonary Infectious Disease Detection from Chest X-ray Images and Medical Records using Deep Learning**" and examined the candidate during open defense. This is, therefore, to certify that the thesis is accepted for partial fulfillment of the requirement of the degree of Master of Science in Computer Science and Engineering.

| _____ | _____ | _____ |
|---|---|---|
| Chairperson | Signature | Date |
| _____ | _____ | _____ |
| Internal Examiner | Signature | Date |
| _____ | _____ | _____ |
| External Examiner | Signature | Date |

Finally, approval and acceptance of the thesis is contingent upon submission of its final copy to the Office of Postgraduate Studies (OPGS) through the Department Graduate Council (DGC) and School Graduate Committee (SGC).

| _____ | _____ | _____ |
|---|---|---|
| Department Head | Signature | Date |
| _____ | _____ | _____ |
| School Dean | Signature | Date |
| _____ | _____ | _____ |
| Office of Postgraduate Studies, Dean | Signature | Date |

# ACKNOWLEDGEMENT

First of all, I praise the alpha and omega almighty God, my savior Jesus Christ for assisting and providing for me throughout my life, the paths I go through, and for not letting me down during my hard times. And for giving me the strength to work on this research. I can only say there is no one like God.

I am extremely thankful to my esteemed Advisor Dr. Mesfin Abebe for his priceless guidance, feedback and support starting from title selection throughout each step of this work.

Dr. Tariku Workineh and Dr. Basliel Yishak, if you were not involved in this work, it would be meaningless and far from real-life practice; I am so thankful for sharing your expertise, time and support in data preparation.

Dr. Samuel Tomelisan (senior radiologist) thank you for your help by labelling radiology images.

Dr. Amanuel Abebe, Dr. Lydia Paulos and Sister Hadas Desta; I sincerely appreciate the assistance you provided me during the data collection phase of this work.

All of the Artificial Intelligence SIG members, thank you for your comments and encouragement!

I am grateful to my family for their unconditional love, prayer, and support. Nathnael Yonas, thank you for providing me with the cloud resource.

# Table of Contents

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF CODE SNIPPETS

# LIST OF ALGORITHMS

# LIST OF EQUATIONS

# LIST OF ACRONYMS

| | |
|---|---|
| Acc | Accuracy |
| AI | Artificial Intelligence |
| API | Application Program Interface |
| ANOVA | Analysis of Variance |
| CAD | Computer-Aided Detection |
| CNN | Convolutional Neural Network |
| COPD | Chronic Obstructive Pulmonary Disease |
| CPU | Central Processing Unit |
| CRNN | Convolutional Recurrent Neural Network |
| DICOM | Digital Imaging and Communication In Medicine |
| DL | Deep Learning |
| DT | Decision Tree |
| EMR | Electronic Medical Record |
| FC | Fully Connected |
| GB | Gigabyte |
| GPU | Graphics Processing Unit |
| GRU | Gate Recurrent Unit |
| GUI | Graphical User Interface |
| HIV/AIDS | Human Immunodeficiency Virus/ Acquired Immunodeficiency Syndrome |
| HTTP | Hyper Text Transfer Protocol |
| ICU | Intensive Care Unit |
| KNN | K Nearest Neighbor |
| LSTM | Long Short Term Memory |
| MLP | Multi-Layer Perceptron |
| MRI | Magnetic Resonance Imaging |
| MRN | Medical Record Number |
| OCR | Optical Character Recognition |
| OPD | Out Patient Department |

| | |
|---|---|
| PTB | Pulmonary Tuberculosis |
| RAM | Random Access Memory |
| RELU | Rectified Linear unit |
| RGB | Red, Green, and Blue |
| SDG | Sustainable Development Goal |
| STN | Spatial Transformer Network |
| SVM | Support Vector Machine |
| TB | Tuberculosis |
| TPU | Tensor Processing Unit |
| WHO | World Health Organization |

# ABSTRACT

*Pulmonary infectious diseases are major public health problems worldwide. These diseases affect lungs and if not diagnosed properly in time, they can become fatal. Chest x-ray images are widely used for pulmonary disease detection and diagnosis. Detection of pulmonary diseases from chest x-ray images is difficult and needs experience due to similar pathological features of the diseases and sometimes misdiagnosis occurs due to this. Several researchers used deep learning and machine learning techniques to solve this problem. These studies used chest x-ray images exclusively to develop pulmonary disease detection models. But using only chest x-ray images can not necessarily lead to accurate disease detection. Medical records are required to interpret chest x-ray images in the appropriate clinical context. This work intends to develop a multi-input pulmonary infectious disease detection model using chest x-ray images and medical records. Concerning this, the medical record dataset is collected from St. Peter's specialized hospital, Alert hospital, and Yekatit 12 hospital medical college. Chest x-ray images are collected from St. Peter's specialized hospital, Alert hospital and Kaggle repository. Data transformation, normalization, and feature selection are among the preprocessing techniques applied to medical record datasets. For the chest x-ray image data, image cleaning, augmentation and normalization preprocessing techniques are performed. Convolutional neural network for image processing and multilayer perceptron for medical record processing are applied to develop the model. Feature level concatenation is performed to join the output feature vectors from convolutional neural network and multilayer perceptron for disease detection model development. For the purpose of comparison, we also developed an image-only and medical record-only model. The image-only model gives an accuracy of 92.68%, the medical record-only model results in 98.72% accuracy, and the joint model accuracy is recorded at 99.61%. Also the joint model scored accuracy of 93.94% in medical expert evaluation.*

***Keywords:*** *Pulmonary infectious disease, Detection, Chest x-rays, Medical record, deep learning*

# CHAPTER ONE

## 1. INTRODUCTION

### 1.1. Background of the Study

Pulmonary infectious diseases are disease that affect the lung. When large microbe such as virus or bacteria enters the lungs and damage, pulmonary infection occurs. Tuberculosis, pneumonia, bronchitis, influenza and asthma are some examples of pulmonary diseases. Moreover, pneumonia and TB[1] are major public health problems worldwide (Cillóniz et al., 2020).

According to a report from WHO, TB is the second top infectious deadly disease next to COVID-19 and above HIV/AIDS. In 2020, 1.2 million people[2] comprising 214,000 people who are HIV positive died from TB. The number of deaths has increased from 2019, this is the result of TB risk factors like undernutrition and poverty are being aggravated by COVID-19 pandemic ("Global Tuberculosis Report," 2021). It is predicted that in 2021 and 2022 the case will be worsened. On top of this developing countries cover more than 95% of death cases ("Global Tuberculosis Report," 2021). Tuberculosis is also one of Ethiopia's most critical public health problems. In 2020, 151,000 TB incidences are recorded. In addition, TB treatment coverage is reported to be 71%, and the remaining 29% of TB cases were not diagnosed. The main causes of TB morbidity and mortality in Ethiopia are co-infections with other diseases like HIV/AIDS and lack of access to early diagnosis[3] and treatment services.

Additionally, in 2020 only 5.8 million of the 9.9 million people who are sick with TB were detected and notified, leading to a gap of 4.1 million cases[4]. This has dropped from 7.1 million diagnoses made in 2019 ("Global Tuberculosis Report," 2021).

[1] https://www.publichealth.hscni.net/directorate-public-health/health-protection/tuberculosis-tb
[2] https://www.who.int/news-room/fact-sheets/detail/tuberculosis
[3] https://www.usaid.gov/ethiopia/global-health/tuberculosis
[4] https://www.who.int/news-room/facts-in-pictures/detail/tuberculosis

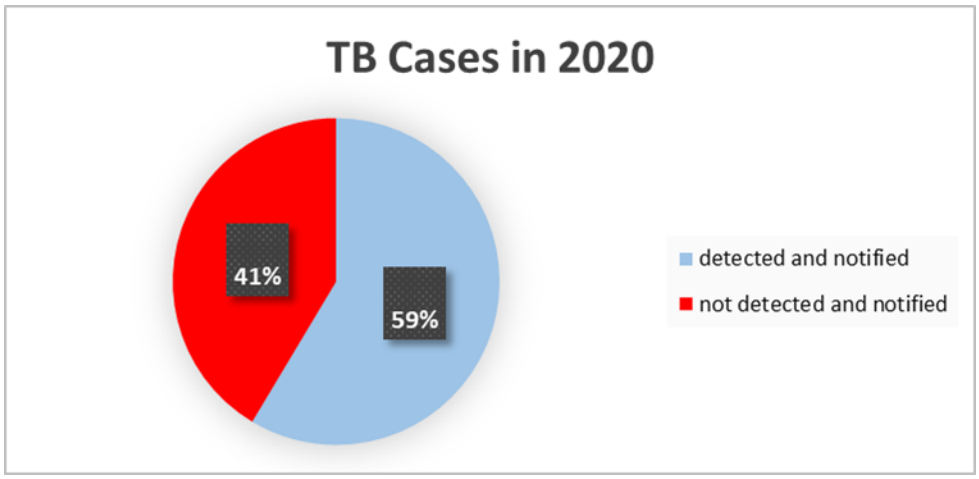



Figure 1: TB Cases In 2020

Mnyambwa et al.'s research (Mnyambwa et al., 2021) on identifying gaps in the detection and diagnosis of TB in East Africa has concluded that TB disease detection and diagnosis gaps are the consequence of increased patient burden, poor laboratory system, and lack of skilled workers. This gap has to be addressed to achieve the SDG target of ending TB by 2030.

Pneumonia, the other pulmonary infectious disease, is the leading reason for the deaths of children globally. It is responsible for 14% of all deaths among children under the age of five. 740,000 children died from pneumonia in 2019[5]. People with pre-existing health problems and adults over the age of 65 are at risk for pneumonia. It can affect either or both lungs. When a healthy individual breathes, the tiny air sacs named alveoli that make up the lungs fill with air. When a person has pneumonia, fluid and pus accumulate in the alveoli, which reduces the ability to take in oxygen and makes breathing painful. Pneumonia is caused by viruses, bacteria, or fungi. Bacterial and viral pneumonia are contagious, while fungal pneumonia is not contagious. Ethiopia is among the sub-Saharan African countries with the highest rate of pneumonia. Even if pneumonia deaths can be prevented by different cost-effective interventions like good nutrition, and immunization, delays in detecting pneumonia contribute to the highest death of children under five (B. B. Abate et al., 2020).

Different medical imaging techniques can be applied for the diagnosis of pulmonary disease. But, due to its accessibility, low radiation dose, and ease of use (Haloi et al., 2018), (Qin et al.,

[5] https://www.who.int/news-room/fact-sheets/detail/pneumonia

2018), (Hooda et al., 2017) chest x-ray is widely used for pulmonary disease diagnosis and health monitoring (Stirenko et al., 2018). Diagnosing pulmonary disease using a chest x-ray image is a difficult task and requires skillfulness and experience. Many pulmonary diseases resemble one another. This creates difficulty in identifying and diagnosing those diseases accurately using chest x-ray. In the diagnosis of tuberculosis, the major task is to differentiate tuberculosis from other pulmonary diseases like pneumonia. Misdiagnosis of these diseases is not uncommon and results in health problems, complications, and even death (Verma et al., 2020).

Traditional x-ray CAD (Computer-aided detection) assistant systems have improved the detection of chest diseases. The advent of artificial intelligence (AI) and the availability of massive amounts of medical images have opened up new approaches for developing deep learning CAD systems in the medical field(Qin et al., 2018).

Researchers apply machine learning and deep learning techniques to develop pulmonary disease detection models. CNN and pretrained models are widely used to develop detection models using chest x-ray images. Deep learning models developed to detect pulmonary disease consider chest x-ray image data solely. These models cannot contextualize the chest x-ray image according to medical history and other medical information in the way that medical practice applies. To achieve a similar goal as in real practice, deep learning models should also use patient medical history in addition to chest x-ray image (S. C. Huang et al., 2020). In this research, deep learning is used to develop a pulmonary infectious diseases detection model from patient medical record and chest x-ray image dataset.

We have chosen TB and pneumonia among pulmonary diseases because, as stated, these diseases are major public health problems in Ethiopia as well as the world. Radiologically, pneumonia and tuberculosis closely resemble each other on chest x-ray images (Verma et al., 2020), (Haloi et al., 2018). Also the misdiagnosis between the two diseases occur more frequently than any other lung disease (*European Respiratory Society Annual Congress 2012*, 2012).

## 1.2. Motivation of the Study

Being healthy is primal to live in life. Living a longer life, productivity, and economic progress can be achieved only when one is healthy. So, preserving people's health is the most important thing. Pulmonary infectious diseases are deadly diseases worldwide. Developing countries like Ethiopia are suffering more from these diseases. Lack of enough health professionals, low economic status, and no advanced technologies are making the problem worse. Due to these, the patient becomes unable to get the proper treatment at the early stage. Not receiving the proper treatment at an early stage leads to more complicated and worse health problems. In addition, radiologists face difficulty in identifying the pulmonary disease from chest x-ray images due to similar features of the disease this results in misinterpretation. This misinterpretation can lead to misdiagnosis and put the patient's life in danger.

Chest x-ray images can show normal lung results in the early stages of pulmonary disease. Such kinds of cases cannot be identified using chest x-ray images alone. Also, different pulmonary diseases can indicate similar symptoms, and identifying the disease using only symptom records is impossible. Furthermore, 20-30% of chest x-ray results are false negative, and 2-5% are false positive (Vock, 1987). To tackle these inaccuracy problems that arise from using patient symptom records or x-ray images exclusively in detecting pulmonary disease, we are motivated to develop a deep learning model that integrates chest x-ray images and patient symptom record data.

## 1.3. Statement of the Problem

Pulmonary infectious disease detection from chest x-ray images is a difficult task due to similar pathological features of the diseases. Radiologists read chest x-ray images by referring to the clinical information provided. But still radiologists make wrong diagnoses due to the complex nature of these features (Haloi et al., 2018). This misdiagnosis can lead to complex health problems even death.

Current studies that apply deep learning for pulmonary disease detection used chest x-ray images solely. The convolutional neural network (CNN) algorithm is the foundation for these works. However these works should have included the patient's medical history and other clinical

information for accurate diagnosis. Deep learning models developed using only chest x-ray image data have a limited impact on clinical practice because an accurate medical diagnosis needs the context of the patient's clinical information. Similar chest x-ray findings with different clinical histories can represent different diseases, two cases that can lead to misdiagnosis if only a chest x-ray image is applied are presented here:

**Case 1:** a chest x-ray image that indicates pneumonia can become correct in an individual with a fever and increased white blood cell count. However, without that supplementary laboratory and clinical information, a comparable chest x-ray finding may represent another pulmonary disease (S. C. Huang et al., 2020).

**Case 2:** a person having TB with HIV positive or additional lung disease like respiratory failure, the chest x-ray image may not show as it is TB rather it looks like other disease like pneumonia (Pinto et al., 2011).

Regarding these cases, the detection of pulmonary disease from a chest x-ray image alone, without supporting medical information is unreliable and cannot distinguish all cases clearly. Different pulmonary diseases can have similar symptoms, and detection of the disease using medical records alone is not applicable.

To fill these gaps in real life medical practice and current deep learning models developed for pulmonary infectious disease detection, the goal of this study is to develop pulmonary infectious disease detection model by incorporating chest x-ray images with medical records with deep learning techniques.

## 1.4. Research Questions

In solving the problems stated above, this research attempts to respond to the following three research questions:

**RQ-1.** How can selected deep learning algorithms be employed to develop a model with more accurate results in pulmonary disease detection?

**RQ2**. How can image augmentation affect the pulmonary infectious disease detection performance of the model?

**RQ3.** Can the performance of the pulmonary infectious disease detection model be improved by combining medical records and chest x-ray images?

## 1.5. Objectives of the Study

### 1.5.1. General Objective of the Study

The general objective of this study is to develop pulmonary infectious diseases detection model from chest x-ray images and medical records using deep learning techniques.

### 1.5.2. Specific Objectives of the Study

To achieve the general objective of this study, the following specific objectives are laid out:

- Review literatures related to pulmonary disease detection.
- Collect medical record and chest x-ray image data.
- Perform medical record data preprocessing.
- Perform chest x-ray image data preprocessing.
- Apply selected deep learning algorithms for model development.
- Concatenate preprocessed features of image and medical record data.
- Measure and Evaluate the performance of the model.
- Develop and deploy a pulmonary diseases detection graphical user interface.

## 1.6. Scope and Limitations of the Study

### 1.6.1. Scope of the Study

The goal of this study is to develop a model that will support medical doctors in making accurate decisions in the detection of pulmonary infectious diseases. Although different mechanisms can be used to detect pulmonary infectious disease, we used chest x-ray image and medical record dataset in this study. Chest x-ray images and medical records are processed and classified using deep learning techniques of convolutional neural networks and multilayer perceptron algorithms. The feature selection of medical records and augmentation of chest x-ray images are also included. This study covered data collection, preprocessing of chest x-ray images, and medical records. This study addresses only the detection of pneumonia and tuberculosis disease among pulmonary infectious diseases. Pneumonia can be classified as bacterial, viral, or fungal

pneumonia, in this work detection of these subdivisions of pneumonia is not addressed due to the unavailability of a sufficient amount of data.

### 1.6.2. Limitation of the Study

Only frontal (poster anterior (PA)) view chest x-ray images are used as image input, lateral view images are not included. We are also limited by time and budget constraints. We used chest x-ray image and structured medical record data to develop the model thus not supporting written notes in natural language is the limitation of this work. Due to unavailability of enough labeled local chest x-ray images we are obliged to include chest x-ray images form Kaggle repository. Not being able to apply the corresponding chest x-ray image of all patients is the limitation of this work.

## 1.7. Significance of the Study

This pulmonary infectious disease detection model by incorporating chest x-ray images and medical record data can help minimize the misdiagnosis problems that stem from models developed using only chest x-ray images or medical record data. The main beneficiaries of this research output are radiologists (medical doctors). The model can be more practically relevant to radiologists (medical doctors) to aid in more accurate detection of pulmonary disease since it incorporates medical record data and chest x-ray images together. But it cannot be used exclusively or fully replace doctors. Next to medical doctors, the patients are the beneficiaries of this work because they can get timely and more accurate disease detection and diagnosis from medical doctors who used the model developed.

This study can be a good road map for those researchers who are willing to work on pulmonary and other disease detection using two or more different types of datasets. This research can help in understanding the steps and challenges of developing a disease detection model using x-ray images and medical records.

## 1.8. Organization of the Thesis

This thesis comprises six chapters. The first chapter is the introductory part of the work. The study's background, motivation, problem statement, research questions, objective, scope, limitations, and significance are all included. The second chapter presents the literature review

and related works that have been previously done in this research area. The third chapter is composed of the methodologies and tools used in this research, the description of the datasets used; and preprocessing and model building techniques are presented. Chapter four presents the proposed solution architecture, selected preprocessing techniques, and algorithms to be implemented. Chapter five delivers the implementation of the proposed solution by including some code snippets. In Chapter six, the results of the implementation and conclusions drawn in light of the results are delivered. The results of preprocessing and augmentation, the performance results of the algorithm, and the medical expert evaluation results are also presented. In the discussion sub-topic, the results and research questions are presented and discussed. The results are interpreted. Finally, the conclusion and future work recommendations are presented.

# CHAPTER TWO

## 2. LITERATURE REVIEW AND RELATED WORKS

### 2.1. Introduction

This chapter constitutes the base literature for this work and related works done previously in the same area. The pulmonary infectious diseases of our focus and their methods of diagnosis are discussed. Machine learning concepts, different machine learning classification algorithms and deep learning are presented as well. Previous work on the detection of pulmonary infectious diseases using deep learning techniques, as well as their gaps, is also presented.

#### 2.1.1. The Lung (Respiratory System)

The lungs are one of a pair of organs in the lower respiratory tract that provide oxygen to the body and eliminate carbon dioxide (*The Lungs - Fact Sheet*, 2018). The main purpose of the lungs is to perform the gas exchange process known as respiration, or breathing. The lungs provide oxygen and keep every other organ in the body functioning by removing carbon dioxide from the body. When the lung function is weakened, the lungs' capacity to exchange gases is lowered. One may not appreciate the value of lung health before becoming sick with lung disease. A lung becoming sick means one has got breathing problems and this makes our body not get enough oxygen to function.

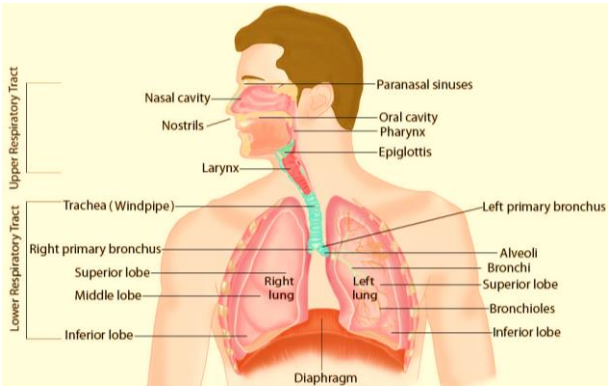



Figure 2: Lung (Respiratory System)[6]

[6] https://quizlet.com/372457089/respiratory-system-diagram/

## 2.2. Pulmonary Infectious Disease

Infectious diseases are diseases caused by viruses, bacteria, or fungus. Pulmonary infectious diseases are infectious diseases that affect the lung. Our airways split into tiny tubes (bronchioles) which terminate in clumps of air sacs known as alveoli. Most of the lung tissue is made up of these air sacs. Pneumonia and tuberculosis are among the lung diseases that harm the alveoli.

### 2.2.1. Tuberculosis

The bacteria known as Mycobacterium tuberculosis causes tuberculosis (TB). This bacteria commonly attacks the lungs, but it can still attack any part of the body, including the brain, kidney, and spine. TB has two conditions: latent TB, or inactive TB, and active TB. Latent TB is when a person has a TB infection but the bacteria is inactive in the body and this is not communicable. Active TB is when a person is sick with TB and this can spread to other people. Coughing for three weeks or more; blood or mucus in the cough; chest pain or difficulty breathing or coughing; unexpected weight loss; fatigue; fever; night sweating; and loss of appetite are all signs and symptoms of active tuberculosis (Leung, 1999). People who have had close contact with a person who has TB infectious disease, people with medical conditions that weaken the immune system such as HIV infection, substance abuse, diabetes mellitus, severe kidney disease, low body weight, and young children are at high risk of developing TB disease (Konstantinos, 2010).

Some patients become ill with TB disease shortly after contracting the infection (within a few weeks), before their immune system is able to combat the TB bacterium. Some people may not get sick for years, but once their immune system deteriorates for other reasons, they may.

### 2.2.2. Pneumonia

Pneumonia is an infection that causes inflammation of the air sacs in one or both lungs. The air sacs could fill up with fluid or pus, resulting in a cough with sputum, fever, and breathing difficulties. Bacteria, viruses, and fungus are among the organisms that can cause pneumonia.

Viral and bacterial pneumonia are contagious, meaning they can be transmitted from one person to another by airborne droplets from coughing or sneezing. Fungal pneumonia does not spread from one person to another. One can be infected by fungal pneumonia from the environment.

Children under the age of 2 and people older than 65 are at high risk of pneumonia Other risk factors include smoking, alcohol consumption, having a chronic illness such as asthma or COPD, cardiovascular disease, people in the intensive care unit (ICU), malnutrition, and a weakened immune system (Laufer, 2013). Pneumonia can have complications like lung abscess, pneumothorax, empyema and pleural effusion (Tuazon, 2005). Pleural effusion is a complication when fluid is accumulated between layers of lung tissues. This fluid can get infected and needs to be drained via a chest tube or removed surgically. In a study conducted about complications of pneumonia in the pediatric population in Oshogbo, Nigeria (Odeyemi et al., 2020), among the participants during the hospital stay of pneumonic pediatrics who were under the condition of hypoxemia, heart failure and anemia occurred in 41.9% and 19.4%, respectively.

## 2.3. Methods of Pulmonary Infectious Disease Detection

People with pulmonary disease (TB or pneumonia) may not have symptoms. In clinics, physicians perform the following techniques to diagnose pulmonary disease in a person: Any of the techniques cannot be used to confirm the presence of a specific disease in a patient alone.

### 2.3.1. Patient History

The first thing the physician does about diagnosis and detection of pulmonary disease is to investigate the medical history of the patient (*Chapter 4 Diagnosis of Tuberculosis Disease*, n.d.). This includes asking the patient what has happened and listen to what they would reply. The following are some questions that could be asked by the physician:

- If the patient is coughing, how long have they been coughing.
- If the cough is dry or productive,
- If breathing is difficult.
- If the patient has close contact with a TB patient or chronic cougher,
- If the patient has previously been diagnosed with TB disease,

- Demographic factors (like age, occupation, country, or origin) If the patient has other medical conditions that suppress the immune system, like diabetes or HIV infection,
- Use of drugs, smoking.
- Recent medications and antibiotics used.

### 2.3.2. Physical Examination

- ***Auscultation:*** After taking the patient's history, the physician will listen to the lungs of the person using a stethoscope. If they have pneumonia, cracking, or wheezy sounds may be heard (Bohadana et al., 2014).
- ***Pulse Oximetry:*** Pulse oximetry is a test to measure the amount of oxygen in the blood[7].
- ***Measure Vital Signs:*** the measurement of the body's fundamental functioning are vital signs. The four primary vital signs are body temperature, respiratory rate, pulse rate, and blood pressure[8].

### 2.3.3. Chest Imaging

Chest x-rays, Computed Tomography (CT), and Magnetic Resonance Imaging (MRI) can be used for chest imaging of pulmonary disease detection (Skinner, 2015). Among these imaging techniques, chest x-rays are the most commonly applied imaging test to detect abnormalities in the lungs (Haloi et al., 2018). Chest x-rays show the fluid in or around the lungs or air surrounding a lung. Chest x-rays of a patient on treatment can also be taken many times to know if the patient's health is getting better or worse.

### 2.3.4. Sputum Test

Sputum tests are used to determine abnormalities in the lungs by testing the sputum that comes from the lung by coughing (Fitz-gerald, 2013). The sputum is placed in the laboratory on a special plate that has the nutrients that help bacteria or other pathogens present in the sputum grow[9]. If the bacteria are grown, the laboratory determines which bacteria cause sickness. This

[7] https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/pulse-oximetry
[8]https://www.hopkinsmedicine.org/health/conditions-and-diseases/vital-signs-body-temperature-pulse-rate-respiration-rate-blood-pressure
[9] https://myhealth.alberta.ca/Pages/HealthInfoToolsDefault.aspx

test is widely used if the patient is suspected to have TB disease. It can also be applied to detect pneumonia and other lung diseases.

