Chest Radiograph Interpretation with Deep Learning

By

Mubashir Ahmad

School of Physics, Engineering and Computer Science

Submitted to the University of Hertfordshire in partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy.

Supervisors

Prof. Farshid Amirabdolahian
Dr. Kheng Lee Koay
Dr. Yi Sun

July 2024

Dedication

This work is dedicated to the Prophet Muhammad (S.A.W), whose unparalleled contributions to knowledge and guidance have illuminated the path for mankind. His teachings and wisdom have been a constant source of inspiration throughout my academic journey. I also dedicate this thesis to my beloved parents, Mr. and Mrs. Capt. Saeed Ahmed. Their unwavering support, both through their prayers and their generous financial and moral support, has been instrumental in my success. Their love and encouragement have given me the strength and determination to persevere and achieve my goals.

Abstract

Chest radiographs are one of the most commonly used diagnostic modalities in healthcare due to their effectiveness in detecting conditions related to the thoracic region. However, the increasing demand for radiological services, coupled with a shortage of radiologists and the potential for diagnostic errors, demands innovative solutions. The integration of artificial intelligence (AI) and deep learning techniques offers a promising approach to support radiology professionals and enhance the diagnostic accuracy and efficiency of the diagnostic process. This thesis addresses the challenge of improving the performance of deep learning models for multi-label classification of chest radiographs using the CheXpert dataset. It focuses on both model-centric and data-centric methods, specifically techniques such as Gaussian Mixture Models (GMM) for relabeling uncertain labels, multi-scale template matching for focused learning, a custom pooling layer for better feature extraction, and Sequential Multi-Label Enrichment (SMLE) for improved detection of coexisting conditions. The primary objective is to enhance the reliability and performance of AI models in chest radiograph interpretation by developing new model- and data-centric methods. This research aims to improve uncertain label handling, overall classification performance, and the detection of multiple coexisting conditions.

The study employs convolutional neural networks (CNNs), DenseNet121, and Swin Transformers to investigate the effects of excluding versus relabeling uncertain labels. It also evaluates the efficacy of a custom pooling layer for classification performance and the effectiveness of the SMLE DenseNet model on co-existing conditions. Significant performance improvements are demonstrated when uncertain labels are appropriately handled using GMM. The custom pooling layer significantly enhances the classification performance of the model, and the SMLE DenseNet model outperforms the baseline DenseNet121 in detecting co-existing conditions. Finally, the study also examines radiologist's shift in perception towards the integration of AI in clinical practice through pre- and post-presentation questionnaires. The findings indicate a generally positive attitude shift towards AI. Despite concerns regarding the deskilling of new radiology professionals, radiologists recognize the potential of AI models to enhance the efficiency of radiology departments and patient care.

This research contributes insights into the practical implementation of AI in medical imaging. It shows that advanced techniques can significantly improve the diagnostic performance and efficiency of AI models. The findings also emphasize the importance of addressing radiologist's concerns to ensure the successful integration of AI into clinical practice.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to the Almighty Allah for His countless blessings and guidance throughout my academic journey. Without His grace and mercy, this achievement would not have been possible. After that, I am profoundly thankful to my supervisor, Professor Farshid Amirabdollahian, for his invaluable guidance, unwavering support, and insightful advice throughout the course of my PhD. His mentorship has been instrumental in shaping my research and academic career. I extend my sincere thanks to my second and third supervisors, Dr. Kheng Lee Koay and Dr. Yi Sun, for their significant contributions and support during my studies. Their expertise and feedback have greatly enriched my research work. I am also grateful to my colleagues in the Robotic Research Group, including Dr. Mohamad Reza Shahabian Alashti, Dr. Vignesh Velmurugan, Dr. Mohammad Hossein Bamorovat Abadi, and Dr. Shadiya Alingal Meethal. Our thought-provoking discussions on the latest advancements in the field have been incredibly inspiring, and their fellowship has provided a stimulating and supportive working environment in the lab. I would also like to acknowledge the support of Professor Martin Hardcastle for his prompt and efficient resolution of the issues I encountered while using the high-performance computing facilities. His assistance has been crucial in the successful completion of my research work.

My heartfelt appreciation goes to my beloved wife, Bushra Fatima, and my lovely daughter, Umaiza Fatima, for their unwavering support and patience throughout this demanding journey. Their love and encouragement have been my constant source of strength and motivation.

Lastly, I would like to thank everyone who has contributed to my academic and personal growth during this journey. Your support and encouragement have been invaluable, and I am truly grateful to each and every one of you.

Author's Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: Mubashir Ahmad **DATE:** 28th June, 2024

Contents

List of Tables

List of Figures

Abbreviations

ACA Automated Compliance Assessment. 22

AI Artificial Intelligence. 9, 10, 23

AL Active Learning. 22

AP Anteroposterior. 3

AUC Area Under the ROC Curve. 41

CNN Convolutional Neural Networks. 9–11, 13, 15, 20, 22, 55, 69

CSV Comma Separated Values. 25

CT Computed Tomography. 9

DL Deep Learning. 9–12, 15, 21, 22, 28, 99

FPR False Positive Rate. xii, 41, 43

GAN generative adversarial networks. 20, 100

GBM Gradient Boosting Machines. 21

GCN Graph Convolutional Networks. 21

GMM Gaussian Mixture Models. 6, 20, 27, 44, 45, 53, 98, 99, 101, 102

ILSVRC Large Scale Visual Recognition Challenge. 2, 11

LSTM Long Short-Term Memory. 22

ML Machine Learning. 9, 21

NLP Natural Language Processing. 25

PA Posteroanterior. 3

ReLU Rectified Linear Unit. 11

ROC Receiver Operating Characteristic. 41

ROI Region of Interest. 28

SGGCN ShuffleGhost Graph Convolutional Network. 22

SMLE-DenseNet Sequential Multi-Label Enrichment DenseNet. 5–8, 38, 78, 97

SQDIFF Sum of Squared Differences. 25, 28, 30

SUS system usability scale. 101

SVM Support Vector Machines. 21

TPR True Positive Rate. xii, 41, 43

ViT Vision Transformers. xi, 10, 13, 14

Glossary

Alveoli Tiny air sacs within the lungs.. 3, 17

Lung Parenchyma Section of the lungs involved in gas transfer.. 17

Chapter 1

Introduction

1.1 Motivation

The field of artificial intelligence (AI) has witnessed significant advancements in the recent years which impact various industries and domains, including healthcare. AI is a collection of diverse range of techniques which also includes machine learning and deep learning. These techniques enable computers to learn from and generalize, based on the data they receive. Addition of AI to the field of Information and Communication Technologies used in different sectors allows for benefiting from significant analytical capabilities and inference possibilities that are provided often in a short time span. For healthcare, AI provides a great potential to change the way of disease diagnosis, treatment, and overall patient care. For example, recent applications of AI in medicine includes areas such as drug discovery and clinical trial design [10, 11]. Similarly, the recent advancements in deep learning [12] which is a sub field of AI, inspired by the structure and functioning of the human brain, have laid the path for more sophisticated applications in medical image analysis.

Medical imaging has been crucial in diagnosing and monitoring various diseases. However, interpreting medical images, particularly for complex modalities like chest radiographs, also known as x-rays, can be challenging due to subtle variations in the anatomical structures and pathology. AI, specifically deep learning, offers a promising solution by automatically detecting important features in a radiograph and provides valuable insights to radiologists by using those features to classify the radiograph as normal or affected. Deep learning algorithms can be trained on large datasets of labeled chest radiograph images, typically annotated by experienced radiologists. These labels are derived through a process that often involves multiple rounds of validation to ensure accuracy. The reliability of these labels is further enhanced by consensus from multiple annotators. The reliable labels enable deep learning models to learn complex patterns and identify abnormalities with high confidence.These technological advancements have significant implications for various fields within radiology.

X-rays are one of the most commonly performed imaging tests, with millions conducted annually. For instance, in England alone, there were approximately 44.9 million diagnostic imaging tests reported in year 2019, with 23.2 million plain x-rays making up a significant portion of these tests [13]. In emergency medicine, x-rays are frequently used to quickly diagnose conditions such as fractures, infections, and lung conditions like pneumonia due to their accessibility and speed. Patients with chronic conditions, such as chronic obstructive pulmonary disease (COPD), heart disease, and osteoporosis, often undergo regular x-rays to monitor disease progression and treatment efficacy. Additionally, radiology plays a critical role in cancer detection and management, with mammography used for breast cancer screening and other forms of radiographic imaging helping to diagnose and assess various cancers. Orthopedics relies heavily on x-rays to diagnose and monitor bone and joint conditions, including fractures, dislocations, and degenerative diseases like arthritis. Dental health also benefits from x-rays, which are used to diagnose and monitor conditions of the teeth and jaw, such as cavities, impacted teeth, and bone loss. These

examples underscore the widespread application and importance of radiographic imaging in modern medicine.

Respiratory diseases, including COPD, influenza, and pneumonia are the third leading causes of death in the United Kingdom following circulatory diseases and cancer [14]. According to a survey by Conor Stewart, the mortality rate from respiratory diseases in the United Kingdom in 2019 was 148.52 per 100k male population and 107.85 per 100k female population [15]. These are pre-covid-19 stats. Chest radiographs are one of the most common diagnostic modality for lungs or chest related conditions. However, it takes much experience for a radiologist to accurately analyse the chest radiograph to detect many chest-related conditions, such as Edema, Cardiomegaly, Atelectasis, Lung Effusion, Consolidation and many more. Moreover, according to a report in 2020 by the Royal college of radiologists, hospitals do not have enough radiologists to keep their patients safe. Also, there will be a 44% shortfall in the UKs clinical radiology workforce by 2025. As a result, in the last two years, National Health Service (NHS) has spent 58% more on outsourcing radiology reports [16]. In addition to the scarcity of radiologists, there is a problem of diagnostic error in radiology reports. According to Michael et al. (2015), worldwide, annually, at least 40 million out of 1 billion radiology reports contain errors made by the radiologists [17].

In response to the growing demand and complexity of chest radiograph interpretation, researchers have started exploring the potential of using deep learning to address the increasing demand of radiologists and to reduce the number of misdiagnoses. The idea of chest radiograph interpretation using deep learning is not new. Much work has been done to detect illnesses from chest radiographs using deep learning techniques [18–21]. For example, a radiologist-level performance was demonstrated for pneumonia classification on the ChestX-ray14 dataset, which boosted further research in this area [22, 23]. Almost all medical imaging, including chest X-ray interpretation work, is based on convolutional neural networks (CNN) [24]. This incline towards CNN is because, it works very well on image datasets. All well-known deep architectures for images, such as AlexNet [4], VGGNet [25], ResNet [6] and DenseNet [1], are CNN based. All of them have outperformed the previous state-of-the-art in ImageNet Large Scale Visual Recognition Challenge (ILSVRC), a massive dataset of images belonging to 1000 classes [26]. Figure1.1 shows DenseNet architecture, which is a deep neural network with direct connections given by equation 1.1. Where L represents the number of layers in the dense block, and the term $\frac{L(L-1)}{2}$ calculates the total number of direct connections between layers. This design maximise information flow between layers and facilitate more efficient feature reuse and gradient propagation. As a result, DenseNet models often achieve higher performance with fewer parameters compared to traditional convolutional neural networks. DenseNet121 has been used as a base backbone architecture in this research.

$$\frac{L(L-1)}{2}(L = NumberOfLayers) \tag{1.1}$$

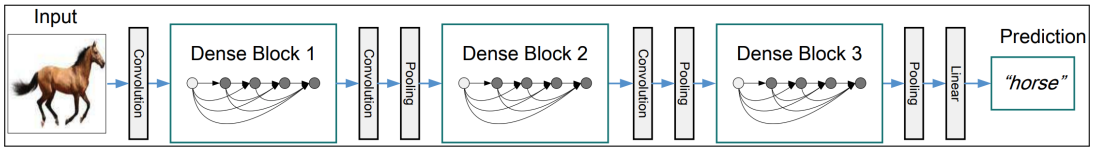


Figure 1.1: DenseNet architecture with three dense blocks, all layers in a dense block receive input from the previous layers. The image is taken from [1]

Despite a considerable progress in classifying chest radiographs and other medical images with deep learning algorithms, researchers are still actively exploring ways to further improve performance. While the primary goal is to accurately classify X-rays for a condition(s) or at least detect the likelihood for the presence of disease, several challenges remain. This research focuses on a number of them, which includes: the presence of a significant number of uncertain labels within chest radiograph datasets such as

CheXpert, the potential variation in the model performance after being trained on the specific radiograph projection (e.g., Posteroanterior (PA), Anteroposterior (AP), and lateral views), exploring different data preprocessing techniques to potentially improve model performance, tweaking model architectures and training procedures to achieve performance improvement on detecting multiple co-existing conditions in a single radiograph, and finally, understanding the perception of radiologists regarding the integration of deep learning models into their clinical workflow. Getting their insights and addressing their concerns are essential for the successful adoption of these technologies in real-world medical practice.

Open datasets and competitions play a vital role in supporting research and innovation in this line of inquiry. The intention was to participate in the CheXpert competition. Because of that, the CheXpert [9] dataset has been used throughout this research. There are 14 labels associated to each image, indicating the presence or absence of a different medical condition. These labels include (No Finding, Enlarged Cardiomediastinum, Cardiomegaly, Lung Opacity, Lung, Lesion, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture and Support Devices). A single radiograph can exhibit multiple coexisting conditions. For example, a radiograph image might simultaneously show signs of "Cardiomegaly" and "Pleural Effusion". Five conditions (Edema, Consolidation, Cardiomegaly, Atelectasis, and Pleural Effusion) were chosen for this research based on their clinical significance and because the CheXpert competition was evaluating these conditions for ranking [9]. Other labels, such as 'Support Devices', 'Enlarged Cardiomediastinum', 'AP/PA', and 'Lung Opacity' were used to pre-process the data for various models training.

1.2 Thoracic Diseases

Based on [2], the ABCDE systematic approach to chest X-ray interpretation involves examining specific parts of the radiograph associated with each alphabet. This method results in a thorough assessment, with radiologists concentrating on the following aspects. Airway, Bones, Cardiac, Diaphragm, and Everything else. Understanding these key features helps radiologists identify and diagnose the five conditions being focused on in this research. Figure 1.2 shows a chest radiograph labeled with key anatomical landmarks essential for a thorough diagnostic interpretation. The labels correspond to the ABCDE systematic approach. Airways, such as the trachea and carina; Bones, including the clavicle, spinous processes, scapula, and both anterior and posterior ribs; Cardiac structures, such as the aortic knob, pulmonary artery, cardiac silhouette, and hilum; Diaphragm, featuring the right and left hemidiaphragms; and Everything Else, including the cardiophrenic and costophrenic angles, and the lung fields with their fissures and pleura.

- A Airways: Trachea, right main bronchus, left main bronchus etc.
- B Bones: Ribs, sternum, Vertebral body etc.
- C Cardiac: Cardiac silhouette and mediastinum etc.
- D Diaphragm: Right hemidiaphragm, left hemidiaphragm, etc.
- E Everything Else: Lungs fissures, Pleura, Costophrenic angles, etc.

Based on this learning, following is a brief description of each of the five conditions present in CheXpert data, specifically how the condition appears on a radiograph.

1.2.1 Edema

Edema is a condition that affect the lungs where the Alveoli, the tiny air sacs responsible for gas exchange, become filled with excess fluid that has leaked from the blood vessels. This accumulation of fluid can occur in either one lung (unilateral) or both lungs (bilateral). When observed through a chest radiograph, edema manifests in different patterns and grades of opacity. These appearances include the characteristic

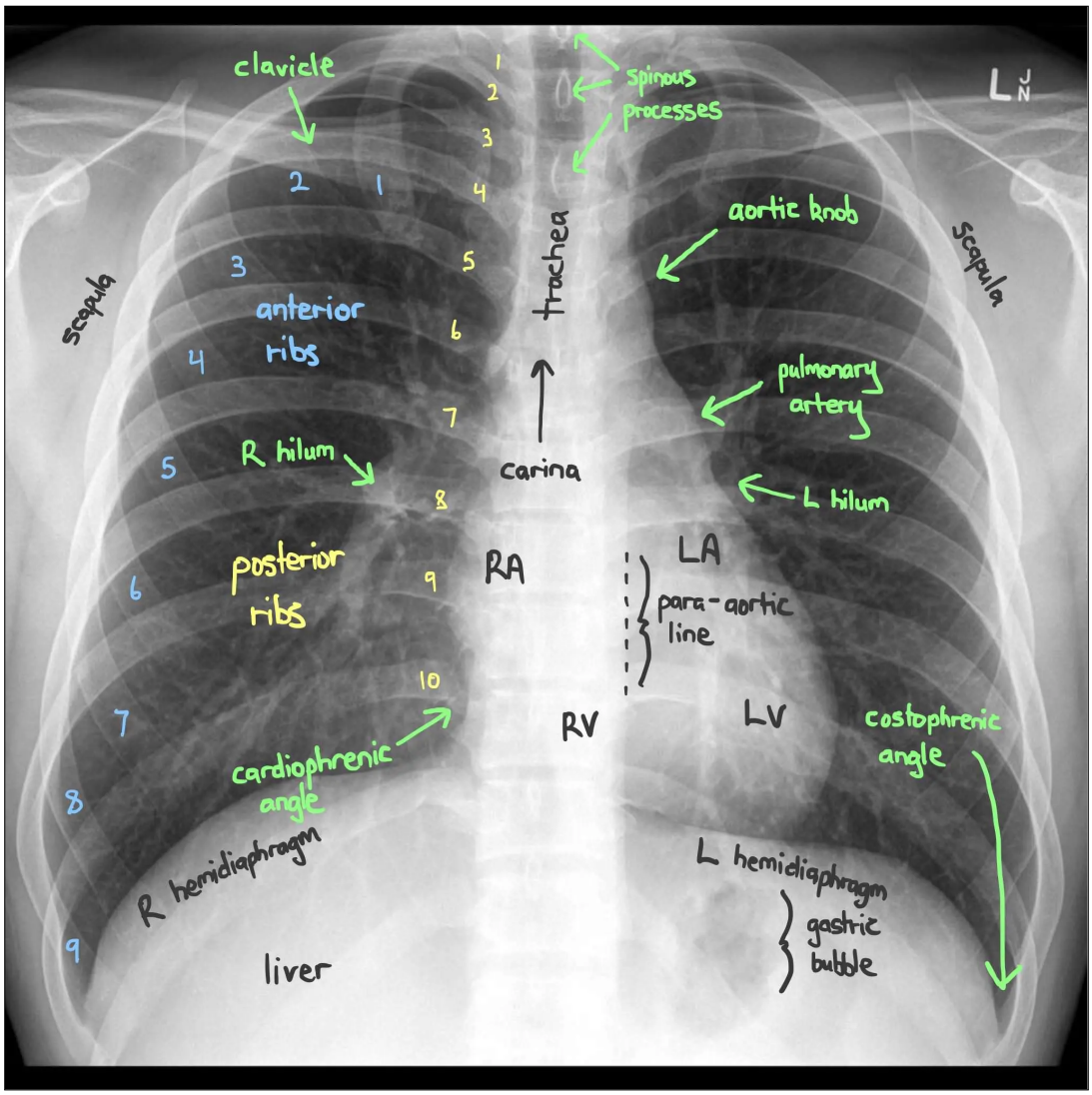


Figure 1.2: Approximation of the radiologist way of looking at the X-ray. Labelled are some of the major things to notice. This picture is taken from [2].

"bat wing" pattern, where the fluid distribution resembles the shape of bat wings, and "curly B lines", which are short and horizontal lines at the lung periphery which represents interstitial fluid. The presence of these patterns helps in diagnosing and assessing the severity of the edema.

1.2.2 Consolidation

The lung becomes consolidated when the small airways, normally filled with air, are occupied with dense material. It can be unilateral or bilateral. It is not always infectious. However, it could be due to a more severe problem such as pneumonia, pulmonary edema and cancer. Consolidation appears on a chest radiograph as dense white patches. The doctor can diagnose the cause depending on the distribution pattern of consolidation. Figure1.3(a) shows a radiograph with consolidation.

1.2.3 Atelectasis

Atelectasis is a lung condition happens due to the collapse of lung tissues which results in a significant reduction in lung volume.The primary reason is the bronchial obstruction. This condition manifests on chest X-rays in many ways, such as a shrunken lung lobe or the presence of plate-like or round opacities. These radiographic features indicate areas where the lung has collapsed which leads to reduced air space and impaired respiratory function. Atelectasis can affect both lungs simultaneously, which is referred as bilateral atelectasis. Figure1.3(b) provides a radiographic example of this condition.

1.2.4 Cardiomegaly

Cardiomegaly, referred to as heart enlargement, occurs when the heart becomes larger than its normal size. This condition can be identified through a chest radiograph, specifically by calculating the cardiothoracic ratio on a posteroanterior (PA) view. The cardiothoracic ratio is a measure used to determine the size of the heart in relation to the total thoracic width. If this ratio exceeds 0.5, it indicates the presence of cardiomegaly. This ratio is given by equation 1.2, that divides the width of the heart by the width of the chest.

$$CardiothroracicRatio = \frac{MaxhorizontalCardiacWidth}{MaxHorizontalThoracicWidth} \tag{1.2}$$

1.2.5 Pleural Effusion

There are two thin layers of tissue: one is around the lungs, and the other is on the inner side of the chest wall. The very little space between the two layers (pleural space), normally filled with some lubrication fluid (pleural fluid), accumulate an abnormally excessive amount of fluid in it, which causes the lung to shrink. It can be unilateral or bilateral. Pleural effusion can be detected on a chest radiograph by noticing the bluntness of costophrenic angles. Please see Figure1.3(c) for a better understanding where the chostophrenic angle is circled in red/green.

1.3 Research Goals

The primary goal of this research is to enhance the performance and reliability of AI-based interpretation of chest radiographs by proposing and utilizing new techniques. These include a GMM-based uncertain label handling and a multi-scale template matching approach as well as integration of a custom pooling layer and using SMLE-DenseNet approach for model training. While the focus of this research is on chest radiographs, the methodologies and techniques developed are designed with a level of generality that allows their application to other similar problems in computer vision and specially medical imaging.