### 2.3.5. TB blood Test

A blood test is another method of detection of TB. A small amount of blood is taken to the laboratory to see if there are TB bacteria in the blood (*Tuberculosis ( TB ) Blood Test ( IGRA )*, n.d.).

## 2.4. Machine Learning Techniques

Artificial intelligence is making problem-solving techniques more accurate and easier from time to time. Machine learning is an artificial intelligence technique that makes systems learn from datasets without programming manually. There are four major types of machine learning algorithms:

- **Supervised Learning:** In supervised learning, the labeled data and target outputs are fed into the system. The supervised system creates a relationship between given data and output. As a result, it can give the possible output for the data given as an input that it has never seen before.
- **Unsupervised Learning:** in unsupervised learning, no output labels or categories are given to the model. But based on the algorithm given it tries to model the relationships.
- **Semi-Supervised Learning** is an amalgam of supervised and unsupervised learning. A small number of labeled data and a large number of unlabeled data can be used.
- **Reinforcement Learning:** This strategy seeks to adopt behaviors that increase reward or decrease risk based on observations acquired from interactions with the environment.

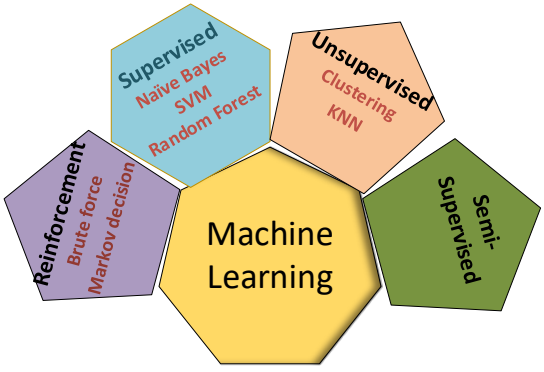



Figure 3: Machine Learning Classifications

### 2.4.1. Supervised Machine Learning for Classification

In supervised machine learning, algorithms learn from labeled data to decide on unlabeled data. Classification is a supervised machine learning technique for determining the class of dependent variables based on one or more independent variables.

#### 2.4.1.1. Support Vector Machine (SVM)

The support vector machine algorithm aims to find a hyperplane in an N-dimensional space that categorizes data points. To sort two data groups by categorization, support vector machines are utilized. The algorithms divide the groupings in to categories based on patterns using lines (hyperplanes) (Chen Junli & Jiao Licheng, 2002). Support vector machines can be applied for both classification and regression tasks.

#### 2.4.1.2. Logistic Regression

Logistic regression is a technique most commonly used to classify binary class problems. It calculates the class membership probability of each class and classifies it into either of the classes (Dreiseitl & Ohno-Machado, 2002). An example of logistic regression can be the binary class of whether a person has pneumonia or not.

#### 2.4.1.3. Decision Tree

A decision tree is a graph that shows choices and their results as a tree. The graph's nodes represent choices, while the graph's edges indicate decision rules or conditions. Each tree is combination of Nodes and branches. Each node represents attributes in a group that needs to be categorized, and each branch indicates a possible value for that node. (F.Y et al., 2017).

#### 2.4.1.4. K-Nearest Neighbor (KNN)

The KNN algorithm considers that similar objects are near to each other. After storing all the available data, a new data item is classified in KNN approach depending on how similar it is to the existing data. It is a lazy learning method because it stores the dataset rather than learning from training, and when it gets new data, it categorizes it into a data class that is very similar to the new data (Cunningham & Delany, 2021). If we have images of chest x-ray and leg x-ray and

if we give a new chest x-ray image, the KNN algorithm finds similarities between the stored datasets and the new dataset and it will classify the new data as a chest x-ray.

#### 2.4.1.5. Naive Bayes

Naive Bayes is a classification technique based on the Bayes theorem and the premise of predictor independence. A Naive Bayes classifier, to put it simply, holds that the existence of one feature in a class has no influence on the existence of any other feature (Y. Huang & Li, 2011). The text categorization field is the primary attention of Naive Bayes. In addition to classification it is employed for clustering (F.Y et al., 2017).

### 2.4.2. Deep Learning for Classification

Deep learning is the part of machine learning that needs a high quantity of data to learn from and gives better accuracy than traditional machine learning algorithms. Deep learning classification techniques can also produce numerical outputs between 0 and 1, but classification algorithms in traditional machine learning approaches are only considered to be either 0 or 1 (Ertam, 2017). Knowing the type of data we are working on and understanding the features of the data is much important in building an appropriate machine learning or deep learning model.

Deep learning's history traces back to 1943 when the MCP (McCulloch and Pitts) model is developed which is considered as the ancestor of the artificial neural model. Since then deep learning passed different development stages like the development of the recurrent neural network in 1986, the discovery of backpropagation in the 1990s and dropouts, an efficient way of training neural networks in 2012 (Haohan Wang, 1956). Deep learning has got popularity in the classification of images, detecting objects, and recognizing faces. In the medical field, classification using deep learning is used to classify different diseases based on medical images.

## 2.5. Recurrent Neural Networks

Recurrent neural networks are extensions of feed-forward neural networks that are mainly used for time series data classification and prediction (Hüsken & Stagge, 2003). Recurrent neural networks form recurrent connection between nodes in the hidden layer. These loops enable RNNs to remember the past, which other algorithms cannot. RNNs can generate outputs by

keeping sequence of track. LSTM (Long Short Term Memory) and GRU (Gated Recurrent Unit) are two of the most popular RNN algorithms for time series prediction. These networks are mainly used for time series data, text data, and audio data.

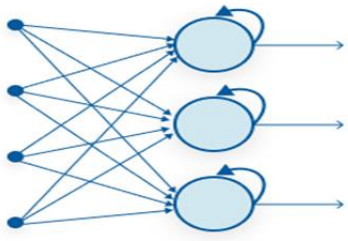


Figure 4: Recurrent Neural Networks[10]

## 2.6. Convolutional Neural Network

CNN is a deep learning algorithm for processing data with a grid pattern that is inspired by the organization of the animal visual cortex and meant to learn spatial hierarchies of features from low-level to high-level patterns. CNNs are the most effective neural networks in computer vision tasks of image recognition and classification. The convolution layer, pooling layer, and fully connected layers are the building blocks of CNN (Patil & Rane, 2021).

***A convolution layer*** is a key element of the CNN architecture which performs feature extraction by combining linear and nonlinear techniques like convolution and activation functions.

***A pooling layer*** performs down sampling operations on the feature maps, minimizing their in-plane dimensionality to establish translation invariance to tiny shifts and distortions and minimize the number of trainable parameters.

***In the Fully Connected Layer (Dense Layer),*** the output feature maps of the convolution or pooling layer are flattened and transformed into a one-dimensional array and linked to other fully connected layers where a learnable weight connects each input to each output.

[10] https://www.analyticsvidhya.com/blog/2022/03/a-brief-overview-of-recurrent-neural-networks-rnn/

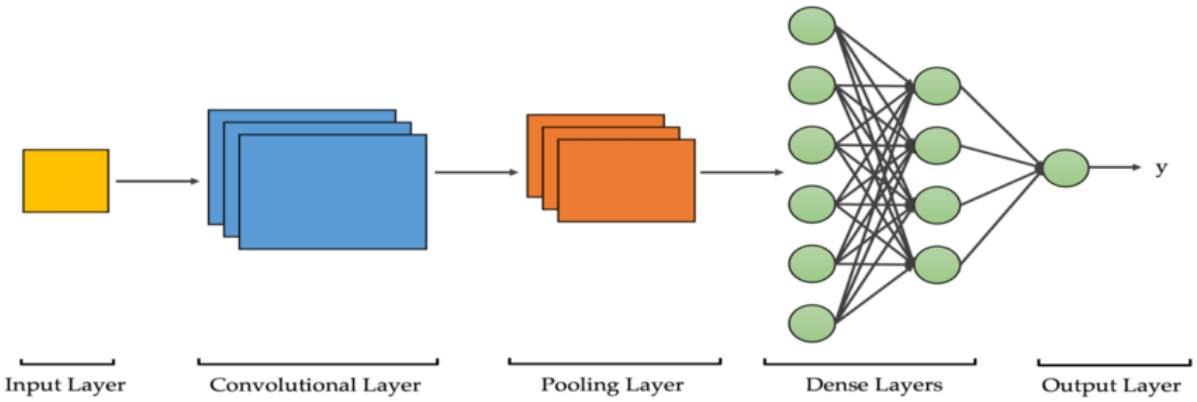



Figure 5: CNN Architecture[11]

## 2.7. Multi-Layer Perceptron

In deep learning, a multi-layered perceptron (MLP) is one of the most common feedforward types of neural network models. A multilayer perceptron (MLP) maps input data to a collection of acceptable outputs. An MLP is made up of numerous layers, each of which is fully connected to the one before it. Except for the nodes of the input layer, the nodes of the layers are neurons with nonlinear activation functions (Zhai et al., 2016). One or more nonlinear hidden layers can exist between the input and output layers. A MLP has a linear activation function and the training of the MLP is performed by using backpropagation, which consists of forward pass and backward pass. In forward pass the expected output according to the given inputs are evaluated. The partial derivatives of the cost function according to various parameters are transmitted back over the network during the backward pass.

[11]https://www.researchgate.net/figure/A-basic-convolutional-neural-network-structure-for-image-classification-Convolutional

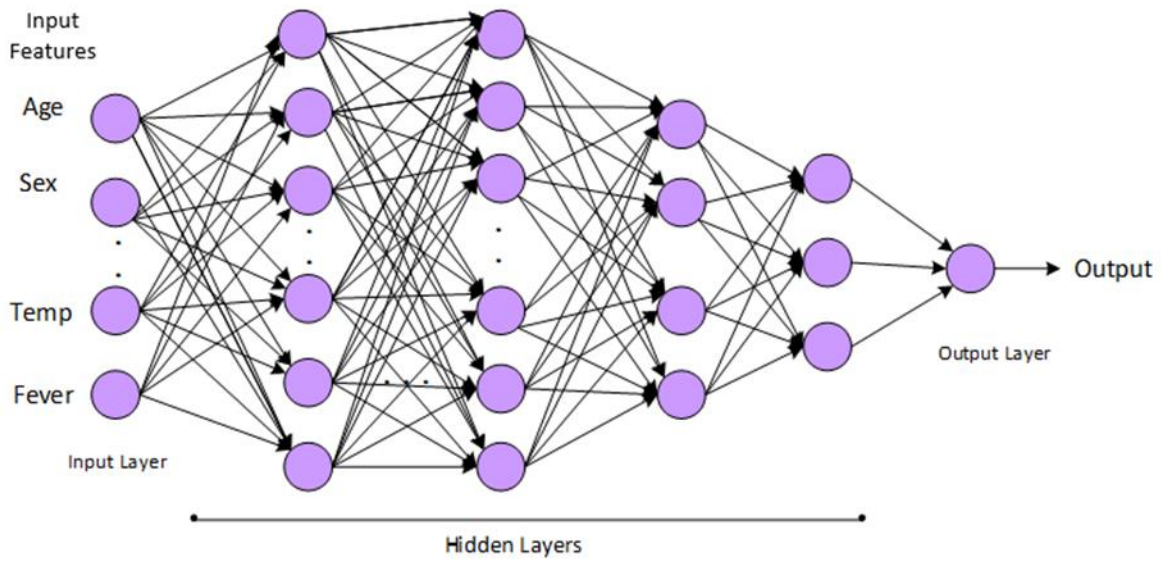



Figure 6: Multi-Layer Perceptron

## 2.8. Image Processing Techniques

### 2.8.1. Image Augmentation

Image augmentation is a method that enhances the quantity of data by adding slightly modified copies of already existing data (Shijie et al., 2017). The training set's diversity, which reduces overfitting, can also enhance the predictive performance of a model. Image augmentation can improve accuracy and provide a flexible model when encountering various shifts in testing or real-world data. Image augmentation techniques include cropping, rotation, zooming, flipping, color range changing, and gray scaling.

- ***Cropping****:* cropping the image to get similar-sized images.
- ***Zooming****:* the image is zoomed at a specified value.
- ***Rotation:*** the image rotated to some degree.
- ***Greyscale conversion:*** converting RGB images to greyscale images to minimize complexity.
- ***Random brightness:*** maximize or minimize the image brightness in a specified range.
- ***Flipping:*** flipping the y-axis image left to right.

- ***Translation:*** shifting the image to different sides. Translation can be done horizontally, vertically, or both horizontally and vertically.
- ***Noise injection:*** adding random noise to the image

## 2.9. Structured Data Feature Selection Methods

Feature selection aims to select some features among the total features and achieve the same accuracy. Three types of feature selection techniques exist in supervised machine learning. Wrapper method, filter method, and hybrid method (Jovic et al., 2015).

- ***Wrapper Method:*** wrapper method evaluates a feature based on a specific machine learning algorithm. It consists of forward feature selection, backward feature selection, and stepwise feature selection methods.
- ***Filter Method:*** filter methods are generic methods that do not incorporate machine learning algorithms for the selection of features. Filter methods have high computational methods for datasets with many features. Chi-Squared, ANOVA, and information gain are examples of filter feature selection methods.
- ***Embedded Method:*** embeds feature selection methods into the model development process. Embedded methods perform automatic selection of features during training. Embedded methods lie between the wrapper method and the filter method in time complexity. Random forest classifiers and decision tree classifiers are among the embedded methods of feature selection.
- ***Hybrid Method:*** the hybrid method combines the best properties of wrappers and filters. First, the filter method minimizes the feature dimension, then the wrapper method selects the best candidate subset based on the ML algorithm.

Before using the collected data, it is important to select important features because not all features contribute equally to the result of classification. Selection and extraction of relevant features reduce overfitting of the model, simplify the model, and reduce the training time it takes. In this work, we used random forest classifiers for feature selection of the medical record dataset.

### 2.9.1. Random Forest Classifier

Random forest classifier algorithms are combinations of decision trees. Random forest classifiers improve the classification accuracy of a single tree by adding bootstrap aggregation and randomization methods during the construction of the decision tree classifier (Chen et al., 2020). Gini impurity or information gain is used as a measure of impurity. Features with lower impurity levels are important, and vice versa. Our medical record dataset contains both categorical and continuous independent variables. Random forest classifiers can handle both categorical and continuous data. We have selected random forest classifiers because of their high accuracy, better generalization, and interpretability.

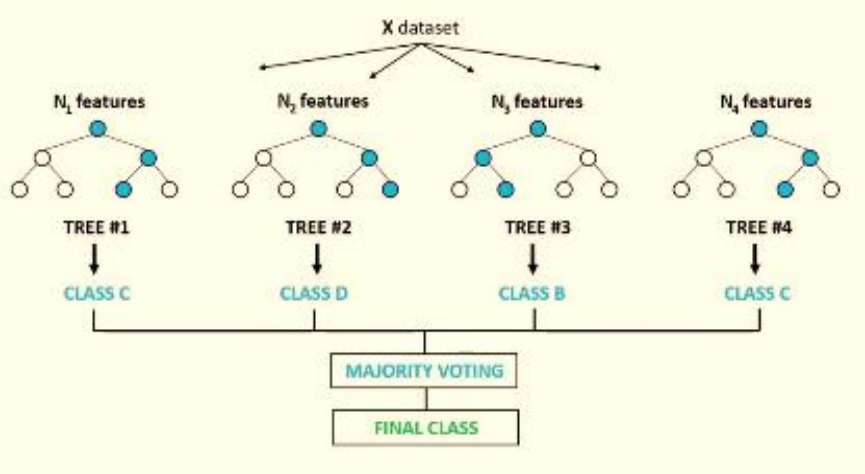



Figure 7: Random Forest Classifier[12]

### 2.9.2. Chi-Squared Test

A chi-squared test is used to compare values of categorical data. It is not possible to make a comparison between continuous data or categorical and continuous data. For the multilayer perceptron algorithm, the chi-squared test is the most effective feature selection algorithm according to Karabulut et al., 2012 (Karabulut et al., 2012). A very low chi-square result indicates a strong correlation between the observed and expected value, whereas a very high chi-square result means there is no relationship or does not fit the data between the observed and expected value.

[12] https://medium.com/analytics-vidhya/random-forest-classifier-and-its-hyperparameters

$$x_c^2 = \sum \frac{(O^i - E^i)}{E^i}$$

Where: c = degree of freedom

O = observed value(s)

E = expected value(s)

Equation 1: Chi-Square Equation[13]

## 2.10. Related Works

This section presents and discusses previous works done to detect pulmonary disease using machine learning and deep learning approaches.

Enes Ayan and Halil M Univer (Ayan & Ünver, 2019) used Xception and Vgg16 CNN models to develop a pneumonia detection model from chest x-rays. The reason for using these two models was that both Xception and Vgg16 have better performance weights on the ImageNet dataset. Transfer learning and fine-tuning approaches were employed to acquire good performance on the limited number of datasets. They have fixed some early layers and trained the model excluding these fixed layers. To avoid overfitting, batch normalization, dropout, and data augmentation methods were applied to different layers of the model. The accuracy of Vgg16 was 0.87% and Xception was 0.82%, but Xception exceeded Vgg16 in the detection of pneumonia cases. This shows that every algorithm has its own capabilities on different datasets. In this work, if the strength of the two models were combined, it would result in a better performing model to detect pneumonia cases.

Rajpurkar et al. (Rajpurkar et al., 2017) developed a model named CheXNet that can detect 14 diseases including pneumonia from chest x-ray images at a level that outperforms radiologist's work. CheXNet is a 121-layered CNN that takes a chest x-ray image and gives the likelihood of pneumonia and indicates portions of the image that are most suggestive of pneumonia. Before being input to the network, the images were downscaled to 224x224 and normalized. In addition, the horizontal flipping image augmentation method is used. The model has been trained on the

[13] https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning

chest x-ray 14 dataset with some modifications in the dataset to detect all 14 diseases. The performance of CheXNet exceeded the average performance of radiologists on the F1 score. The authors stated their limitations as using only the frontal chest x-ray images and being unable to include patient history in addition to chest x-ray images. In this study, we want to confront that detecting many diseases on a single model can lead to the wrong detection of diseases that are not found.

Verma et al. (Verma et al., 2020) pointed out that pulmonary infectious diseases are difficult to diagnose (Rahman et al., 2020) because they mimic each other. As a result of this, the patient will suffer from misdiagnosed treatment. A case of a patient with tuberculosis misdiagnosed as pneumonia that resulted in death is presented. This paper aimed to propose a framework to classify PTB, bacterial pneumonia, and viral pneumonia from chest x-ray images using image processing techniques. The Neural network classifier and Inception V3 were used for the analysis of images and image embedding respectively. For training purposes, the Shenzhen chest x-ray dataset and the China dataset on pneumonia are used. Rescaling, rotation, width and height shift, shear range, zoom range, and horizontal flipping augmentation techniques are applied. The author stated that taking the depth of the image in addition to the width and length of the x-ray image has improved the performance of the model. Not including patient history is the undeniable limitation of this work.

Aiming to improve CNN's low performance on rotated and tilted image orientations Bharati et al. (Bharati et al., 2020) proposed a new hybrid deep learning algorithm technique called VDSNet which combines CNN, VGG, data augmentation, and spatial transformer network (STN). This work conducted binary classification using chest x-ray images and x-ray view position, records of gender, and age of the patient. The developed model performed at a validation accuracy of 73%. In this work, no image augmentation technique is utilized. The use of patient age and gender attributes has improved the accuracy of the model. The model developed only detects whether the lung disease is found or not; it does not identify which lung disease is detected. This dragged the usability and practicality of the model one step backward since in real life medical diagnosis, only the detection of lung disease does not directly lead to the treatment of the disease and requires examining the x-ray again to identify the specific disease.

Hooda et al. (Hooda et al., 2017) proposed a deep CNN-based method for TB detection from chest x-ray image data. The reason behind developing this research was that traditional CAD systems use handcrafted features, but deep learning automatically extracts and learns the best features. A 19-layered simple deep CNN architecture has been developed. The proposed architecture consists of convolutional, Relu, FC, and dropout layers. Chest x-ray images used in this work are taken from Shenzhen and Montgomery public datasets by reducing the dimensions to 224x224 size. A total of 800 chest x-ray images were used, of which 600 images for training and 200 images for validation. Adam, SDG, and momentum optimizers were compared and the Adam optimizer outperformed, with overall accuracy and validation accuracy obtained of 94.73% and 82.09%, respectively. No image augmentation technique was performed in this research. Even if more images are used than in other works compared, no significant performance improvement is recorded. We believe that applying image augmentation and including patient history data would increase the accuracy of the model.

Haloi et al. (Haloi et al., 2018). This work has developed a pneumonia and TB classification model by modifying the architecture of a 21-layered deep network for traffic sign classification by one of the authors M Haloi (Haloi, 2015). The model developed is a 211-layered deep CNN model. The dataset used was a collection of four publicly available chest x-ray datasets. A partial attention mechanism was used on inception classifier feature maps. Disagreement values between the model and radiologists are used as feedback to optimize the network parameters. A pre-trained model on the chest x-ray 14 dataset and data augmentation are used to improve the performance of the model. Data augmentation techniques of random flips to the left or right, up or down, and changing the pixel values randomly are applied in this work. The performance of this model exceeded that of CheXNet (Rajpurkar et al., 2017) in detecting pneumonia cases. Overexposed chest x-rays applied for the development of the model have affected the performance of the model. The overlapping domain of pneumonia and tuberculosis is considered in this work; the model can identify a chest x-ray image that contains both pneumonia and tuberculosis pathological features. This model only uses x-ray images to detect pneumonia and TB; no associated medical history of the patient is used.

Khatibi et al. (Khatibi et al., 2020) proposed a two-step automatic method of pneumonia and TB detection using a stacked ensemble classifier. In the first step of the model, the symptoms

and demographic factors of 191 patients are imputed and features are selected by using stacked ensemble classifiers; the first step classification is also performed here. On the second step, the chest x-ray reports and laboratory results from the first step are applied to the stacked classifier again. The ensemble classifiers applied are AdaBoost, random forest, and gradient boosting. Single classifiers like decision tree (DT), SVM, linear regression, and KNN are also applied to compare their performance. Another aim of developing a two-step model is to save the time of waiting for laboratory results and radiographic reports, in which this work can first show the result based on demographic information and patient symptoms in the phase one output. In this study, the radiology reports are used instead of the radiology images. This makes the developed system dependent on the decisions made by the radiologists directly, and this leads to biased results and variation in the views of the radiologists.

Muse Ibrahim (Muse Ibrahim, 2020) applied region based convolutional neural network to develop computer aided diagnosis model for TB. The author used microscopic sputum smear images to develop the TB detection CAD model. Mask R-CNN, Hungarian algorithm and hard example mining are implemented consecutively. Mask R-CNN is used for instance segmentation, Hungarian algorithm is implemented to identify misclassified predictions and hard example mining is also used to provide hard examples. The author did not apply any patient history in addition to the microscopic sputum images.

## 2.11. Summary of the Related Works

Table 1: Summary of the Related Works

| *S.no* | *Author /Year* | *Objective* | *Algorithm Used* | *Gap* |
|---|---|---|---|---|
| 1 | (Ayan & Ünver, 2019) | ➢ Improve computer-aided diagnosis system in detecting pneumonia | Xception<br>Vgg16 | ➢ Used only chest x-ray image for pneumonia diagnosis |
| 2 | (Rajpurkar et al., 2017) | ➢ Detect 14 diseases including pneumonia from chest x-rays in more accuracy than radiologists | DenseNet | ➢ Used only chest x-ray images for pneumonia detection<br>➢ Detecting many diseases on a single model may lead to false disease detection (Tilve et al., 2020) |
| 3 | (Verma et al., 2020) | ➢ Classify between pulmonary tuberculosis, bacterial pneumonia, viral pneumonia from chest x-ray images | Neural Network Classifier<br>Inception v3 | ➢ Applied only chest x-ray images for pneumonia and TB detection |
| 4 | (Bharati et al., 2020) | ➢ Develop a framework for better detection of lung disease for abnormal image orientations. | CNN<br>Vgg | ➢ Only age and gender attributes are used, and no other medical records are included.<br>➢ Do not differentiate between the diseases detected.<br>➢ No image augmentation technique is utilized. |

| | | | | |
|---|---|---|---|---|
| 5 | (Hooda et al., 2017) | ➢ Develop TB detection model using Deep CNN from chest x-ray image | CNN | ➢ Only chest x-ray images are used.<br>➢ No image augmentation technique is utilized. |
| 6 | (Haloi et al., 2018) | ➢ Develop pneumonia and tuberculosis detection model from chest x-ray images | Deep CNN | ➢ Patient record data is not considered in the detection model.<br>➢ Overexposed chest x-ray images are included. |
| 7 | (Khatibi et al., 2020) | ➢ Two step decision support system for pneumonia and TB identification | Stacked ensemble classifiers (KNN, LR,SVM, DT, RF) | ➢ The chest x-ray interpretations are used images are not processed directly<br>➢ Small number of data is used |
| 8 | (Muse Ibrahim, 2020) | ➢ Developing Computer aided TB diagnosis model by using region based convolutional neural network | Region based CNN | ➢ No medical record data is included |

From the related works presented in this section, we can infer that almost all of them used chest x-ray images to detect one or more pulmonary diseases. The works have applied CNN as the basis of the work and different pre-trained models are also utilized. Horizontal flip and vertical flip are not medically recommended data augmentation techniques used in these works because they can change the position of organs and produce unrealistic chest x-ray images (Elgendi et al., 2021). The works presented only utilized chest x-ray images for pulmonary disease detection. Only one study tried to include patient information like age and gender with the view position of the chest x-ray image, but this work failed to distinguish the disease. The last work by Khatibil et al. applied medical records and radiology reports to classify pneumonia and TB. This work does not directly use the radiology image, so it is highly biased by the decision written by the radiologist on the radiology report. Generally, the use of chest x-ray images alone for the detection of pulmonary disease fails the medically right diagnosis scheme and can lead to under-diagnosis or misdiagnosis, while the use of medical records alone is also not sufficient for proper medical diagnosis. So, it is crucial to include the patient's medical record along with the chest x-ray image for accurate and proper detection of disease using deep learning techniques. Moreover, this research work seeks to fill these gaps.

# CHAPTER THREE

## 3. MATERIALS AND METHODS

### 3.1. Chapter Overview

In this chapter, the methodology of the study is presented. The proposed workflow, the sources of datasets and the preprocessing steps are described. Moreover, the model evaluation methods are also presented.

### 3.2. Research Design

This research design falls under quantitative experimental research because we carry out experiments and manipulate independent variables to find out the effect on the dependent variable (Kamolson, 2007). Data collection, data preprocessing, development of the mathematical model, and workflow of the research are the main decisions to perform the research.

A specific workflow system is designed to carry out this research, as shown in *figure 8*. The workflow steps started with identifying the research area and then exploring the research done on the identified research area to identify the research gap. Identifying the research gap is followed by formulating the problem statement, developing research questions and writing a research proposal. After writing the research proposal, the needed data is collected from different sources. The data collected is raw and needs preprocessing. We train the model by employing the selected algorithm after preprocessing the data. After model training, model evaluation follow. Finally we develop a graphical user interface and deploy the solution to make it convenient to use.

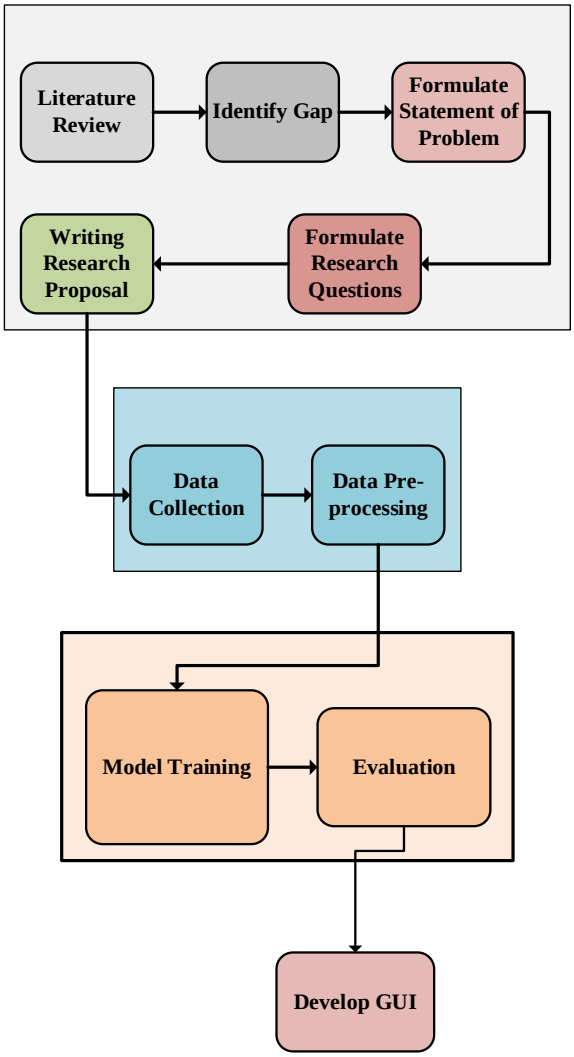



Figure 8: Work flow Diagram of the Research

### 3.2.1. Data Collection

Deep learning and machine learning algorithms learn the patterns and the relationships between dependent and independent variables. They use the datasets as a reference to make predictions and detections on the new data that is given. In this research, we have used two types of datasets: medical records and chest x-ray images. The medical record dataset is collected from St Peter's specialized hospital, Alert hospital, and Yekatit 12 hospital medical college. The chest x-ray image dataset is collected from St. Peter's specialized hospital. Due to the unavailability of enough local chest x-ray image data, we used datasets from the publicly available Kaggle repository in addition to St. Peter's specialized hospital.

#### 3.2.1.1. Chest X-ray Image Dataset

Chest x-ray images of associated patient medical records are collected from St. Peter's specialized hospital filmless radiology system. This radiology reading and storage system contains x-ray images since August 2021 at the time of our visit. By using the Medical Record Number (MRN) of the patients diagnosed with pneumonia or TB, 482 images are collected from the system. The dimensions of local chest x-ray images range from 2832x2836 to 720x635, while datasets collected online have dimensions in the range of 512x512 to 2564x2519.

In similar manner chest x-ray images are also collected from Alert hospital by taking the MRN numbers of the patients diagnosed as TB to the filmless radiology system. At the time of our visit the system was renewed and do not contain sufficient amount of chest x-ray images. Among the medical records we collect only 33 of them found the associated chest x-ray image of the patient.

Table 2: Local Chest X-ray Image Sources

| | **Pneumonia** | **TB** | **Other** |
|---|---|---|---|
| **Alert hospital** | 0 | 33 | 0 |
| **Yekatit 12 hospital** | 0 | 0 | 0 |
| **St. Peter's Hospital** | 152 | 143 | 154 |
| **Total** | **152** | **176** | **154** |

The chest x-ray images collected are DICOM (Digital Imaging and Communication in Medicine) formatted images, which contain metadata to describe the image. DICOM is created to make images and associated data easier and safer to exchange and to avoid confusion (Mustra et al., 2008). The images in the system contain metadata about patient name, patient ID, age, sex, image number, hospital name and others. The zoomed view of information on the four edges of a single chest x-ray image is shown in figure 9 below.

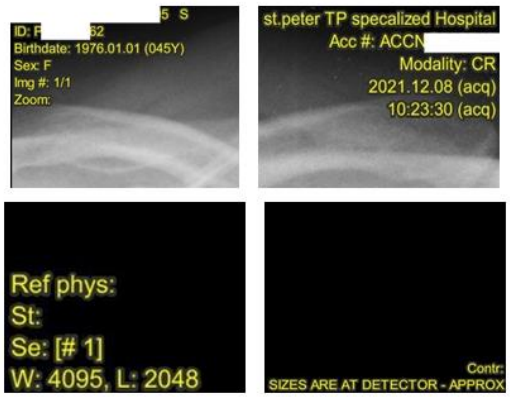



Figure 9: The Four Edges of a Single Local Chest X-ray Image

An additional 2463 chest x-ray images are collected from the Kaggle repository. ***The Tuberculosis (TB) chest x-ray database***[14]*:* The largest database of chest x-ray images for TB positive cases as well as normal images has been developed by a team of researchers from Qatar and Dhaka universities together with collaborators from Malaysia and Hamad Medical Corporation. This database contains 700 publicly available TB images, 3500 normal images, and 2800 TB images that can be downloaded on the NAID TB portal[15] after signing an agreement. From this database, we have collected 695 TB chest x-ray images.

The other database that we have collected pneumonia and normal chest x-ray images is the Mendely data version 2 on the ***Kaggle repository chest x-ray images (pneumonia)***[16] (Kermany et al., 2018). This database contains 5863 chest x-ray images with normal and pneumonia categories (Kermany et al., 2018). These chest x-rays were screened for quality, then low quality and unreadable images were removed and labeled by two expert physicians. From this database, we have collected 922 pneumonia and 846 normal chest x-ray images.

Table 3: Chest x-ray Dataset Description

| | **Local** | **Kaggle** | **Total** |
|---|---|---|---|
| **Pneumonia** | 152 | 922 | 1074 |
| **TB** | 176 | 695 | 871 |
| **Other** | 154 | 846 | 1000 |
| **Total** | 482 | 2463 | 2945 |

[14] https://www.kaggle.com/datasets/tawsifurrahman/tuberculosis-tb-chest-xray-dataset
[15] https://tbportals.niaid.nih.gov/
[16] https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia

#### 3.2.1.2. Medical Records Dataset

A patient medical record dataset is a dataset that comprises the symptoms of the patient, vital signs, and demographic and diagnosis information. The dataset was collected from St. Peter's specialized hospital, Alert hospital, and Yekatit 12 hospital medical college.

St. Peter's specialized hospital is a governmental hospital established in June 1961 GC (D. Abate et al., 2012). The hospital offers a variety of services, especially TB diagnosis and treatment services. ALERT (All Africa Leprosy and Rehabilitation Training Center) hospital was built in 1934 GC as zenebe werk memorial hospital by the Sudan interior mission (Mariam, 1986). The hospital focuses on serving as a specialized medical care and research and training center focused on leprosy, TB, HIV/AIDS, and other infectious diseases to improve the wellbeing of the community. Yekatit 12 Hospital was founded in 1923 GC as Haile Selassie I Hospital. After the Graziani massacre on February 19, 1937 GC (Yekatit 12 in the Ethiopian calendar), for consecutive three days, the hospital was named Yekatit 12 hospital as a memorial to thousands of lost civilian lives. In 2011 GC medical college opened. The hospital began implementing the EHR system in four phases in 2012 GC and is now 95 percent successful (Kitesa & Mamo, 2021). The collected data ranges from 2017 to 2022 GC of those hospital patients. The inclusion criteria of the data collection were patients who were with pneumonia or TB diagnosis. The collected data ranges from 2017 to 2022 GC of those hospital patients. The inclusion criteria of the data collection were patients who had pneumonia or TB diagnosis. The patient history folder contains a lot of information about the patient, but all of the information is not relevant to our goal. The selection of the important features of the patient data was performed by three physicians (medical doctors).

In the data collection from St. Peter's Specialized Hospital, the OPD register of OPD 83, OPD 22, OPD 21, OPD 24, and OPD 2 were used to filter out the (MRN) of the patients who were diagnosed with TB or pneumonia. The second step was taking those selected MRN numbers to the card room. The card room is a place where the patient record folder of any diagnosis in any year can be found. The data needed was collected according to the features selected by the physicians in csv format. At Alert hospital, the MRN of the patients diagnosed with TB is taken from the TB clinic unit TB register and tuberculosis patient referral register. Using the MRN numbers recorded, the patient folders are taken out from the shelf in the card room and recorded

with the needed information for each patient. In Yekatit 12 hospital medical college, the patient information is taken from the Electronic Medical Record (EMR) system. The patient data is collected from Medical A, Medical B, Medical C, Pediatric A, and Pediatric B classes in the EMR system. We have prepared csv dataset called *other* based on the normal ranges of WBC, PR, RR and temperature based on age range under the assistance of medical doctors. Other datasets contain normal medical record ranges and normal chest x-ray images. This class helps in creating boundaries between the two classes and normal findings by letting the model have knowledge of normal chest x-ray images and medical records.

Table 4: Medical Record Dataset

| | **pneumonia** | **TB** | **Total** |
|---|---|---|---|
| **Alert Hospital** | 0 | 328 | |
| **Yekatit 12 hospital** | 712 | 8 | |
| **St. Peter's hospital** | 362 | 535 | |
| **Total** | **1074** | **871** | **1945** |

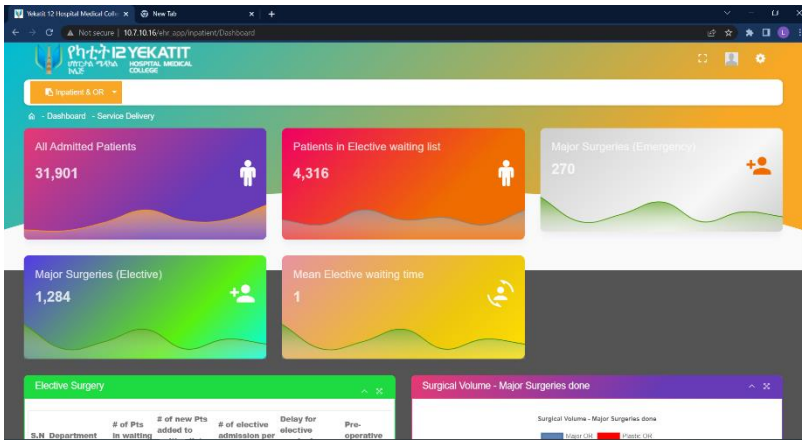



Figure 10: Sources of Medical Record Dataset
(EMR System and Card Room)

The number of medical records dataset is similar to the chest x-ray image dataset. Meaning, one row of medical records corresponds to a single chest x-ray image. The dataset distribution ratio of the three classes is shown below in figure 11.

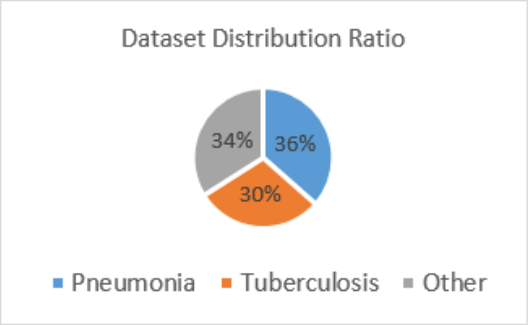



Figure 11: Medical Records Dataset Distribution Ratio

The distribution of the dataset based on gender is almost proportional. It is shown below on the graph.

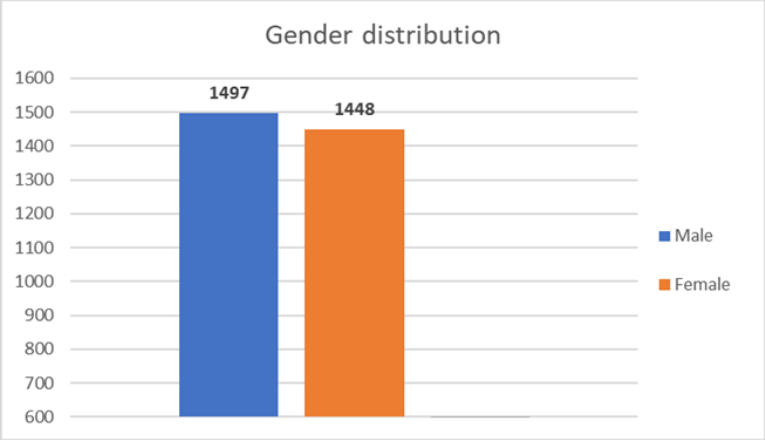



Figure 12: Gender Distribution of the Dataset

## 3.3. Data Preprocessing Techniques

Data preprocessing is taking raw data and making it in an understandable format for computers, machine learning, and deep learning models. Data preprocessing is a crucial step in any deep learning model development because if the data is not prepared well, the result will not be as good as it has to be. In this work, two different types of datasets are used. So, data preprocessing is performed separately according to the nature of the data.

### 3.3.1. Chest X-ray Image Preprocessing

After collecting chest x-ray images from local and online sources image preprocessing is performed. Image cleaning and image augmentation is performed.

#### 3.3.1.1. Image Cleaning

As mentioned in ***section 3.3.1*** the local chest x-ray images we have collected are in DICOM format and contain Meta data (textual information). To meet our goal and detect the pulmonary disease well, the textual information needs to be removed. This information can be removed by using OCR (Morales, 2020). The OCR is a combination of a convolutional neural network and a recurrent neural network, which is a convolutional recurrent neural network (CRNN). CRNN detects combinations of letters and characters from images. We apply the OCR library and clean the texts in the chest x-ray image.

#### 3.3.1.2. Image Augmentation

Image augmentation is a technique of modifying existing images in order to produce additional images for the model training process and enhance the model's performance (Khosla & Saini, 2020). Put differently, it is the method of artificially maximizing and varying the dataset that can be used to train deep learning models. Image augmentation can also be applied to get a variety of images to enhance the performance of the model when facing new images without increasing the number of datasets. When we perform augmentation of medical images, we need to be aware that we are not performing augmentation that alters the result of the classification and that creates ambiguity in identifying the class of the model.

### 3.3.2. Medical Record Data Preprocessing

After collecting the medical record dataset in csv format, data preprocessing involves different techniques as described below. Different units are converted to same unit then data cleaning and transformation is performed.

#### 3.3.2.1. Converting to Similar Units

At the start of medical record data preprocessing, the column having records of different units will be converted to a similar unit. In the age column, ages are recorded in years and months for kids under the age of one. These records are converted to years. Cough duration has been recorded as days, weeks, and months of duration. Months and days are converted to weeks.

#### 3.3.2.2. Data Cleaning and Transformation

In the data cleaning preprocessing step, wrong values are cleaned and substituted. Also, the data is checked for outliers. The dataset collected in tabular form contains different data types. Ordinal encoding, label encoding, and one-hot encoding are used to transform data into a computer-understandable format.

#### 3.3.2.3. Feature Selection

Not every independent variable contributes to the outcome of the dependent variable or detection of the model. Selection of relevant features contributes to the reduction of training time and improves the performance of the model. In ***section 2.9*** different feature selection techniques are discussed. The random forest classifier is suitable and better performing technique for structured data having a mixture of continuous and categorical datatype features. So we performed feature selection by using random forest classifier. According to the output of the random forest classifier the less relevant features are removed.

## 3.4. Algorithms for Pulmonary Infectious Disease Detection Model

We have tabular and image datasets; we need to process those datasets and join them for classification. Supervised deep learning algorithms are selected to perform the task. We choose deep learning algorithms over machine learning because deep learning algorithms can do deep feature extraction by themselves without losing the important information in the data.

### 3.4.1. Functional Neural networks

In sequential neural networks the weights of the previous layer are learned instead of the function itself. But in functional neural networks the neural function itself is learned by suppressing the weights (Castillo, 1998). The input layer of the functional neural networks holds the input dimension of the dataset. Also the output in one unit can serve as the input to next unit or input in the output layer. In functional networks information flows in only forward direction from input layer to the output layer. The final output of these networks can be outputs from one or more neurons but in sequential neural networks the final output is the output of a single neuron

(last neuron). In this work we apply the functional neural network structures of selected neural networks to develop our model.

### 3.4.2. Convolutional Neural Network (CNN)

Different neural networks can be used for image classification, but CNN is the field of choice. Convolutional neural network algorithms are the main deep learning algorithms for image detection and processing. In this research, chest x-ray image processing is developed using CNN.

### 3.4.3. Multi-Layer Perceptron (MLP)

Machine learning algorithms perform well on tabular data but in order to make suitable for the model concatenation and deployment we preferred using neural network (Kadra et al., 2021). A multi-layer perceptron is a type of artificial neural network that can be used for classification and regression tasks. Complex neural networks cannot perform well on tabular data. Multilayer perceptron can perform well on tabular data classification. Due to its simplicity, good performance on tabular datasets, and good achievements in the medical field (Chaitrali S. & Sulabha S., 2012), MLP is selected to process medical record tabular dataset in this work.

## 3.5. Design and Development Tools

To develop this research various design and development tools are used. A design tool is used to draw flowchart and diagrams in different sections of this paper. Different development environments and libraries are also employed to build up this research. These design and development tools are briefly presented in this section.

### 3.5.1. Design Tool

- ***Microsoft Visio Professional 2016:*** Microsoft Visio Professional 2016 is a simple and free to use drawing tool which contains different shapes to draw various kinds of diagrams and flowcharts. Microsoft Visio Professional 2016 is used as a design tool for designing diagrams and flow charts in this research.

### 3.5.2. Development Tools

- ***Tensor flow***[17]*:* A complete open-source platform for machine learning applications. It is a symbolic math toolkit that carries out several operations targeted at deep neural network training and inference using dataflow and differentiable programming.
- ***Keras***[18]: A Python-based deep learning API running on top of the Tensorflow machine learning platform for implementing neural networks.
- ***Google Drive***[19]***:*** Online file access and storage made possible via a free cloud-based storage service. The datasets are stored on google drive to perform data preprocessing and model training on google Colab.
- ***Google Colab***[20]***:*** A cloud-based Jupyter notebook online platform that needs no setup to use while offering free access to computation resources such as GPUs. The model will be trained on Colab due to the free access to GPU and the ability to integrate Keras and tensor flow libraries.
- ***Matplotlib***[21]***:*** Open-source plotting library in python for data visualization and 2D plotting. Accuracy and loss plots, confusion matric plots and other plots for data visualization are carried out by using the Matplotlib library.
- ***Anaconda***[22]***:*** Open-source python platform with thousands of free open source packages and libraries. Anaconda is easily manageable for deploying any project.

## 3.6. Evaluation Methods

Evaluating the developed model is important to emphasize the performance and identify the strengths and weaknesses. Identifying the performance helps in improving the performance and avoiding actions that decrease the performance in the future. In this work, algorithmic evaluation and senior radiologists (doctors) evaluation methods are used.

[17] https://www.tensorflow.org/
[18] https://keras.io/
[19] https://www.google.com/drive/
[20] https://colab.research.google.com/
[21] https://matplotlib.org/
[22] https://www.anaconda.com/

### 3.6.1. Algorithmic Evaluation Method

#### 3.6.1.1. Cross-Validation with Confusion Matrix

K-fold cross-validation is a well-liked evaluation metric for classification algorithms. The count of folds, the count of instances in the fold, and the repetition of k-fold cross validation affect its performance (Wong, 2015). So, carrying out experiments on the common values of 3, 5, and 10 folds in the K fold cross-validation technique is an unquestionable task. These 3, 5, and 10 values of K will partition the data into 3, 5, and 10 divisions respectively and use the train and test sets interchangeably. In this study, we apply k-fold cross validation by applying the common k values and selecting the best performing k value by employing a confusion matrix. A confusion matrix is a table that depicts the performance of a classification model. A confusion matrix contains the numerical values of the following four elements:

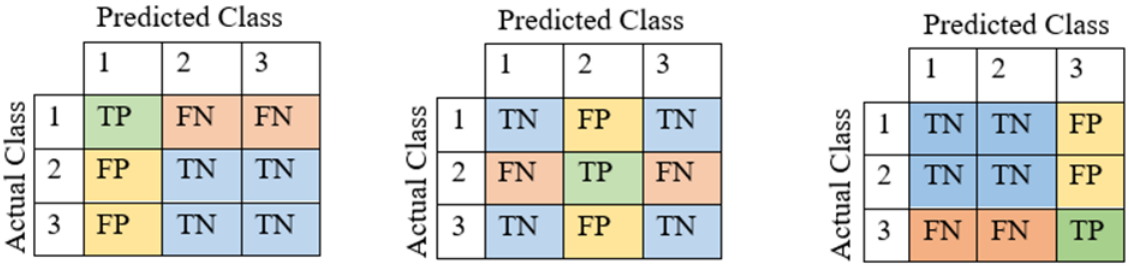



Figure 13: Confusion Matrix for Class 1, 2 and 3 Respectively

True Positive (TP) = both actual and predicted values are positive

True Negative (TN) = both actual and predicted values are negative

False Positive (FP) = actual values are negative and predicted values are positive

False Negative (FN) = actual values are positive and predicted values are negative

***The accuracy*** of the model is the proportion of correctly identified results out of all identified results it can be calculated from the confusion matrix as:

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN}$$

Equation 2: Accuracy Performance Measure

A machine learning algorithm learns from the data given to make decisions. To figure out how well the developed model performed, it has to be evaluated using machine learning evaluation metrics. We cannot depend only on the accuracy evaluation technique because it has limitations in the evaluation and discrimination process and is less informative (Hossin & Sulaiman, 2015). Therefore, in this work, different evaluation metrics are applied to evaluate the performance of the model in addition to the accuracy metric.

***Precision:*** the proportion of the truly positive results out of all positive results.

$$\text{Precision} = TP/TP + FP$$

Equation 3: Precision Performance Measure

***Recall:*** the proportion of the truly positive results out of all actual results.

$$\text{Recall} = TP/TP + FN$$

Equation 4: Recall Performance Measure

***F1-score:*** takes precision and recall in order to reach at a single measure.

$$\text{F1-score} = 2\ Precision * Recall/(Precision + Recall)$$

Equation 5: F-score Performance Measure

### 3.6.2. Senior Radiologist (Medical Doctor) Evaluation

A medical doctor's evaluation is the evaluation made by the medical doctor by using the data in their hand. After comparing the decision made by the model with the decision of the senior radiologist, the accuracy of the model can be calculated.

# CHAPTER FOUR

## 4. PROPOSED SOLUTION

### 4.1. Chapter Overview

In this chapter, the proposed solution for pulmonary infectious disease detection is presented. The proposed model architecture, selected feature selection technique, medical record data preprocessing techniques, and chest x-ray image preprocessing techniques proposed are also presented along with diagrams and descriptions.

### 4.2. Proposed Solution Architecture

In this work, functional convolutional neural network and multilayer perceptron are used to develop the pulmonary infectious disease detection model.

The medical record dataset collected in the csv file contains the chest x-ray image file name in a column called 'IMG path'. Excluding this IMG path column, medical record preprocessing is performed by applying different data preprocessing techniques. The chest x-ray image dataset is also preprocessed and loaded according to the path found in the IMG path column of the csv file. In functional neural networks, the input layer takes the input dimension of the data. The multilayer perceptron's input layer takes the number of features of the medical records dataset. Also, the input layer of the CNN holds the input dimension of the chest x-ray images, which is 256x256x1.

Then, according to the input dimension of the medical record, the multilayer perceptron processes the medical record dataset. The output of the multilayer perceptron is a one-dimensional vector. The CNN also processes the chest x-ray image. The flattening layer is applied at the end of CNN to get a one-dimensional feature vector output. The one-dimensional feature vector outputs from CNN and MLP are concatenated using a concatenation layer. The concatenated feature vector becomes a one-dimensional vector and its length is the sum of the lengths of feature vectors applied for concatenation. The concatenated feature vector is then fed to the fully connected layers for further processing and output, providing the final classification result.