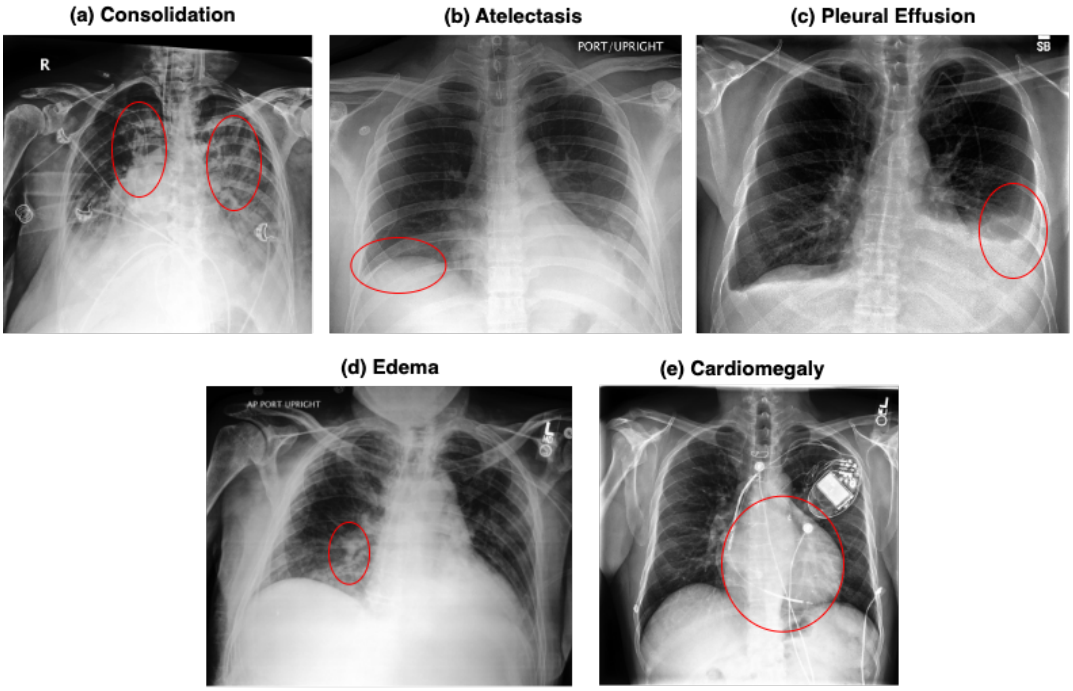


Figure 1.3: Images from CheXpert dataset. (a) Consolidated areas are encircled. (b) We can see right upper lobe atelectasis. (c) The right costophrenic angle is blunt. (d) Edema can be seen beside the heart. (e) The heart size is bigger than the 50% of the thoracic width.

For instance, the GMM-based uncertain label handling technique can be employed in any domain where the dataset contains a significant number of uncertain or noisy image labels. The multi-scale template matching approach and custom pooling layers are also applicable in scenarios that requires both, low and high intensity feature detection and preservation, which are common in various object detection and classification tasks. Moreover, the SMLE-DenseNet approach can be adapted to other complex image classification problems, such as identifying co-existing multiple objects (e.g., cat, car, tree) in a single image. This adaptability is due to the model's inherent design that efficient feature extraction, which are important for computer vision applications.

1.3.1 Hypothesis and Research Questions

The central hypothesis of this research is:

"The integration of advanced techniques significantly enhances the performance of deep learning models in chest radiograph classification and positively influences the perception and trust of radiology professionals."

The investigation of this hypothesis involves addressing the following four research questions.

RQ1: How does GMM based uncertain label handling influence model performance?
Handling uncertainty in medical imaging data remains a significant challenge. The CheXpert dataset, which is being used in this research, carries a vast number of uncertain labels. Several approaches have been attempted to handle this issue. For example, some studies have relabeled all the uncertain labels as either 1 or 0, or relabeled them using a CNN model that had been trained on the rest of the samples. Others have assigned these uncertain labels random values or discarded all samples with uncertain labels [9,27]. This work investigates the use of semi-supervised relabeling with Gaussian Mixture Models (GMMs). GMMs provide a probabilistic approach to cluster

samples and help in dealing with uncertainty by grouping uncertain samples into various clusters, each corresponding to a specific class.

RQ2: What is the impact of custom pooling layers and multi-scale template matching on chest radiograph classification performance?
Traditional pooling layers, such as max pooling, typically utilized in CNNs, tend to drop low value pixels to reduce the size of the feature maps [28], which could become crucial for correct diagnostic performance in the case of chest radiographs. For this, very few solutions have been proposed in the past, such as using adaptive average pooling to reduce the feature map to the desired shape [29], and Use of 3x3 max pooling to consider a wider pooling window for chest radiograph classification [30]. This work proposed a custom pooling layer, which amalgamate max pooling and min pooling to retain a wide range of features. Moreover, simple template matching has been used to preprocess the radiograph data [27]. This work introduced multi scale template matching technique to accurately extract the thoracic regions on radiographs. The integration of these two approaches is set to improve the performance of deep learning models in chest radiograph classification.

RQ3: How does the SMLE-DenseNet approach improves the detection of multiple co-existing conditions in radiographs?
The CheXpert dataset contains more than two hundred thousand radiographs, with over 30% of them containing multiple coexisting conditions. Several studies have trained various models on the CheXpert dataset and reported model performance either as a whole or by individual condition separately [31–33]. However, no particular study has focused on investigating how these trained models perform in detecting multiple coexisting conditions. This research addresses this gap by evaluating model performance specifically on samples with coexisting conditions and proposing a new training approach to enhance the detection of multiple coexisting conditions.

RQ4: How does the presentation of deep learning capabilities in chest radiograph interpretation influence the perception and trust of radiology professionals?
Understanding the perception of radiology professionals towards AI is very important, as they are the primary users of these technologies in clinical settings. Many studies have assessed the attitudes and knowledge of radiologists regarding AI integration into their workflows [34–50]. However, none of these studies have employed a pre-and post-presentation questionnaire to evaluate shift in perception after engaging with and reviewing the performance capabilities of AI models on chest radiograph classification. This work presents this unique approach aimed to provide radiology professionals the opportunity to see the capabilities of AI models and then measure the shift in perception using post presentation questionnaire.

1.4 Key Contributions

Research highlighted in this thesis comprises of a series of sequential approaches related to data preparation, and model development, with the primary goal of highlighting their influence on the performance of the classification models. The following aspects can be highlighted as main contributions of this work:

- To improve the label quality and model robustness, a semi-supervised relabeling technique is developed using Gaussian Mixture Models (GMMs) to probabilistically handle uncertain labels in the CheXpert dataset .

- To enhance the performance of deep learning models, a custom pooling layer is proposed that combine max pooling and min pooling for better preservation of high and low-intensity features in chest radiographs.

- To improve the data quality, an improvised template matching (multi-scale) technique is applied to eliminate irrelevant areas in CheXpert radiographs, which improve the models performance by focusing on relevant thoracic regions.
- Assessed model performance specifically on samples with multiple coexisting conditions, a previously underexplored area, and proposed a new training approach (SMLE-DenseNet) to enhance the detection of multiple coexisting conditions.
- Conducted a unique study involving pre-and post-presentation questionnaires to measure the shift in perception of radiology professionals after engaging with and evaluating the performance capabilities of AI models, which provides insights into the acceptance and trust of AI amoung radiology professionals.

1.5 Organisation of the Thesis

Following the **Introduction (Chapter 1)**, the **Background and Literature Review (Chapter 2)** delves into existing studies related to artificial intelligence, deep learning, computer vision, medical imaging and specifically AI for chest radiograph interpretation. It ends with a review on studies related to the perception of radiology professionals towards AI. This Chapter highlights the gaps in current research and sets the stage for the contributions of this thesis.

The **Methodologies (Chapter 3)** details the data sourcing, including the CheXpert dataset, and explains the techniques used for uncertain label handling, custom pooling layers, and multi-scale template matching along with other model-centric and data-centric methodologies. This Chapter also describes the experimental setup and the evaluation metrics employed to assess model performance. Subsequently, the **Uncertainty in Labels and Radiograph Projections (Chapter 4)** explores an approach to handle uncertain labels and introduces a semi-supervised relabeling technique using Gaussian Mixture Models (GMMs) which addresses the **RQ1**. The **Efficacy of Custom Pooling Layer (Chapter 5)** presents the superior performance of custom pooling layer along with the effectiveness of the multi-scale template matching. This addresses **RQ2**. The **Sequential Multi Label Enrichment (Chapter 6)** evaluates model performance on samples with multiple coexisting conditions and proposes a new training approach to improve detection, addresses **RQ3**.

The **Radiologist's Perception of AI in Chest Radiology (Chapter 7)** presented a study using a unique pre and post-presentation questionnaires to gauge the shift in perception among radiology professionals after engaging with AI models. This addresses **RQ4**. The **Discussion and Conclusion (Chapter 8)** presents the conclusion and analyze the findings, also summarizes the key contributions of the research and suggests directions for future studies. The thesis ends with a comprehensive list of references followed by the appendices.

Chapter 2

Background and Literature Review

This Chapter provides a comprehensive background and literature review which is structured to address the key research questions identified in Chapter 1. It begins with an overview of Artificial Intelligence (AI) and focus on deep learning and its applications across various sectors including healthcare. The Chapter then delves into computer vision and explains the foundational backbone architectures that underpin modern computer vision techniques. Following this, it explores the advancements in medical imaging, and emphasize on the significant role of deep learning in diagnosing and treating diseases. The Chapter also discuss five specific conditions relevant to chest radiographs and highlight both data-centric and model-centric methods to enhance classification performance. Finally, it covers the literature capturing the perception of radiology professionals towards AI integration, to understand the practical implications and acceptance of these technologies in clinical practice. This structured approach to study background work aligned with the research questions, provides a clear and focused review of existing literature and methodologies.

2.1 Artificial Intelligence

Deep learning (DL), a subfield of machine learning (ML), is a branch of artificial intelligence AI. The simulation of human intelligence in machines which enables them to think and act like humans, is referred to as Artificial Intelligence (AI). AI encompasses various techniques, some of which involve learning from data provided, such as machine learning and deep learning, while others may rely on rule-based systems, heuristics, or predefined algorithms that do not necessarily learn from data [51].The learning of an AI system is very much inspired by a human brain. By continuously experiencing similar situations the AI systems learns the pattern and adapt to the data [52–54]. In the past decade DL algorithms have shown great contributions across various sectors. This include retail and E-commerce, deep neural networks have enhanced the shopping experience by providing personalised product recommendations to the shopper, helped in understanding the customers sentiments to improve the product offerings and enhanced the ability to maintain the optimal inventory levels by accurately forecasting the demand [55–57].

Furthermore, DL has been used extensively in the field of heath care. Such as early disease diagnosis by detecting and grading diabetic retinopathy from retinal images and analyzing genomic data to predict interactions between drugs and their biological targets. The DL based models benefit from vast amounts of biological data to identify potential new drugs and understand genetic factors which influence disease and accelerate the drug discovery process for personalized medicine [58, 59]. In addition to that, DL has been extremely helpful for medical imaging application such as the diagnosis of Alzheimer and Parkinson disease with neuroimaging using convolutional neural networks (CNNs), identifying spinal deformities and measure parameters related to sagittal alignment which aids the diagnosis and monitoring of spinal disorders and diagnosis of various diseases from the thoracic region by analyzing the chest radiographs and computed tomography CT scans.

2.2 Computer Vision

Computer vision comes under the broader field of AI, focuses on enabling machines to interpret and understand visual information from the world, similar to human vision. By utilizing algorithms and models, computer vision allows systems to analyze and process images and videos to recognize patterns, objects, and even behaviors. DL has significantly advanced the capabilities of computer vision through the development of complex neural networks that can automatically learn and extract features from raw visual data. One of the most influential advancements in this domain is the CNN, which mimics the human visual processing system by using multiple layers to progressively extract higher-level features from input images. Architectures such as LeNet, AlexNet, VGG, ResNet, DenseNet and more recently, EfficientNet and Vision Transformers (ViTs), have revolutionized image classification, object detection, and segmentation tasks. These state of the art architectures are pushing the boundaries of what computer vision applications can achieve in various fields, from healthcare to autonomous driving cars.

2.2.1 Backbone Architectures

A backbone architecture refers to the foundational part of a model which is responsible for extracting features from an image. It acts like the workhorse to analyze the input image and identify patterns, shapes, and other visual components. Most of the famous backbone architectures are CNN based. The earlier layers of the architectures extract the basic features such the vertical and horizontal edges while the later part identify the more complex features such as the face of a dog in case of cat and dog classification [3]. Following are details of the popular backbone architectures.

2.2.1.1 LeNet

This is one of the earliest CNN based architecture. It has played a pivotal role in the evolution of DL. Initially designed for handwritten digit recognition. The architecture consists a series of layers that include convolutional layers, pooling layers, and fully connected layers. The typical LeNet-5 architecture includes two convolutional layers (C1 and C3), two subsampling layers (S2 and S4), and three fully connected layers (C5, F6, and the output layer) [3]. Before the output layer, it employs "Gaussian connections", particularly in the way convolutional layers and fully connected layers are structured and initialized. The term "Gaussian connections" can be particularly attributed to the weight initialization strategy where weights are drawn from a Gaussian (normal) distribution. This initialization is important for breaking symmetry and enabling efficient learning during the training process. Figure 2.1 shows the details of this architecture.

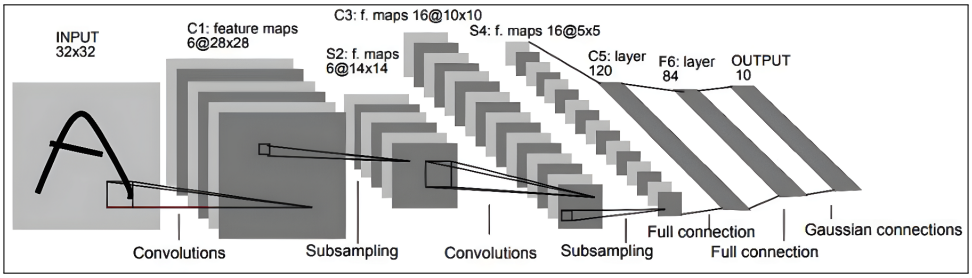


Figure 2.1: Diagram of the LeNet architecture. This diagram is taken from [3].

The significance of LeNet lies in its pioneering use of convolutional layers, which are now a fundamental component of modern DL models. These layers are designed to automatically and adaptively learn spatial hierarchies of features from input images and reduces the need for manual feature extraction. By combining convolutional and subsampling layers, LeNet captures spatial relationships and patterns in image data. This design not only enhanced the accuracy of image classification tasks but also influenced the architecture of subsequent models like AlexNet, VGG, and ResNet. The input image, say 32x32

pixels, first processed through the C1 layer, which applies six convolutional filters, each produces a 28x28 feature map. The S2 layer then performs average pooling to reduce the feature maps to 14x14. This process gets repeated in the C3 and S4 layers which further extracts and condense features. In the final stages, the C5 layer applies 120 convolutional filters followed by the F6 fully connected layer with 84 neurons. This brings to the output layer which classifies the input image into one of the ten digits (0-9). This flow of data through the network enables LeNet to learn and recognize complex patterns in images and makes it a cornerstone in the field of computer vision.

2.2.1.2 AlexNet

AlexNet was designed to tackle the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), and its success popularized the use of deep CNNs for image classification tasks. It was build upon LeNet's foundation and pushes the boundaries of depth, filter size strategies, and training techniques. The architecture consists of eight layers, which includes five convolutional layers followed by three fully connected layers. The network also introduced the use of Rectified Linear Unit (ReLU) activations, the dropout layer for regularization, and the data augmentation techniques to improve the models performance and generalization ability [4]. It achieved state of the art performance in ILSVRC competition and encouraged further research and development in DL.

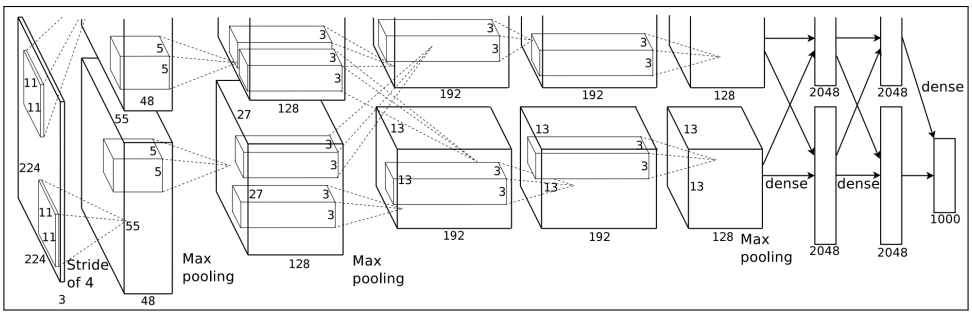


Figure 2.2: Diagram of the AlexNet architecture. This diagram is taken from [4].

The working mechanism of AlexNet can be understood through its architecture diagram shown in Figure 2.2. The input to the network is a 224x224 RGB image, which first goes through the convolutional layer with 96 filters of size 11x11 with a stride of 4. This produce feature maps of reduced spatial dimensions. This is followed by a max-pooling layer that further reduces the spatial size. This process is repeated with various filter sizes and depths in the subsequent convolutional layers to capture complex features at a later stage of the model. After the convolutional layers, the output is flattened and passed through three fully connected layers. The first two fully connected layers contain 4096 neurons each and use dropout for regularization. Finally, using the softmax activation in the last layer output probabilities for the 1000 classes of the ImageNet dataset. This approach allows AlexNet to learn hierarchical representations of the input data which makes it highly effective for image classification tasks.

2.2.1.3 VGG16

Unlike its predecessors, VGG focused on the impact of depth on a network's performance. It consists of 16 convolutional layers which is significantly deeper as compared to LeNet and AlexNet. This increased depth allowed VGG16 to learn more complex features from images. It stacked many 3x3 convolutional layers with small filter sizes one after another to achieve a similar effect compared larger filters and maintained the computational efficiency. It also employed ReLU activation functions after each convolutional layer to introduce non-linearity which improves the network's ability to learn complex patterns [25]. It achieved state-of-the-art performance on image classification tasks. However, the depth

also came with drawbacks. The large number of parameters made it computationally expensive to train and use. Figure 2.3 shows the architectural diagram of VGG16 network.

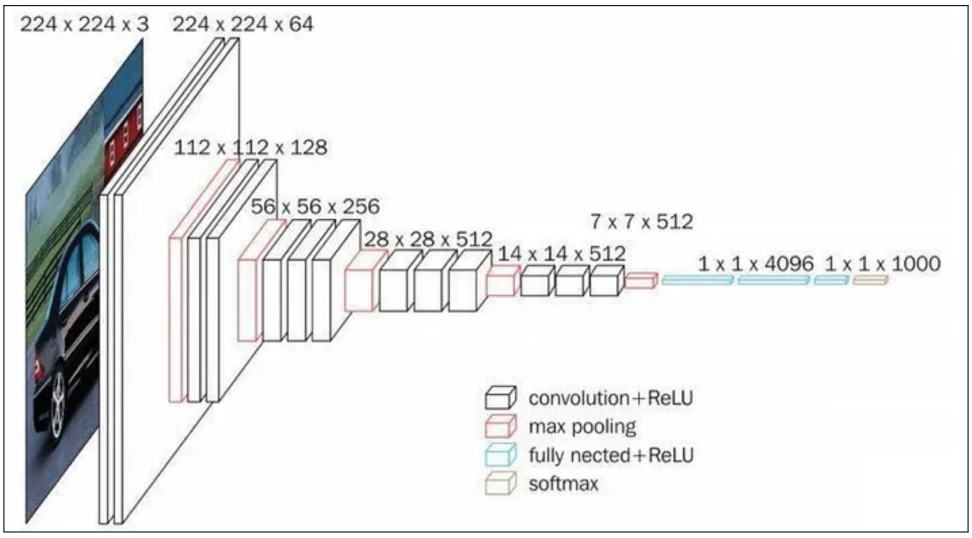


Figure 2.3: Diagram of the VGG16 architecture. This diagram is taken from [5].

2.2.1.4 ResNet

Residual Networks, or ResNets, introduced a new concept in DL for computer vision tasks. Unlike previous architectures like VGGNet, AlexNet, and LeNet, which stacked convolutional layers directly, ResNets used residual blocks. These blocks contain skip connections that add the input of the block directly to its output and then pass it through activation functions. This very simple addition addresses vanishing gradient problem in training very deep neural networks. In deeper networks, the signal from earlier layers can become too weak to have an impact on the training of later ones. Residual blocks solves this by allowing the gradient to flow directly through the identity connection which makes the earlier layers relevant throughout the training process [6]. This gives ResNets the ability to achieve much greater depths as compared to previous architectures. Figure 2.4 shows a 34 layer ResNet with skip connections.

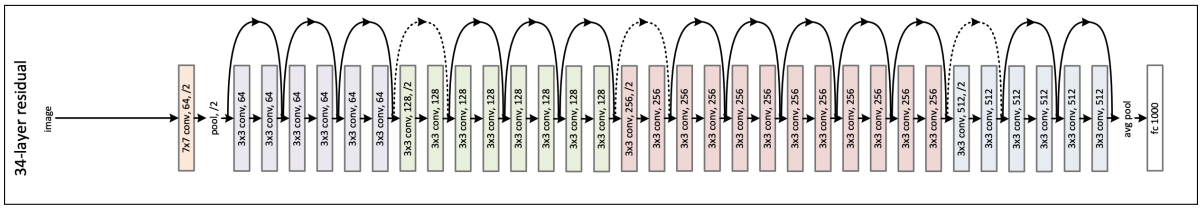


Figure 2.4: Diagram of the ResNet architecture. This diagram is taken from [6].

2.2.1.5 DenseNet

DenseNets takes a different approach to DL architecture compared to VGGNet, AlexNet, LeNet, and even ResNets. Unlike these architectures that rely on sequential layers where each layer receives input only from the one before it, DenseNets have dense connections. This dense connectivity is achieved through a concept called dense blocks. Each layer receives the outputs of all preceding layers within that dense block as additional input, along with its normal input from the previous layer. Figure 2.5 shows a 5 layer dense block where input of each layer is connected to the output of all previous layers. This enables each layer to learn from the combined knowledge of previous layers. Additionally, DenseNets do feature

concatenation to combine the outputs from each layer instead of passing them through activation functions one after another [1]. This allows the network to maintain information from all layers throughout the training process. The reusing of features by DenseNet makes it well suited for task with limited training data.