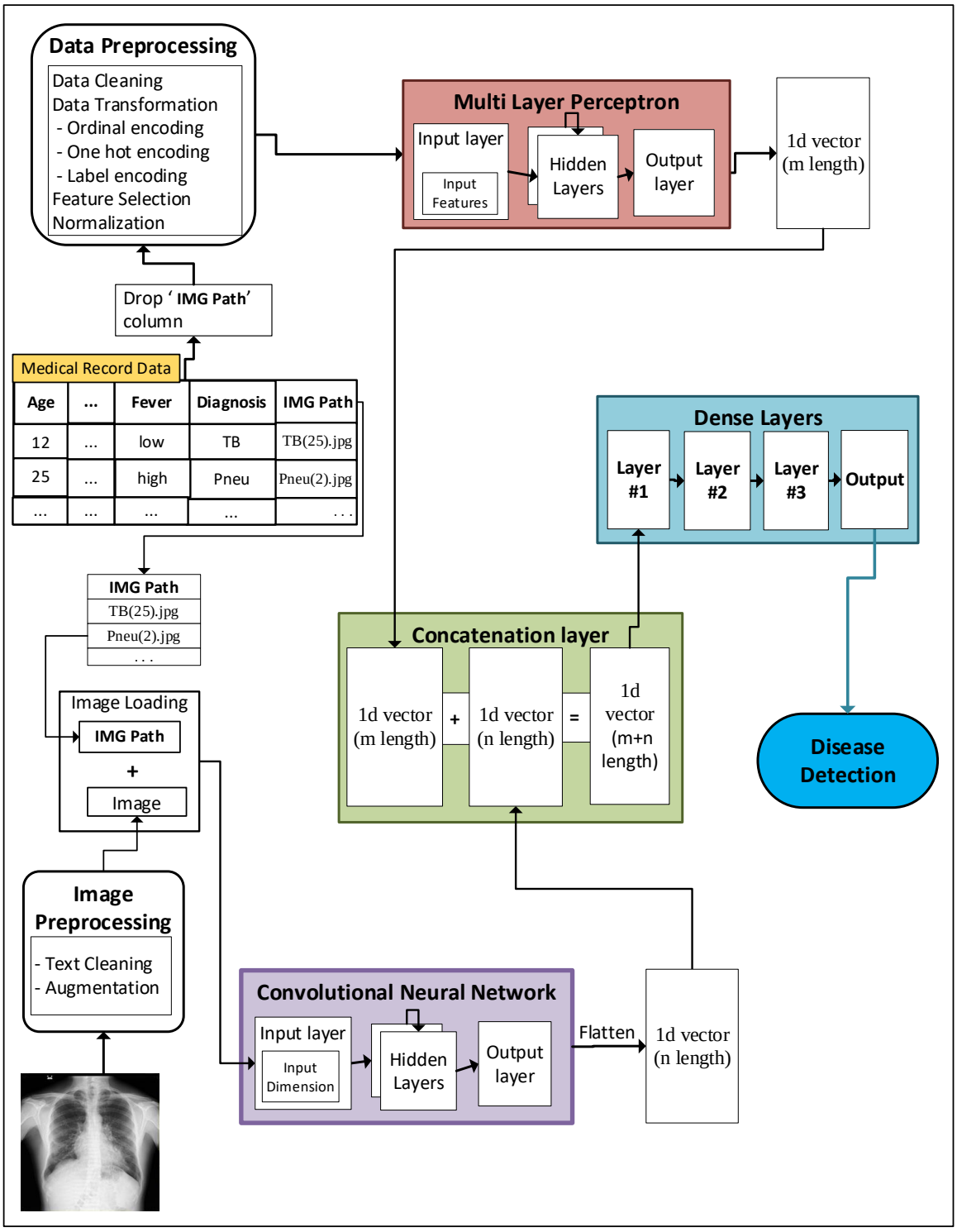



Figure 14: Proposed Solution Architecture

## 4.3. Chest X-ray Image Pre-Processing

The chest x-ray images collected are RGB images with 3 channels. These images are converted to greyscale to minimize complexity and reduce computational requirements. In image preprocessing, Image cleaning and image augmentation techniques are applied to the chest x-ray images. Meta information on the local chest x-ray images is cleaned using OCR and then, image augmentation is performed. Figure 14 shows the sample local chest x-ray images before preprocessing.

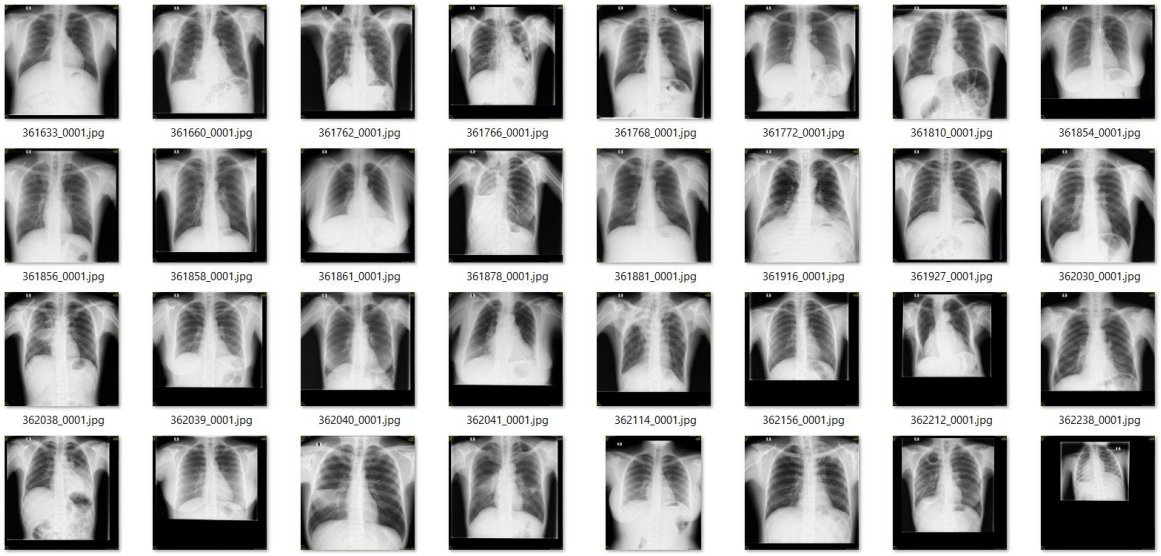



Figure 15: Sample Local Chest X-ray Images before Preprocessing

### 4.3.1. Image Cleaning

Local chest x-ray images include textual information about the patient and the hospital name, which is removed using OCR to reduce image noise because it is irrelevant to our work.

### 4.3.2. Image Augmentation

After cleaning, image augmentation techniques are applied to the chest x-ray images. According to (Sirazitdinov et al., 2019) rotation, horizontal flip, and random brightness augmentation techniques applied to chest x-ray images have shown better performance gain on the model. Rotation between -5 and 5, an equal scaling in x and y axis, and translation (height shift and width sift) are stated as medically recommended augmentation techniques, and horizontal flip is not recommended medically because it can lead to non-physiological image (Elgendi et al.,

2021). Pivoting from these two papers, we selected rotation between -5 and 5, random brightness, height shift, and width shift augmentation techniques to perform on our chest x-ray image dataset. We apply image augmentation to only get variety of versions of the original dataset without increasing the dataset size. We do not increase the size of the dataset because if the number of chest x-ray images and medical record rows is not equal, concatenation cannot be performed. Half of the chest x-ray images collected from Kaggle and all of the local images are taken to perform image augmentation.

**Algorithm: Image Augmentation**
**Input:** m number of chest x-ray images of a class
**Output:** Augmented m number of chest x-ray images with size 512x512
**Begin**
Batch size =1
Convert to greyscale
image= image/255 // normalization
Apply augmentation techniques(different combinations)
    Random brightness=(0.3, 0.9)
    Width shift = 0.2
    Height Shift = 0.2
    Rotation= -5 degree
**End**

Algorithm 1: Image Augmentation

## 4.4. Medical Record Data Pre-Processing

Our medical record dataset contains 2945 rows with 18 columns. We perform different structured data preprocessing techniques to make the dataset ready for model building. Dataset features with their symbols used in this work along with their data types are presented in table 3 below.

Table 5: Features with Their Data Types

| **Feature** | **Symbol** | **Data Type** |
|---|---|---|
| Age | Age | Continuous |
| Sex | Sex | Categorical(Nominal) M - 0 F - 1 |
| Temperature | Temp | Continuous |
| Respiratory Rate | RR | Continuous |
| Pulse Rate | PR | Continuous |
| Cough duration in weeks | Cough_D | Continuous |
| Cough type | Cough_T | Categorical(Ordinal)<br>0 - dry 1 - whitish<br>2 - yellowish 3 - bloody |
| White Blood Cell count | WBC | Continuous |
| Chest Pain | CP | Categorical(Nominal) 0- no 1- yes |
| Shortness of Breath | SOB | Categorical(Nominal) 0- no 1- yes |
| Oxygen Saturation | SPO2 | Continuous |
| Fever | Fever | Categorical(Ordinal)<br>0 - no 1 - low 2 - high |
| Night Sweat | NS | Categorical(Nominal) 0- no 1- yes |
| History of TB | HTB | Categorical(Nominal) 0- no 1- yes |
| Contact with Chronic Cougher | Contact | Categorical(Nominal) 0- no 1- yes |
| History of Smoking | HS | Categorical(Nominal) 0- no 1- yes |
| Weight Loss | WL | Categorical(Nominal) 0- no 1- yes |
| HIV test result | HIV | Categorical(Nominal) 0 - P 1 - N |
| Diagnosis | Diagnosis | Categorical(Nominal)<br>0 - other 1 - pneumonia 2 - TB |

### 4.4.1. Data Cleaning

***Unit Conversion:*** we converted different units used to record values of a column to one unit. In the age column, the ages of children below age 1 are recorded in months. Such records are converted to years by dividing the month by 12. The cough duration of patients is also recorded

by days, weeks, months, and years. Days, months, and years are converted to weeks. Days are divided by 7, months are multiplied by 4.345 for an approximate result, and years are multiplied by 52.143. We have checked for outliers, and outliers are removed.

### 4.4.2. Data Transformation

We transform the medical record dataset categorical features based on their data types shown in table 3. We apply categorical encoding, label encoding, and one-hot encoding data transformation techniques. The data transformation flow chart is shown in figure 16.

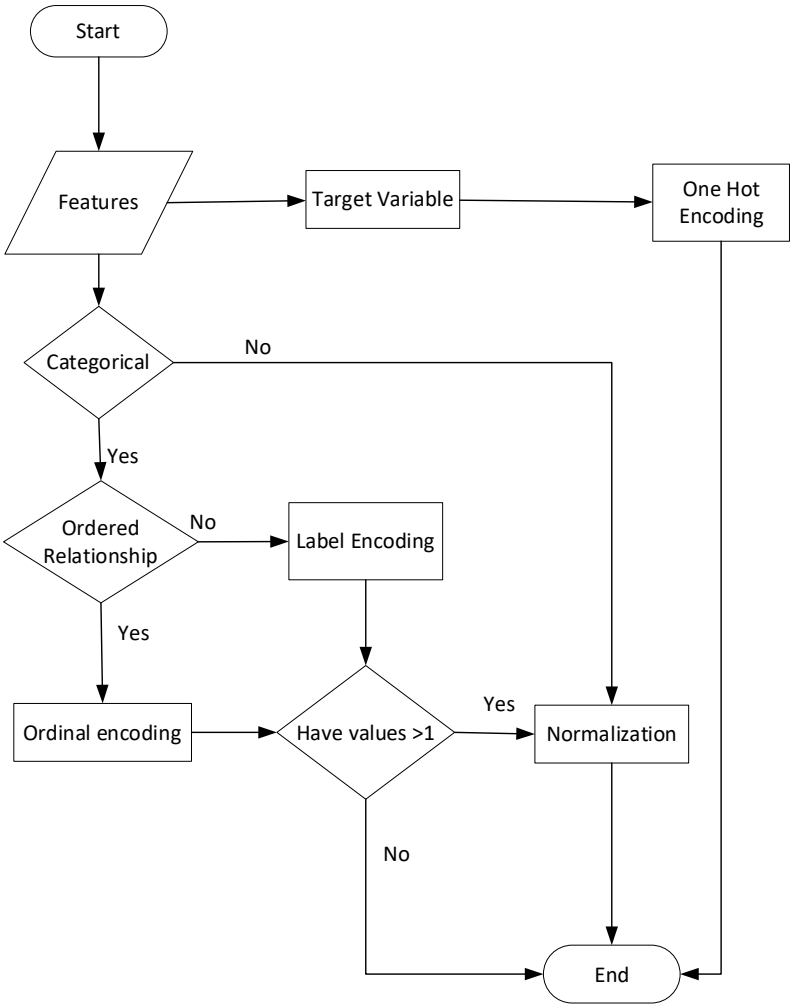



Figure 16: Data Transformation Flowchart

- **Ordinal encoding:** for the categorical data types that have ordered relationships, we applied ordinal encoding. Cough type and fever are ordinal categorical features where ordering of values matters.

- **Label encoding:** Label encoding is applied to encode categorical values that do not have an ordered relationship. Label encoding is used for features such as sex, chest pain, shortness of breath, night sweat, history of tuberculosis, contact with a chronic cougher, history of smoking, weight loss, and HIV test results.
- **One hot encoding:** One hot encoding is used to encode each output value into a new column by creating dummy variables. We applied one hot encoding to the target variable diagnosis to let the model identify each class more clearly.

### 4.4.3. Data Normalization

Data normalization allows the values to have a common scale without distorting the variations in values. We perform the normalization technique by using the absolute maximum scaler technique for columns having values greater than 1. The absolute maximum scaler scales the data points by dividing them by the absolute maximum value of the features. Normalization techniques are used on age, temperature, respiratory rate, pulse rate, cough duration, white blood cell count, oxygen saturation, fever, and cough type. After normalization, every value scales between 0 and 1.

$$d(scaled) = \frac{d}{\max(|d|)}$$

Equation 6: Absolute Maximum Scaler

### 4.4.4. Feature Selection

After the data was cleaned, we employed feature selection to select features that contributed to improving the model's performance. A random forest classifier is used as a feature selection technique. A random forest classifier develops a forest of individual decision trees and makes a decision based on the matched results. The random forest classifier outputs the importance of each feature to the target classes. We become able to identify the more relevant features and less relevant features. As a result, we selected features that have more importance to the model and removed less important features.

## 4.5. Algorithms for Proposed Model Development

Convolutional neural networks and multilayer perceptron are applied to develop a pulmonary infectious disease detection model.

### 4.5.1. Multi-Layer Perceptron

The multilayer perceptron is the processing algorithm for the medical record data. First, the input dimension of the medical record data is defined in the input layer of MLP. The first hidden layer contains 256 neurons with an activation function relu and takes the output vector of the input layer. The second hidden layer has 128 neurons with the same relu activation function as the previous one. The output vector of the second hidden layer is fed to the third hidden layer of 64 neurons. The output vector of the third hidden layer is fed to the output layer with 32 neurons, and the output of this layer, a 1 dimensional feature vector, is the output of this network. Now, the output feature vector of the MLP is ready for concatenation with the feature vector coming from CNN.

**Algorithm: Multilayer Perceptron**

**Input:** I with 17 medical record features

**Output:** 1 dimensional vector

**Begin**

Input layer: $I = 17\ features$

First hidden layer

$f_i^1 = ((a_j^1 = \sum_{i=1}^{256}\ w_j^1 + b_i^1,\ h_i^1 = max(0, a_j^1))(I))$

second hidden layer

$f_i^2 = ((a_j^2 = \sum_{i=1}^{128}\ w_j^2 + b_i^2,\ h_i^2 = max(0, a_j^2))(f_i^1))$

third hidden layer

$f_i^3 = ((a_j^3 = \sum_{i=1}^{64}\ w_j^3 + b_i^3,\ h_i^3 = max(0, a_j^3))(f_i^2))$

output layer

$f_i^4 = ((a_j^4 = \sum_{i=1}^{32}\ w_j^4 + b_i^4,\ h_i^4 = max(0, a_j^4))(f_i^3))$

**End**

Algorithm 2: Multilayer Perceptron

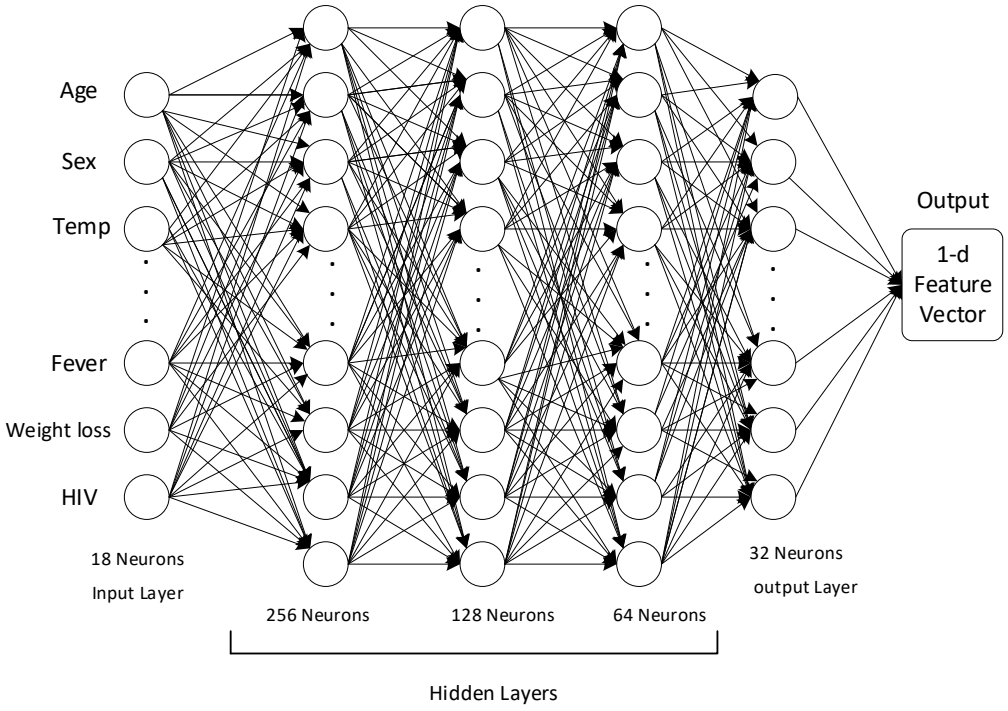



Figure 17: MLP Architecture

### 4.5.2. Convolutional Neural Network (CNN)

A convolutional neural network is used for the processing of chest x-ray image data. A 12 layer convolutional neural network with 5 additional layers of flatten and fully connected is utilized as an image processing pipeline.

- **Input Layer:** the input layer of a convolutional neural network takes the input dimension of the image.
- **Convolutional Layer:** has neurons that connect the sub-parts of the input image. The convolutional layer performs convolution calculation to the input image and calculates the convolution of each image with each filter. The convolution layer used in our CNN model is a 2D convolution layer having a filter size of 256, 128, 64 and 32 with activation function relu. This 2D convolution layer uses a kernel of size (7, 7) that slides over our 2D image input and creates different 2D feature matrices from a single 2D matrix.
- **Pooling Layer:** the pooling layer follows the convolution layer. It reduces the size of the image that outputs from the convolution layer and reduces the number of parameters, reduces model complexity and improves the efficiency of the model (Sarvamangala & Kulkarni, 2022). The type of pooling we used is max pooling, which takes the maximum

output. Pool size (2, 2) is employed. This means a 2 x 2-dimensional filter slides over feature maps and takes the maximum element from the features it covers. We selected max pooling because it makes the representation invariant to small translations in the input. This means the output will not be easily affected by a small number of translations detected on the input, like shifting, rotation, or scaling of the image.

- **Dropout Layer:** During training, the dropout layer randomly empties some neurons and passes other neurons to the next layer. But during testing, the dropout layer passes all neurons (Yasaka et al., 2018). This prevents overfitting.
- **Activation Function:** The relu activation function replaces negative values received as inputs by zeros and returns values greater than zero itself. We used the Relu activation function in the input and hidden layers. The LeakyReLU Activation function is also applied in the hidden layer. By adding a small slope, the LeakyReLU activation function prevents negative functions from becoming zero (Ding et al., 2018).
- **Flatten Layer:** converts the 2 dimensional output tensor of convolution to a one-dimensional vector to be input to the next layer. The 2- dimensional output tensor is flattened and becomes a 1- dimensional feature vector for concatenation with the output feature vector of the MLP.
- **Fully Connected Layer:** after the 2- dimensional tensors of convolutional layers are flattened. The flattened 1- dimensional feature vector is fed to the fully connected layers. Fully connected layers learn the non-linear combinations of high-level features that the convolutional layer produces.

**Algorithm: Convolutional Neural Network**

**Input:** 256x256x1

**Output:** 1 dimensional feature vector

**Begin**

$\mathrm{I} = 256\mathrm{x}256\mathrm{x}1//$ Input layer

//Convolution layer

$o[i,j]^1 = ((a_{ij}^1 = \sum_{i=1}^{256} \sum_{j=1}^{256} K[7,7]C[i-u, j-v], h_{ij}^1 = max(0, a_{ij}^1))(I))$

$o[i,j]^2 = (y_{kij} = max\ x_{kpq})(o[i,j]^1)$ //max pooling layer

$o[i,j]^3 = (r = 1-p)o[i,j]^2$//dropout layer

$o[i]^{12} = (Flatten)(o[i,j]^{11})$

//fully connected layers

$o[i]^{13} = ((a_i^{13} = \sum_{i=1}^{512} w_j^{13} + b_i^{13},\ h_i^{13} = max(0, a_i^{13}))(o[i]^{12}))$

$o[i]^{14} = (s = 1 - p)(o[i]^{13})$

$o[i]^{15} = ((a_i^{15} = \sum_{i=1}^{128} w_j^{15} + b_i^{15},\ h_i^{15} = max(0.1a_i^{15}, a_i^{15}))(o[i]^{14}))$

$o[i]^{16} = (flatten)(o[i]^{15})$

**End**

Algorithm 3: Convolutional Neural Network

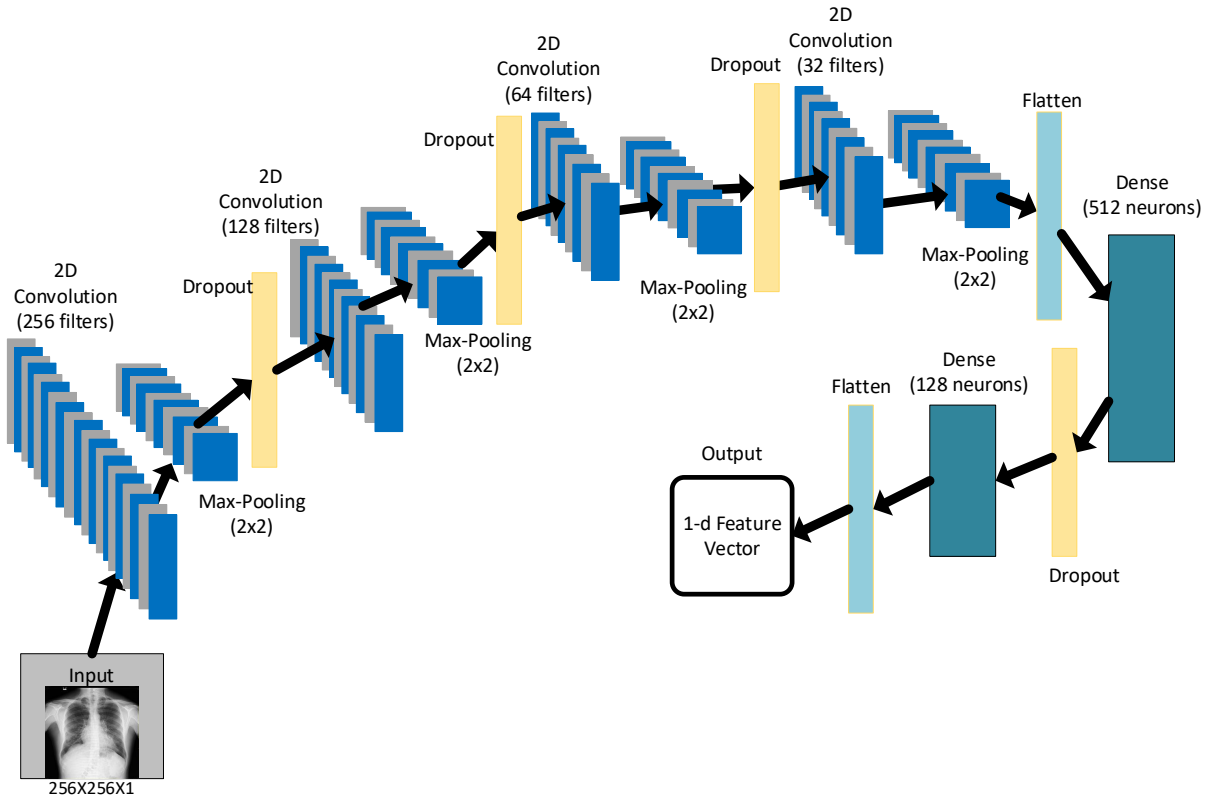



Figure 18: CNN Architecture

### 4.5.3. Concatenation Layer

We define a concatenation layer to concatenate the output feature vectors of CNN and MLP. The feature concatenation strategy is shown in figure 18. The length of the features from CNN and MLP can be equal or not. But the number of datasets of medical records and chest x-ray images should be equal. Because concatenation concatenates the features of a single data row to the corresponding chest x-ray image features found at the same index position In other words,

the medical record information of a single person becomes concatenated with the associated single chest x-ray image. No medical record of a single person can be joined with another chest x-ray image with a different index position. There is a one-to-one correspondence between the two datasets.

**Algorithm: Concatenation of feature vectors**

**Input** $I_1$:1 dimensional feature vector (from CNN) $(a_1, a_2 , , a_m)$

$I_2$:1 dimensional feature vector (from MLP) $( b_1, b_2 , , b_n)$

**Output:** $I_c$: $m + n$ length concatenated 1 dimensional vector

**Begin:**

$(a_1, a_2 , , a_m) + ( b_1, b_2 , , b_n)$ // concatenation

$(a_1, a_2 , , a_m, b_1, b_2 , , b_n)$// output

**End**

Algorithm 4: Concatenation of Feature Vectors

### 4.5.4. Fully Connected Classifier Network

After concatenation, we added fully connected layers to make the model learn the features of the concatenated feature vector and define the detection output layer. The addition of fully connected hidden layers to the detection model increases the capacity of the model to learn the important features (Nguyen et al., 2018) extracted by CNN and MLP for the specific medical record row and associated chest x-ray image.

***The Softmax activation function*** is used as the activation function in the output layer since we have multiple classes to be classified. The exponential values from the output layer are calculated and divided by the sum of the exponential values to normalize and then convert to probabilities, which sums up to 1.

$$softmax(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}}$$

Equation 7: Softmax Activation Function

The classification outputs of the model are three; pneumonia, TB, and other. Therefore, the output layer has 3 nodes and applies a softmax activation function for the probability detection output of the three classes.

**Algorithm: Fully Connected Pulmonary Disease Classification Network**

**Input:** $I_c$: $m + n$ length concatenated 1 dimensional vector

**Output:** Detection probability of pneumonia, tuberculosis and other

**Begin**

fully connected layer one

$o_i^1 = ((a_j^1 = \sum_{i=1}^{256} w_j^1 + b_i^1,\ h_i^1 = max(0, a_j^1))(I_c))$

fully connected layer two

$o_i^2 = ((a_j^2 = \sum_{i=1}^{128} w_j^2 + b_i^2,\ h_i^2 = max(0, a_j^2))(o_i^1))$

fully connected layer three

$o_i^3 = ((a_j^3 = \sum_{i=1}^{64} w_j^3 + b_i^3,\ h_i^3 = max(0, a_j^3))(o_i^2))$

Output fully connected layer

$o_i^4 = ((z_j = \sum_{k=1}^{3} w_j^4 + b_i^4,\ softmax(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{3} e^{z_j}})(o_i^3))$

**End**

Algorithm 5: Fully Connected Pulmonary Disease Classifier Network

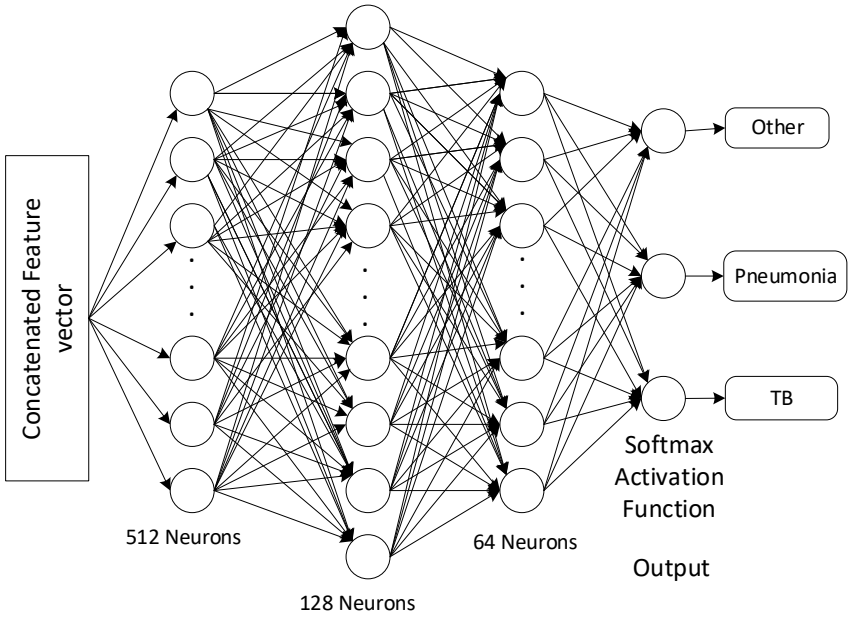



Figure 19: Fully Connected Pulmonary Disease Classifier Architecture

## 4.6. Prototyping

Following the development of the pulmonary infectious disease detection model by applying the selected best-performing combinations of the development choices, we deploy the model by using a flask server. Flask server uses the web pages as HTTP clients to pass requests and display responses. The developed model will detect the disease after it is provided with the requested medical record and associated chest x-ray image. The detection process does not require an internet connection and it is easy and flexible to use. The deployment architecture is provided below.

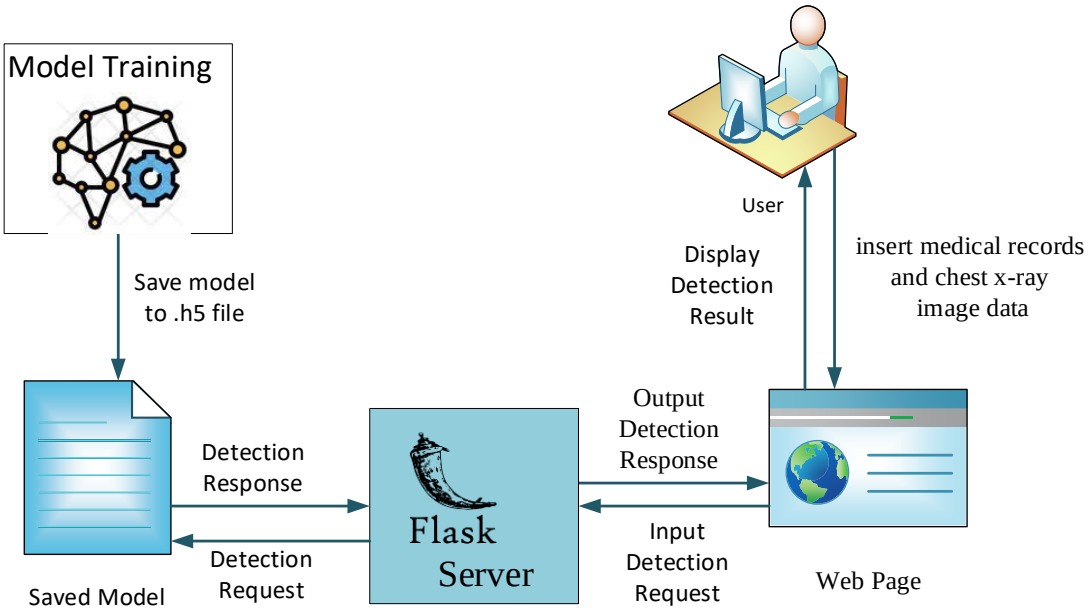



Figure 20: Model Deployment Architecture

# CHAPTER FIVE

# 5. PROPOSED SOLUTION IMPLEMENTATION

## 5.1. Chapter Overview

This chapter presents step by step implementation of the proposed solution starting from the environment setup. Medical record preprocessing and chest x-ray image preprocessing joint chest x-ray image and medical record model implementations are presented.

## 5.2. Environment Setup

The environments used in the implementation of this thesis are

- A laptop computer with 8GB RAM, Windows 10 Pro, and an Intel(R) Core i7 processor running at 1.80GHz.
- Google Colab free and Google Colab Pro for the implementation of the system. The free version has 12 GB of RAM and 107GB of disk space with free GPU and TPU offers for limited usage time in 24 hours. Colab-free is used for the preprocessing of medical records and image augmentation. The Colab Pro is utilized for text cleaning of chest x-ray images using Keras OCR and for training our model since it provides increased RAM, which is 25.46GB, faster GPU, and more usage time than the free version.

The Python programming language is used to develop the disease detection model. It is a simple and convenient programming language for the machine learning environment.

The following libraries are used for model development and deployment:

Table 6: Libraries Used for the Model Development

| **Libraries** | **Version** | **Utilization** |
|---|---|---|
| Tensorflow | 2.8.0 | Backend |
| Keras | 2.8.0 | Deep learning API |
| Matplotlib | 3.2.2 | Plotting library for data visualization |

| | | |
|---|---|---|
| Numpy | 1.21.6 | Work with arrays |
| Pandas | 1.3.5 | Data analysis and manipulation |
| Scikit learn | 1.0.2 | Data modeling |
| Keras OCR | 0.9.1 | Text removing library of Keras |
| Seaborn | 0.11.2 | Python package for data visualization |
| Flask server | 1.1.1 | Run HTTP requests on private networks |

## 5.3. Implementation of Medical Record Preprocessing for MLP Input

The first task in preprocessing the medical record dataset is importing the csv file. While importing the dataset, we used the **shuffle()** function to mix our dataset in order not to create bias in the model while training.

```
from google.colab import drive
drive.mount('/content/drive')
```

```
#importing the dataset
original_csv = pd.read_csv('/content/19-HW-B-ALL/New-19-MOOdified.csv', header=0)
original_csv = shuffle(original_csv)
original_csv.head()
```

Code Snippet 1: Importing csv File

We have assigned each chest x-ray image to each row by adding the 'IMG' column which contains the specified image file name with file extension in each row. Each row in the csv file contains the medical record of a single patient with corresponding chest x-ray image.

Figure 19 illustrates a sample view of the dataset with medical records and a chest x-ray image file name.

| | Age | Sex | Temp | RR | PR | Cough_D | Cough_T | WBC | CP | SOB | SPO2 | Fever | NS | HTB | Contact | HS | WL | HIV | Diagnosis | IMG |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **2646** | 35.0 | M | 38.8 | 30.0 | 81 | 4.30 | yellowish | 11.50 | yes | yes | 90.0 | low | yes | no | no | no | yes | P | TB | tb (573).png |
| **874** | 65.0 | M | 36.0 | 16.0 | 95 | 0.14 | dry | 11.00 | no | no | 95.0 | no | no | no | no | no | no | N | Other | other (875).png |
| **1794** | 0.3 | M | 36.5 | 32.0 | 110 | 0.42 | dry | 8.37 | yes | yes | 100.0 | high | no | no | no | no | no | N | Pneumonia | pneumonia (795).png |
| **297** | 86.0 | M | 37.0 | 20.0 | 77 | 0.14 | dry | 10.00 | no | no | 99.0 | no | no | no | no | no | no | N | Other | other (298).png |
| **1076** | 55.0 | M | 36.5 | 30.0 | 114 | 2.00 | whitish | 8.37 | yes | yes | 93.0 | high | no | no | no | no | no | N | Pneumonia | pneumonia (77).png |

Figure 21: Imported Medical Record Dataset Sample View

### 5.3.1. Data Cleaning

Data cleaning on the medical record dataset is performed by checking for null and missing values by using the **isnull()** function of Python. Yes and yes with the same lowercase spelling but one yes with a single space in front will be counted as two different yeses, and such discrepancies can be identified using the. **describe ()** function. This function checks how many choices there are on a single column. **Value_counts()** shows how many values are counted for each choice in a column. The dataset was also checked for outliers using **boxplot([column name])** and checked for data imbalance. Code snippet 2 shows the syntax of the mentioned data cleaning process.

```
#check for NaN row (row with missing value) in the imported data
print("The count of NaN values in the data:",original_csv.isnull().sum().sum())
print("Is there a NaN value in the dataset:",original_csv.isnull().values.any())
```

```
original_csv['Column_Name'].value_counts()
```

```
original_csv['Column_Name'].describe()
```

Code Snippet 2: Data Cleaning Implementation

The column containing the image path, 'IMG', is not relevant for the medical record preprocessing for the MLP input. So we drop this column during the preprocessing for the MLP input. We created a csv file called **MLP_csv**.

```
# delete the IMG coloumn since it is not important for the MLP
mlp_csv = original_csv.drop("IMG", axis=1, inplace = False)
```

Code Snippet 3: Dropping the Image Column

### 5.3.2. Data Transformation

We perform data transformation on the imported medical record dataset to make it more suitable for the algorithms and the model we develop. Category encoders are imported for the conversion of columns that are ordinal where the ordering of values matters. Then the dictionary for the encoding of the ordinal columns is created. Cough type and fever are the parameters that need ordering of the values as shown below in code snippet 4. Then the encoded columns are transformed using **encoder.fit_transform**.

```
!pip install Category_encoders
```

```
#ordinal encoding
import category_encoders as ce
```

```
#create dictionary to use for ordinal encoding
dictionary = [{'col': 'Cough_T', 'mapping':{'dry':0, 'whitish':1,
                                            'yellowish':2, 'bloody':3}},
              {'col': 'Fever', 'mapping':{'no':0, 'low':1, 'high':2}}]
encoder = ce.OrdinalEncoder(cols='Fever', mapping=dictionary)
encoder = ce.OrdinalEncoder(cols='Cough_T', mapping=dictionary)
```

```
mlp_csv =encoder.fit_transform(mlp_csv)
```

Code Snippet 4: Ordinal Encoding Implementation

Label encoding is performed for categorical columns where ordering between values is not necessary. Code snippet 5 shows label encoding.

```
from sklearn.preprocessing import LabelEncoder
```

```
#Label encoding
label_encoder = LabelEncoder()
mlp_csv.Sex =label_encoder.fit_transform(mlp_csv.Sex)
mlp_csv.HIV =label_encoder.fit_transform(mlp_csv.HIV)
mlp_csv.CP=label_encoder.fit_transform(mlp_csv.CP)
mlp_csv.SOB =label_encoder.fit_transform(mlp_csv.SOB)
mlp_csv.NS =label_encoder.fit_transform(mlp_csv.NS)
mlp_csv.HTB =label_encoder.fit_transform(mlp_csv.HTB)
mlp_csv.Contact =label_encoder.fit_transform(mlp_csv.Contact)
mlp_csv.HS =label_encoder.fit_transform(mlp_csv.HS)
mlp_csv.WL =label_encoder.fit_transform(mlp_csv.WL)
```

Code Snippet 5: Label Encoding

One hot encoding is done to transform the dependent variable diagnosis. **Pd.get_dummies** creates dummy variables that are equal to the number of classes we have, so three dummy variables are created. The created dummy variable uses 0 and 1 values to indicate the absence or presence of a category.

```
from sklearn.preprocessing import OneHotEncoder
```

```
# applying one hot encoding to the target variable
oneHotEndcode = pd.Series(cnn_csv['Diagnosis'])
encodeDiagnosis = pd.get_dummies(oneHotEndcode)
```

Code Snippet 6: One Hot Encoding

### 5.3.3. Data Normalization

After transforming non-numerical categorical values to numbers by using category encoding and label encoding, normalization is performed. We perform normalization for columns that have values greater than 1. Normalization enables our algorithms to be less sensitive to higher numbers by making all values between 0 and 1. We used the absolute maximum scaling technique, which divides each value by the maximum value of the variable.

```
# Normalization
column = ['Age', 'Temp','RR','PR', 'Cough_D', 'WBC','SPO2','Fever','Cough_T']
mlp_csv[column[0]] = mlp_csv[column[0]] /mlp_csv[column[0]].abs().max()
mlp_csv[column[1]] = mlp_csv[column[1]] /mlp_csv[column[1]].abs().max()
mlp_csv[column[2]] = mlp_csv[column[2]] /mlp_csv[column[2]].abs().max()
mlp_csv[column[3]] = mlp_csv[column[3]] /mlp_csv[column[3]].abs().max()
mlp_csv[column[4]] = mlp_csv[column[4]] /mlp_csv[column[4]].abs().max()
mlp_csv[column[5]] = mlp_csv[column[5]] /mlp_csv[column[5]].abs().max()
mlp_csv[column[6]] = mlp_csv[column[6]] /mlp_csv[column[6]].abs().max()
mlp_csv[column[7]] = mlp_csv[column[7]] /mlp_csv[column[7]].abs().max()
mlp_csv[column[8]] = mlp_csv[column[8]] /mlp_csv[column[8]].abs().max()
```

Code Snippet 7: Normalization

### 5.3.4. Feature Selection

Each feature in the dataset does not contribute equally to the prediction of the dependent variable. Keeping this in mind, we performed feature selection by using a random forest classifier. We imported the necessary libraries and classes that are required to perform the feature selection.

```
# get the importance of the resulting features.
importances = rf_w.feature_importances_
# create a data frame for visualization.
final_df = pd.DataFrame({"Features": X_train.columns, "Importances":importances})
final_df.set_index('Importances')

# sort in ascending order to better visualization.
final_df = final_df.sort_values('Importances')

# plot the feature importances in bars.
plt.figure(figsize=(10,3))
plt.xticks(rotation=45)
sns.barplot(x="Features",y= "Importances", data=final_df)
```

Code Snippet 8: Feature Selection Implementation

## 5.4. Implementation of Chest X-ray Image Processing for CNN Input

### 5.4.1. Image Cleaning

The local chest x-ray images are cleaned from texts using the Keras OCR library. After installing the OCR module, the necessary libraries are imported and the midpoint for the mask that will remove the text is defined. After defining the midpoint of the four edges x1, y1, x2, and y2 the line mask is defined. We loaded a batch size of images to be cleaned by the OCR. The images are loaded from the drive according to their path in a csv file. The OCR checks every part of the image for text and removes the text found on the image. To save the cleaned images we need to convert the images to the RGB format to avoid color inversion. The cleaned images will then be saved to the specified output path.

```
def midpoint(x1, y1, x2, y2):
    x_mid = int((x1 + x2)/2)
    y_mid = int((y1 + y2)/2)
    return (x_mid, y_mid)
```

```
#importing the dataset
test_csv = pd.read_csv("/content/Cc-other/Atest-other.csv", header=0)
test_csv.head()
```

```
#loading images in to one array
train_image = []
for i in tqdm(range(test_csv.shape[0])):
    imgPath = '/content/Cc-other/'+test_csv['IMG'][i]
    imgOutPath = '/content/Cc-other/CLEANED-REMAIN-OTHER/'+test_csv['IMG'][i]
    print(imgPath)
    editedImage = inpaint_text(imgPath,pipeline)
    img_rgb = cv2.cvtColor(editedImage, cv2.COLOR_BGR2RGB)
    cv2.imwrite(imgOutPath,img_rgb)
```

Code Snippet 9: Image Cleaning Implementation

### 5.4.2. Image Augmentation

Image augmentation is performed independently from the detection model. To perform image augmentation, first we need to import the necessary libraries and the image data generator. The original images collected are RGB images. The local images are cleaned using OCR, and the images collected from Kaggle are merged and image augmentation is performed. After importing the images, the images are converted to greyscale and normalized. Image augmentation is performed on Google Colab free using **the ImageDataGenerator** class. The ImageDataGenerator class accepts input images and performs augmentation according to the specified augmentation techniques. Images collected are of different dimensions, as mentioned in ***section 3.3.1;*** therefore, the sizes of all images are converted to the dimension of 512x512. The augmentation operation that needs to be performed is stated as shown in the code snippet 10. Finally, the augmented images are saved to the specified path.

```
from keras.preprocessing.image import ImageDataGenerator
```

```
#augmentation techniques to be applied
data_gen_args = dict(rescale=1./255,
                     height_shift_range=0.2,
                     width_shift_range=0.2,
                     brightness_range=(0.3,0.9),
                     rotation_range= -5,
                     fill_mode='constant' )
myGenerator = trainGenerator(5,'TB','image',data_gen_args,save_to_dir = "/content/TB/HW-B-R-TB")
```

Code Snippet 10: Image Augmentation Implementation

### 5.4.3. Chest X-ray Image Loading

From the original medical record, csv dataset diagnosis and image columns are extracted and named as **cnn_csv** to load the image dataset according to the paths specified in the image (IMG

path) column. The images are loaded with a target size of (256, 256, 1). The loaded images are then converted to a Numpy array.

```
# taking the Diagnosis and the image column
cnn_csv =original_csv[['Diagnosis', 'IMG']]
cnn_csv.head()
```

```
#loading images in to one array
train_image = []
for i in tqdm(range(cnn_csv.shape[0])):
    img = image.load_img('/content/drive/MyDrive/W-R-ALL-✅/'
    +cnn_csv['IMG'][i],target_size=(256,256,1), color_mode = "grayscale")
    img = image.img_to_array(img)
    img = img/255
    train_image.append(img)
X_img = np.array(train_image)
```

Code Snippet 11: Loading Chest X-ray Images

## 5.5. Defining B-Input Functional Model

Having imported and preprocessed medical records and chest x-ray images, we define and implement the proposed bi-input functional model. Our model has two inputs medical record, dataset array and chest x-ray image array.

### 5.5.1. Multi-Layer Perceptron

We define multi-layer perceptron to accept and process our csv dataset array. The visible layer is the input layer and takes the number of features of the medical record dataset. The output of the previous layer will be fed to the next layer in the functional neural network.

```
# medical record processing MLP model
visible1 = Input(shape=(n_features))
model1 = Dense(256, activation='relu')(visible1)
model1 = Dense(128, activation='relu')(model1)
model1 = Dense(64, activation='relu')(model1)
model1 = Dense(32, activation='relu')(model1)
```

Code Snippet 12: Multi-Layer Perceptron

### 5.5.2. Convolutional Neural Network

The image processing network is the convolutional neural network, as shown in code snippet 13. The first input layer, visible2, holds the input dimension of the image and passes to the next layer.

```
# Image input CNN model
visible2 = Input(shape=(input_dim))
model2 = Conv2D(filters=256,strides= 1, kernel_size=(3, 3), activation="relu")(visible2)
model2 = MaxPooling2D(pool_size=(2, 2))(model2)
model2 = Dropout(0.5)(model2)
model2 = Conv2D(filters=128, kernel_size=(3, 3), activation="relu")(model2)
model2 = MaxPooling2D(pool_size=(2, 2))(model2)
model2 = Dropout(0.5)(model2)
model2 = Conv2D(filters=64, kernel_size=(3, 3), activation="relu")(model2)
model2 = MaxPooling2D(pool_size=(2, 2))(model2)
model2 = Dropout(0.5)(model2)
model2 = Conv2D(filters=32, kernel_size=(3, 3), activation="relu")(model2)
model2 = MaxPooling2D(pool_size=(2, 2))(model2)
model2 = Flatten()(model2)
model2 = Dense(512, activation='relu')(model2)
model2 = Dropout(0.5)(model2)
model2 = Dense(128, activation='LeakyReLU')(model2)
model2 = Flatten()(model2)
```

Code Snippet 13: Convolutional Neural Network

### 5.5.3. Concatenation Layer

After providing the two arrays to the two defined algorithms, the outputs of the two algorithms are concatenated as shown in code snippet 14. Model1 is the output array from the MLP, and Model2 is the output array of the CNN.

```
# concatenate the two models (MLP + CNN)
merge = concatenate([model1, model2])
```

Code Snippet 14: Concatenation Layer

### 5.5.4. Fully Connected Classification Layer

Next to the concatenation of the two output arrays, we pass the concatenated array through fully connected layers for classification. We used the softmax activation function as we have multi-class output and defined output class numbers as 3. The final model takes the inputs of the MLP model, visible1 (input shape) and the input of the CNN model, visible2 (input dimension) as inputs and takes the concatenated output array as an output, as shown below in the code snippet. Then we compile and fit the data to the model.

```
# fully connected layers
merge = Dense(256, activation='relu')(merge)
merge = Dense(128, activation='relu')(merge)
merge = Dense(64, activation='LeakyReLU')(merge)
output =Dense(3, activation='softmax')(merge)
model = Model(inputs=[visible1, visible2], outputs=output)
```

Code Snippet 15: Fully Connected Classification Layer

### 5.5.5. Model Compile and Fit

Finally, we compile the model using **'model.compile'** and fit the final model using the **'model.fit'** function.

```
# compile the model
model.compile(loss='categorical_crossentropy', optimizer= 'adam', metrics=['accuracy'])
#train the model
model.fit([train_x_csv,train_x_img], train_y, batch_size=batch_size,verbose=1, epochs=epochs,
          validation_data=([test_x_csv,test_x_img],test_y))
```

Code Snippet 16: Compiling and Fitting the Model

The output of the MLP is a 32 length one dimensional feature vector while the output of the CNN is a 128 length one dimensional feature vector. Concatenating the two feature vector in concatenation layer results in a 160 length one dimensional feature vector. The parameters of the model are shown in table 7 below.

Table 7: Model Parameters

(a) MLP parameters

| MLP | Neuron | Activation | Output |
|---|---|---|---|
| Input | 256 | | 17 |
| Dense1 | 128 | Relu | 256 |
| Dense2 | 64 | Relu | 128 |
| Dense3 | 32 | Relu | 32 |

(b) CNN parameters

| CNN | Filter | Kernel | Activation | Output |
|---|---|---|---|---|
| Input | | | | 256x256x1 |

| | | | | |
|---|---|---|---|---|
| Conv1 | 256 | 7x7 | Relu | 250x250x25 |
| Max-Pool1 | | | | 125x125x25 |
| Dropout1 | | | | 125x125x25 |
| Conv2 | 128 | 7x7 | Relu | 119x119x12 |
| Max-Pool2 | | | | 59x59x128 |
| Dropout2 | | | | 59x59x128 |
| Conv3 | 64 | 7x7 | Relu | 53x53x64 |
| Max-Pool3 | | | | 26x26x64 |
| Dropout3 | | | | 26x26x64 |
| Conv4 | 32 | 7x7 | Relu | 20x20x32 |
| Max-Pool4 | | | | 10x10x32 |
| Flatten1 | | | | 3200 |
| Dense4 | 512 | | Relu | 512 |
| Dropout4 | | | | 512 |
| Dense5 | 128 | | LeReLU | 128 |
| Flatten2 | | | | 128 |

(c) Fully connected network parameters

| | **Neuron** | **Activation** | **Output** |
|---|---|---|---|
| **Concatenate** | | | 160 |
| Dense6 | 256 | relu | 256 |
| Dense7 | 128 | relu | 128 |
| Dense8 | 64 | LeReLU | 64 |
| Output | 3 | softmax | 3 |

## 5.6. Model Testing and Evaluation

Model testing and evaluation are performed by k-fold cross validation, in which we interchange k values as 3, 5, and 10 to use the better scoring k value to train our final detection model. The

confusion matrix is applied to calculate the accuracy, precision, recall, and f1 score of the model from the matrix values.

```
def evaluate_both_model(data_x_img,data_x_csv, data_y):
    k_fold = KFold(3, shuffle=True, random_state=1)

    predicted_targets = np.array([])
    actual_targets = np.array([])

    for train_ix, test_ix in k_fold.split(data_x_img):
        train_x_img, train_y, test_x_img, test_y = data_x_img[train_ix],
        data_y[train_ix], data_x_img[test_ix], data_y[test_ix]
        train_x_csv, train_y_csv, test_x_csv, test_y_csv = data_x_csv[train_ix],
        data_y[train_ix], data_x_csv[test_ix], data_y[test_ix]

        #define input models
        #compile model
        #fit the data to the model

        # Predict the labels of the test set samples
        predicted_labels = model.predict([test_x_csv,test_x_img], verbose=0)
        predicted_labels = predicted_labels.argmax(axis=-1)
        predicted_targets = np.append(predicted_targets, predicted_labels)

        test_label = test_y_csv.argmax(axis=-1)
        actual_targets = np.append(actual_targets, test_label)

    return predicted_targets, actual_targets
```

Code Snippet 17: Model Testing and Evaluation Implementation

## 5.7. Detection Model Implementation GUI

The pulmonary infectious disease detection model is deployed by using a flask server. Anaconda is used to install the flask server and necessary libraries. We developed an HTML interface and used a flask server to deploy the model by saving our model in a .h5 file. We created the file called app.py which contains the flask server module and saved the model. Necessary operations will be performed here, like converting days, months, and years of cough duration to similar values in weeks as we did in the preprocessing of the input medical record data in ***section 4.3.2.1.*** The incoming data will be normalized by dividing the maximum value of the parameter as we did on our model in ***section 5.3.3.*** The image that will be uploaded for detection will be converted to greyscale and shaped to the dimension of (256, 256, 1). The model will be reconstructed and detect the data received. Then the detected class will be displayed.

# CHAPTER SIX

## 6. RESULT AND DISCUSSION

### 6.1. Chapter Overview

The results of the implementation presented in previous chapter are delivered in this chapter. Data preprocessing results of both image and medical records, and the outcome of the proposed model are presented. Results of feature selection and image augmentation along with diagrams and tables are delivered. Also, a discussion about the outcome and successes with the expert evaluation is presented.

### 6.2. Pulmonary Infectious Disease Detection

The results of the implementation presented in the previous chapter are presented in this chapter. Data preprocessing results of both image and medical records and the outcome of the proposed model are presented. The results of feature selection and image augmentation, along with diagrams and tables, are delivered. Also, a discussion about the outcome and successes with the expert evaluation is presented.