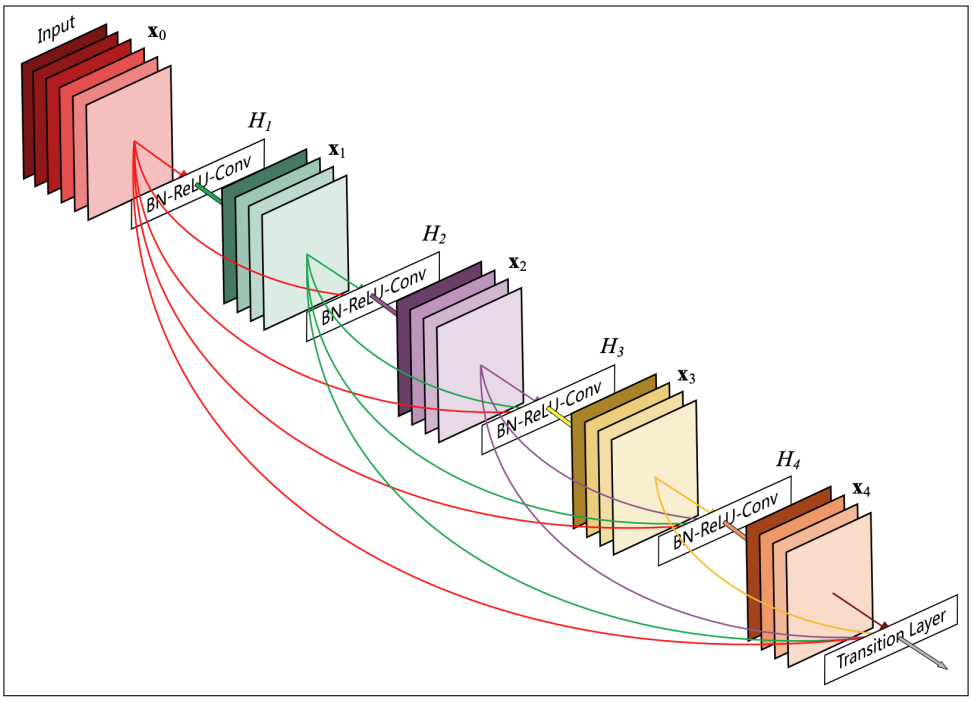


Figure 2.5: Diagram of a five layer dense block. This diagram is taken from [1].

2.2.1.6 Vision Transformer

The Vision Transformer (ViT) represents a huge shift from the traditional CNN based architectures of computer vision. Unlike CNNs, that rely on convolutional filters to extract features from images, ViTs leverage the power of transformer architectures, which were originally designed for natural language processing. At its core, a ViT divides an image into a grid of smaller patches. These patches are then flattened and converted into vectors very much similar to how words in a sentence are represented. It then use positional encodings to track the spatial relationships between these patches. This kind of relationship tracking is missing in CNN based models, while it gives ViT the ability to understand how different parts of the image relate to each other [7]. The encoders consist of multiple stacked layers with multi-head self-attention mechanism. This mechanism allows the ViT to attend to specific parts of the processed image data (patches) and learn how relevant they are to each other. Through this process, the ViT builds a global understanding of the image by giving attention to relationships between different patches. This is something not directly achieved by the local filtering operations in CNNs. Figure 2.6 shows the working of a ViT. However, ViTs also come with drawbacks like higher computational costs compared to CNNs.

2.2.1.7 Swin Transformer

The Swin Transformer is an improvised version of (ViT)s explained in the section 2.2.1.6. Unlike (ViTs), Swin Transformer employs a hierarchical structure that gives the ability to efficiently handle high-resolution images. Similar to ViT, the architecture starts by splitting the input image into non-overlapping patches. However, it introduces a novel technique known as shifted windows, which allows self-attention within local windows while also enable cross-window connections. This approach helps

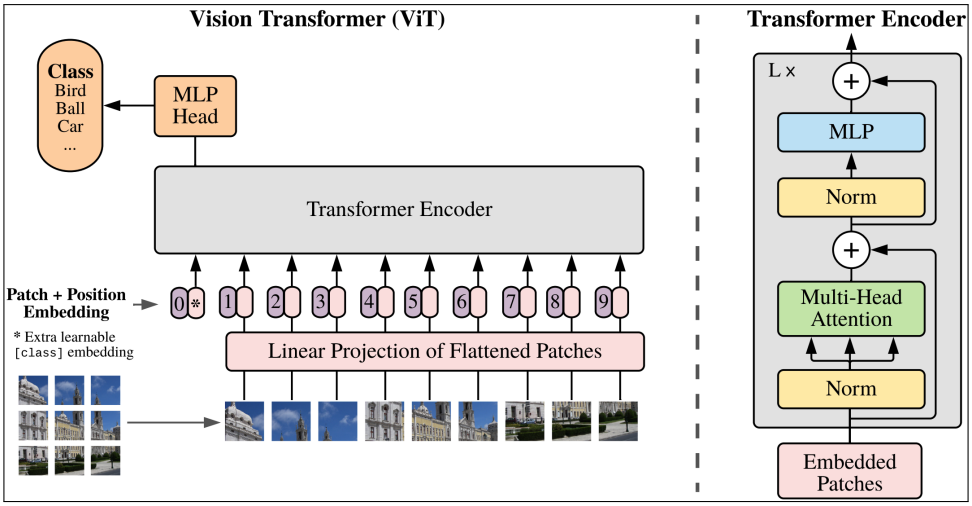


Figure 2.6: Diagram of a vision transformer with image patches and their positional embedding. The individual components of a transformer encoder is also given. This diagram is taken from [7].

to maintain the computational complexity at a manageable level and make it linearly scalable with image size [8]. There are various benefits of using the Swin Transformer. For a complete understanding of the image, the swin architecture captures both local and global context by integrating local self attention with the shifted windows mechanism. This method preserves the spatial hierarchies and fine-grained pixel relationships in visual data which allows it for precise modeling of complex scenes. Additionally, the hierarchical design of Swin Transformer generates multi-scale feature maps, which are beneficial for dense prediction tasks like object detection and segmentation. The efficiency and scalability of Swin Transformer, combined with its ability to maintain pixel-to-pixel relationships through local self-attention, make it a powerful tool for a wide range of computer vision applications. Figure 2.7 shows a comparison of swin transformer and the ViT.

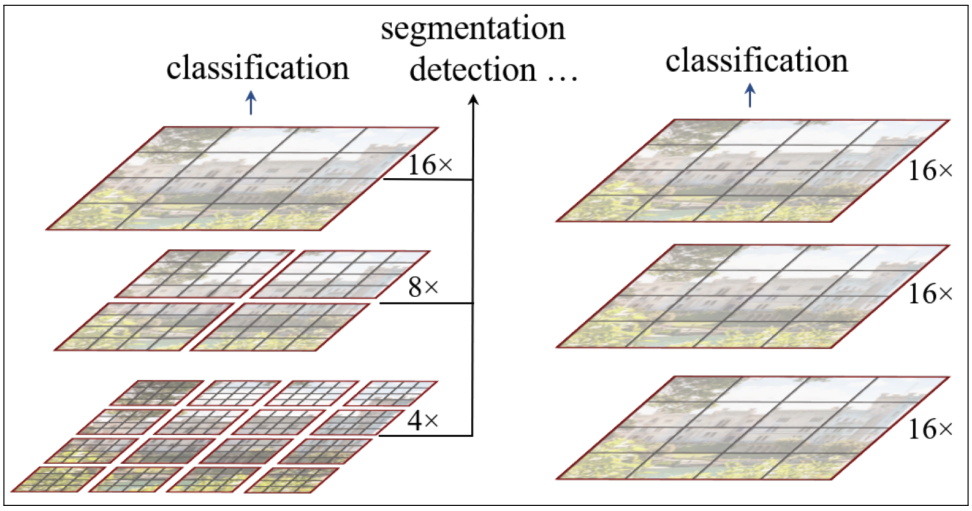


Figure 2.7: Diagram of the hierarchical feature maps of swin transformer on the left. The gray boxes are the image patches while the red boxes are the windows, the windows gets bigger by merging the patches as the model go deeper. On the right, the single feature map of the ViT is visible. This diagram is taken from [8].

2.3 Medical Imaging

In the past decade, DL for computer vision has emerged as a revolution in the field of medical imaging. It has shown a great efficiency and accuracy in detection, diagnosis, and treatment of various diseases. The mind blowing ability of computer vision algorithms to automatically identify the relevant features in the images give researchers the confidence to apply them on extremely specialised medical image interpretation tasks. The application of DL in medical imaging cover several areas, which includes diabetic retinopathy, cancer, brain disorders, and chest radiology. Previous work have shown that each of them benefits from the ability of these models to identify subtle patterns that may be invisible to the human eye. This section delves into the advancements and previous work on the above four critical applications of DL in medical imaging.

Diabetic retinopathy is a diabetes complication that affects the eyes, specifically the retina, which is the light-sensitive tissue at the back of the eye. The early detection and management of diabetic retinopathy are extremely important because it can cause permanent blindness [60]. The Detection of diabetic retinopathy involves regular eye exams and blood sugar control. To treat this condition, laser therapy, injections, or surgery are used to prevent further vision impairment [61]. Various studies have shown the detection of diabetic retinopathy by training CNN-based models using retinal fundus images [62] [63]. Others have developed models with multi-self-attention mechanisms to improve the relevant feature extraction [63], or used Ultra-Wide Field Scanning Laser Ophthalmoscope Images and applied transfer learning techniques to improve performance [64, 65]. Moreover, DL models have been developed to diagnose various types of cancers. Studies have demonstrated that computer vision has achieved a good level of accuracy in the early detection of skin and breast cancer [66, 67]. Furthermore, DL models are equally capable of detecting brain disorders, such as the detection of Parkinson's disease by using spirals and meanders filled out in forms, then using computer vision to match the spirals and meanders with the template to detect the presence of the disease [68]. Other brain disorders can also be detected by analyzing MRI scans with computer vision models [69].

2.4 Chest Radiology

DL and computer vision are revolutionizing chest radiology. CNN have demonstrated exceptional capabilities in identifying and diagnosing a variety of thoracic diseases from chest radiographs. By training on vast datasets of labeled radiographs, these models learn to recognize patterns and anomalies that may indicate conditions such as pneumonia, lung cancer, tuberculosis, and COVID-19. This advancement promises to augment radiologist's diagnostic capabilities and provide quicker and often more accurate interpretations of chest radiographs. The use of DL in chest radiology is driven by its ability to learn from large volumes of data. Traditional methods require substantial time and expertise, whereas DL models can process hundreds of radiographs in a fraction of the time [70]. For instance, the CheXNet algorithm, a CNN-based model, has shown capability to detect pneumonia from chest X-rays with a level of accuracy comparable to that of expert radiologists [22]. This level of performance aids not only in diagnosing diseases but also in prioritizing cases that need urgent attention, thereby improves patient outcomes.

Deep Learning (DL) models require vast amounts of high-quality labeled data to perform effectively. To meet this demand, various large datasets have been released, with NIH ChestX-ray14 [23], CheXpert [9], and MIMIC-CXR [71] being among the most significant. The CheXpert dataset, in particular, is extensively used in the medical imaging research community to train and validate models due to its highly reliable labels. This dataset includes labels not only for disease conditions but also additional labels such as support devices (medical apparatus attached to the patient), enlarged cardiomediastinum, and Lung Opacity, which are important for defining hierarchies among the conditions, and effective data preprocessing.

The CheXpert dataset, developed from a large collection of chest radiographs, includes over 224,000 images from more than 65,000 patients. Radiographs are labeled using advanced natural language processing (NLP) techniques on the associated radiology reports which provide labels for 14 different observations including various chest conditions and other clinically relevant findings. In this thesis, CheXpert [9] served as the base work. The five conditions (atelectasis, cardiomegaly, consolidation, edema, and pleural effusion) were chosen due to their relevance to the CheXpert competition and their significant clinical importance. These conditions are not only prevalent in clinical settings but also critical for accurate diagnosis and treatment, which makes them ideal focal points for this research. Focusing on these conditions defines a targeted approach in improving the diagnostic capabilities of DL models. Figure 2.8 provides a screenshot of the data labels as represented in the CheXpert dataset. It shows the five conditions annotated in the dataset and provides a visual representation of the labeling structure. Each row in the table represents an image file path along with binary labels for five conditions. For example, the labels indicate the presence (1), absence (0) or uncertain (-1) of conditions such as cardiomegaly, edema, consolidation, atelectasis, and pleural effusion. The subsequent sections will delve into the explanation of the five clinically significant conditions and provide a brief description of each condition and the signs to spot it on a chest radiograph.

Path	No Finding	Cardiomegaly	Edema	Consolidation	Atelectasis	Pleural Effusion
CheXpert-v1.0-small/train/patient00055/study1/view1_frontal.jpg	1	0	0	0	0	0
CheXpert-v1.0-small/train/patient00055/study3/view1_frontal.jpg	0	0	0	0	1	0
CheXpert-v1.0-small/train/patient00056/study1/view1_frontal.jpg	0	0	-1	0	1	0
CheXpert-v1.0-small/train/patient00056/study4/view1_frontal.jpg	0	1	1	0	0	1
CheXpert-v1.0-small/train/patient00056/study3/view1_frontal.jpg	0	0	0	-1	0	1
CheXpert-v1.0-small/train/patient00057/study1/view1_frontal.jpg	1	0	0	0	0	0

Figure 2.8: Screenshot of the CheXpert dataset labels showing paths to image files and their associated labels (positive (1), Negative (0), Uncertain (-1)) for various conditions, such as No Finding, Cardiomegaly, Edema, Consolidation, Atelectasis, and Pleural Effusion.

Moreover, the medical imaging research community has proposed various model-centric as well as data-centric methods to improve the classification performance of DL models in detecting various diseases on chest radiographs. These studies cover a range of techniques, which are also discussed in this Chapter. By leveraging detailed, reliable data from the CheXpert dataset and focusing on clinically significant conditions, this work contributes to the broader efforts to enhance the accuracy and efficiency of medical diagnostics through advanced methods.

2.4.1 Clinically Significant Conditions

2.4.1.1 Edema

Edema refers to the abnormal accumulation of fluid in tissues, it occurs in the lungs due to heart failure or other conditions. On a chest radiograph, edema can be identified by the appearance of hazy areas and fluid lines, which are the representation of fluid accumulation in the alveolar and interstitial spaces. This condition can lead to significant breathing difficulties and requires medical attention. Edema is a common finding in patients with congestive heart failure and can also occur due to acute respiratory distress syndrome, renal failure, or fluid overload. The radiographic signs of pulmonary edema include Kerley B lines, peribronchial cuffing, and a bat-wing pattern of alveolar edema [72, 73]. Figure 2.9 illustrates an example of Edema indicated by the hazy opacities in the lower lung fields and increased density around the central lung region near the heart.

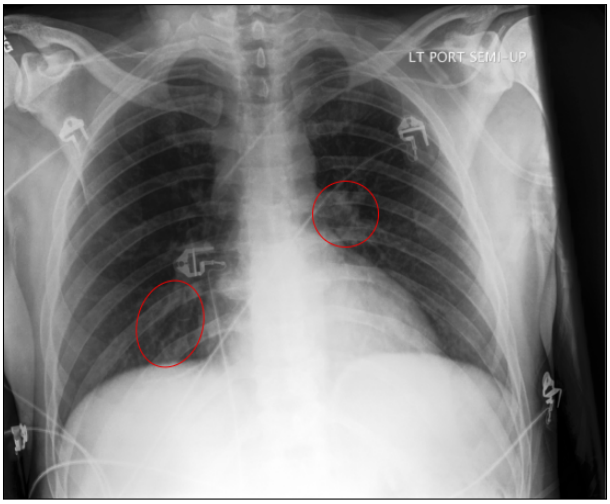


Figure 2.9: Example of pulmonary edema in a chest X-ray (encircled in red). This image is taken from CheXpert [9]

2.4.1.2 Consolidation

Consolidation occurs when the Alveoli in the lungs fill with fluid, pus, blood, cells, or other substances which leads to a solidification that can be detected on a chest radiograph as white patches. This is often caused by pneumonia, where infectious agents attack the Lung Parenchyma, which leads to an inflammatory response and fluid accumulation. Consolidation can also result from non-infectious causes such as pulmonary hemorrhage, lung cancer, or aspiration of gastric contents. Clinically, patients with consolidation may present with symptoms such as cough, fever, shortness of breath, and chest pain. The detection of consolidation on a chest radiograph is important for further management and treatment. Figure 2.10 shows a typical chest radiograph with consolidation. The mid to lower zone of the right lung shows increased opacity with darker linear structures.

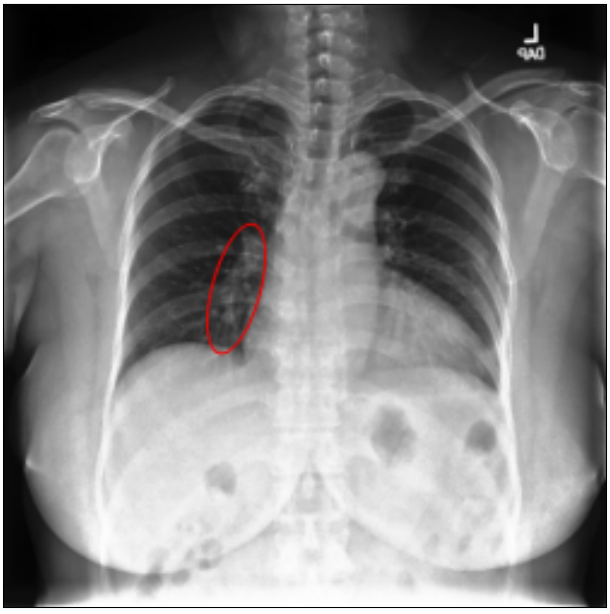


Figure 2.10: Example of lung consolidation in a chest X-ray (encircled in red). This image is taken from CheXpert [9].

2.4.1.3 Cardiomegaly

Cardiomegaly is the enlargement of the heart, which can be easily identified on a chest radiograph by an increased cardiothoracic ratio. This condition can result from various cardiovascular diseases, including hypertension, cardiomyopathy, and valvular heart disease. The enlargement of the heart can lead to impaired cardiac function and symptoms such as shortness of breath, fatigue, and swelling in the legs and abdomen. On a radiograph, cardiomegaly is diagnosed when the transverse diameter of the heart exceeds 50% of the thoracic diameter. This finding is used to further investigate and determine the underlying cause to suggest appropriate treatment. Figure 2.11 illustrates a chest radiograph containing cardiomegaly. The left heart border extends further than usual, indicates the presence of cardiomegaly.

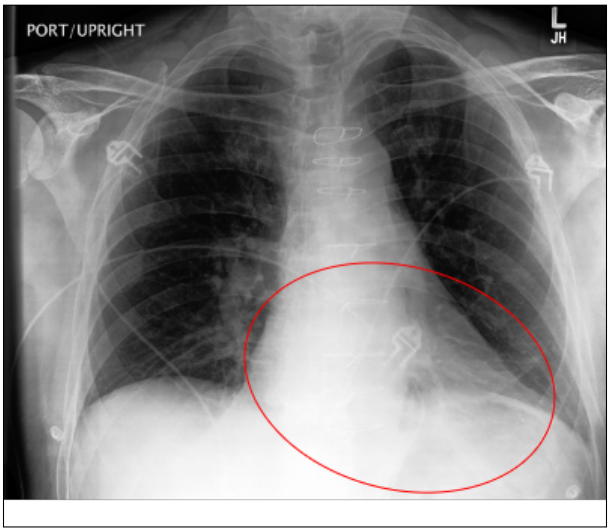


Figure 2.11: Example of cardiomegaly in a chest X-ray (encircled in red). This image is taken from CheXpert [9]

2.4.1.4 Atelectasis

Atelectasis is the collapse of a lung which result in reduced gas exchange. It appears as a region of increased density on a chest radiograph. Atelectasis can occur due to obstruction of the airways or compression of the lung by external structures including other factors. Clinically, patients with atelectasis may present with symptoms such as difficulty breathing, cough, and chest pain. The condition is commonly seen in postoperative patients, individuals with chronic lung diseases, and those with tumors obstructing the airways. The recognition of atelectasis on chest radiographs is essential for quick intervention to re-expand the affected lung tissue. Figure 2.12 provides an example of a radiograph with atelectasis. The increased opacity in the right lower lung field suggests the presence of atelectasis.

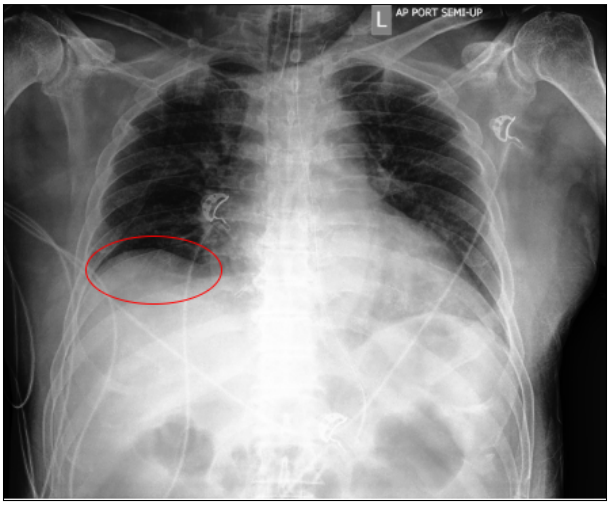


Figure 2.12: Example of atelectasis in a chest X-ray (encircled in red). This image is taken from CheXpert [9]

2.4.1.5 Pleural Effusion

Pleural effusion is the accumulation of excess fluid between the layers of the pleura outside the lungs. It can be seen on a chest radiograph as an area of opacity at the lung base. Pleural effusion can result from various conditions, which includes congestive heart failure, infections (such as tuberculosis or pneumonia etc). Patients with pleural effusion may present with symptoms such as chest pain, cough, and shortness of breath. The diagnosis and management of pleural effusion often involve imaging studies, thoracentesis (to remove and analyze the accumulated access fluid), and addressing the underlying cause. Figure 2.13 shows an example of pleural effusion. Here, the increased opacity at the base of the left lung and bluntness of the costophrenic angle indicates the presence of pleural effusion.

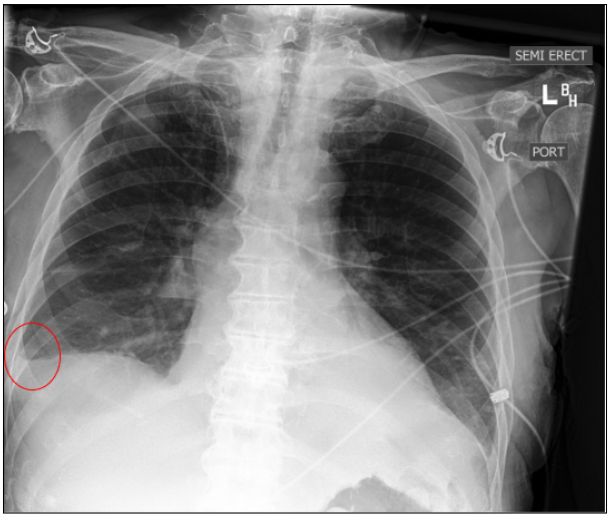


Figure 2.13: Example of pleural effusion in a chest X-ray (encircled in red). This image is taken from CheXpert [9]

2.5 Data-Centric Methods in Chest Radiology

The data-centric term refers to the data preprocessing techniques to enhance the quality and utility of the dataset itself to improve model performance. One common technique is image normalization, which adjusts the pixel intensity values of radiographs to a standard scale. This helps in reducing variability

due to differences in exposure and contrast and leads to a more consistent and standardised inputs for the model [74]. By standardizing the image data, normalization facilitates better convergence during training and improves the overall accuracy of the model in classifying chest radiographs [75, 76]. However, a potential limitation of this technique is that, improper normalization can sometimes lead to the loss of important image features, which might adversely affect the models performance.

Moreover, image cropping and resizing are also widely used preprocessing techniques in chest radiograph classification. Cropping involves removing unnecessary parts of an image to focus on the region of interest (usually the lung area), while resizing adjusts the dimensions of the image to a fixed size which is suitable for the model input. These techniques help in reducing computational complexity and ensures that the model processes only the relevant parts of the image [77, 78]. The main benefit is that it eliminates irrelevant information and maintains the focus on important areas. Doing the image cropping by hand can be labour intensive. To do this automatically, some studies have employed template matching technique to crop the best matching patch of the image [79]. An improper cropping or use of wrong template might exclude essential diagnostic features or distort the image, which could lead to reduced model performance. This work employed a multi scale template matching technique to remove the non thoracic areas in the CheXpert radiographs.

In addition to that, another very popular data preprocessing technique called data augmentation, is prevalent in chest radiograph classification tasks. It involves the application of various transformations to the existing images to generate new diverse samples [3]. Techniques such as rotation, horizontal/vertical flipping, scaling, and adding noise can significantly increase the size and variability of the training data, which helps in preventing overfitting and improving the generalization capability of the model. In the context of chest radiographs classification, almost every study has employed data augmentation which helps to simulate different patient positions and makes the model more robust to variations [27, 80, 81]. Nevertheless, the excessive or inappropriate augmentation can introduce features that are not present in real-world radiographs which misleads the model.

Another data-centric approach is synthetic data generation by creating artificial chest radiograph images using techniques such as generative adversarial networks (GANs) [82]. This approach can be particularly useful in scenarios where obtaining a large annotated dataset is challenging. Synthetic images can help fill gaps in the dataset, especially for underrepresented classes, which is very common in medical images where most of the images comes from the healthy people. This improve model's generalisation capability. Many studies have used GAN for chest radiograph classification tasks and improved models performance [83, 84]. A challenge with using synthetic data is to make sure that the generated images are realistic and representative of actual clinical scenarios. Which requires experienced radiologist's review of the generated data samples. If the synthetic data lacks authenticity, it can lead to poor model generalization, reduce performance, or cause the model to rely on spurious correlations (the Clever Hans effect [85]), which ultimately results in inaccurate predictions in a real-world task.

Relabelling uncertain labels is another important data-centric technique to address the issue of the presence of uncertain labels in the dataset. The CheXpert dataset used in this thesis also has a significant number of uncertain labels. However, relabelling is a labor-intensive process and requires domain expertise, which can be a significant limitation, especially for large datasets like CheXpert. Researchers employed a number of techniques to remove the uncertainty of the labels by assigning 1 or 0 to all uncertain labels, relabelling using CNNs or assigning random values [9, 27]. This work presents a semi supervised (GMM) based technique to remove the uncertainty from the CheXpert dataset. [86].

2.6 Model-Centric Methods in Chest Radiology

The model-centric term refers to improvements in terms of changes in model architectures and training procedures. Researchers have employed a range of model-centric techniques to enhance the performance of various DL models for classifying chest radiographs. Given the data-hungry nature of DL, many studies have utilized pretrained models through a technique called transfer learning [21, 27, 32]. In this approach, the model is initially trained on a large dataset to learn basic image features, and then fine-tuned on the chest radiograph dataset to learn domain-specific features. This method comparatively reduces the amount of domain-specific data required, as the model already possesses a foundational understanding of basic image features. Additionally, it allows for quicker convergence and often leads to improved performance in specific tasks. However, transfer learning can be limited by the need for a sufficiently similar pretrained model and may not perform well if the pretrained models domain is too different from the target domain.

Ensemble methods have also been extensively used to boost overall performance in chest radiograph classification. In ensemble learning, multiple models are trained on the same dataset and then tested individually on each test radiograph. The final prediction is derived through majority voting, where the prediction agreed upon by most models is chosen as the final outcome [87–89]. This approach is beneficial because each model in the ensemble may capture different aspects of the data which leads to a more comprehensive and nuanced generalisation. As a result, ensemble methods produce more robust and accurate predictions by integrating diverse perspectives and interpretations from each model. However, the main limitation of ensemble methods is their increased computational cost and complexity, as they require training and maintaining multiple models, which can be resource-intensive and may not be feasible for all medical imaging applications.

Furthermore, various hybrid models have been proposed in the literature for chest radiograph interpretation. Hybrid models for chest X-ray combine different approaches to leverage their individual strengths and improve overall performance. For instance, features are extracted from chest radiographs using CNN-based models, and these deep features can then be used to train traditional ML models such as Support Vector Machines (SVM), Random Forests, or Gradient Boosting Machines (GBMs) for final classification [90–92]. This combination of a good feature extractor with a robust classifier provides improved generalization, works well with small datasets, improves classification performance, and reduces computational cost. By utilizing the strengths of both DL and traditional ML models, hybrid models offer an effective solution for chest X-ray classification. Nonetheless, hybrid models can suffer from the challenge of integrating systems and making sure that the feature extraction and classification stages are optimally aligned.

In addition to these approaches, many studies have proposed different loss functions to better guide the gradient and tackle data imbalance issues, which are very common in medical imaging tasks. These loss functions include focal loss, weighted loss, and multi-label softmax loss functions. Focal loss, for example, focuses on hard-to-classify samples by reducing the relative loss for well-classified examples and thus addresses class imbalance [93]. Weighted loss assigns different weights to classes based on their frequency which ensures that minority classes receive more attention during training [94]. Multi-label softmax loss functions are designed to handle cases where multiple conditions can be present in a single radiograph. These tailored loss functions help enhance model performance, address class imbalance, and provide more accurate and reliable predictions for multi label chest radiographs [95]. However, the implementation of these specialized loss functions can sometimes complicate the training process and may require careful tuning to achieve the desired performance improvements.

Moreover, hierarchical learning algorithms and Graph Convolutional Networks (GCNs) have been introduced to address the limitations of traditional classification schemes in chest radiology, particularly in capturing the complex correlations and hierarchical features among diseases. And it has shown substantial

performance improvement on CheXpert dataset. An advancement in this area is the ShuffleGhost Graph Convolutional Network (SGGCN), which combines a CNN with GCNs to improve diagnostic performance and computational efficiency. By using the SGNet-101 backbone, built with ShuffleGhost Blocks, SGGCN extracts high-quality image features with low computational cost [96]. Similarly, Active Learning AL technique is applied on CheXpert dataset to detect support devices (one of the 14 labels) attached to the patient. This AL mechanism works by selectively using the most relevant data for training and reduces the need for extensive annotated datasets [97]. Automated Compliance Assessment ACA is another effective approach to selecting high quality images of CheXpert radiographs, while training deep learning models for chest radiology. By extracting deep features from a patient position, [98] effectively identify non-compliant radiograph images which helps the deep learning model to avoid confusing scenarios.

Recent research has proposed several ways to improve chest X-ray classification models. One approach combines CNNs with LSTM layers. This CNN-LSTM model captures both spatial and sequential patterns in the images which enhance the classification of lung diseases like COVID-19 and pneumonia [99]. Another method is feature disentanglement, which helps models focus on disease-specific features instead of dataset-specific ones, thereby improving their generalization capabilities [100]. These advancements show that strategic changes in model architecture can improve chest X-ray classification. To build on these improvements, this thesis proposes a DenseNet121 model with a custom pooling layer, detailed in Chapters 3 and 5. Additionally, a new training method is introduced to enhance the detection of multiple coexisting conditions detailed in Chapter 6.

2.7 Effectiveness of Deep Learning Models in Chest Radiology

Deep learning models have shown significant promise in enhancing the perfromance and efficiency of chest radiology assessments. These models, particularly CNNs, have been trained on extensive datasets like CheXpert to interpret chest radiographs and often achieves performance levels comparable to or exceeding those of human radiologists. The integration of these models into clinical practice aims to support radiologists by providing second opinions to detect abnormalities with high confidence, and reduce diagnostic errors. For instance, [101] has shown that radiologist's performance in interpreting chest radiographs significantly improves when assisted by deep learning models . Moreover, DL models have demonstrated particularly high performance in detecting specific conditions such as pneumothorax, opacity, nodule, and fracture. [102] reported that DL models achieved expert-level performance in detecting clinically relevant abnormalities like edema, fibrosis, mass, pneumonia, and pneumothorax on ChestX-ray14 dataset with high AUC values. However, the performance can vary based on the prevalence of conditions in the training data. Models trained on diverse datasets performed well on new data from the same sites but showed reduced effectiveness on external datasets which indicates the potential limitations in generalizability of these models [23].

Furthermore, in comparing the performance of DL models with that of human radiologists, several studies have highlighted the impressive capabilities of these models. For instance, the CheXNeXt model achieved performance comparable to radiologists in detecting multiple pathologies and even outperformed radiologists in some cases such as atelectasis [103]. Regarding emergency department settings, a study shows that DL models have shown high diagnostic performance and improved the sensitivity of radiology resident's evaluation [104]. This shows the assistive capability of DL models in high-pressure environments where quick and accurate decisions are crucial. Deep learning models have also shown potential for specific applications beyond routine diagnostic tasks. For example, they have effectively detected conditions such as pneumoconiosis and tuberculosis on chest radiographs, with high AUC values and sometimes outperformed the certified radiologists [105]. Additionally, these models can assess changes in findings over serial radiographs, although their specificity for categorizing specific findings remains limited [106]. Despite these strengths, challenges remain, including the need

for large annotated datasets and the potential for domain gaps affecting transfer learning performance. It is also extremely important to continue refining these technologies to address their limitations. More importantly, their reliability and generalizability in diverse clinical settings will be key to their successful integration into routine clinical practice to enhance acceptance among radiologists which ultimately improve patient care.

2.8 Perception of Radiology Professionals

The integration of AI in radiology has been met with varying levels of acceptance and skepticism among radiology professionals and other medical disciplines. A survey revealed that while radiologists generally view AI as a beneficial tool that can enhance diagnostic accuracy and efficiency, there is an underlying fear that other medical disciplines, such as surgeons and medical students, may leverage AI to encroach upon traditional radiology roles [34]. This concern expresses a broader anxiety within the radiology community about the potential for AI to disrupt established professional boundaries and job security. Furthermore, a study conducted a systematic review alongside a cross-sectional survey which finds that physicians and medical students display a spectrum of acceptance towards AI in clinical settings. While some accept AI's potential to streamline workflows and improve patient outcomes, others are unsure about its reliability and the extent to which it might replace human expertise [35]. This mixed sentiment highlights the need for comprehensive education and training in AI applications to make sure that medical professionals can effectively integrate these technologies into their practice without fear of obsolescence.

The concerns of radiologists are also mirrored in other studies, such as [36] reported that many radiologists acknowledge the advantages of AI, particularly in enhancing diagnostic accuracy and reducing workload. However, they also express significant concerns about the loss of professional autonomy and the ethical implications of relying on AI for critical diagnostic decisions. Similarly, [45] highlighted that while there is optimism about AI's potential to revolutionize radiology, there are substantial hurdles to its widespread adoption, including technical challenges, lack of standardized protocols, and the need for rigorous validation and regulatory frameworks. These findings are further confirmed by surveys that focuses on medical student's perceptions of AI in radiology. Studies like [37] and [38] indicate that while AI is seen as a progressive force that can enhance the appeal of radiology as a career choice, there is also a prevailing concern about its impact on job availability and the nature of radiological work. The concern among future radiologists about AI's role highlights a gap in current educational curriculum, which must evolve to address these concerns and equip new professionals with the necessary skills to work alongside AI effectively. Collectively, these studies suggest a cautious optimism towards AI in radiology with concerns that must be addressed to foster confident and competent integration of AI in medical practice.

However, there is a gap in the literature, which is the lack of studies that examine how perceptions might shift after radiology professionals directly engage with AI models, in practical diagnostic scenarios such as interpreting chest radiographs. Current research, including that of [36] and [45], primarily focuses on the static perceptions and theoretical concerns of radiologists without considering the dynamic aspect of experiential learning and real-world application of AI technologies. This gap leaves a critical question unanswered: does hands-on experience with AIs diagnostic capabilities influence radiology professional's acceptance and trust in these tools?. This thesis addresses this gap by evaluating the shift in perception among radiology professionals before and after they interact with AI models specifically designed for detecting conditions on chest radiographs. By providing direct exposure to AIs performance in a clinical context, this research aims to offer a clearer understanding of how practical experience can mitigate concerns and potentially transform concerns into confidence and acceptance. This approach not only fills a gap in the current literature but also provides actionable insights for designing educational and training programs that could facilitate smoother integration of AI in clinical radiology.

Chapter 3

Methodologies

3.1 Introduction

Robust research in chest radiograph interpretation demands a multifaceted experimental approach. This Chapter outlines the various experimental designs employed in this thesis. The hardware and software specifications necessary for replication are detailed, along with a thorough explanation of data analysis and preprocessing protocols.

Both data-centric and model-centric methods are encompassed within the experiments, including strategies such as the effective relabelling of uncertain data samples, comparative analysis of different chest radiograph projections, and investigations into various deep learning architectures from basic convolutional neural networks (CNNs) to cutting-edge DenseNet121 and transformer models. Furthermore, a weighted loss function is defined for optimised handling of imbalanced data in CheXpert, a custom pooling layer is implemented for improved feature selection, and a multistage training technique is employed to enhance multi-condition detection. Benefits derived from data augmentation and transfer learning are also explored in this Chapter.

3.2 Hardware and Software Setup

All deep learning models were developed and trained in a powerful high performance computing cluster at the University of Hertfordshire, using four NVIDIA A100 80GB PCIe GPUs (with compute capability of 8.0, which refers to a classification used by NVIDIA to indicate the features supported by their GPUs [107]), offering a total of 316,012 MB of GPU memory, used in a mirrored strategy. The software environment consisted of Ubuntu Linux CentOS/RHEL 7.x with TensorFlow 2.6.2 as the deep learning framework. Python 3.6.8 served as the programming language. The key libraries and their versions included Keras (2.6.0), Keras-Preprocessing (1.1.2), NumPy (1.19.5), Pandas (1.1.5), scikit-learn (0.24.2), and SciPy (1.5.4). CUDA 11.4 enabled GPU acceleration.

Furthermore, Models were trained for 100 epochs with an early stopping mechanism based on validation AUC (area under the ROC curve) after a patience of 10 epochs, and the weights corresponding to the best validation AUC were selected. Training used a batch size of 32, binary cross-entropy/weighted loss, and the Adam optimizer with a learning rate of 1e-4.The training time of a CNN model ranges between 12 to 72 hours based on the size of the network and the dataset.

3.3 Data-Centric Methods

This research uses the CheXpert dataset, a large-scale collection of 223,414 chest radiographs obtained from 65,240 patients at Stanford Hospital between October 2002 and July 2017. This dataset is partic-

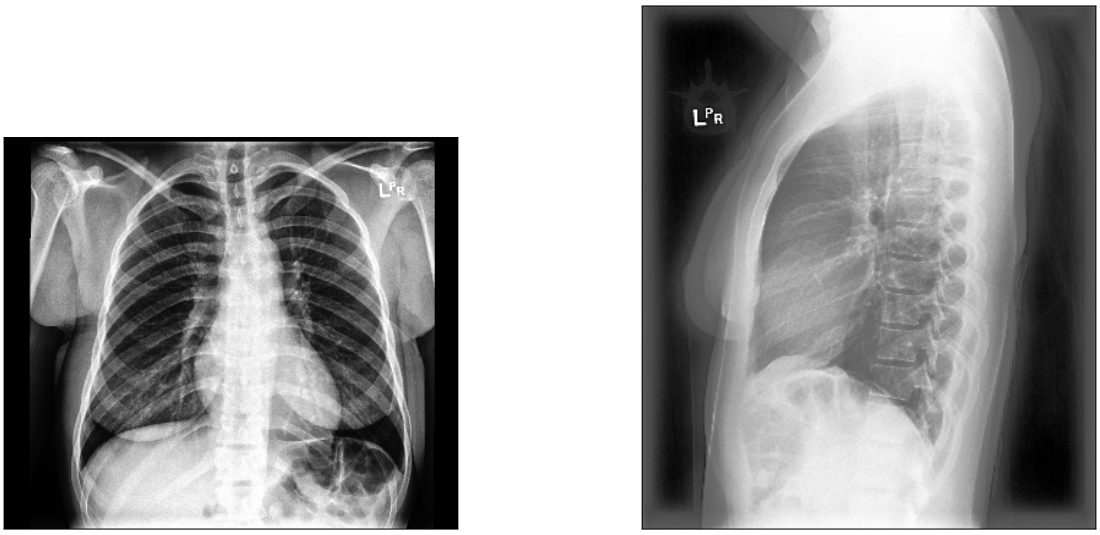
ularly valuable, as each image is associated with up to 14 labels that indicate the presence or absence of various medical conditions [9]. After obtaining access for research purposes through the CheXpert website [108]. The dataset was downloaded using the provided link. For this research, five clinically significant conditions were prioritized: Edema, Consolidation, Cardiomegaly, Atelectasis, and Pleural Effusion [9]. Additionally, labels such as Support Devices, Enlarged Cardiomediastinum, and Lung Opacity facilitate hierarchical analysis of labels and essential data preprocessing for model training.

The CheXpert dataset labeling process is notable for its handling of uncertainty. Labels are extracted from radiology reports using a natural language processing (NLP) system, designed to identify positive, negative, and uncertain mentions of pathologies [9]. This uncertainty labeling reflects real-world challenges in the interpretation of medical images and creates opportunities to develop robust AI models that can handle ambiguity.

3.3.1 Data Acquisition and Filtering

The downloaded data zip file contained a hierarchical directory structure. The radiograph images were organized first by patient number and then by study number (for patients with multiple radiographs). Within each study folder, images were named with a "view1," "view2," etc. prefix, followed by the projection (frontal or lateral). For example, a frontal chest radiograph for patient 01805, study 1, would be placed in the directory as "patient01805/study1/view1_frontal.jpg." Figure 3.1 shows examples of frontal and lateral projection radiographs from the CheXpert dataset.

As the dataset was provided as a source for an international competition, the dataset also includes three comma separated values (CSV) files named "train," "valid," and "test," which define the image splits for model development. The training set contains 223,414 images, the validation set contains 235 images, and the test set contains 669 images. Within each CSV file, the first column provides the absolute path to the corresponding radiograph within the image directory. This structure simplifies the process of loading and preprocessing the radiographs during model training and testing.



(a) Frontal (PA) Projection (371×320)

(b) Lateral (Side) Projection (320×369)

Figure 3.1: Frontal and lateral radiograph images from CheXpert dataset.

Initially, all radiograph images were matched with a template image and calculated the squared difference (SQDIFF) between the template and the image. Then all images with the highest SQDIFF were visually inspected to identify and remove low-quality or corrupted (images in which thoracic region is not visible) ones. A total of 17 images were excluded due to factors such as broken pixels or the absence

Uncertain Labels	One	Two	Three	Four	Five
Number of Samples	64085	26388	3439	258	10

Table 3.1: Shows the number of samples with one or multiple uncertain labels.

of relevant thoracic anatomy. Figure 3.2 illustrates two examples of such rejected images. Next, the labels within the CSV files were examined. These labels use the following scheme: 0 indicates the absence of a condition, 1 indicates its presence, and -1 indicates uncertainty. To maintain consistency across the train, valid, and test sets, all missing entries (originally empty cells) were filled with 0, indicating the absence of the corresponding condition.

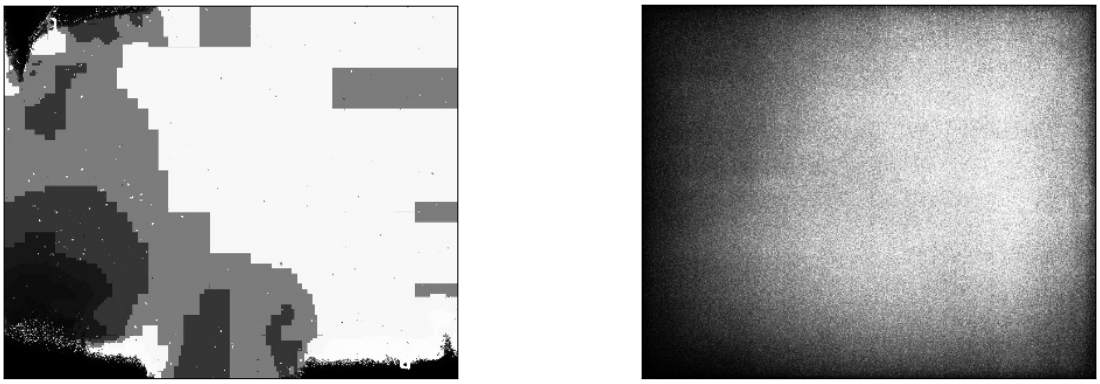


Figure 3.2: Rejected poor quality radiograph images from CheXpert dataset.