#### 6.2.1. Chest X-ray Image Preprocessing Results

Chest x-ray images show abnormalities in the lung and other respiratory systems. Due to their availability and affordability, chest x-ray images are widely used for the detection and diagnosis of pulmonary infectious diseases. Our CNN layer input uses the chest x-ray image for the detection of pulmonary disease. Proper image preprocessing contributes to the good performance of the model. According to this, we have performed proper image preprocessing and augmentation. Since the textual information about the chest x-ray image is not necessary for our model, the text is removed using OCR. Sample results of cleaned images are shown below.

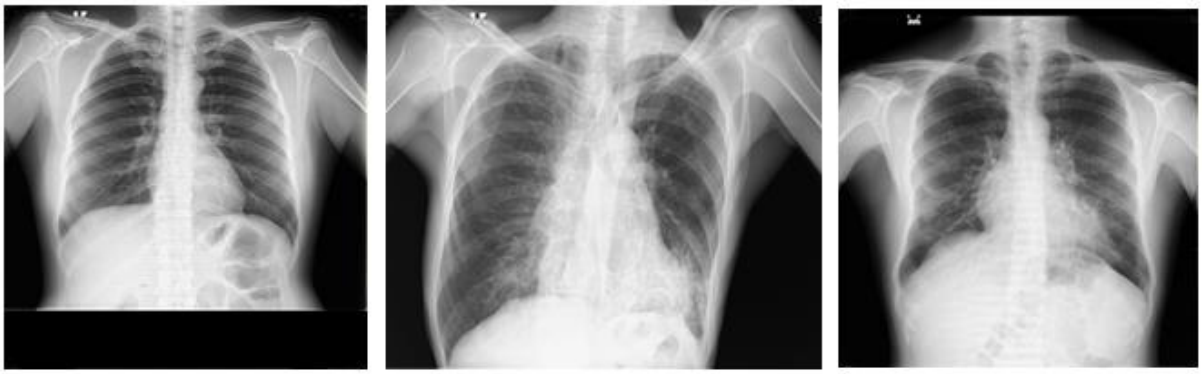


Figure 22: Sample Pneumonia Text Cleaned Images

#### 6.2.1.1. Augmentation Results

Chest x-ray images are augmented according to augmentation techniques applied in literature as described in section 4.3.2. All images are converted to greyscale, and images are resized to 512x512. Rescaling or normalization is also performed.

Augmentation can allow us to have a variety of images that will help to develop a better performing model. We converted all images to grayscale before performing other augmentation techniques. Next, we performed 6 experiments to augment the chest x-ray image dataset, including conversion to greyscale with no other augmentation technique applied. Height shift and width shift of 0.2 each, random brightness range of (0.3, 0.9), rotation of -5 degrees.

Before applying any other image augmentation technique, all chest x-ray images are converted to greyscale and no other augmentation technique is used in this experiment. An accuracy of 98.77% is obtained in this experiment.

**Experiment 1:** In the first experiment, height shift, width shift, and brightness augmentation techniques are applied to the chest x-ray images. The accuracy of this experiment is 98.70%.

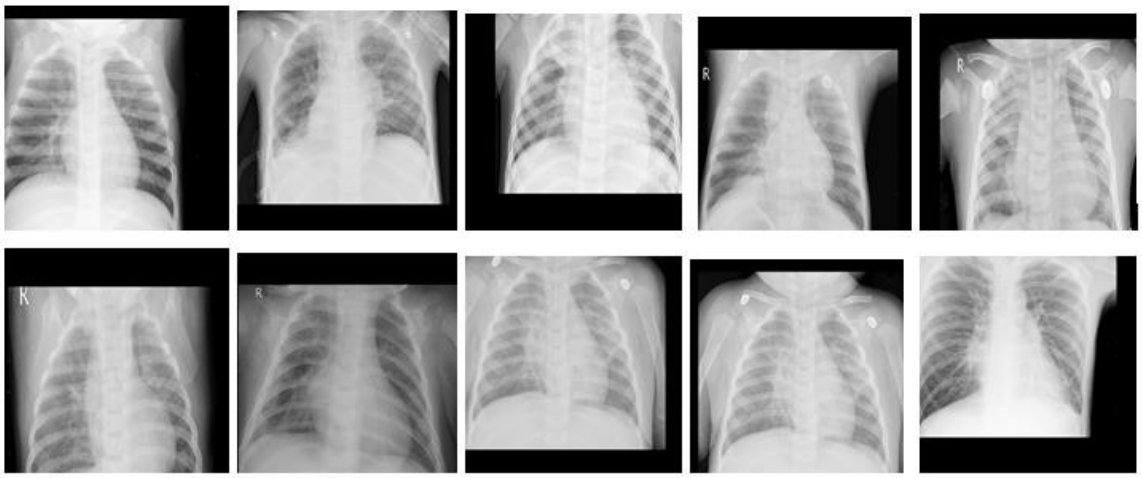


Figure 23: Sample Augmentation Results of Experiment 1

**Experiment 2:** In the second experiment, height shift and brightness augmentation techniques are applied to the chest x-ray images. The accuracy of the experiment is 98.87%.

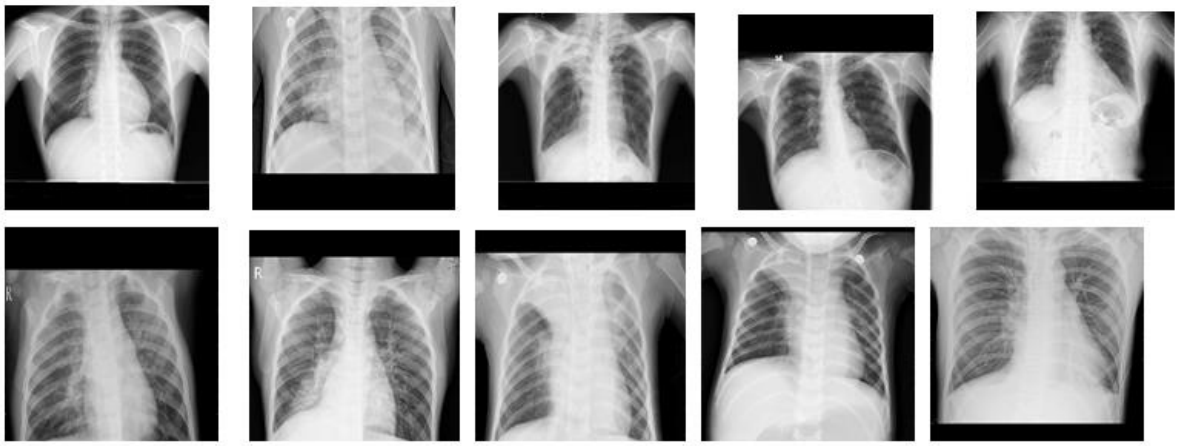


Figure 24: Sample Augmentation Results of Experiment 2

**Experiment 3:** In this experiment, random brightness and rotation augmentation techniques are applied and 98.98% accuracy is achieved.

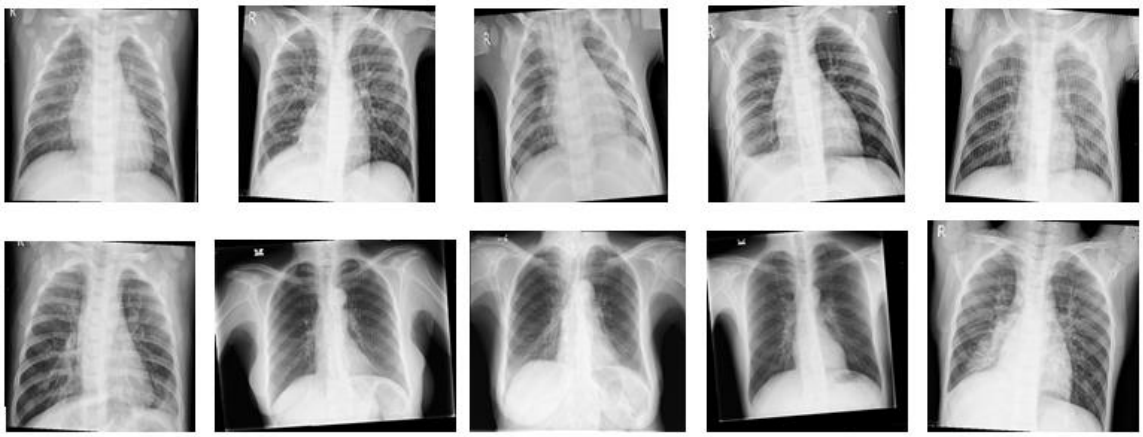


Figure 25: Sample Augmentation Results of Experiment 3

**Experiment 4:** In experiment 4, all of the four augmentation techniques (height and width shift, rotation, and brightness) are applied to the chest x-ray image dataset. An accuracy of 98.94% is obtained.

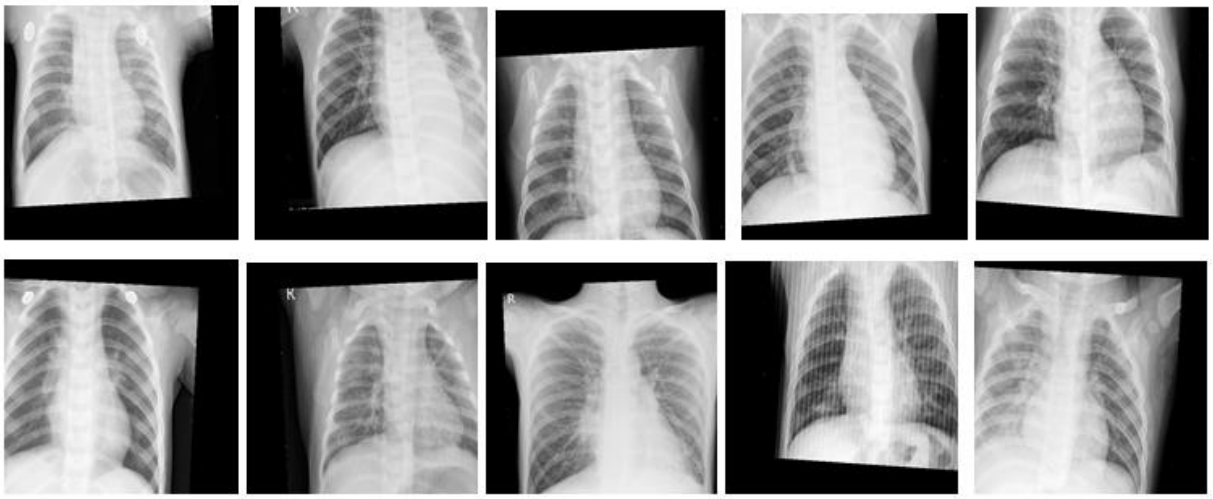


Figure 26: Sample Augmentation Results of Experiment 4

**Experiment 5:** In this experiment, height shift, width shift, and rotation augmentation techniques are applied to the chest x-ray image dataset and an accuracy of 98.98% is obtained.

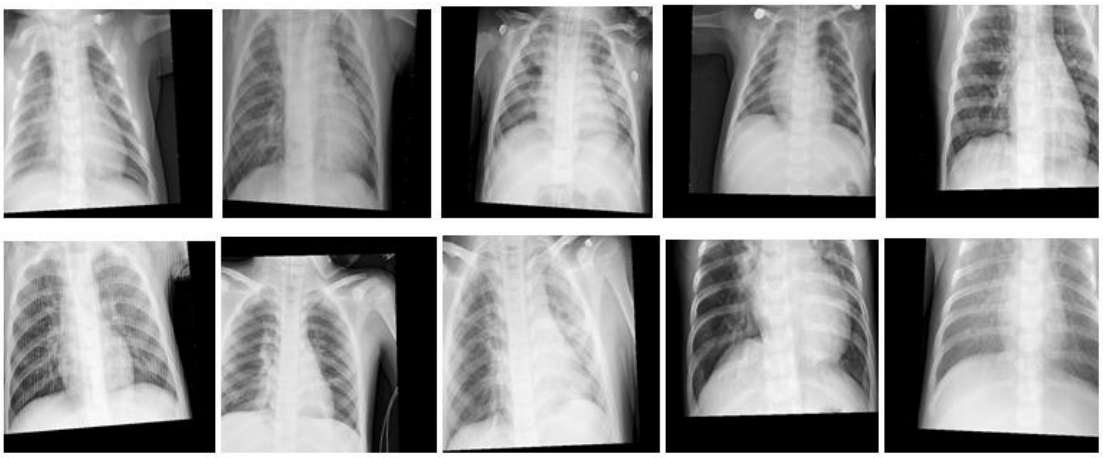


Figure 27: Sample Augmentation Results of Experiment 5

**Experiment 6:** In this experiment, rotation and width shift augmentation techniques are applied and an accuracy of 99.08% is obtained.

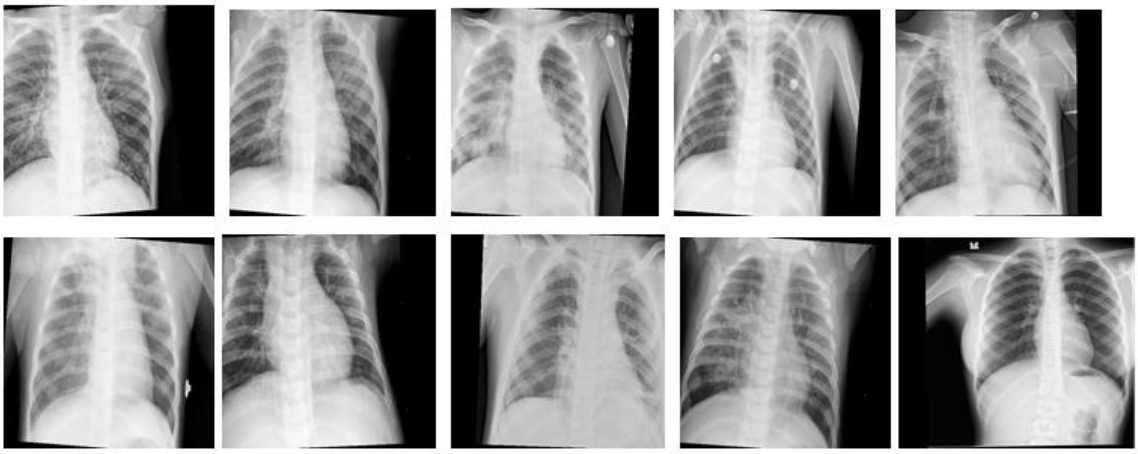


Figure 28: Sample Augmentation Results of Experiment 6

The experiments and augmentation techniques performed with accuracy are summarized in table 8.

Table 8: Augmentation Techniques Applied and Results Summary

| Experiment | Brightness | Rotation | Height Shift | width Shift | Accuracy |
|---|---|---|---|---|---|
| 1 | √ | | √ | √ | 98.70% |
| 2 | √ | | √ | | 98.87% |
| 3 | √ | √ | | | 98.98% |
| 4 | √ | √ | √ | √ | 98.94% |
| 5 | | √ | √ | √ | 98.98% |
| 6 | | √ | | √ | **99.08%** |

By comparing the experiments performed based on their accuracy, the images augmented with experiment 6 with an accuracy of 99.08% are selected to develop the pulmonary infectious disease detection model.

### 6.2.2. Medical Record Preprocessing Result

Data preprocessing is the most crucial part of any work that uses a dataset. The medical record dataset is transformed using label encoding, category encoding, and one hot encoding. The transformed dataset is then normalized using the maximum absolute scaling technique. The medical record of the patient is preprocessed and features are selected using a random forest classifier. The transformed and normalized dataset is shown in figure 29 below.

| Age | Sex | Temp | RR | PR | Cough_D | Cough_T | WBC | CP | SOB | SPO2 | Fever | NS | HTB | Contact | WL | HIV | Diagnosis |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.754902 | 0 | 0.891089 | 0.138298 | 0.324873 | 0.002258 | 0.000000 | 0.171875 | 0 | 0 | 0.98 | 0.0 | 0 | 0 | 0 | 0 | 0 | Other |
| 0.098039 | 1 | 0.915842 | 0.308511 | 0.451777 | 0.002258 | 0.000000 | 0.187500 | 0 | 0 | 0.95 | 0.0 | 0 | 0 | 0 | 0 | 0 | Other |
| 0.470588 | 1 | 0.898515 | 0.265957 | 0.517766 | 0.140161 | 0.666667 | 0.171875 | 1 | 1 | 0.90 | 0.5 | 1 | 0 | 0 | 1 | 0 | TB |
| 0.725490 | 0 | 0.915842 | 0.191489 | 0.380711 | 0.002258 | 0.000000 | 0.078125 | 0 | 0 | 0.95 | 0.0 | 0 | 0 | 0 | 0 | 0 | Other |
| 0.392157 | 1 | 0.900990 | 0.319149 | 0.578680 | 0.032258 | 0.333333 | 0.078125 | 1 | 1 | 0.94 | 1.0 | 1 | 0 | 0 | 0 | 0 | Pneumonia |

Figure 29: Data Preprocessing and Normalization Result

### 6.2.2.1. Feature Selection Result

Medical record data feature selection is performed by a random forest classifier, and the plot of the feature importance to the classification is shown in figure 30. As shown in the figure below, the feature importance is set from high to no importance. According to the results obtained, history of TB and sex have very little feature importance, while contact with chronic cougher and history of smoking have no importance. In the medical context, we cannot remove the feature of contact with chronic cougher because it is important to know when deciding the diagnosis of a patient, especially for children. Therefore we only removed the history of smoking. The accuracy before and after feature selection is discussed below.

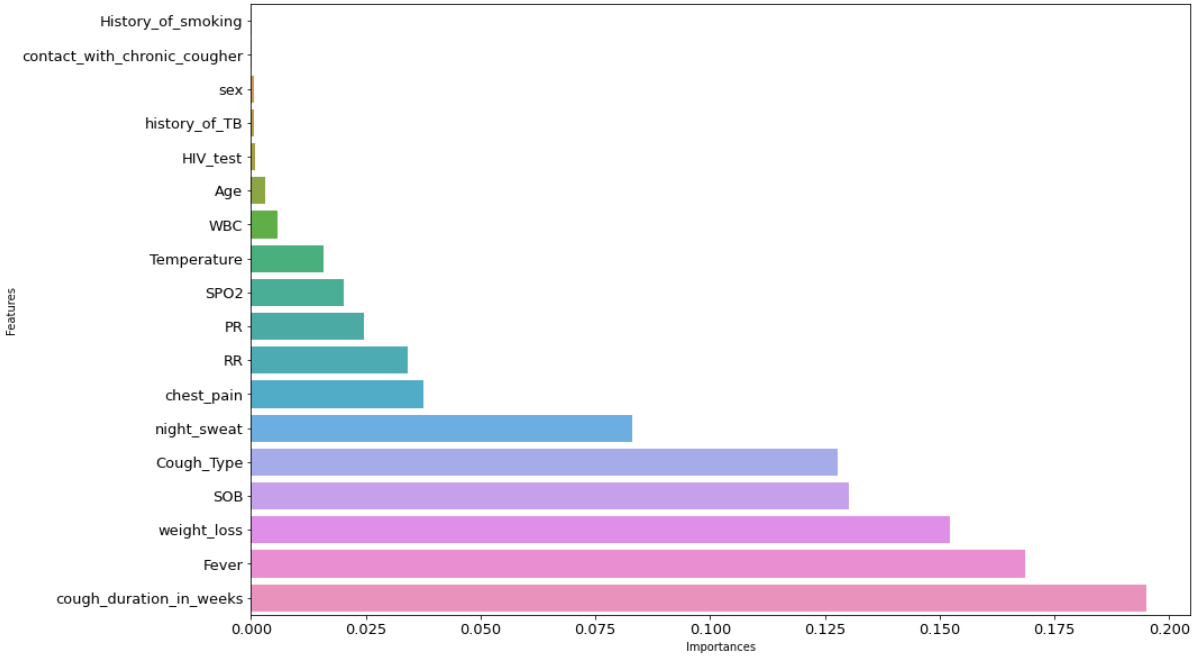



Figure 30: Feature Selection Result

We trained our model using k value 3 cross-validation with 30 batch size and 25 epochs before and after feature selection, and the results of instances classified on both models are shown in figures 31 and 32.

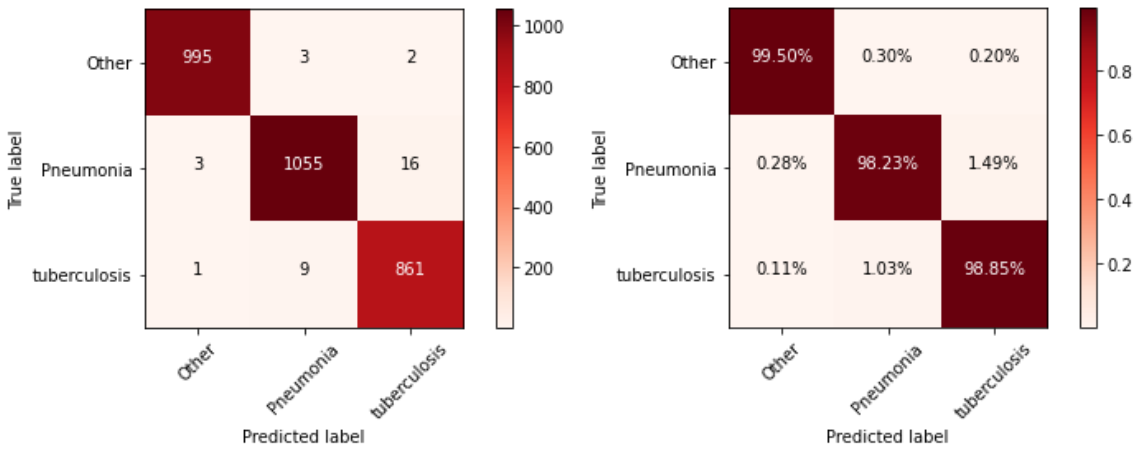



Figure 31: Confusion Matrix Results before Applying Feature Selection

Before applying feature selection as shown in figure 31 above, the pneumonia class has 19 instances misclassified as false negative. On the confusion matrix of features selected below on figure 32, this misclassification has decreased to 6.

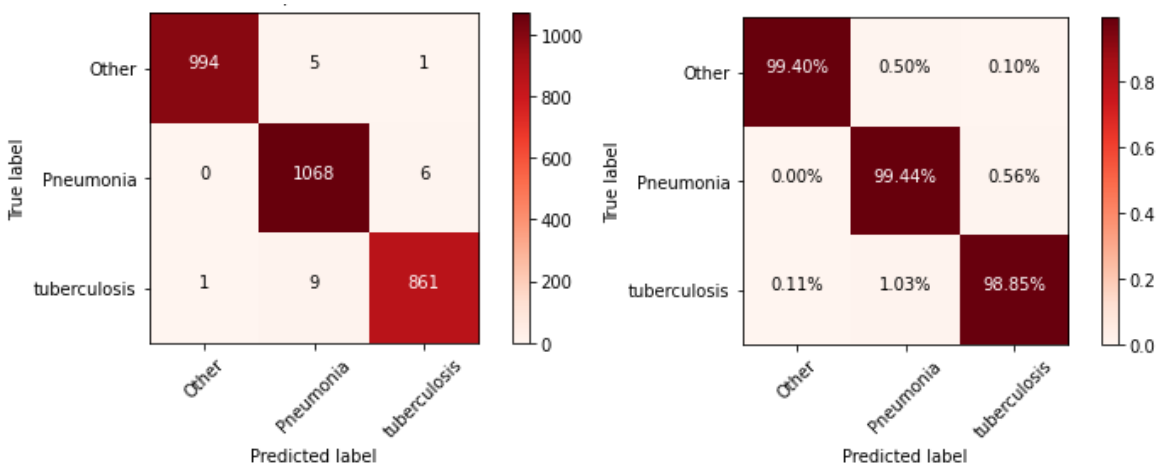



Figure 32: Confusion Matrix Results after Applying Feature Selection

The average accuracy of the model increased by 0.31% from 99.19% to 99.50% as a result of applying feature selection.

Table 9: Feature Selection Effect on the Performance of the Model

| Input Dimension | Accuracy |
|---|---|
| 18 | 99.19% |
| **17** | **99.50%** |

### 6.2.3. Pulmonary Infectious Disease Classifier Result

#### 6.2.3.1. Hyper Parameter Tuning Result

Hyper parameter tuning is essential in controlling the behavior of our model. We perform hyper-parameter tuning to minimize training time and remove useless parameters. A manual hyper-parameter tuning technique is applied. The parameter combinations we used and the results obtained are shown in table 7 below.

Table 10: Hyper Parameters Combination for Tuning

| Task | Dropout | Batch size | Epoch | Activation function | Learning rate | Kernel size | Acc (%) | F1score (%) |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.25 | 32 | 100 | Relu | 0.001 | 3x3 | 99.36 | 99.03 |
| 2 | 0.1 | 60 | 50 | Relu | 0.001 | 3x3 | 99.40 | 99.11 |
| 3 | 0.5 | 40 | 60 | Relu | 0.01 | 5x5 | 99.27 | 98.97 |
| 4 | 0.1 | 30 | 50 | LeakyReLU | 0.0001 | 5x5 | 99.25 | 98.87 |
| **5** | **0.25** | **40** | **80** | **Relu** | **0.0001** | **7x7** | **99.49** | **99.24** |
| 6 | 0.2 | 30 | 60 | LeakyReLU | 0.001 | 3x3 | 95.58 | 93.42 |
| 7 | 0.1 | 30 | 50 | Relu | 0.001 | 7x7 | 99.29 | 98.97 |
| 8 | 0.5 | 30 | 60 | LeakyReLU | 0.0001 | 7x7 | 99.33 | 99.00 |
| 9 | 0.25 | 40 | 80 | Relu | 0.001 | 5x5 | 99.43 | 99.14 |
| 10 | 0.25 | 40 | 80 | Relu | 0.0001 | 3x3 | 99.27 | 98.90 |

By applying the combination of hyper parameters shown above, we found the average accuracy and average F1 score for each combination. The combination indicated in task 5 has shown slightly better performance than other combinations. So we have selected the combination of hyper parameters indicated in task 5, dropout of 0.25, batch size of 40, with 80 epochs and relu activation function and learning rate of 0.0001 and kernel size of 7x7. The selected combination is used to develop our pulmonary infectious disease detection model.

#### 6.2.3.2. Result of Applying Different K-Fold Values

Using medical records along with chest x-ray images, the proposed model classifies the disease detected as pneumonia, TB, or other. One way of improving the performance of the algorithms is by using k-fold cross validation. We fed the preprocessed chest x-ray image and medical record to the algorithms and trained those using different K fold values. The results of applying different K-fold values are laid out.