As explained above, the CheXpert dataset associates 14 labels with each radiograph, representing various conditions related to the lungs and the thoracic region. However, this research focuses on a specific subset: five clinically significant conditions plus the "No Finding" label. Since the original "No Finding" label reflects the absence of all 14 conditions, removing eight labels necessitates its redefinition. To ensure consistency within the scope of this research, the revised "No Finding" label is assigned '1' only if none of the selected five conditions is present. If at least one of these conditions is detected, the label assigned to 'No Finding' is '0'. This relabeling process was essential to align the data with the chosen research focus.

3.3.2 Relabeling of Uncertain Labels

An uncertain label means that the condition may or may not exist. This needs to be addressed. The CheXpert training set contains a significant amount of uncertain labels. Of the total of 223,414 samples, 64,085 contain at least one uncertain label. It is 28.6% of the dataset. Moreover, some samples have multiple uncertain labels, further increasing the uncertainty. Table 3.1 shows the number of samples that contain at least one, two, three, four, or five uncertain labels. To better understand this issue at the conditions level, Table 3.2 shows a detailed positive and uncertain label count for each of the five conditions. Interestingly, for Atelectasis, the number of uncertain labels is greater than the positive labels. Similarly, for consolidation almost double uncertain labels are present as compared to its positive labels.

As this is a multi-label problem, the presence of one condition can impact the appearance of another coexisting condition on the radiograph. To understand how the labels are structured, please see Figure 2.8. Also the manifestation of chest conditions on radiographs varies significantly with the cause of the condition, the health status of an individual, and the progression of the disease [109] [110]. To address the presence of uncertain labels in the CheXpert dataset, a semi-supervised approach using Gaussian

Conditions	Positive Labels	Uncertain Labels
Edema	38,569	12,984
Cardiomegaly	16,737	8,087
Atelectasis	24,224	33,739
Consolidation	9,689	27,742
Pleural Effusion	56,932	11,628

Table 3.2: Shows the number of positive and uncertain labels per condition.

Mixture Models (GMMs) was employed. GMM is a probabilistic model that assumes that data points are generated from a mixture of a finite number of gaussian distributions with unknown parameters [111]. The probability density function is given by equation (3.1). Where x represents the features extracted from the chest radiograph images. The features are derived from the raw pixel values of the images, which are first loaded, normalized, and flattened into one-dimensional vectors. This way, each image is represented as a single high-dimensional vector (e.g., a 224x224 image is represented by a 50,176-element vector), which serves as the feature vector for clustering by the GMM. wi is the mixture weights to tell the importance of each Gaussian component and $g(x|\mu_i, \Sigma_i)$ represents the Gaussian densities of the components. GMM were chosen over other clustering methods because they are superior at modeling complex data distributions and identifying distinct clusters within them [112]. The working principle of GMM makes it suitable to capture the various manifestations of chest conditions and make tighter clusters that reflect the distinct appearances of conditions on radiographs, even when multiple conditions coexist. As GMMs group data points into clusters based on their similarity. This can allow it to be used for the assignment of probabilities to uncertain labels.

$$p(x|\lambda) = \sum_{i=1}^{M} w_i g(x|\mu_i, \Sigma_i) \tag{3.1}$$

A separate GMM model was trained for each of the five conditions using only samples with positive labels. The models were optimized using the expectation maximization (EM) algorithm. The EM algorithm iteratively estimates the parameters (means, covariances, and weights) of the Gaussian distributions [113]. The optimal number of clusters for each condition was determined using the silhouette score method. A range of potential cluster numbers were evaluated (e.g., 10, 50, 100, 200, 300, 500, 600), and for each cluster number, a separate GMM was fitted to the data, with cluster assignments predicted accordingly. The silhouette score, which measures how similar each data point is to its assigned cluster compared to other clusters, was computed for each configuration. Specifically, the silhouette score for a data point is calculated by considering the mean intra-cluster distance (the average distance to other points within the same cluster) and the mean nearest-cluster distance (the average distance to points in the nearest neighboring cluster). This is given by equation 3.2. Where $s(i)$ refers to the silhouette score of a data point i, $a(i)$ is the mean intra-cluster distance, and $b(i)$ is the mean nearest-cluster distance [114]. This results in a score that ranges from -1 to 1 and a higher value indicates better-defined clusters. The cluster number that gave the highest silhouette score was identified as the optimal choice for that condition.

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{3.2}$$

Based on that, the number of clusters for each condition was determined for training are as follows: 100 for Consolidation, 500 for Pleural Effusion, 200 for Cardiomegaly, 300 for Atelectasis, and 350 for Edema. Upon convergence, the trained models were used to assign labels to uncertain labels of each condition. Each uncertain sample was evaluated by the corresponding trained GMM model, which calculated the likelihood of the sample belonging to each (positive) cluster. A threshold of 0.5

was applied to this likelihood value: if the likelihood of an uncertain sample being part of any cluster exceeded 0.5, it was relabeled as positive (1) for that condition; otherwise, it was assigned a negative (0) label. This approach offers an improvement over previous methods that either exclude these samples or assigned them positive(1)/negative(0) labels, respectively [9] [27]. Experiments were conducted with (after relabelling) and without uncertain samples to evaluate the impact on model performance. The results were also published in a conference paper [86].

3.3.3 Data Preparation for Each Projection

The CheXpert dataset offers a valuable resource for the development of deep learning (DL) models for chest radiograph interpretation, comprising frontal and lateral (This is a side view of the chest) projections. The frontal projection itself is further categorised into anteroposterior (AP) (this view is taken with the X-ray source in front and patient's back against the detector) and posteroanterior (PA) (this is the standard frontal view, like looking straight through the chest views). To facilitate a comprehensive study of how the performance of deep learning models varies between projections, distinct CSV files were carefully created for training, validation, and test sets. This involved filtering the original CSV files, using the "Frontal/Lateral" and "AP/PA" columns to ensure precise segregation of the different projections.

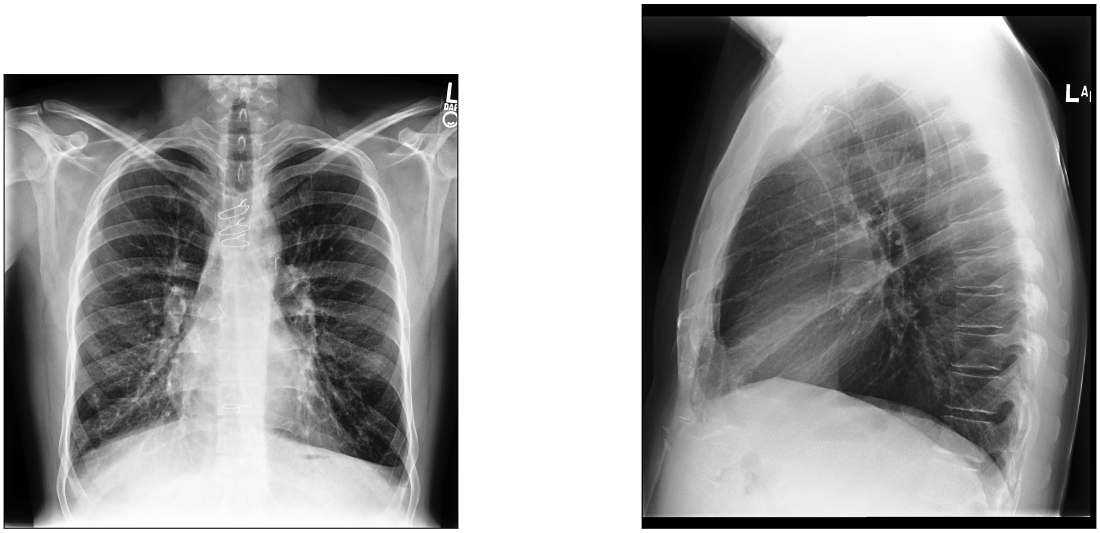
The creation of these projection-specific datasets allowed for a systematic series of experiments. Different deep learning models were independently tested on each projection, along with various model-centric techniques. The goal was to gain insight into how factors such as radiograph orientation impact the performance of the model. To offer a contribution to the ongoing advancement of AI within the medical imaging domain, the findings of this in-depth analysis were published in a conference paper [86].

3.3.4 Multi-scale Template Matching

Within the CheXpert dataset, chest radiographs often include areas beyond the clinically relevant thoracic region. These extraneous elements could potentially introduce noise during model training. To address this challenge, multi scale template matching was used to isolate the regions of interest [115]. Figure 3.3 illustrates examples of frontal chest radiographs from the dataset prior to template matching. Since manual cropping of the vast number of images was impractical, this process was automated with carefully selected templates.

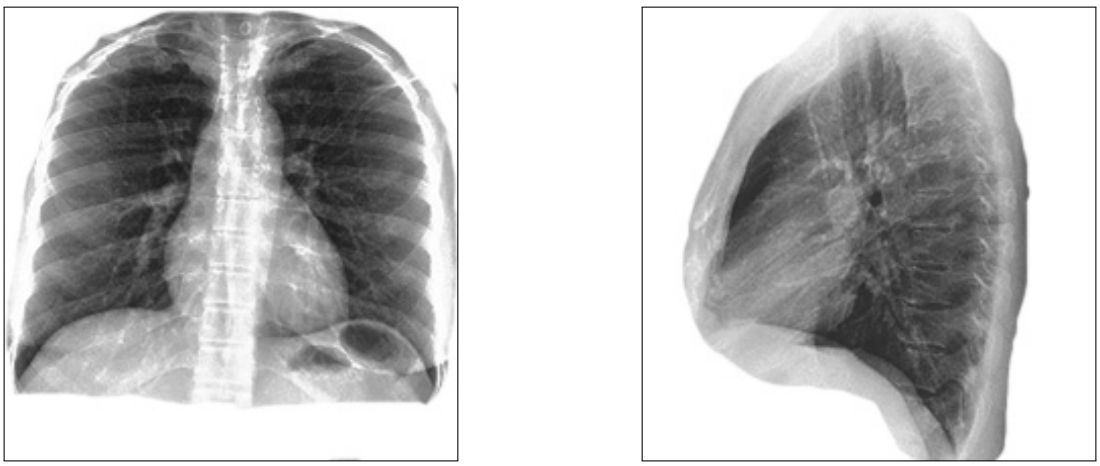
To ensure an optimal representation of the dataset, high-quality template images were carefully selected from the training data, specifically considering gender, age groups (0-20, 21-40, 41-60, 61-80, and above 80) and radiograph views (frontal, lateral). One template image was manually chosen for each combination of gender, age group, and view. High-quality in this context is defined by the clear appearance of the lung and thoracic region so that the anatomical structures are well-visible. All chosen images were resized to 224x224 pixels for standardisation. To further refine the template matching process, the selected templates were carefully edited in Photoshop software to isolate the region of interest (ROI) within the thoracic area and remove the background to reduce chances of false matches. This step improves the focus of the template matching algorithm on the most clinically relevant anatomical structures. Figure 3.4 shows one of the frontal and lateral view template used in this study.

Furthermore, instead of relying on a fixed-scale template applied by [27], a multiscale approach was used. Template images were iteratively rescaled within an empirically chosen range of 0.6 to 1.0. This range allows the template to "search" for the best thoracic region matching at different sizes within the original radiograph images. At each scale, the algorithm slides the template across the image in overlapping patches. For each patch, the sum of squared differences (SQDIFF) is computed as a similarity



(a) Frontal view. (b) Lateral view.

Figure 3.3: Frontal and lateral radiographs from CheXpert dataset before applying multi-scale template matching.



(a) Frontal view. (b) Lateral view.

Figure 3.4: Frontal and lateral view template images.

metric given by 3.3.

$$R(x,y) = \sum_{x',y'} (T(x',y') - I(x+x',y+y'))^2 \tag{3.3}$$

Here R is the result, T is the template image, I is the original image, x and y are the width and height of the image; lastly, x' and y' are the width and height of the template, respectively. Smaller (SQDIFF) values indicate a greater visual match between the patch and the template. This patch is then extracted and used for subsequent model training to ensure that the model is generalised only in clinically relevant areas. Figure 3.5 shows the same images as in Figure 3.3 after going through the multi-scale template matching process. The result clearly shows the effectiveness of this process as it successfully removes the irrelevant areas.

(a) Frontal view.

(b) Lateral view.

Figure 3.5: Frontal and lateral view images after multi-scale template matching applied.

3.3.5 Data Augmentation

Deep learning models are known for their ability to extract complex patterns from large datasets. However, its success is highly dependent on the quality and quantity of training data available. This dependence on data poses a challenge in scenarios where large, well-annotated datasets are difficult to acquire. The medical imaging domain is a prime example in which curating a sufficient volume of accurately labelled images can be both time-consuming and expensive.

The CheXpert dataset, while extensive, exemplifies the challenging characteristic of multi-label classification task. As explained earlier, in this research, a single radiograph has multiple associated labels. Despite the size of the dataset, the number of images representing specific combinations of positive conditions may still be limited, which could hinder the model's ability to learn subtle, yet important, distinctive features.

Here, data augmentation plays a crucial role. Data augmentation techniques generate a diverse set of transformed samples from the original dataset, simulating real-world variations that the model might encounter. By artificially expanding the training set, these techniques help the model generalise better, which helps to improve its performance on unseen data. Moreover, data augmentation is particularly valuable for multi-label tasks where specific label combinations may be under-represented in the original dataset.

In order to expand the dataset and improve the robustness of the model, several augmentation techniques were applied in this research. Normalisation techniques included the 'sample-wise centre'

(subtracts the image mean for brightness invariance) and 'sample-wise std normalisation' (standardises the pixel value distribution for contrast invariance). To aid in convergence, rescaling by 1.0/255.0 brought the pixel values into the 0-1 range. Geometric transformations included a rotation range (up to 7 degrees) for orientation invariance, a shear range (up to 20%) for robustness to positioning variations, and a zoom range (up to 20% zoom) for scale invariance. Lastly, 'horizontal flip' randomly mirrored images to increase the dataset size and reduce left-right bias. Figure 3.6 shows the original image as well as the transformed images after random application of the above-mentioned augmentation techniques.

Data augmentation is applied directly to each batch of training and validation data as they are fed into the model during training. This means that for each batch of images. The specified augmentation techniques (rotations, flips, etc.) are randomly applied to each individual image. This creates variations of the original images. In each epoch of the training, either the original or the randomly transformed version of that image becomes the part of the training batch. This exposes the model to a wider range of variations and improves its ability to generalise to unseen data.

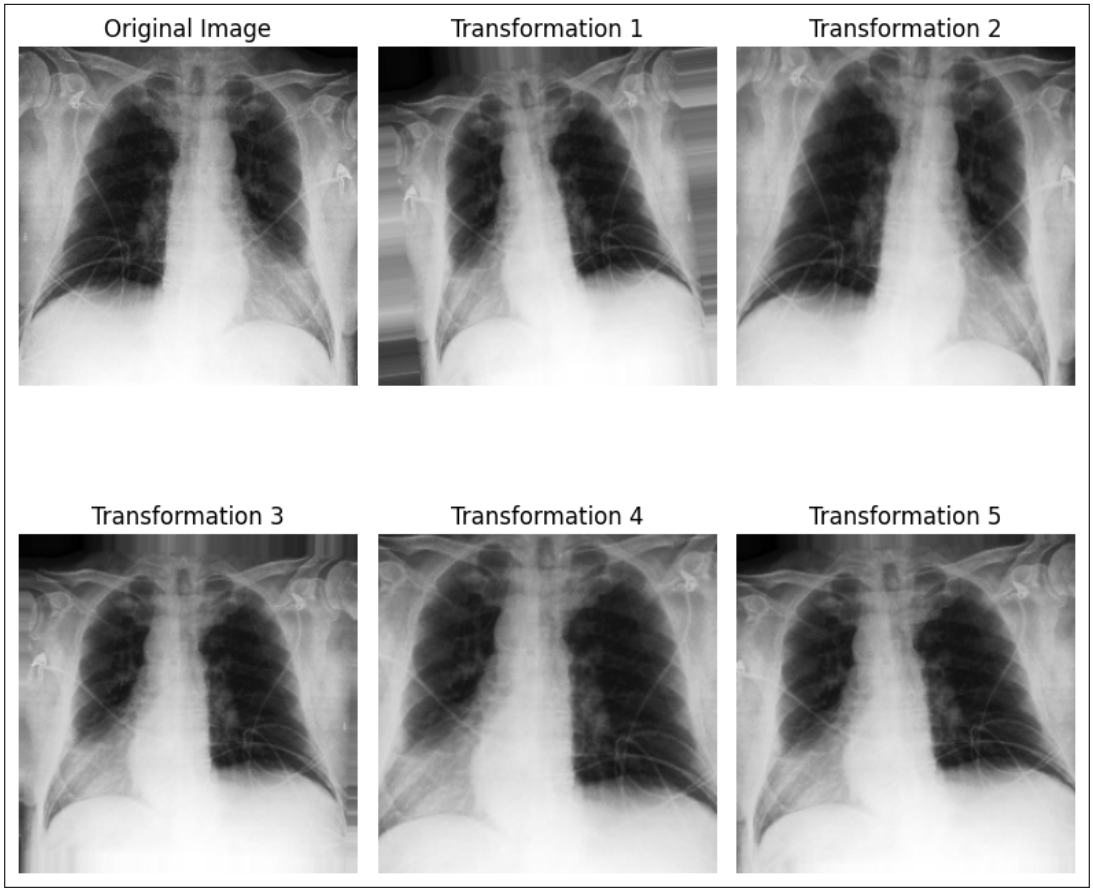


Figure 3.6: The original radiograph image from CheXpert and radomly applied augmentation transformations on it.

3.4 Model-centric Methods

This section explores methodological advancements employed in the field of model architectures and training strategies for the classification of multilabel chest radiographs. A multifaceted approach is adopted by taking advantage of established techniques such as transfer learning and convolutional neural networks (CNNs) alongside custom-designed components. These improvements are intended to enhance the model's ability to accurately identify and classify multiple pathologies present in a single

chest radiograph.

Firstly, the efficacy of transfer learning is investigated by fine-tuning pretrained models on the CheXpert dataset. This approach allows for the capitalisation of knowledge gained from the vast ImageNet dataset which accelerates the learning process and improves feature extraction capabilities. A weighted loss function designed to address the inherent class imbalance challenges associated with multi-label classification tasks. In addition, a custom pooling layer is introduced to pool the most promising features and stack them together.

Finally, two training strategies are presented. The first, Sequential Multi Label Enrichment DenseNet (SMLE-DenseNet), utilises a multistage DenseNet architecture with a progressive label enrichment scheme. This approach is designed to mitigate the problem of low performance in detecting multiple conditions simultaneously. The second strategy incorporates a self-attention mechanism with a Swin Transformer architecture that allows the model to focus on crucial regions of the chest radiograph with a long-lasting pixel-to-pixel relationship and improve classification for some subtle pathologies.

3.4.1 Transfer Learning

Transfer learning is a highly effective model-centric technique designed to enhance the performance of deep learning models. Instead of training a model from scratch, which is both time-consuming and resource-intensive, transfer learning leverages a pre-trained model. These models, trained on huge datasets such as ImageNet [26], have already learned to identify and extract rich and general features from images, such as horizontal and vertical edges. This pre-existing knowledge forms a strong basis for further training on specific tasks which leads to faster convergence and often superior performance [4]. In transfer learning, the initial step involves selecting a model pre-trained on a large, diverse dataset. Models like DenseNet121 [1] have undergone training on datasets such as ImageNet which enables them to develop a deep understanding of various image features. These models capture fundamental patterns and structures that are commonly present across different types of images. When these pre-trained models are repurposed for new tasks, the early layers, which detect basic features, are often retained, while the later layers can be fine-tuned to adapt to the specific features of the new task. This process can significantly reduce the amount of data and computational power needed for training.

The efficiency of transfer learning lies in its ability to apply learned features from one domain to another, hence called transfer learning. By starting with a model that has already mastered basic visual concepts, the need for vast amounts of labeled data for the new task is minimized. The pre-trained model's initial layers are usually frozen to keep their learned features, while the final layers are re-trained to cater to the new task. If the target dataset is significantly different from the dataset used for pre-training, then all layers of the pre-trained model needs to be fine tuned on the new dataset. Transfer learning not only speedup the training process but also enhances the model's ability to generalize better to new data. It allows the transfer of knowledge, reduce the redundancy and optimize the learning process for specific applications.

To adapt DenseNet121 to the CheXpert dataset, the original, task-specific fully connected layers at the end of the network were first removed. These were replaced with a new 6-node dense layer specifically designed for multi-label classification problem. This modification allowed the model to learn mappings unique to CheXpert data while still leveraging the powerful feature extraction abilities developed during pretraining on ImageNet. Two fine-tuning strategies were explored to further refine the model [116]. Initially, the pre-trained DenseNet121 layers were kept frozen and trained only the newly added layers, to preserve the learnt features. Subsequently, the entire network was fine-tuned, including the DenseNet121 layers. Given the significant differences in the types of data in ImageNet (natural images) and CheXpert (medical radiographs), this overall fine-tuning allows the model to adapt effectively to the domain-specific characteristics of medical images.

3.4.2 Weighted Loss Function

A loss function plays a pivotal role in image classification tasks and serves as a crucial feedback mechanism and guiding the learning process of a deep neural network. During training, the model receives an image as input and produces a set of predictions probabilities, which represents its confidence in the image belonging to each potential class. The loss function then quantifies the error between these predicted probabilities and the ground truth labels associated with the image. This error value acts as a signal that directs the model's optimisation process.

Through backpropagation [117], the weights within the network are iteratively adjusted to minimise the difference between predictions and targets. Here, the choice of loss function directly influences how the model learns. For instance, binary cross-entropy loss is commonly used in image classification to encourage confident predictions for the correct classes while penalising incorrect ones. By progressively minimising the loss, the model learns to extract increasingly discriminative features from the images, to enhance its ability to accurately classify unseen images. In essence, the loss function acts as a compass, continuously pointing the model towards more optimal representations of image data and ultimately leading to improved classification performance.

Binary cross-entropy loss [118] is designed for binary classification problems where each data point can only belong to one of two mutually exclusive classes. Measures the dissimilarity between the predicted probability distribution (how confident the model is in its prediction for each class) and the true label distribution (where the true label is either 0 or 1). It is given by the equation 3.4.

$$L(y, \hat{y}) = -\left(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\right) \tag{3.4}$$

Where:

- **y** is the true label (either 0 or 1).
- $\hat{\mathbf{y}}$ is the predicted probability that the output is 1.

It works by penalising confident but incorrect predictions. When the true label y is 1, and the predicted probability $\hat{y}$ is close to 1 (high confidence), the loss is close to zero. However, if the predicted $\hat{y}$ is low (low confidence), the loss increases significantly. The same applies when the true label y is 0: a predicted $\hat{y}$ close to 0 produces a low loss, while a $\hat{y}$ close to 1 incurs a high loss.

While binary cross-entropy loss is effective for binary tasks, its direct application in multi-label settings faces limitations. In a multi-label classification problem like the one this thesis focusses on, each radiograph can belong to multiple conditions simultaneously. Since each label can be either 0 or 1 for a sample, ensuring a perfectly balanced dataset (where each class has an equal number of samples) is nearly impossible. This inherent class imbalance can lead to sub-optimal results with standard binary cross-entropy. The loss function will penalise misclassifications of minority and majority classes equally, which potentially hinder the model's ability to learn effectively from under-represented conditions.

To address this issue, a custom weighted loss function [119] is employed. This loss function is designed to calculate the loss for each condition individually, with their subsequent summation. Thus, using equation 3.4 yields equation 3.5, which represents the comprehensive loss function. Importantly, when calculating the loss for each condition, the class imbalance is carefully considered. To accommodate this, class weights are determined separately for labels 0 and 1 within each condition.

$$L(y, \hat{y}) = L(y, \hat{y}_{ed}) + L(y, \hat{y}_{ca}) + L(y, \hat{y}_{co}) + L(y, \hat{y}_{at}) + L(y, \hat{y}_{pe}) \tag{3.5}$$

Where:

- $\hat{y}_{ed}$ are true labels of Edema .

- $\hat{y}_{ca}$ are true labels of Cardiomegaly .
- $\hat{y}_{co}$ are true labels of Consolidation.
- $\hat{y}_{at}$ are true labels of Atelectsis.
- $\hat{y}_{pe}$ are true labels of Pleural Effusion.

The weight of the class for each condition is calculated by using the frequency of the label. To calculate the frequencies of positive and negative labels, the total number of samples is first determined. This is achieved by counting the number of samples. Next, to compute the frequency of positive classes, the occurrences of each condition are summed up. This count is then divided by the total number of samples to obtain the proportion of samples that belong to each positive class.

On the contrary, to determine the frequency of negative classes, the count of positive classes is subtracted from the total number of samples. This count is also divided by the total number of samples to yield the proportion of samples that are classified as negative. The resulting positive and negative class frequencies then multiplied to the negative and positive losses, respectively. Redefining the equation 3.4 by including the weights gives the equation 3.6. The python code for defining weighted loss function is given in Appendix A (Weighted Loss Function).

$$L(y, \hat{y}) = -\left(w_n * (y \log(\hat{y})) + w_p * ((1-y)\log(1-\hat{y}))\right) \tag{3.6}$$

Where:

- w_n is the positive class weight.
- w_p is the negative class weight.

3.4.3 Custom Pooling Layer

Pooling layers serve as a crucial element within convolutional neural networks (CNNs) and offer several advantages [120]. Their primary function lies in downsampling feature maps to reduce their spatial dimensions. This reduction leads to decreased computational costs and helps mitigate overfitting within the network. The pooling layers accomplish this downsampling by summarising the information within a receptive field of the feature map into a single representative value. Figure 3.11 illustrates examples of the max-pooling operations.

Various types of pooling operations are found in CNNs, with max pooling reigning as the most prevalent choices. Max pooling operates by selecting the maximum value within the pooling window as the representative output. Less frequently employed pooling operations include stochastic pooling, in which elements are randomly selected based on a probability distribution, and mixed pooling, which offers a hybrid approach that blends aspects of multiple pooling operations. The pooling window size and stride are important hyperparameters. A commonly used configuration involves a 2x2 window size with a stride of 2.

Max-pooling often works better than other pooling methods in CNNs. This is because it focusses on the highest activations within the pooling window and highlights the highest features (in terms of pixel values) found by the convolutional layers. The exact reason for why max pooling works so well is not yet fully understood [121]. The general idea is that the strongest activations picked up by max pooling likely point to the most important and distinctive features of an image. Max-pooling also helps the model work better even if the input image is slightly shifted or changed. This makes the model stronger overall. Figure 3.7 shows the maxpooling operation on a chest radiograph focussing on the costophranic angle

of the left lung. It is important to check the bluntness of the costophrenic angle of the lungs to detect pleural effusion [122].

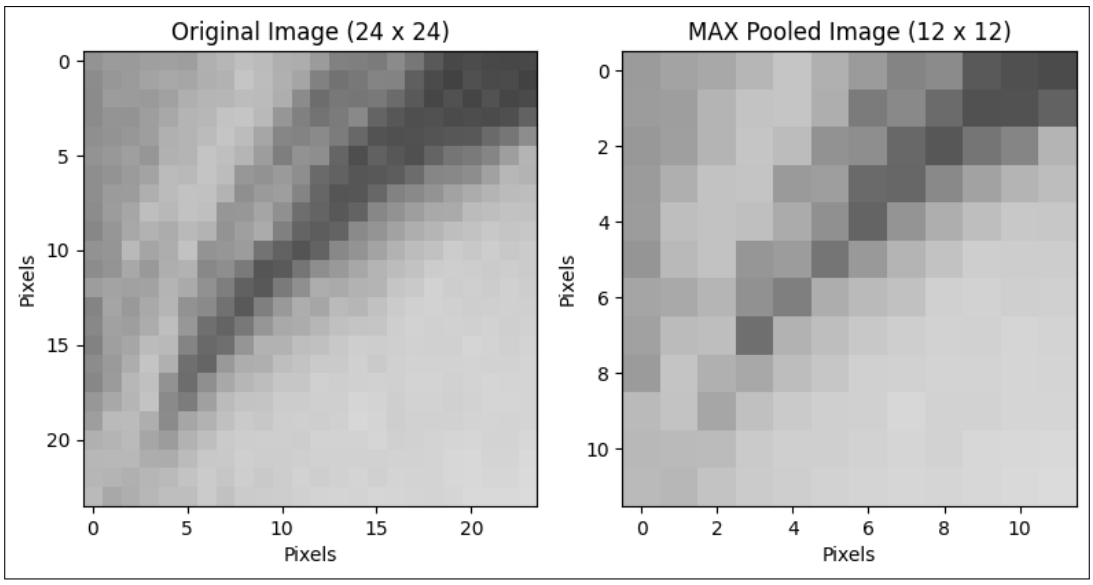


Figure 3.7: Max-pooling operation Performed on costophrenic angle.

To gain a deeper understanding, max-pooling has been extensively investigated in this thesis because its suitability for all image types warrants consideration. The approach of extracting only the highest value elements demonstrates effectiveness in general image classification, likely because the highest-valued pixels often represent a key feature within a small image region. However, this paradigm may not hold equally for radiograph images, where the entire spectrum of white and grey shades carries interpretive significance for diagnosis [72]. Subtle low-intensity shadows could represent discriminative features indicating a specific condition. Due to its mechanics, max pooling risks overlooking these crucial features, especially in feature map regions dominated by high pixel values [123]. Figure 3.8 illustrates this potential drawback and demonstrate how a discriminative feature is lost after a single max pooling operation, in situations where the important feature is a bunch of low value pixels (black) in a mostly high-value (white) area. This highlights the need to carefully evaluate the suitability of max-pooling in the context of radiograph images.

One potential strategy to address the limitations of max pooling involves selecting the minimum value from each feature map region rather than the maximum. This theoretical approach underpins the concept of 'min pooling'. To evaluate this hypothesis, min pooling was implemented by first inverting the feature map's values (multiplying by -1), thereby transforming the highest values into the lowest. Subsequently, standard max pooling was applied and the resulting feature map was multiplied by -1 to restore the original sign convention. This technique provides a potential alternative to max pooling by successfully preserving the discriminative feature. Although min-pooling can be useful in some cases, it also has its inherent weaknesses. Recall the example in Figure 3.8, where the important feature consisted of low value pixels surrounded by mostly high-valued pixels. Min pooling is great for that. But what if the important feature is a bunch of high-value pixels in a mostly low-value pixel area. In that situation, min pooling works just as poorly as max-pooling previously did, and the advantage is lost. But here max-poling can work perfectly. Figure 3.9 shows how max pooling is doing the job perfectly in retaining the discriminative feature in this scenario.

To illustrate the distinct ways in which min pooling and max pooling preserve discriminative features, let us compare their effects. Min pooling has been applied to the same image used in Figure 3.7, with

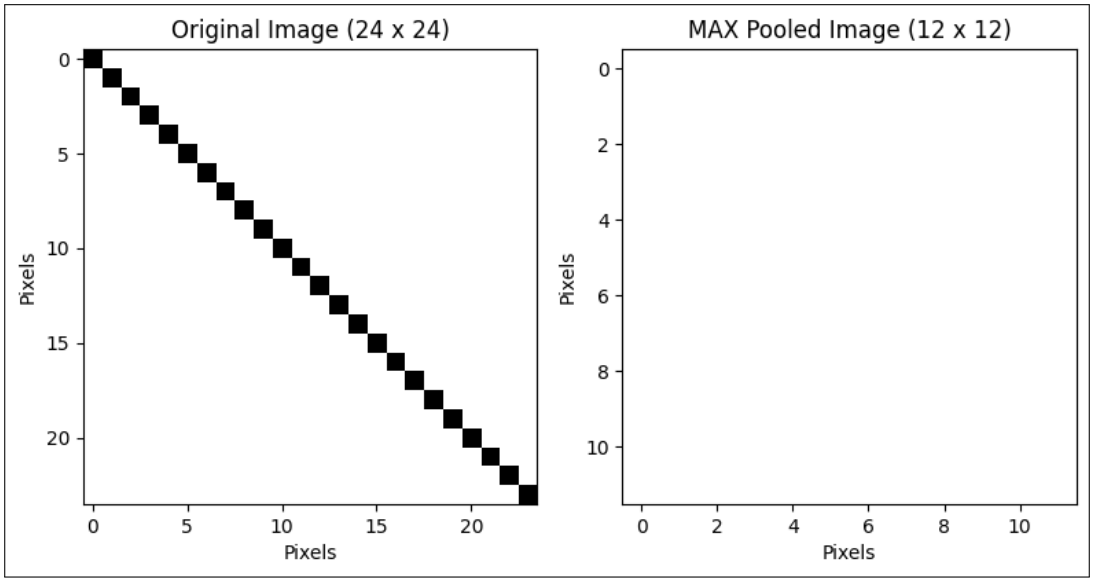


Figure 3.8: Discriminative feature disappreas after max pooling operation .

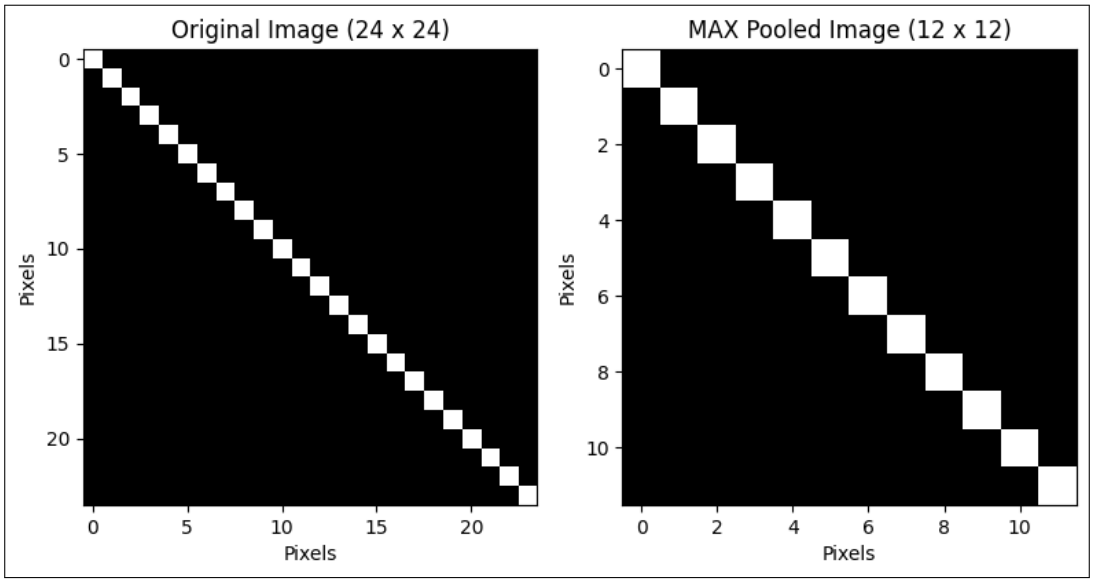


Figure 3.9: Discriminative feature preserved after Max pooling operation .

the results shown in Figure 3.10. Close examination reveals that each technique emphasises a different set of discriminative features. In the case of min-pooling, the costophrenic angle is more visible due to its composition of low pixel values. This aligns with the previous discussion on why min-pooling excels in such scenarios.

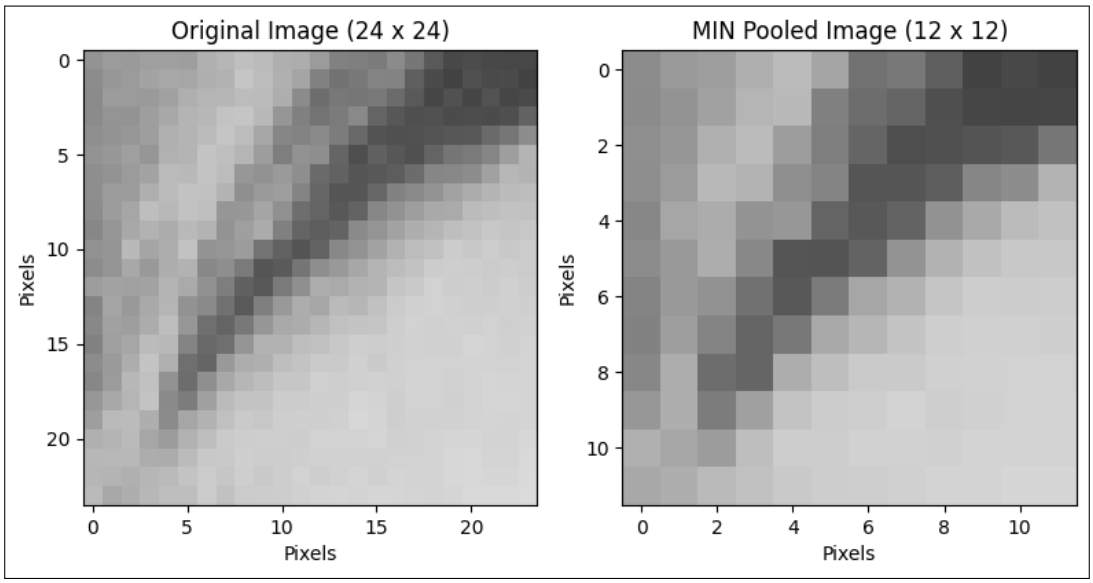


Figure 3.10: Min-pooling operation Performed on costophrenic angle.

Based on the insights gained from this discussion, a custom pooling layer was developed to harness the advantages of both max and min pooling. This novel layer operates by independently applying max pooling and min pooling to the same feature map and subsequently concatenate the resulting feature maps along the channel dimension. For instance, if the input feature map has dimensions $H \times W \times C$ (height, width, and channels), the max pooled and min pooled feature maps will each have dimensions $H' \times W' \times C$. These two feature maps are then concatenated along the channel dimension which results in a final feature map of dimensions $H' \times W' \times 2C$. This approach allows the model to extract significant features with even greater efficacy. Experiments that compare the custom pooling layer to standard max pooling demonstrate its enhanced performance in Chapter 5 of this thesis. Figure 3.11 provides a visual representation of the custom pooling operation. The python code for the custom pooling operation is given in Appendix A (Custom Pooling) of this thesis.

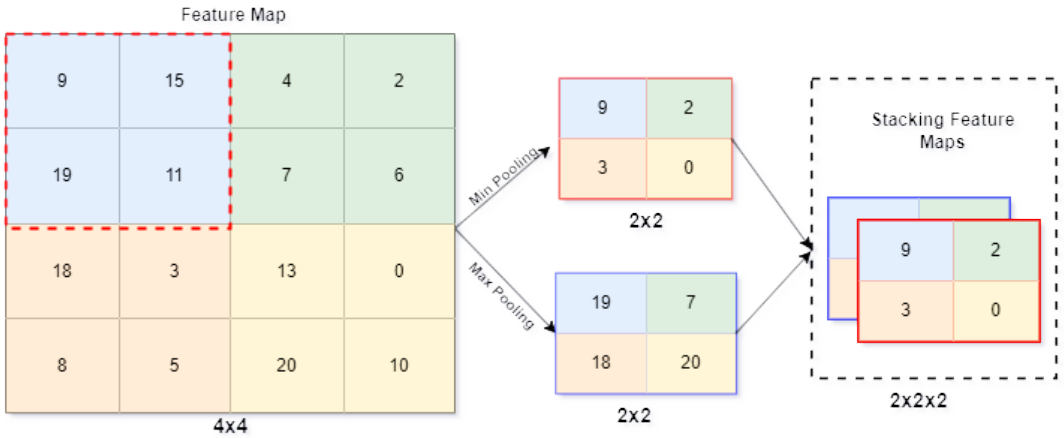


Figure 3.11: Custom-pooling operation with min and max pooling

Training Data Subset	Number of Samples	% of Total Training Samples
One Condition	84,876	44.53
Two Conditions	46,716	30.63
Three Conditions	17,278	11.33
Four Conditions	3,327	2.18
Five Conditions	275	0.18

Table 3.3: Number of samples in the training set with the number of positive conditions.

3.4.4 Sequential Multi Label Enrichment (SMLE-DenseNet)

Multilabel classification presents unique challenges, especially within the medical imaging domain. In the CheXpert dataset, accurately identifying multiple co-occurring conditions within a single image is complicated by the decreasing frequency of samples as the number of positive conditions per sample increases. This imbalance, where easily classified single-condition samples dominate the dataset, hinders a single model's ability to learn and generalise effectively. Table 3.3 illustrates this distribution within the training set.

Furthermore, the co-existence of multiple conditions on a single radiograph changes the appearance of distinctive features and poses a significant challenge for both human experts and deep learning models [124]. When multiple conditions overlap, radiographic features can become distorted and lead to possible misinterpretations of the model. For example, a deep learning model trained to detect pneumonia could focus primarily on the presence of consolidations (dense white opacities). When these consolidations are obscured or mixed with the haziness caused by pulmonary edema, the model's ability to confidently identify pneumonia can falter [125]. Similarly, a model designed to detect atelectasis may struggle if the expected linear opacities and lobar collapse are partially masked by the enlarged heart silhouette in cardiomegaly [126]. Therefore co-existing conditions introduce novel or atypical appearances, the models may struggle to distinguish between subtle variations and true signs of a single disease. Figure 3.12 shows two CheXpert radiographs with coexisting conditions.

(a) Cardiomegaly with Atelectasis.

(b) All five conditions present.

Figure 3.12: Radiographs with co existing conditions.

To address this, a novel method is proposed entitled Sequential Multi-Label Enrichment (SMLE-DenseNet). This approach combines a strategic multistage training procedure with progressive expansion of the model's architecture. The core principle is to train different sections of the model with a different subset of dataset outlined in Table 3.3, to give the model a greater capacity to learn their complex patterns. SMLE-DenseNet modifies the standard DenseNet121 architecture by incorporating an additional 4-layer

convolutional block within each of its 5 stages. Figure 3.13 shows the new conv block (in dotted black line) added in each stage. This block receives input from the last dense block of the DenseNet121 network.

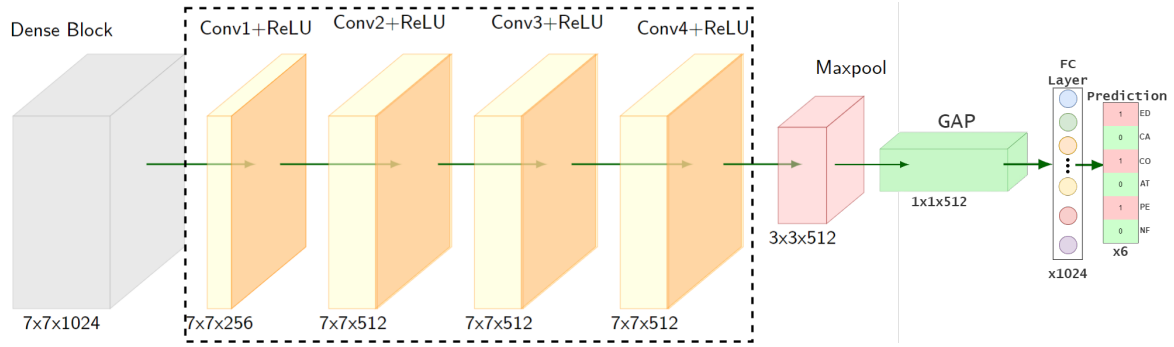


Figure 3.13: A block of 4 convolutional layers (in dotted line) followed by a maxpool, global average pool and fully connected layers. The output layer is a 6-bit vector representing the presence of the following conditions: CA (Cardiomegaly), CO (Consolidation), AT (Atelectasis), PE (Pleural Effusion), ED (Edema), and NF (No Finding).

In stage 1 of the SMLE-DenseNet process, the dataset comprising samples with only one positive label is utilised to fine-tune a pre-trained DenseNet121 model. This initial stage establishes a strong baseline for recognising the core radiograph features relevant to each of the five conditions. Subsequent stages involve adding blocks of convolutional layers (four layers, one with 256 nodes, and the rest with 512 nodes each) to the existing architecture. Importantly, the number of trainable layers progressively increases at each stage. For example, in stage 2, all layers except the last 25 are frozen. Datasets in stages 2 and above are carefully curated to contain samples with an increasing number of positive conditions (Table 3.3) (e.g., stage 2 uses samples with exactly two positive conditions). The final stage (stage 5) involves training with samples having exactly five positive conditions. This sequential training, combined with layer-wise control during fine-tuning, forces the model to learn increasingly complex patterns associated with multiple co-existing conditions at the last stages of training. Upon completion of the training process, the full SMLE-DenseNet model is evaluated on CheXpert test set to measure its performance on the multi-label classification task. Figure 3.14 shows the full configuration of SMLE-DenseNet architecture.