**K=3:** when we apply a k value of 3, the dataset is divided into three classes; two of the classes are applied for training and the remaining one class is used for validation. We have got a relatively better score in other class than in the Pneumonia and TB classes.

Table 11: Classification Result for K=3

| **Class** | **Precision (%)** | **Recall (%)** | **F1-score (%)** | **Accuracy (%)** |
|---|---|---|---|---|
| Pneumonia | 98.61 | 99.44 | 99.02 | 99.28 |
| TB | 99.30 | 98.62 | 98.95 | 99.38 |
| Other | 99.59 | 99.30 | 99.44 | 99.62 |
| **Average Accuracy** | | | | **99.42** |

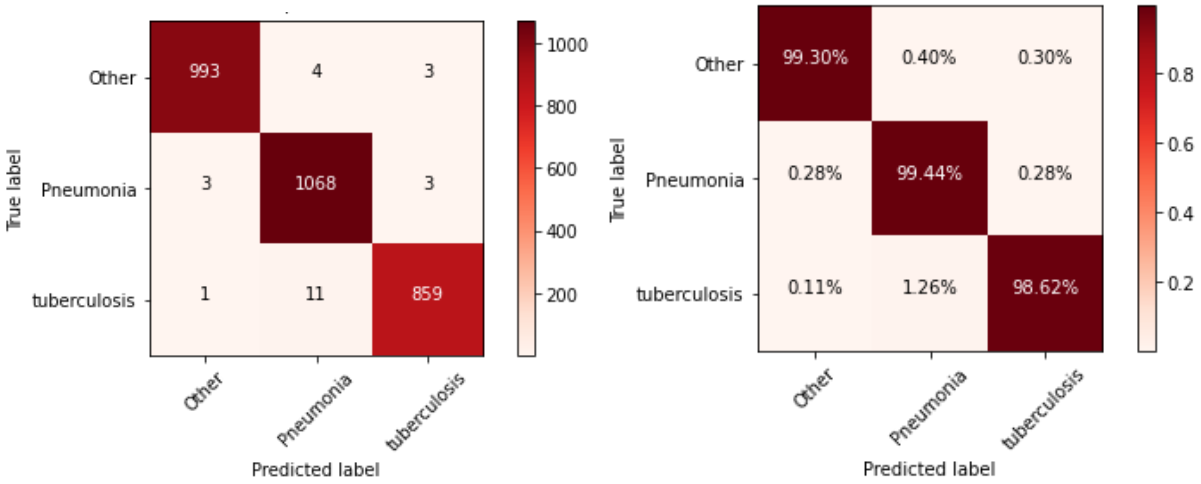



Figure 33: Confusion Matrix Result for K=3 with Numbers and Percentage

**K=5:** when applying the 5 value of k, the whole dataset is split into five classes, and model fitting is performed five times by swapping the test set. We recorded an average accuracy score of 99.45%, which is slightly better than the value recorded on k = 3 fold.

Table 12: Classification Result for K=5

| **Class** | **Precision (%)** | **Recall (%)** | **F1-score (%)** | **Accuracy (%)** |
|---|---|---|---|---|
| Pneumonia | 98.70 | 99.34 | 99.01 | 99.28 |
| TB | 99.19 | 98.85 | 99.01 | 99.42 |
| Other | 99.69 | 99.3 | 99.49 | 99.66 |
| **Average Accuracy** | | | | **99.45** |

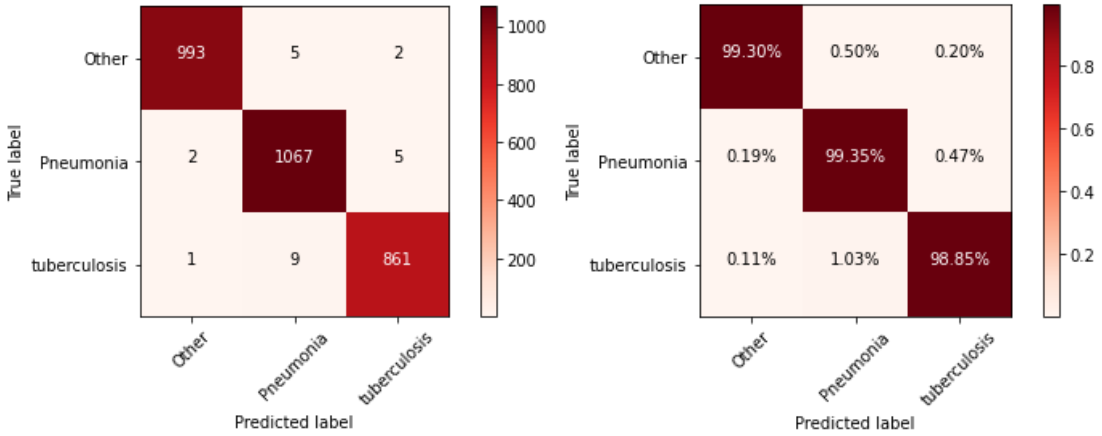



Figure 34: Confusion Matrix Result for K=5 with Numbers and Percentage

**K=10:** the whole dataset is partitioned into 10 sets and the model will be trained 10 times by swapping the training and test sets, using 90% of the data for training and the remaining 10% for testing. We have got better average accuracy on this training than 3 and 5 folds.

Table 13: Classification Result for K=10

| **Class** | **Precision (%)** | **Recall (%)** | **F1-score (%)** | **Accuracy (%)** |
|---|---|---|---|---|
| Pneumonia | 99.07 | 99.44 | 99.25 | 99.52 |
| TB | 99.53 | 99.19 | 99.35 | 99.62 |
| Other | 99.69 | 99.4 | 99.54 | 99.69 |
| | | | **Average accuracy** | **99.61** |

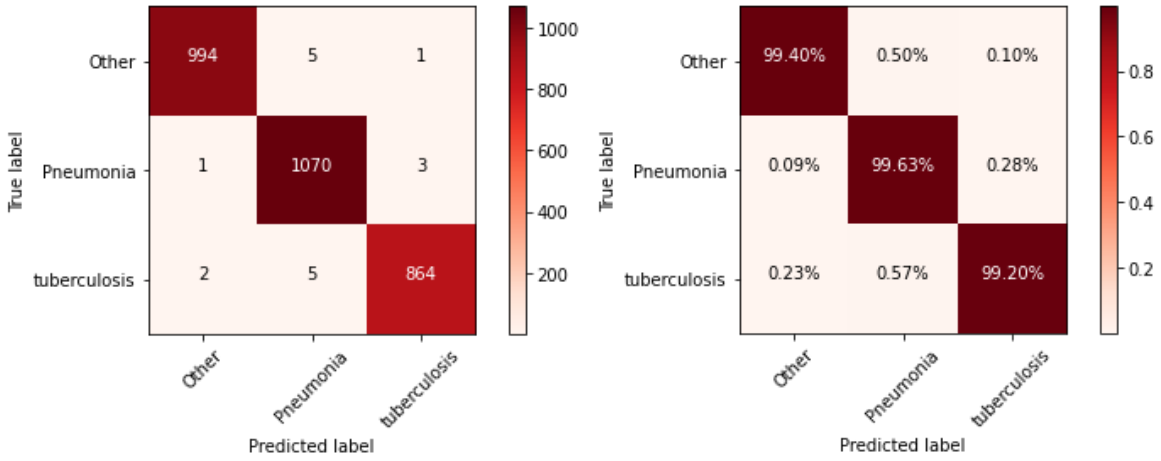


Figure 35: Confusion Matrix Result for K=10 with Numbers and Percentage

The training and validation accuracy and loss plot trained on 10 fold is shown below in figure 36.

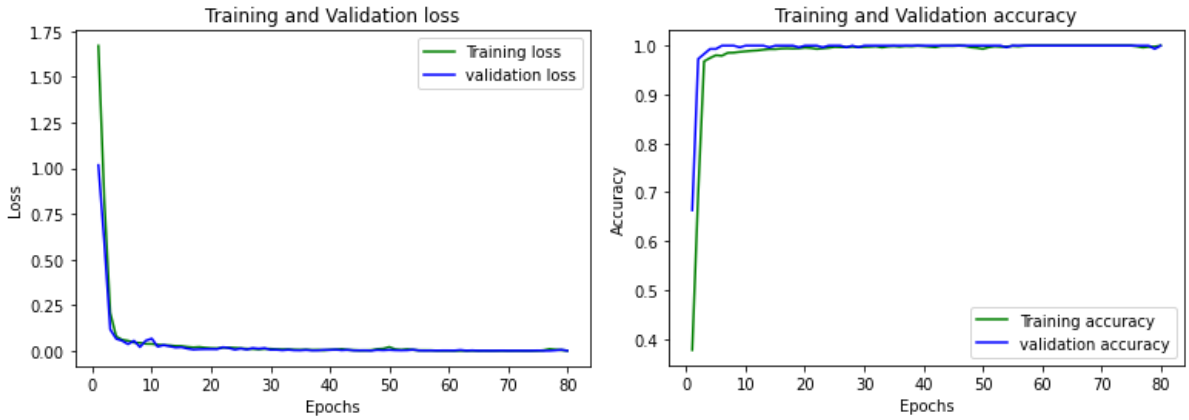



Figure 36: Accuracy and Loss Curve Results for K=10

By comparing the values obtained from k values of 3, 5, and 10 we have selected a k = 10 fold value to develop our final model.

For the sake of comparison of performance, we have also developed image-only and medical record-only models. We used CNN and MLP sequential models, respectively. To develop the image-only model, a similar number of neurons and layers as the CNN of the functional model is used. We applied L2 regularization to increase the performance of the model. The Adam optimizer with a learning rate of 0.0001. The LeakyReLU activation function is also used in the hidden layer of the model.

To develop the medical record-only model, we used similar layers as in the MLP of the functional model. The result of the MLP-only model confusion matrix is shown below in figure 37.

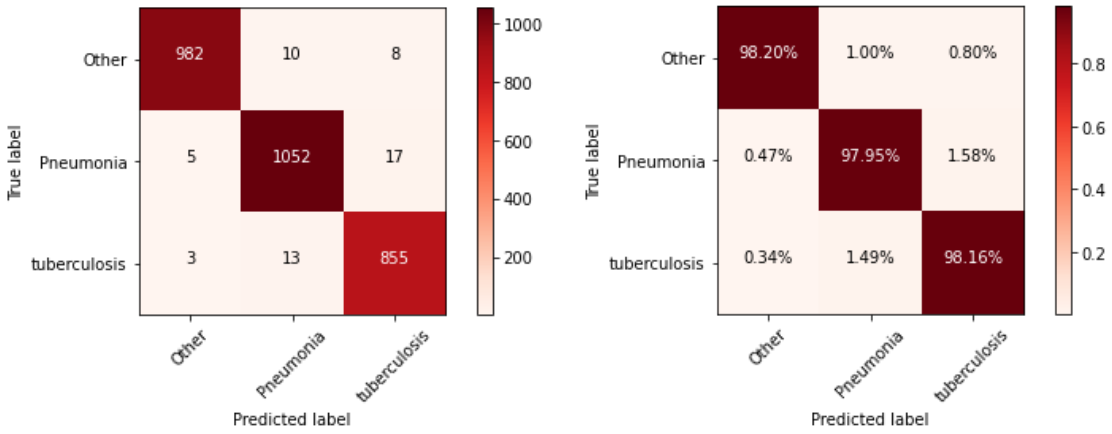



Figure 37: MLP Sequential Model Confusion Matrix Result with Numbers and Percentage

The confusion matrix that resulted from fitting the CNN sequential model is shown in figure 38.

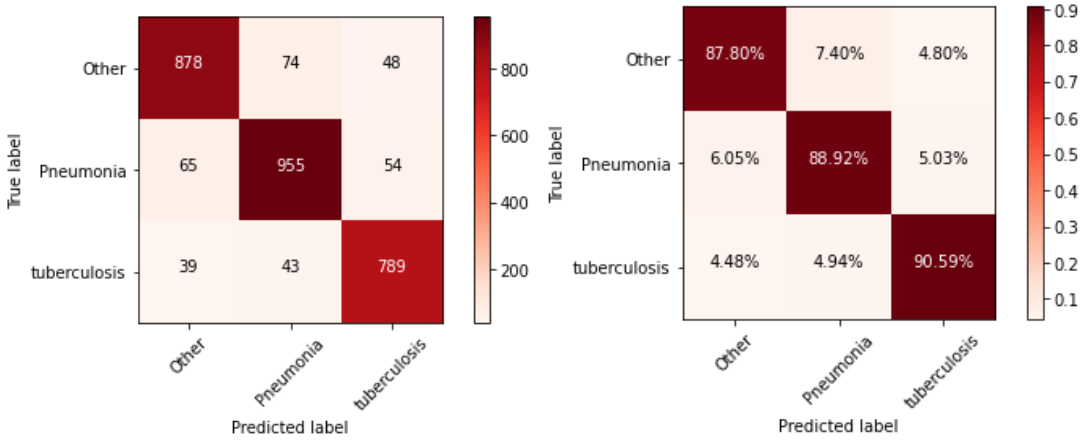



Figure 38: CNN Sequential Model Confusion Matrix Result with Numbers and Percentage

Table 14: Comparison of the Three Models

| Developed Model | MLP Only Model | CNN Only Model | MLP+CNN Model |
|---|---|---|---|
| Accuracy (%) | 98.72 | 92.68 | **99.61** |

As shown in table 14, the performance of the joint model exceeded the single models of image and medical record.

### 6.2.4. Senior (Medical Expert) Radiologist Evaluation Result

The developed model is evaluated by the senior radiologist with new data from medical records and associated chest x-ray images of patients. The developed model is evaluated by comparing the cases evaluated by a senior radiologist at St. Peter's specialized hospital and the model with those image and medical record cases. The results are shown in table 15.

Table 15: Medical Expert Evaluation Result

| **Case** | **Radiologist's Evaluation** | **Proposed Model's Detection** | |
|---|---|---|---|
| 1 | TB | TB | 1.00 |
| 2 | TB | TB | 1.00 |
| 3 | TB | Pneumonia | 0.394 |
| 4 | Pneumonia | pneumonia | 1.00 |
| 5 | Pneumonia | Pneumonia | 1.00 |
| 6 | Other | Other | 1.00 |
| 7 | Pneumonia | Pneumonia | 1.00 |
| 8 | Pneumonia | Pneumonia | 1.00 |
| 9 | TB | TB | 1.00 |
| 10 | Pneumonia | Pneumonia | 1.00 |
| Average accuracy | | | **0.9394** |

The results of the disease detection using the model we developed have accurately detected 9 cases out of 10. But the one case with a red color is not detected accurately by the model. Our model has scored 93.94% on detection performance on a new test dataset, and this is a good achievement.

## 6.3. Discussion

Previous works that applied deep learning to develop pulmonary disease detection model (Ayan & Ünver, 2019), (Rajpurkar et al., 2017), (Verma et al., 2020), (Hooda et al., 2017), (Haloi et al., 2018) applied only chest x-ray images to develop models. These works missed the right medical diagnosis procedure of including the patient's medical history for the proper disease detection from chest x-ray images. Khatibi et al. (Khatibi et al., 2020), who applied machine learning techniques to detect pneumonia and TB from patient symptoms and radiology reports, the models' performance is highly dependent on the decision made by the radiologist on the chest x-ray image. In other words, the medical images are not directly detected by the algorithms, and this creates bias. Even though the work of Bharati (Bharati et al., 2020) tried to use patient age, gender, and view position of the image information in addition to the chest x-ray image, this work failed to detect the specific disease; rather, it only shows whether the disease is detected or not.

Therefore, the first aim of this research is to develop a pulmonary infectious disease detection model by incorporating patient medical records and chest x-ray images together. The detection enables the identification of pneumonia and TB disease. We have collected local medical record data from alert hospital, Yekatit 12 hospital medical college, and St. Peter's specialized hospital, and chest x-ray images are collected from St. Peter's specialized hospital and online Kaggle repository. After data collection, we prepared a structured dataset with three target classes having 2945 rows with 18 columns of local medical record and corresponding chest x-ray image datasets as stated in ***section 3.3.***

We applied different preprocessing techniques for the two datasets and employed CNN and MLP algorithms for the model development. Feature level concatenation is performed to concatenate the 1 dimensional feature vector outputs from CNN and MLP.

In ***table 8,*** different combinations of height shift, width shift, rotation, and random brightness chest x-ray image augmentation techniques are applied, and width shift and rotation showed better performance than other augmentation technique combinations employed. As implemented in ***section 5.3.4,*** a random forest feature selection technique is applied to the medical record dataset. Then, the features importance is plotted in ***figure 30***. According to the plot, one feature

with no contribution to the performance of the model is dropped and 17 features are used to develop the model we proposed. This has improved the overall performance of our model, as shown in ***table 9.*** From this, we can infer that the feature selection of the features we apply has a contribution to the performance of the model we develop.

The combinations of the best augmentation result and feature selection are further exposed to hyper parameter tuning, and this has improved the performance of the model. By applying the values of the hyper parameters that gave us a better result, we evaluated our model with 3, 5 and 10 k-fold values of the k-fold cross validation technique. As shown in ***Table 13,*** the result of the 10-fold value got a more honorable performance than the corresponding 3 and 5-fold values. Considering all these results, the augmentation techniques of width shift and rotation tuned-up hyper parameters and 10-fold cross-validation are chosen to develop our pulmonary infectious disease detection model. It is seen that the combination of chest x-ray and clinical information outperformed the use of chest x-ray images only. As shown in ***Table 14,*** the joint use of medical records and chest x-ray images gives the best detection result. This is because the model uses both the medical record and chest x-ray image at the same time to decide on the detection of the pulmonary infectious disease. At last, we develop our pulmonary infectious disease detection model and deployed. Medical doctor evaluation is performed. As shown in ***table 15,*** in the evaluation our model has scored 93.94% by correctly detecting 9 out of 10 cases. To sum up, our pulmonary infectious disease detection model has performed well in the algorithmic and medical doctor evaluation.

The significance of the results of this research is that it makes real the implementation of pulmonary infectious disease detection from chest x-ray images and medical records, which resembles real-life medical diagnosis disease detection techniques. In addition, the joint model is more accurate and reliable than the image-only or medical record-only deep learning models. These things make the work worthy.

The research questions posited in ***section 1.4*** that this research aimed to answer, are presented with their answers below.

***RQ-1:*** *How can selected deep learning algorithms be employed to develop a model with more accurate results in pulmonary disease detection?*

The selected deep learning algorithms to develop pulmonary infectious disease detection were CNN and MLP. By applying feature-level concatenation, selected algorithms give more accurate result in the detection of pulmonary infectious diseases. First we get the 1 dimensional feature vector output from the MLP. Again, by flattening the convolution output of the CNN, we get another 1 dimensional feature vector. Then we perform feature level concatenation of the two 1 dimensional feature vectors. After we perform concatenation of feature vectors, we further make the fully connected layers to learn the concatenated feature vector and perform disease detection output. Therefore, by performing feature level concatenation, we made the two algorithms give better performance in the detection of the disease. In addition, we have proved that this concatenation has outperformed the utilization of CNN or MLP algorithms separately for disease detection.

***RQ2.*** *How can image augmentation affect the pulmonary infectious disease detection performance of the model?*

The image augmentation techniques applied in this work are different combinations of width shift, height shift of 0.2 each, rotation to -5 degree, random brightness range of 0.3 to 0.9.

Five of six augmentation technique combinations we applied have shown to increase the accuracy of the model than the original images. The remaining one augmentation technique combination applied has resulted in a decrease in the model's performance. Specifically, the performance of the experiment that employed the width shift and rotation techniques has exceeded all other combinations in accuracy performance. In general, we conclude that applying augmentation techniques to medical images can improve the performance of the model which performed feature level concatenation with the medical records.

***RQ-3:*** *Can the performance of the pulmonary infectious disease detection model be improved by combining medical records and chest x-ray images?*

Yes, the combination of medical records with chest x-ray images has improved the pulmonary infectious disease detection performance compared to using only chest x-ray images. The

addition of medical records to the chest x-ray images was improved by using proper preprocessing techniques on the images and medical records. Another task performed on the medical record dataset is checking the importance of features and selecting the features that contribute to the model performance. When combining medical records and chest x-rays, there are more features to learn from than a single medical record or chest x-ray data. This enhances the model's ability to properly classify cases.

As far as we discovered, there is no pulmonary infectious disease detection model developed to detect pulmonary infectious disease using chest x-ray images and medical records beyond the age and gender of the patient. We have used our own local dataset of medical records and also a mixture of online and local chest x-ray images. Due to this, we are unable to compare this work to other works of pulmonary disease detection models directly.

# CONCLUSION AND FUTURE WORK

## Conclusion

In this study, we have developed a pulmonary infectious disease detection model by using chest x-ray images and medical records. To achieve our goal, we have come through different steps like data preprocessing, feature selection, and image augmentation. Deep learning algorithms of convolutional neural networks and multilayer perceptron are also implemented in this thesis. The output feature vectors of the two algorithms are concatenated by feature level concatenation. In the feature selection process we performed, we found that close contact with the chronic cougher parameter has no importance to the detection model. But medically, this feature is important in diagnosing coughers, more specifically children, as well as other people at any age. If anyone is in close contact, living or working with a person infected with TB disease, there is a high chance of being infected with TB. Our decision on the feature selection could be a good model for others to not rely only on algorithmic outputs but to consider real-life practice too. Even if we believe we have done enough experiments to answer the research questions, this study also needs more work by using more medical record datasets and labeled local chest x-ray images. We have proven that the joint use of medical records and chest x-ray images improves the performance of the image-only or medical record-only models. This study has a great deal of value to the medical field since it closely resembles how radiologists read and make decisions about the appearance of chest x-ray images in real-world medical diagnosis systems. The output of this study aids the medical doctors (radiologists) in making accurate and timely disease detection from chest x-ray image and medical records. The patients are also the beneficiaries of the output of this study in which they get accurate diagnosis of the disease in time. The model developed eliminates the inaccuracy problems found on the models developed by using only chest x-ray images or medical records.

## Future Work

Pulmonary infectious disease detection from chest x-ray image is difficult since the diseases have similar pathogens and they highly mimic each other. Pneumonia and TB highly mimic on chest x-ray image and need clinical information to make the right diagnosis. We tried to fill this gap with this work. Due to the unavailability of labeled data, the chest x-ray images we used are only the frontal views. Mixing the lateral view images could make the model more flexible and

decide on chest x-ray inputs from the lateral view more accurately. In the future, the use of more local medical records and chest x-ray images from different hospitals is recommended to develop such disease detection models. While developing a TB detection model, for better accuracy and treatment we recommend classifying the status of TB detected as TB infection with no disease; TB infection, clinically active**;** or TB infection, clinically inactive. It is also recommended to classify pneumonia with its subclasses. In the future applying Generative adversarial networks for the chest x-ray image augmentation is also recommended.

* * *

This research work fund was granted by Adama Science and Technology University.

**ASTU/SM-R/506/22**

# REFERENCES


Abate, B. B., Kasie, A. M., Reta, M. A., & Kassaw, M. W. (2020). Neonatal sepsis and its associated factors in East Africa: a systematic review and meta-analysis. *International Journal of Public Health*, *65*(9), 1623–1633. https://doi.org/10.1007/s00038-020-01489-x

Abate, D., Taye, B., Abseno, M., & Biadgilign, S. (2012). Epidemiology of anti-tuberculosis drug resistance patterns and trends in tuberculosis referral hospital in Addis Ababa , Ethiopia. *BMC Research Notes*, *5*(1), 1. https://doi.org/10.1186/1756-0500-5-462

Ayan, E., & Ünver, H. M. (2019). Diagnosis of pneumonia from chest X-ray images using deep learning. *2019 Scientific Meeting on Electrical-Electronics and Biomedical Engineering and Computer Science, EBBT 2019*, 0–4. https://doi.org/10.1109/EBBT.2019.8741582

Bharati, S., Podder, P., & Mondal, M. R. H. (2020). Hybrid deep learning for detecting lung diseases from X-ray images. *Informatics in Medicine Unlocked*, *20*, 100391. https://doi.org/10.1016/j.imu.2020.100391

Bohadana, A., Izbicki, G., & Kraman, S. S. (2014). *Fundamentals of Lung Auscultation*. https://doi.org/10.1056/NEJMra1302901

Castillo, E. (1998). Functional Networks. *Neural Processing Letters*, *7*(3), 151–159. https://doi.org/10.1023/A:1009656525752

Chaitrali S., D., & Sulabha S., A. (2012). a Data Mining Approach for Prediction of Heart Disease. *Nternational Journal of Computer Engineering and Technology (IJCET), ISSN 0976 – 6367(Print), ISSN 0976 – 6375(Online) Volume 3, Issue*, *3*(Issue 3, October-December), 30–40. http://ssrn.com/abstract=2175569

*Chapter 4 Diagnosis of Tuberculosis Disease* (pp. 75–108). (n.d.).

Chen Junli, & Jiao Licheng. (2002). *Classification mechanism of support vector machines*. 1556–1559. https://doi.org/10.1109/icosp.2000.893396

Chen, R. C., Dewi, C., Huang, S. W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, *7*(1).

https://doi.org/10.1186/s40537-020-00327-4

Cillóniz, C., Menéndez, R., García-Vidal, C., Péricas, J. M., & Torres, A. (2020). Defining Community-Acquired Pneumonia as a Public Health Threat: Arguments in Favor from Spanish Investigators. *Medical Sciences (Basel, Switzerland)*, *8*(1), 6. https://doi.org/10.3390/MEDSCI8010006

Cunningham, P., & Delany, S. J. (2021). K-Nearest Neighbour Classifiers-A Tutorial. *ACM Computing Surveys*, *54*(6). https://doi.org/10.1145/3459665

Ding, B., Qian, H., & Zhou, J. (2018). Activation functions and their characteristics in deep neural networks. *2018 Chinese Control And Decision Conference (CCDC)*, *3*(2), 1836–1841. https://doi.org/10.1109/CCDC.2018.8407425

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, *35*(5–6), 352–359. https://doi.org/10.1016/S1532-0464(03)00034-0

Elgendi, M., Nasir, M. U., Tang, Q., Smith, D., Grenier, J. P., Batte, C., Spieler, B., Leslie, W. D., Menon, C., Fletcher, R. R., Howard, N., Ward, R., Parker, W., & Nicolaou, S. (2021). The Effectiveness of Image Augmentation in Deep Learning Networks for Detecting COVID-19: A Geometric Transformation Perspective. *Frontiers in Medicine*, *8*(March), 1–12. https://doi.org/10.3389/fmed.2021.629134

Ertam, F. (2017). Data classification with deep learning using tensorflow. *2nd International Conference on Computer Science and Engineering, UBMK 2017*, 755–758. https://doi.org/10.1109/UBMK.2017.8093521

*European Respiratory Society Annual Congress 2012*. (2012). 7–8.