3.4.5 Self Attention with Swin Transformer

Traditional image classification approaches predominantly use convolutional neural networks (CNNs). Although effective, CNNs exhibit inherent limitations due to their localised convolutional operations. This focus can hinder their ability to effectively model long-range dependencies within images, which is an important aspect for understanding complex relationships in medical imagery such as chest radiographs. To address these shortcomings, the field has turned to attention-based transformer models for images [7].

Transformer models excel by incorporating a self-attention mechanism. Unlike the localised focus of convolutions, attention allows the model to dynamically weigh the importance of different regions within an image (in this case, image patches) and establish relationships across the entire input at each layer. Figure 3.15 highlights the key difference between CNNs and transformers, CNNs rapidly lose long-range pixel relationships as information propagates through the network, while transformers maintain these relationships throughout. This mechanism resembles the diagnostic process employed by radiologists. When examining a chest radiograph, they carefully compare and analyse various regions, such as the lungs, to identify abnormalities or patterns indicative of disease conditions. Transformer

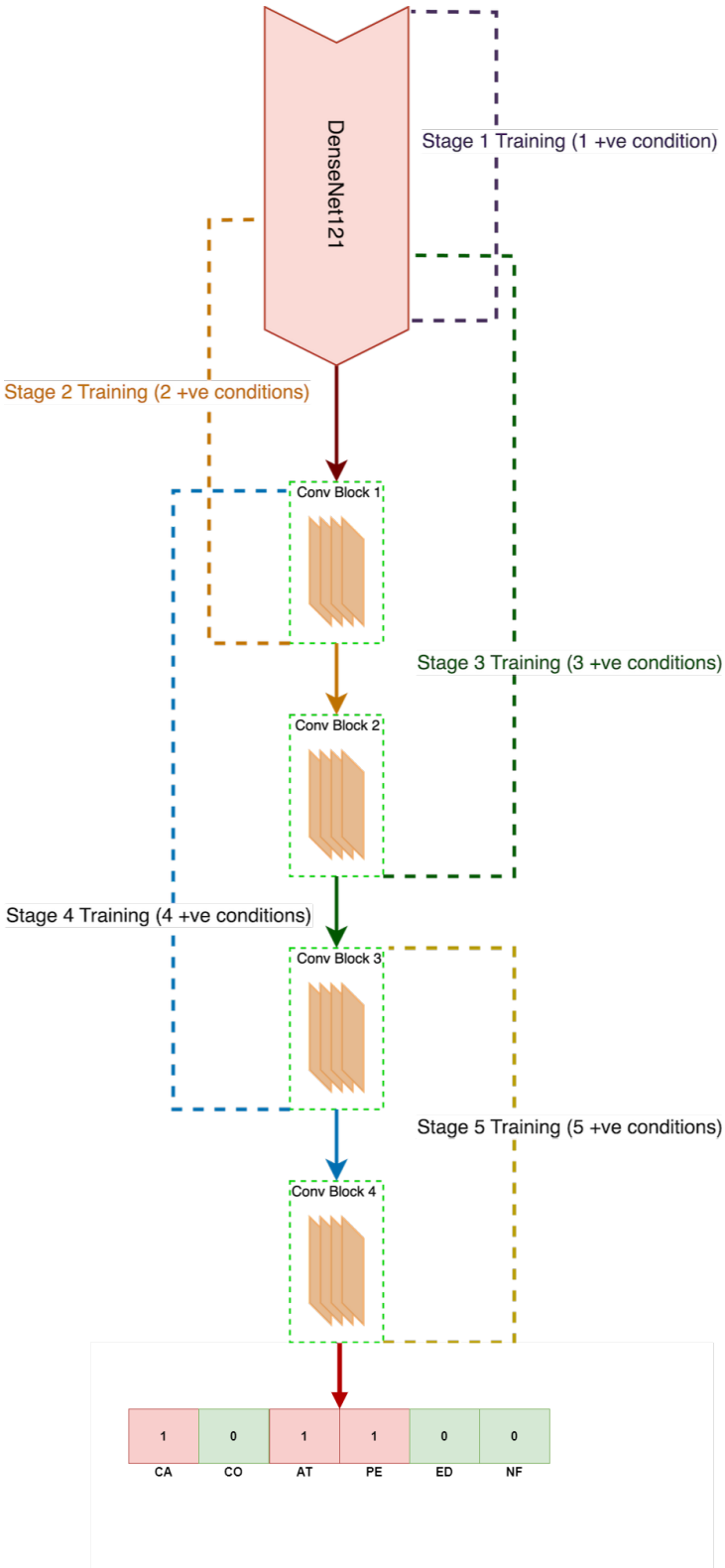


Figure 3.14: SMLE DenseNet architecture with training stages covering the model layers mentioned with dotted lines. The last layer is a 6-bit vector representing the presence of the following conditions: CA (Cardiomegaly), CO (Consolidation), AT (Atelectasis), PE (Pleural Effusion), ED (Edema), and NF (No Finding).

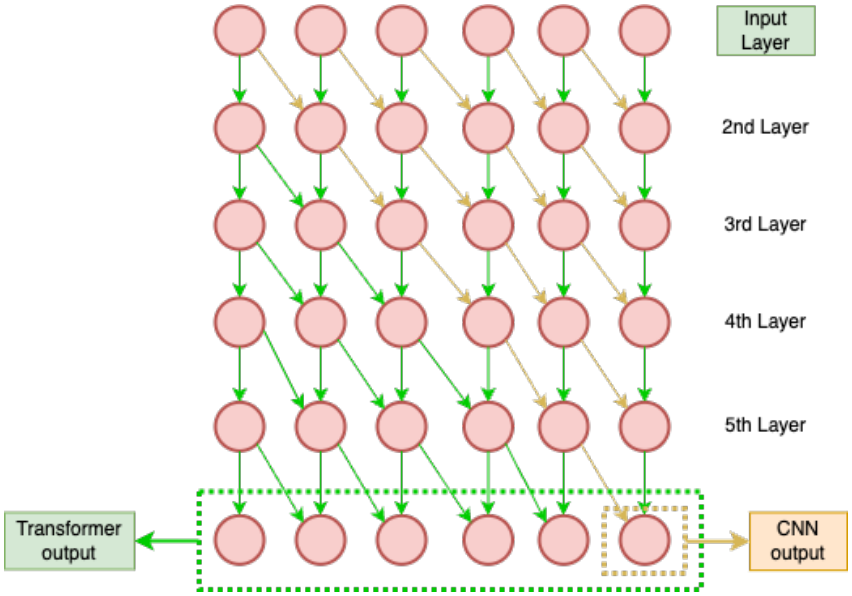


Figure 3.15: CNN vs transformers long range pixel to pixel relationship. Arrows indicate the flow of context, Orange for CNN and Green for Transformers.

models intrinsically emulate this contextual analysis. Given the potential of attention-based models, Swin Transformer based architecture was employed for classifying multi-label chest radiographs. The Swin Transformer offers advantages through its hierarchical design and the shifted window scheme that enhance its ability to capture fine-grained details and global relationships [8]. Figure 3.16 shows the 3 swin block architecture utilised in this research.

3.5 Evaluation Metrics

While doing classification, the models try to predict categories such as positive (1) vs. negative (0) in the case of each condition in this thesis. It is paramount to know how well the model performs. The challenge of measuring model performance in the case of single-label classification is much lower as compared to multilabel classification. The Area Under the Receiver Operating Characteristic (AUC-ROC) curve is a vital tool for this [127]. Think of it as a report card summarising the model's ability to differentiate classes. The AUC score tells the likelihood that the model will correctly assign a higher score to a randomly selected positive example than to a randomly selected negative example. The models in this research assign probabilities to each of the labels, which is why the AUC is better than accuracy in terms of describing the true picture of the performance of the model [128].

The ROC curve is the visual heart of the AUC-ROC metric. It is a graph where **Y-axis**: True Positive Rate (TPR) given by equation 3.7 : Also called sensitivity or recall, this reveals how good the model is at finding the true positives within the data (e.g., accurately labelling a condition as positive). **X-axis**: False Positive Rate (FPR) given by equation 3.8: This shows how often the model makes mistakes, classifying negatives as positives (e.g., incorrectly labelling a condition as positive). Ideally, the ROC curve hugs the top left corner of the graph. This means a high TPR (lots of true positives) and a low FPR (few mistakes). A perfect classifier would score an AUC of 1.0, while random guessing would get an AUC of 0.5.

The TPR calculates the proportion of actual positive cases the model correctly predicted.

$$TPR = TP/(TP + FN) \tag{3.7}$$

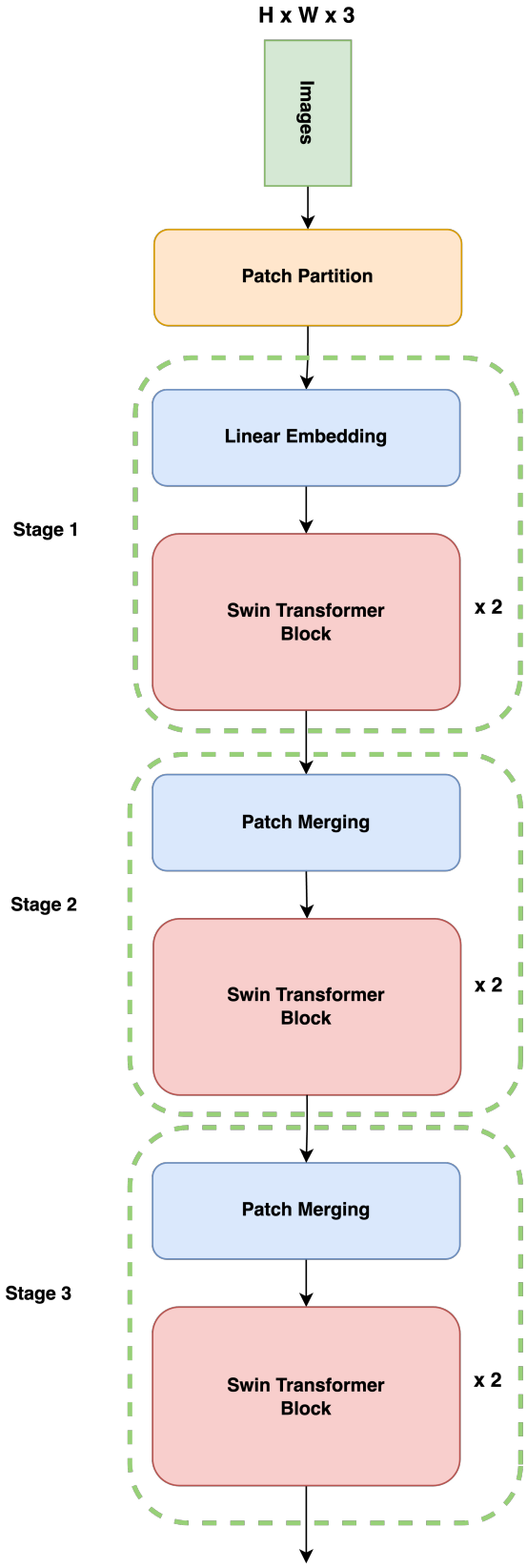


Figure 3.16: Swin Transformer network with initial patch partition and linear embedding, and three Swin Transformer blocks. Patch merging occurs in stages 2 and 3.

Where:

- TP are the True Positives (Correctly classified positives).
- FN are the False Negatives (Incorrectly classified negatives).

The False Positive Rate (FPR) shows the proportion of actual negative cases the model wrongly labeled as positive.

$$FPR = FP/(TN + FP) \tag{3.8}$$

Where:

- FP are the False Positives (Incorrectly classified positives).
- TN are the True Negatives (Correctly classified negatives).

In practice AUC-ROC is particularly useful when the distribution of labels is imbalanced (e.g., many more true positive labels for one condition as compared to the others). It Provides a more holistic view of model performance than simple accuracy. By using AUC-ROC, The classification thresholds can be adjusted to find the sweet spot between finding true positives and minimising false alarms. Figure 3.17 shows an ROC curve plot with AUC = 0.94. A high true positive rate can be observed as the curve is skewed towards the upper left corner.

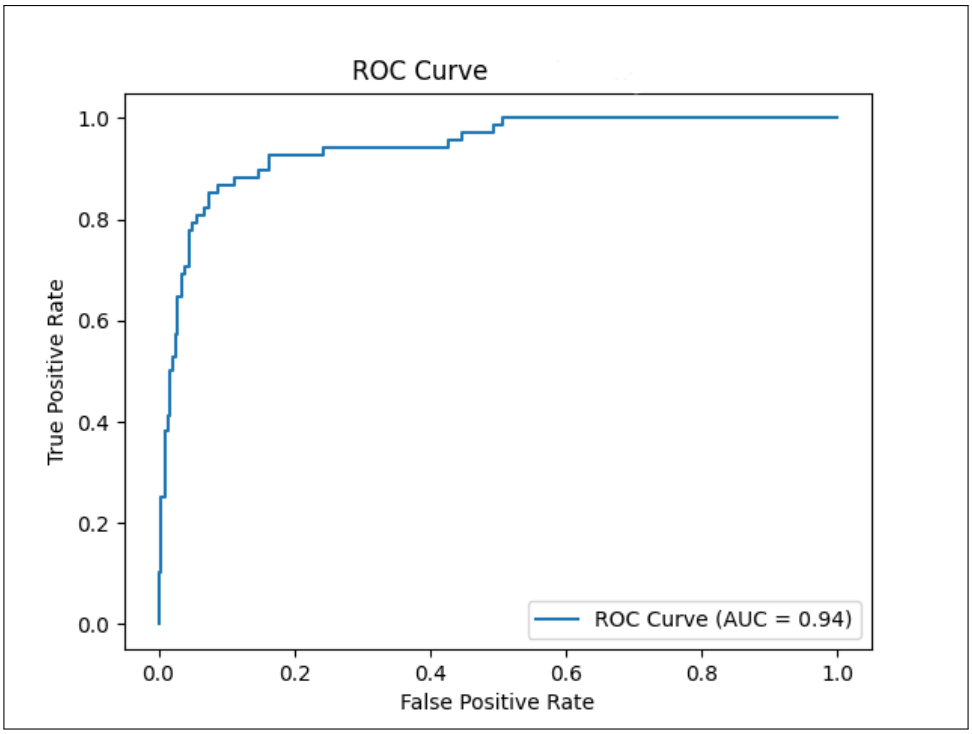


Figure 3.17: AUC-ROC curve between TPR on Y-axis and FPR on X-axis.

Chapter 4

Uncertainty in Labels and Radiograph Projections

4.1 Introduction

This Chapter explores two strategies to improve the performance of deep learning models trained for multi-label classification of chest radiographs. The focus is on experiments that address key challenges in medical image classification: the handling of uncertain labels and the impact of different radiograph views. These experiments leverage the CheXpert dataset, which is a large collection of chest radiographs.

The first experiment addresses the issue of uncertain labels, which is a significant problem in the CheXpert dataset due to its reliance on an NLP system to extract labels from radiograph reports. This introduces ambiguities and potential errors. To investigate the impact of these uncertainties, deep learning models were trained with and without uncertain samples. First, uncertain samples were excluded and three models were trained: starting with a very basic 4-layered CNN model, a DenseNet121, and a 6-layer Swin transformer. Performance was evaluated on the CheXpert test set. Next, uncertain samples were relabelled using a Gaussian Mixture Model (GMM; see section 3.3.2 for details), added back into the training set, and the models were retrained and re-evaluated.

The second experiment investigates the performance of deep learning models across three primary chest radiograph projections: posteroanterior (PA), anteroposterior (AP), and lateral. Variations in view can significantly impact the visibility of anatomical structures and pathologies. This experiment trains and evaluates models separately on each view, which provide insights into the strengths and limitations of each projection for optimal model performance. These findings can inform clinical decision-making and the development of more robust chest radiograph classification models.

4.1.1 Hypothesis: Hyp 01

The hypothesis for this Chapter is as follows:

The implementation of GMM-based uncertain label handling on CheXpert dataset, significantly improves the classification performance of deep learning models in chest radiograph interpretation.

4.2 Experiment 1: Uncertain Label Handling

4.2.1 Preprocessing

For this experiment, two distinct preprocessing approaches were applied. First, samples containing uncertain labels were identified and excluded from the training dataset. This was achieved using a filter

Condition	Positive Labels (1)	Negative Labels (0)	Uncertain Labels (-1)
Edema	38,569	96,925	12,984
Cardiomegaly	16,737	118,757	8,087
Consolidation	9,689	125,805	27,742
Atelectasis	24,224	111,270	33,739
Pleural Effusion	56,932	78,562	11,628

Table 4.1: Positive Labels (Training Set), Negative and Uncertain label distribution per condition. Note that the samples may be shared across different conditions due to the presence of coexisting conditions.

Condition	Number of Clusters
Edema	350
Cardiomegaly	200
Consolidation	100
Atelectasis	300
Pleural Effusion	500

Table 4.2: Number of clusters used for each condition to train a condition specific GMM model.

that separated all samples with even a single uncertain label (-1) to establish a baseline to assess the impact of relabelling of uncertain labels. The total number of training samples are 223,414, out of which 64,085 contains uncertain labels. After the filtering, the total number of training samples left were 159,329. The label distribution for each condition, including Positive Labels, Negative Labels, and Uncertain Labels, is shown in Table 4.1. Note that the samples may be shared across different conditions due to the presence of coexisting conditions.

The uncertain samples were relabelled using a Gaussian Mixture Model (GMM) method, as detailed in Section 3.3.2. A total of five GMM models were trained one for each of the conditions (Edema, Cardiomegaly, Consolidation, Atelectasis and Pleural Effusion) using samples with only the positive labels for that condition. The number of samples used to train each GMM model is given under "Positive Labels" in table 4.1. Table 4.2 lists the number of optimal clusters used for training each condition's GMM. After relabelling, the samples were then added to the previously used training set, and the dataset was shuffled. Table 4.3 provides the label distribution across five conditions after assigning labels with GMM.

4.2.2 Models

Three distinct deep learning architectures were employed in this experiment: a 4-layer convolutional neural network (CNN) shown in Figure 4.1, a state-of-the-art DenseNet121, and a 6-layer Swin Transformer shown in Figure 3.16. This inclusion of models with varying levels of complexity and architectural

Condition	Positive Labels	Negative Labels
Edema	54,365	98,107
Cardiomegaly	25,790	126,682
Consolidation	24,395	128,077
Atelectasis	42,442	110,030
Pleural Effusion	80,859	71,612

Table 4.3: Label distribution per condition after relabelling the uncertain labels with GMM and adding them to the training set.

approaches was chosen to make sure the robustness of the observed findings. The 4-layer CNN served as a baseline to see whether a very simple CNN based model can benefit from relabelling. In contrast, the DenseNet121 was chosen because of its proved effectiveness and frequent use in chest radiograph classification, specially on CheXpert dataset [9, 27]. Lastly, to determine whether the impact of relabelling extends to more recent and powerful transformer based architectures, a Swin Transformer, which is known for its strong performance in image recognition tasks was chosen. By analysing the results across these models, this experiment sought to determine whether the effects of uncertain label handling are generalisable or model-specific.

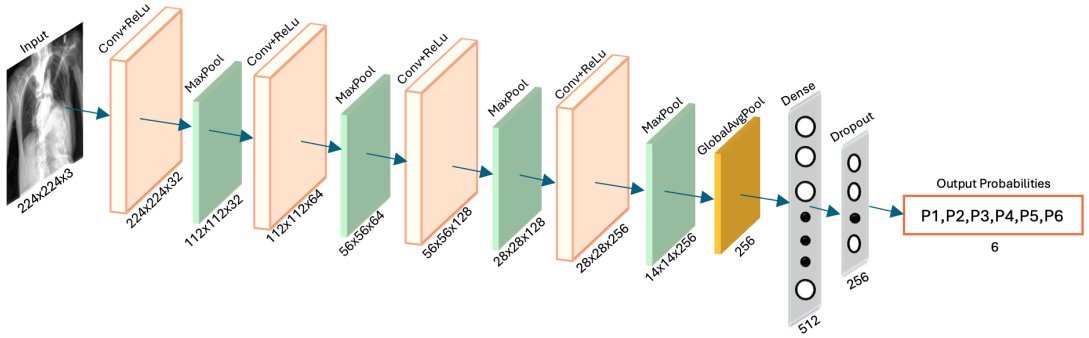


Figure 4.1: A block of 4 convolutional layers and one maxpool layer.

4.2.3 Training and Evaluation

Models were trained using the Adam optimizer across all experiments. Image data was resized to 224 x 224 pixels for uniformity. Max pooling was used for the pooling operations in the architectures. To improve generalization and boost the performance of the DenseNet121, transfer learning was employed by taking pre-trained DenseNet121 model on ImageNet dataset.

Furthermore, Data augmentation techniques, as outlined in Section 3.3.5, were applied to introduce variability. These augmentations were applied dynamically during each training epoch using the ImageDataGenerator class from Keras, so each image in the training set was presented to the model in various augmented forms over different epochs. This on-the-fly augmentation process means that for each batch, a different augmentation techniques (mentioned in 3.3.5) could be applied to the same image to make a rich and diverse set of training examples. This mechanism does not increase the number of images per epoch to keep the model efficient. Moreover, an early stopping mechanism was implemented to monitor the validation AUC for 10 epochs. This helps to prevent overfitting by stopping training if the validation performance does not improve for a consecutive 10 epochs. The weighted loss function described in Section 3.4.2 was used in all experiments. To account for potential variability, each experiment was repeated five times.

The Swin Transformer model was configured with the following hyperparameters: a patch size of (2,2), a dropout rate of 0.03, 8 attention heads, an embedding dimension of 64, an MLP layer size of 256, and query-key-value (QKV) bias set to True. An attention window size of 4 and a shift size of 1 were selected. To align with other models, the initial image size was set to 224. All other general training parameters are detailed in Section 3.2.