F.Y, O., J.E.T, A., O, A., J. O, H., O, O., & J, A. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal of Computer Trends and Technology*, *48*(3), 128–138. https://doi.org/10.14445/22312803/ijctt-v48p126

Fitz-gerald, M. (2013). *Sputum Microscopy The handbook*.

Global tuberculosis report. (2021). In *World Health Organization*. https://apps.who.int/iris/rest/bitstreams/1379788/retrieve

Haloi, M. (2015). *Traffic Sign Classification Using Deep Inception Based Convolutional Networks*. 11–15. http://arxiv.org/abs/1511.02992

Haloi, M., Rajalakshmi, K. R., & Walia, P. (2018). *Towards Radiologist-Level Accurate Deep Learning System for Pulmonary Screening*. http://arxiv.org/abs/1807.03120

Haohan Wang, B. R. (1956). On the Origin of Deep Learning. *Journal of the Fisheries Research Board of Canada*, *13*(3), 303–308. https://doi.org/10.1139/f56-020

Hooda, R., Sofat, S., Kaur, S., Mittal, A., & Meriaudeau, F. (2017). Deep-learning: A potential method for tuberculosis detection using chest radiography. *Proceedings of the 2017 IEEE International Conference on Signal and Image Processing Applications, ICSIPA 2017*, 497–502. https://doi.org/10.1109/ICSIPA.2017.8120663

Hossin, M., & Sulaiman, M. . (2015). a Review on Evaluation Metrics for Data Classification Evaluation. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, *5*(2), 1–11.

Huang, S. C., Pareek, A., Seyyedi, S., Banerjee, I., & Lungren, M. P. (2020). Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *Npj Digital Medicine*, *3*(1). https://doi.org/10.1038/s41746-020-00341-z

Huang, Y., & Li, L. (2011). Naive Bayes classification algorithm based on small sample set. *CCIS2011 - Proceedings: 2011 IEEE International Conference on Cloud Computing and Intelligence Systems*, *2*, 34–39. https://doi.org/10.1109/CCIS.2011.6045027

Hüsken, M., & Stagge, P. (2003). Recurrent neural networks for time series classification. *Neurocomputing*, *50*, 223–235. https://doi.org/10.1016/S0925-2312(01)00706-8

Jovic, A., Brkic, K., & Bogunovic, N. (2015). A review of feature selection methods with applications. *2015 38th International Convention on Information and Communication*

*Technology, Electronics and Microelectronics (MIPRO)*, *112*(May), 1200–1205. https://doi.org/10.1109/MIPRO.2015.7160458

Kadra, A., Lindauer, M., Hutter, F., & Grabocka, J. (2021). *Well-tuned Simple Nets Excel on Tabular Datasets*. *NeurIPS*. http://arxiv.org/abs/2106.11189

Kamolson, S. (2007). Fundamentals of quantitative research Suphat Sukamolson, Ph.D. Language Institute Chulalongkorn University. *Language Institute*, 20. http://www.culi.chula.ac.th/e-Journal/bod/Suphat Sukamolson.pdf%5Cnhttp://isites.harvard.edu/fs/docs/icb.topic1463827.files/2007_Sukamolson_Fundamentals of Quantitative Research.pdf

Karabulut, E. M., Özel, S. A., & İbrikçi, T. (2012). A comparative study on the effect of feature selection on classification accuracy. *Procedia Technology*, *1*, 323–327. https://doi.org/10.1016/j.protcy.2012.02.068

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M. K., Pei, J., Ting, M., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., … Zhang, K. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, *172*(5), 1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010

Khatibi, T., Farahani, A., & Sarmadian, H. (2020). Proposing a two-step Decision Support System (TPIS) based on Stacked ensemble classifier for early and low cost (step-1) and final (step-2) differential diagnosis of …. *ArXiv Preprint ArXiv:2009.02316*. https://arxiv.org/abs/2009.02316

Khosla, C., & Saini, B. S. (2020). Enhancing Performance of Deep Learning Models with different Data Augmentation Techniques: A Survey. *Proceedings of International Conference on Intelligent Engineering and Management, ICIEM 2020*, 79–85. https://doi.org/10.1109/ICIEM48762.2020.9160048

Kitesa, G., & Mamo, E. (2021). *Effect of Electronic Health Records on Patient Satisfaction and Waiting Time at Selected Hospitals , in Addis Ababa*. 1–17.

Konstantinos, A. (2010). Testing for tuberculosis. *Australian Prescriber*, *33*(1), 12–18. https://doi.org/10.18773/austprescr.2010.005

Laufer, P. (2013). Practice Gap. *Pediatrics in Review*, *34*(10), 439.

Leung, A. N. (1999). Pulmonary tuberculosis: The essentials. *Radiology*, *210*(2), 307–322. https://doi.org/10.1148/radiology.210.2.r99ja34307

Mariam, W. K. (1986). Historical outline of ALERT. *Leprosy Review*, *57*(SUPPL. 1), 9–25. https://doi.org/10.5935/0305-7518.19860039

Mnyambwa, N. P., Philbert, D., Kimaro, G., Wandiga, S., Kirenga, B., Mmbaga, B. T., Muttamba, W., Najjingo, I., Walusimbi, S., Nuwarinda, R., Okelloh, D., Semvua, H., Ngocho, J., Senkoro, M., Stephen, O., Castelnuovo, B., Wilfred, A., Mgina, E., Sanga, C., … Ngadaya, E. (2021). Gaps related to screening and diagnosis of tuberculosis in care cascade in selected health facilities in East Africa countries: A retrospective study. *Journal of Clinical Tuberculosis and Other Mycobacterial Diseases*, *25*, 100278. https://doi.org/10.1016/j.jctube.2021.100278

Morales, F. (2020). *keras Ocr Documentation*.

Muse Ibrahim, I. (2020). *Developing a computer-aided diagnosis model for TB using region-based convolutional neural network*. *November*.

Mustra, M., Delac, K., & Grgic, M. (2008). Overview of the DICOM standard. *Proceedings Elmar - International Symposium Electronics in Marine*, *1*(October 2008), 39–44.

Nguyen, L. D., Lin, D., Lin, Z., & Cao, J. (2018). Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation. *Proceedings - IEEE International Symposium on Circuits and Systems*, *2018-May*, 3–7. https://doi.org/10.1109/ISCAS.2018.8351550

Odeyemi, A. O., Oyedeji, A. O., Adebami, O. J., Odeyemi, A. O., & Agelebe, E. (2020). Complications of pneumonia and its associated factors in a pediatric population in Osogbo, Nigeria. *Nigerian Journal of Paediatrics*, *47*(4), 318–323.

https://doi.org/10.4314/njp.v47i4.4

Patil, A., & Rane, M. (2021). Convolutional Neural Networks: An Overview and Its Applications in Pattern Recognition. *Smart Innovation, Systems and Technologies*, *195*, 21–30. https://doi.org/10.1007/978-981-15-7078-0_3

Pinto, L. M., Shah, A. C., Shah, K. D., & Udwadia, Z. F. (2011). Pulmonary tuberculosis masquerading as community acquired pneumonia. *Respiratory Medicine CME*, *4*(3), 138–140. https://doi.org/10.1016/j.rmedc.2010.11.004

Qin, C., Yao, D., Shi, Y., & Song, Z. (2018). Computer - aided detection in chest radiography based on artificial intelligence : a survey. *BioMedical Engineering OnLine*, 1–23. https://doi.org/10.1186/s12938-018-0544-y

Rahman, T., Khandakar, A., Kadir, M. A., Islam, K. R., Islam, K. F., Mazhar, R., Hamid, T., Islam, M. T., Kashem, S., Mahbub, Z. Bin, Ayari, M. A., & Chowdhury, M. E. H. (2020). Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization. *IEEE Access*, *8*, 191586–191601. https://doi.org/10.1109/ACCESS.2020.3031384

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). *CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning*. 3–9. http://arxiv.org/abs/1711.05225

Sarvamangala, D. R., & Kulkarni, R. V. (2022). Convolutional neural networks in medical image understanding: a survey. *Evolutionary Intelligence*, *15*(1), 1–22. https://doi.org/10.1007/s12065-020-00540-3

Shijie, J., Ping, W., Peiyi, J., & Siping, H. (2017). Research on data augmentation for image classification based on convolution neural networks. *Proceedings - 2017 Chinese Automation Congress, CAC 2017*, *2017-Janua*(201602118), 4165–4170. https://doi.org/10.1109/CAC.2017.8243510

Sirazitdinov, I., Kholiavchenko, M., Kuleev, R., & Ibragimov, B. (2019). Data augmentation

for chest pathologies classification. *Proceedings - International Symposium on Biomedical Imaging*, *2019-April*(Isbi), 1216–1219. https://doi.org/10.1109/ISBI.2019.8759573

Skinner, S. (2015). *Guide to thoracic imaging*. *44*(8), 558–563.

Stirenko, S., Kochura, Y., Alienin, O., Rokovyi, O., Gordienko, Y., Gang, P., & Zeng, W. (2018). Chest X-Ray Analysis of Tuberculosis by Deep Learning with Segmentation and Augmentation. *2018 IEEE 38th International Conference on Electronics and Nanotechnology, ELNANO 2018 - Proceedings*, 422–428. https://doi.org/10.1109/ELNANO.2018.8477564

*The Lungs - Fact Sheet*. (2018). *7*(1). https://lungfoundation.com.au/wp-content/uploads/2018/09/Factsheet-The-Lungs-Jul13.pdf

Tilve, A., Nayak, S., Vernekar, S., Turi, D., Shetgaonkar, P. R., & Aswale, S. (2020). Pneumonia Detection Using Deep Learning Approaches. *International Conference on Emerging Trends in Information Technology and Engineering, Ic-ETITE 2020*, 1–8. https://doi.org/10.1109/ic-ETITE47903.2020.152

Tuazon, A. O. (2005). *Complications of Pneumonia in Children*.

*Tuberculosis ( TB ) Blood Test ( IGRA )*. (n.d.). 64975.

Verma, D., Bose, C., Tufchi, N., Pant, K., Tripathi, V., & Thapliyal, A. (2020). An efficient framework for identification of Tuberculosis and Pneumonia in chest X-ray images using Neural Network. *Procedia Computer Science*, *171*(2019), 217–224. https://doi.org/10.1016/j.procs.2020.04.023

Vock, P. (1987). *Frequent errors in the interpretation of chest x-ray films*. Schweizerische Medizinische Wochenschrift. https://pubmed.ncbi.nlm.nih.gov/3616585/

Wong, T. T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, *48*(9), 2839–2846. https://doi.org/10.1016/j.patcog.2015.03.009

Yasaka, K., Akai, H., Kunimatsu, A., Kiryu, S., & Abe, O. (2018). Deep learning with convolutional neural network in radiology. *Japanese Journal of Radiology*, *36*(4), 257–272. https://doi.org/10.1007/s11604-018-0726-3

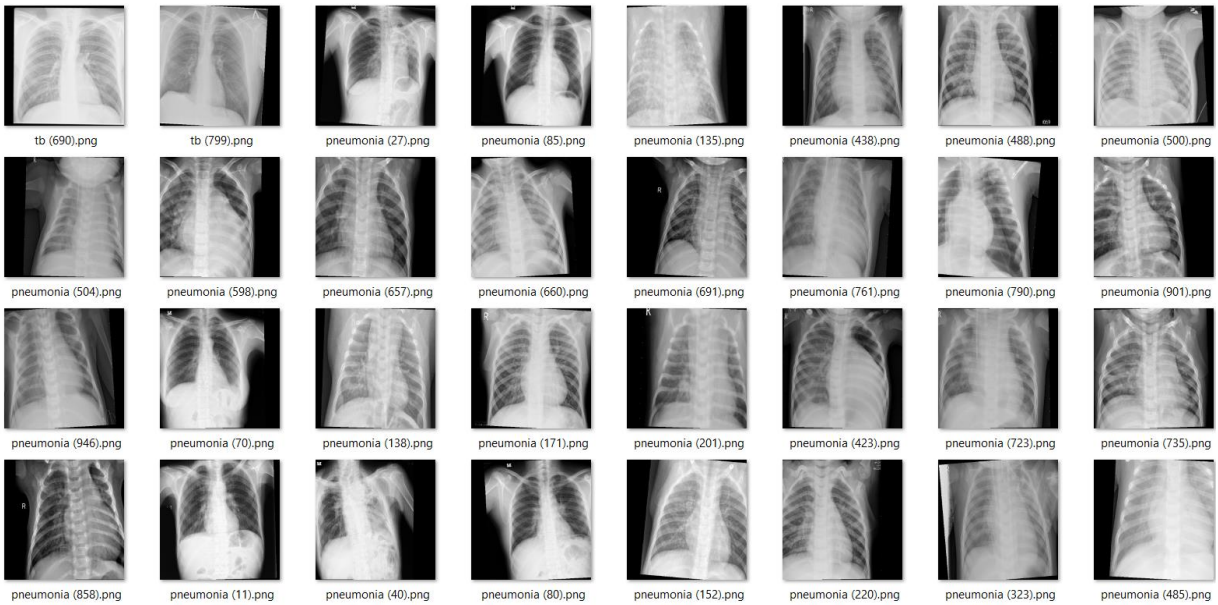
Zhai, X., Ali, A. A. S., Amira, A., & Bensaali, F. (2016). MLP Neural Network Based Gas Classification System on Zynq SoC. *IEEE Access*, *4*(c), 8138–8146. https://doi.org/10.1109/ACCESS.2016.2619181

# APPENDIX

## Appendix A: Sample Medical Record Dataset

| | Age | Sex | Temp | RR | PR | Cough_D | Cough_T | WBC | CP | SOB | SPO2 | Fever | NS | HTB | Contact | HS | WL | HIV | Diagnosis |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 50 years | M | 36.3 | 18 | 92 | 5 weeks | yellowish | 6.2 | yes | yes | 90 | low | yes | no | no | no | yes | N | TB |
| 3 | 47 years | M | 37 | 15 | 64 | 2 days | dry | 11 | no | no | 97 | no | no | no | no | no | no | N | Other |
| 4 | 62 years | M | 37.9 | 26 | 108 | 4 weeks | dry | 6 | yes | yes | 84 | high | yes | yes | yes | yes | no | N | Other |
| 5 | 23 days | F | 36 | 49 | 137 | 2 days | dry | 4 | no | no | 97 | no | no | no | no | no | no | P | Other |
| 6 | 70 years | M | 36.2 | 24 | 116 | 4 days | dry | 8.37 | yes | yes | 94 | low | no | no | no | no | no | P | Pneumonia |
| 7 | 72 years | F | 36 | 16 | 80 | 10 days | dry | 4.7 | yes | yes | 95 | high | no | no | no | no | no | N | Pneumonia |
| 8 | 9 years | F | 36 | 23 | 57 | 2 days | dry | 12 | no | no | 98 | no | no | no | no | no | no | N | Other |
| 9 | 9 months | F | 36.1 | 30 | 130 | 3 weeks | yellowish | 8.08 | yes | yes | 90 | low | yes | no | no | no | yes | N | TB |
| 10 | 17 years | M | 36.4 | 30 | 114 | 4 days | dry | 5.9 | yes | yes | 93 | high | no | no | no | no | no | N | Pneumonia |
| 11 | 1 year | F | 36.2 | 38 | 132 | 5 days | dry | 8.37 | yes | yes | 93 | high | no | no | no | no | no | N | Pneumonia |
| 12 | 18 years | M | 36 | 17 | 78 | 2 days | dry | 8 | no | no | 97 | no | no | no | no | no | no | N | Other |
| 13 | 23 years | M | 38 | 20 | 121 | 3 weeks | whitish | 8.37 | yes | yes | 94 | high | yes | no | no | no | no | N | Pneumonia |
| 14 | 3 months | M | 36.6 | 36 | 170 | 3 days | dry | 8.37 | yes | yes | 99 | high | no | no | yes | no | no | N | Pneumonia |
| 15 | 55 years | M | 36.3 | 24 | 80 | 3 months | yellowish | 8.08 | yes | yes | 90 | low | yes | no | no | no | yes | P | TB |
| 16 | 3 years | F | 37.2 | 28 | 102 | 1 month | yellowish | 8.08 | yes | yes | 98 | low | yes | no | no | no | yes | N | TB |
| 17 | 20 days | F | 36 | 23 | 112 | 4 days | dry | 8.37 | yes | yes | 98 | high | no | no | no | no | no | N | Pneumonia |
| 18 | 35 years | M | 36.3 | 25 | 106 | 2 month | yellowish | 6.2 | yes | yes | 90 | low | yes | no | no | no | yes | P | TB |
| 19 | 15 years | M | 37 | 14 | 88 | 2 days | dry | 10 | no | no | 98 | no | no | yes | no | no | no | N | Other |
| 20 | 35 years | M | 36.3 | 19 | 121 | 2 weeks | yellowish | 9.5 | yes | yes | 93 | low | yes | no | no | no | yes | P | TB |
| 21 | 53 years | F | 37 | 22 | 98 | 1 week | dry | 14.4 | yes | yes | 93 | high | no | no | no | no | no | N | Pneumonia |
| 22 | 1 month | M | 36.2 | 38 | 118 | 1 week | whitish | 8.08 | yes | yes | 96 | low | yes | no | yes | no | no | N | TB |
| 23 | 38 years | M | 36.3 | 25 | 156 | 2 month | yellowish | 4.3 | yes | yes | 61 | low | yes | no | no | no | yes | P | TB |
| 24 | 14 years | F | 36 | 14 | 67 | 2 days | dry | 11 | no | no | 100 | no | no | no | no | no | no | N | Other |
| 25 | 59 years | M | 37 | 16 | 88 | 2 days | dry | 10 | no | no | 99 | no | no | no | yes | no | no | N | Other |
| 26 | 10 years | M | 37 | 29 | 89 | 2 days | dry | 12 | no | no | 95 | no | no | no | no | no | no | N | Other |
| 27 | 2 months | F | 37 | 32 | 138 | 2 days | dry | 10 | no | no | 97 | no | no | no | no | no | no | P | Other |
| 28 | 24 years | M | 36.3 | 25 | 102 | 1 month | yellowish | 8.08 | yes | yes | 90 | low | yes | no | yes | no | yes | N | TB |
| 29 | 1 month | M | 36 | 34 | 132 | 2 days | dry | 13 | no | no | 96 | no | no | no | no | no | no | P | Other |
| 30 | 3 months | F | 36.1 | 63 | 129 | 3 days | whitish | 10.24 | yes | yes | 99 | high | yes | yes | no | no | no | N | TB |
| 31 | 7 years | F | 36.4 | 36 | 136 | 1 week | dry | 8.37 | yes | yes | 93 | high | no | no | no | no | no | N | Pneumonia |
| 32 | 39 years | M | 36.4 | 30 | 114 | 1 month | whitish | 3.6 | yes | yes | 93 | high | yes | yes | no | no | no | N | Pneumonia |

## Appendix B: Sample Chest X-ray Image Dataset

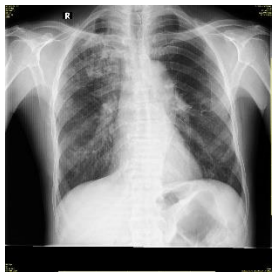


## Appendix C: Medical Record Dataset Features Description

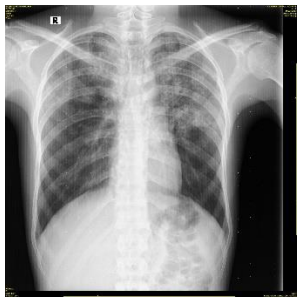
| Feature | Symbol | Description |
|---|---|---|
| Age | Age | The patient's age |
| Sex | Sex | The patient's gender |
| Temperature | Temp | The body temperature of the patient |
| Respiratory rate | RR | The number of breaths the patient takes in a minute |
| Pulse Rate | PR | The number of times the heart beats per minute |
| Cough duration | Cough_D | How long does the cough lasts |
| Cough type | Cough_T | Type of cough the patient experiences related to sputum. |
| White Blood Cell count | WBC | The number of white blood cells in the blood |
| Chest Pain | CP | Pain in any area of the chest |
| Shortness of Breath | SOB | Difficulty in breathing |
| Oxygen Saturation | SPO2 | Amount of oxygen in the blood |
| Fever | Fever | Increase in body temperature |
| Night Sweat | NS | Sweating during sleep |
| History of TB | HTB | Previously diagnosed as TB |
| Contact with Chronic Cougher | Contact | Contact with the person coughs for long time or TB patient, either living together or working together |
| Weight Loss | WL | Reduction of weight due to illness |
| HIV test result | HIV | If a patient have HIV virus or not |
| Diagnosis | Diagnosis | The detected disease |

## Appendix D: Test Cases for Evaluation

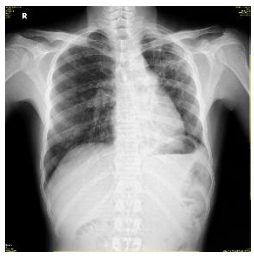
**Case 1:** a 60 years old male with temperature 36.5, respiratory rate 25, pulse rate 102, 2 months of cough duration, whitish sputum cough type, 7.5 count of white blood cell, oxygen saturation 90, have chest pain, shortness of breath and night sweat, no fever, no history of TB, no contact with chronic cougher, have weight loss, negative HIV test result.



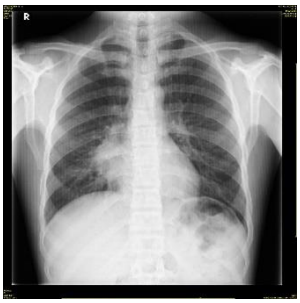
**Case 2:** a 29 years old male with temperature 36.2, respiratory rate 30, pulse rate 109, oxygen saturation 90, 4 months of cough duration, yellowish sputum cough type, 5.3 count of white blood cell, have chest pain, shortness of breath and night sweat, low fever, no history of TB, no contact with chronic cougher, have weight loss, negative HIV test result.



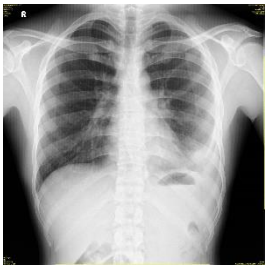
**Case 3:** a 24 years old male with temperature 36, respiratory rate 20, pulse rate 100, oxygen saturation 85, 2 weeks of cough duration, whitish sputum cough type, 5.2 count of white blood cell, have chest pain and shortness of breath, no fever, no night sweat, no history of TB, no contact with chronic cougher, have weight loss, negative HIV test result.



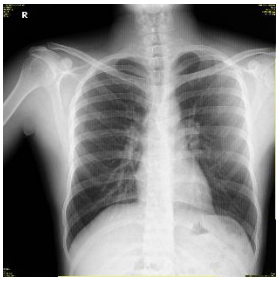
**Case 4:** a 30 years old male with temperature 37, respiratory rate 36, pulse rate 145, oxygen saturation 94, 1 week of cough duration, whitish sputum cough type, 8.4 count of white blood cell, have chest pain and shortness of breath, high fever, have night sweat, no history of TB, no contact with chronic cougher, have weight loss, Negative HIV test result.



**Case 5:** a 25 years old male with temperature 37, respiratory rate 19, pulse rate 66, oxygen saturation 97, 2 weeks of cough duration, dry cough type, 5.4 count of white blood cell, no chest pain, have shortness of breath, high fever, have night sweat, no history of TB, no contact with chronic cougher, no weight loss, Negative HIV test result.

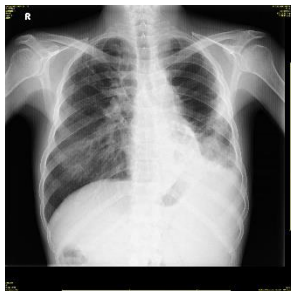


**Case 6:** a 28 years old male with temperature 36.3, respiratory rate 25, pulse rate 102, oxygen saturation 90, 2 weeks of cough duration, dry cough type, 7.0 count of white blood cell, no chest pain, no shortness of breath, no fever, no night sweat, no history of TB, no contact with chronic cougher, no weight loss, Negative HIV test result.

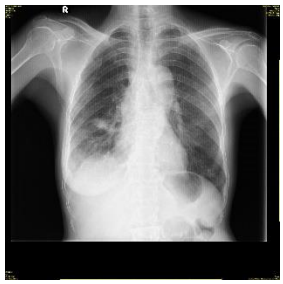


**Case 7:** a 28 years old male with temperature 36, respiratory rate 33, pulse rate 124, oxygen saturation 95, 3 weeks of cough duration, dry cough type, 8.4 count of white blood cell, have

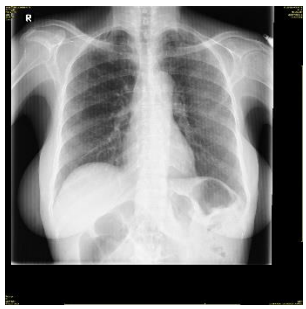
chest pain, have shortness of breath, low fever, no night sweat, no history of TB, no contact with chronic cougher, no weight loss, Negative HIV test result.



**Case 8:** a 60 years old female with temperature 37.5, respiratory rate 26, pulse rate 115, oxygen saturation 94, 2 weeks of cough duration, whitish cough type, 8.2 count of white blood cell, have chest pain, have shortness of breath, high fever, no night sweat, no history of TB, no contact with chronic cougher, no weight loss, Negative HIV test result.

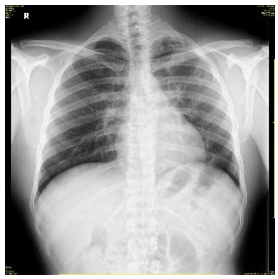


**Case 9:** a 50 years old female with temperature 36.5, respiratory rate 25, pulse rate 104, oxygen saturation 92, 3 months of cough duration, yellowish cough type, 7.4 count of white blood cell, have chest pain, have shortness of breath, no fever, have night sweat, have a history of TB, no contact with chronic cougher, have weight loss, positive HIV test result.



**Case 10:** a 26 years old male with temperature 37.2, respiratory rate 25.5, pulse rate 102, oxygen saturation 93, 1 month of cough duration, dry cough type, 7.7 count of white blood cell, have

chest pain, have shortness of breath, low fever, no night sweat, no history of TB, no contact with chronic cougher, no weight loss, Negative HIV test result.



## Appendix E: Pulmonary Infectious Disease Detection GUI