4.2.4 Results and Discussion

For a comprehensive grasp of the true implications of relabelling uncertain labels using GMM, each model was independently trained and tested five times. Each time, the model was trained from scratch and then tested on the same test dataset. Given that machine learning models involve stochastic processes

Models	Without Uncertain		With Uncertain	
	AUC	SD	AUC	SD
4 layer CNN	0.756 0.748 0.759 0.771 0.766	0.009	0.778 0.781 0.783 0.783 0.79	0.004
densenet121	0.79 0.791 0.796 0.798 0.798	0.003	0.81 0.81 0.81 0.81 0.811	0.0007
transformer	0.803 0.805 0.804 0.806 0.813	0.001	0.826 0.828 0.829 0.826 0.832	0.001

Table 4.4: Average Area Under the Curve (AUC) and Standard Deviation (SD) across five runs for each experiment, comparing results with and without uncertain labels.

such as random initialization of weights and random data shuffling, the results can vary slightly with each training session. By conducting five training and testing cycles, it was aimed to assess the consistency and robustness of the model's performance. Table 4.4 presents the results of each experiment with the average AUC across all conditions.

Figure 4.2 shows the plot of this experiment results. The AUC plot demonstrates the effectiveness of addressing uncertain labels through relabelling. In most disease categories, models trained with relabelled uncertain samples (models with "UN" in their name) achieved higher AUC compared to their counterparts trained without uncertain samples. This suggests that incorporating relabelled uncertain data leads to improved model performance across diverse architectures, including a simple 4-layer CNN based model, a deep DenseNet121 model, and a Swin Transformer model. The improvements, though sometimes small, are consistent across different models and conditions. Relabeling uncertain samples helps the models to learn more robust features, which leads to better classification performance.

In order to assess the effectiveness of Transformer and Transformer-UN models for assigning multiple labels to chest radiographs, ROC curves and confusion matrices were generated for each model. The ROC curve for the Transformer model, trained on a dataset without re-labeled uncertain samples, is presented in Figure 4.3. Similarly, Figure 4.4 depicts the confusion matrix corresponding to the best performing Transformer model trained on the dataset excluding uncertain samples.

The best Transformer-UN achieved better overall AUC of 0.85 on the ROC curve. This can be visualised in Figure 4.5. The improvement in performance was observed across most conditions, with the exception of edema which remained at an AUC of 0.84. This signifies a stronger ability of the Transformer-UN model to distinguish between positive and negative labels. Further evidence supporting this is provided by the confusion matrix presented in Figure 4.6. The matrix shows an increase in the number of true positives for conditions like Atelectasis and No Finding. In addition, there is a decrease in both false positives and false negatives across various categories. These improvements contribute to the overall increase in AUC observed for the Transformer-UN model.

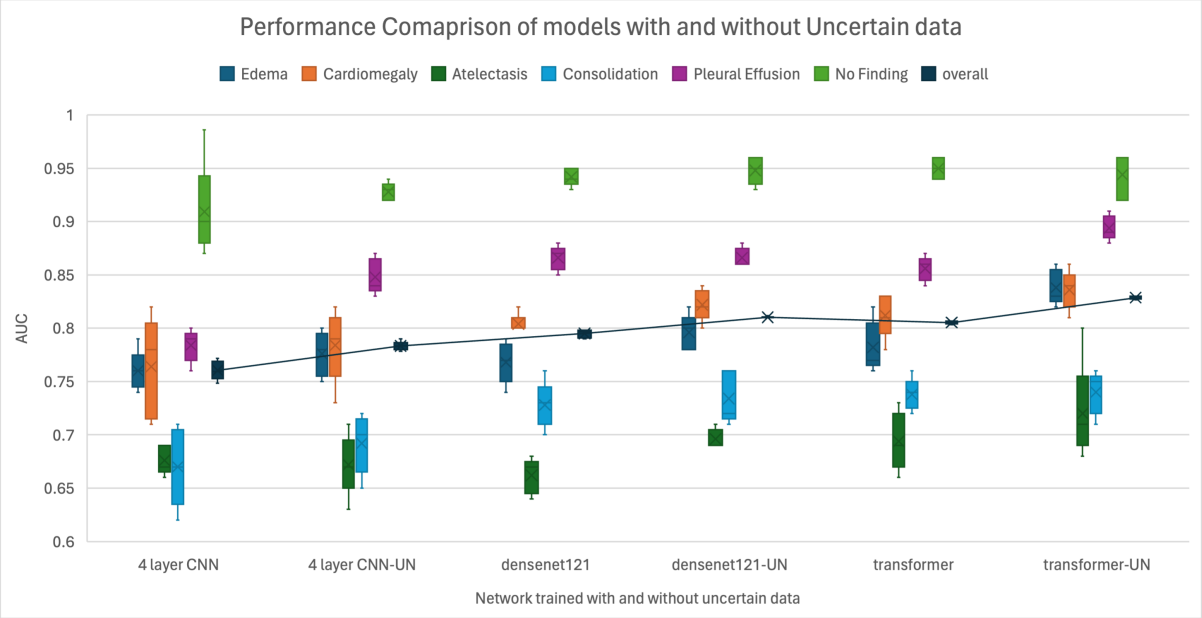


Figure 4.2: Condition wise performance comparison across all models trained with and without uncertain labels.

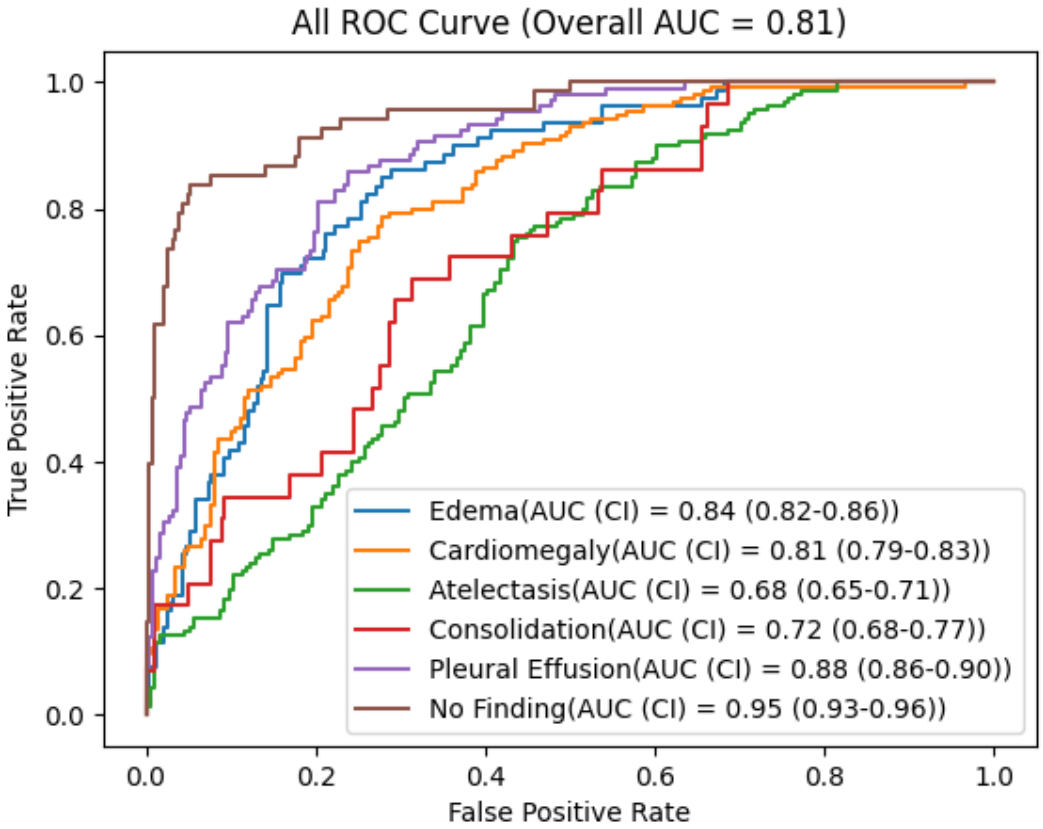


Figure 4.3: ROC curve for all conditions on the test data for Transformer model trained on dataset excluding the uncertain samples.

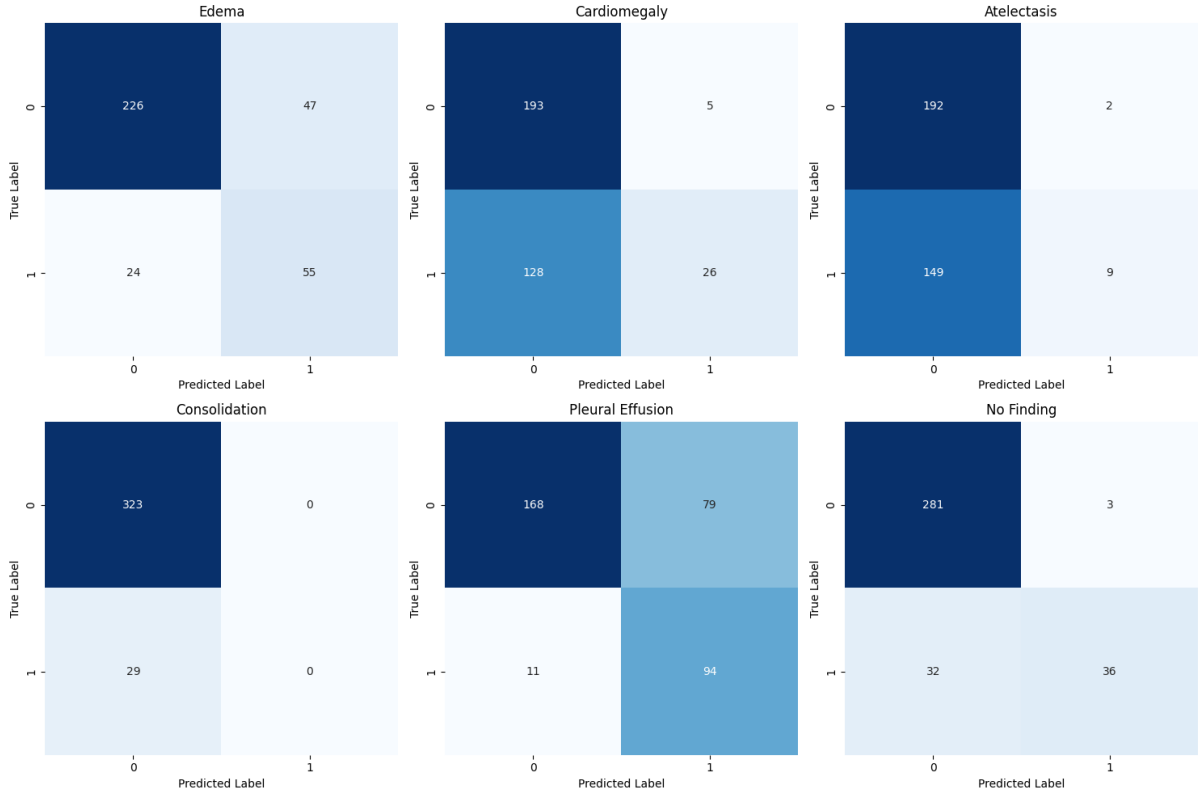


Figure 4.4: Individual confusion matrices for all conditions classified by transformer model trained on dataset with out uncertain samples.

Analysis of variance (ANOVA) found a significant difference between different techniques on AUC scores ($F(5, 24) = 139.123$, $p < 0.001$). Subsequent post-hoc comparison tests further highlight these differences. The results indicated that all pairwise comparisons among the techniques produced statistically significant mean differences in AUC scores ($p < 0.05$). In particular, the highest mean difference was observed between the "transformer-UN" and "4-Layer-CNN" techniques, with a mean difference of 0.06813 and a 95% confidence interval ranging from 0.0769 to 0.0594. This performance improvement can be attributed to the role of relabelling in reducing the negative impact of uncertain labels. By employing a Gaussian Mixture Model (GMM) for relabelling, the experiment effectively transformed these ambiguous labels into more confident predictions. These relabelled data points, integrated back into the training set to provide the models with additional informative examples. The models leveraged these relabelled samples to strengthen their internal representation of the data that leads to a more robust and accurate classification capability, particularly for categories that may have been under-represented in the original dataset.This experiment provides reliable evidence that relabelling uncertain labels with GMM offers a practical strategy to enhance model performance. It is particularly beneficial when dealing with datasets containing a significant portion of uncertain labels, and it appears to be generalisable across various deep learning architectures.

4.3 Experiment 2: Impact of Radiograph Projections

Deep learning models play an increasingly important role in analyzing chest radiographs for diagnosing thoracic pathologies. However, their performance can be influenced by factors such as the projection of the radiograph. This experiment addresses this gap by comparing the performance of deep learning models on three key radiograph projections: Anteroposterior (AP), Posteroanterior (PA), and lateral.

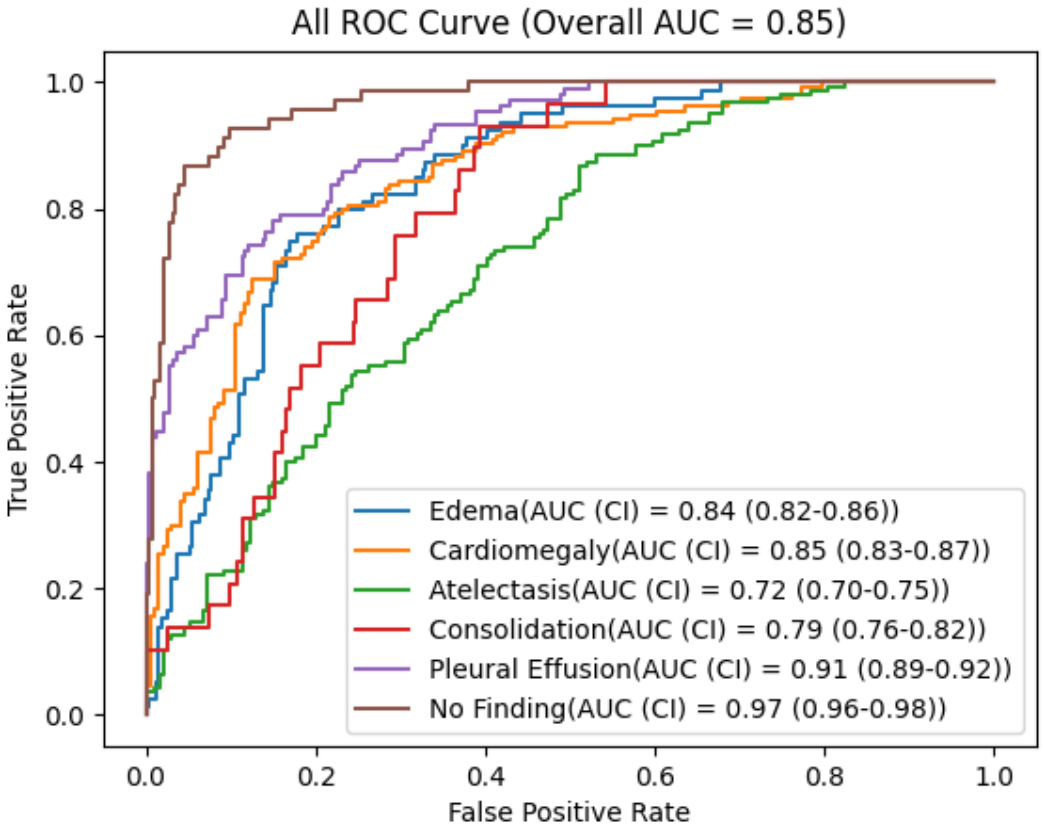


Figure 4.5: ROC curve for all conditions on the test data for Transformer model trained on dataset including the relabelled uncertain samples.

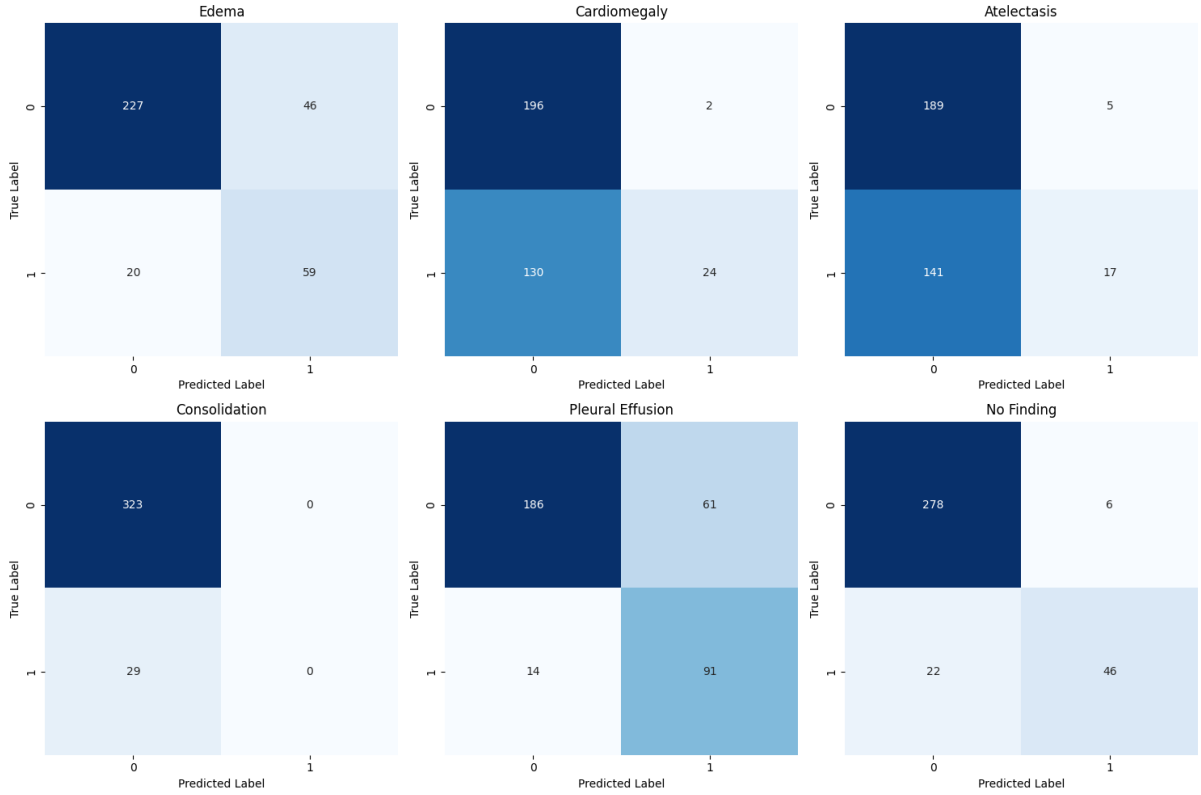


Figure 4.6: Individual confusion matrices for all conditions classified by transformer model trained on dataset including the relabelled uncertain samples.

Chest radiographs are routinely acquired in multiple projections to provide a comprehensive view of the thoracic anatomy. The PA projection, where the X-ray beam travels from back to front of the patient, is considered the gold standard due to its superior visualization of the lung fields and heart size. In contrast, the AP projection, obtained with the X-ray source positioned in front of the patient, may be used when patients are unable to stand for a PA view. However, the AP projection can magnify the heart size and obscure portions of the lung due to superimposed structures. The lateral projection, where the X-ray beam travels from side to side, offers a unique perspective on the chest wall, aorta, and posterior mediastinum, which can be helpful in diagnosing specific conditions like pleural effusions or masses.

This comparative analysis of deep learning model performance across AP, PA, and lateral projections aims to identify potential projection-specific biases, which contribute to the development of robust models that can effectively analyze radiographs regardless of projection. Furthermore, the findings can guide radiologists in selecting the most appropriate projection and highlight the potential benefits of incorporating deep learning models into existing radiograph interpretation workflows.

4.3.1 Preprocessing

The CheXpert dataset was used, and separate datasets were created for the AP, PA, and lateral groups. This was done on the basis of the AP/PA label in the training set CSV file. The creation of separate subsets of the data ensured that the models were trained and evaluated on projection-specific data. Data augmentation was also applied at the time of training. Similar to experiment 1, the augmentation techniques listed in section 3.3.5 were applied dynamically during each training epoch using the ImageDataGenerator class from Keras, to present the training data to the model in various augmented forms over different epochs. Table 4.5 shows the number of training samples for each projection. As the number of samples for each projection was not equal. The AP dataset contained significantly more samples compared to the PA and

Projection	Number of Samples
Posteroanterior (PA)	29,420
Anteroposterior (AP)	161,590
Lateral	32,387

Table 4.5: Number of training samples in CheXpert dataset corresponding to each projection.

lateral. To ensure a fair comparison between the models, AP and lateral training sets were down-sampled to match the size of the PA dataset.

4.3.2 Models

Three separate DenseNet121 models were fine-tuned, one for each projection (AP, PA, and lateral). The selection of DenseNet121 was based on its superior performance on chest radiograph classifications tasks and because of the fact that multiple previous studies have considered it as the base network for CheXpert classification [9, 27]. Each model was trained with a batch size of 32 for 100 epochs. An early stopping mechanism was implemented to monitor the validation AUC for 10 epochs. This helps to prevent overfitting by stopping training if the validation performance does not improve for a consecutive 10 epochs.

To ensure accurate loss calculation during training, the weighted loss function described in Section 3.4.2 was used in all experiments. To account for potential variability coming from several factors, each experiment was repeated 10 times. These factors include variability introduced by down-sampling the dataset, the random nature of data augmentation techniques (such as rotation, zoom, shear, and horizontal flips), the randomness in the initialization of model weights and the order in which the training data is presented to the model due to shuffling of the data. This allows for a more robust evaluation of the model's performance on each projection.

4.3.3 Training and Evaluation

Following a similar approach to experiment 1, all models across the various experiments were trained using the Adam optimizer. To ensure consistency in the training data, all images were resized to a uniform size of 224 x 224 pixels. To leverage pre-existing knowledge and enhance the performance of the DenseNet121 model, transfer learning was employed. This technique utilises the weights learnt by a pre-trained DenseNet121 model on the vast ImageNet dataset.

Furthermore, data augmentation techniques, as detailed in Section 3.3.5, were applied to artificially introduce variations in the training data. This way the data mimicks the real-world scenarios and prevent the model from overfitting to the specific training set. To monitor for overfitting, an early stopping mechanism was implemented. This mechanism tracks the validation AUC (Area Under the Curve) metric and halts training if there's no improvement in validation performance for 10 consecutive epochs. Additionally, to account for potential variability and ensure robust results, each experiment was repeated ten times. The weighted loss function, as described in Section 3.4.2, was consistently used throughout all experiments.

4.3.4 Results and Discussion

By training a state-of-the-art deep learning algorithm DenseNet121 on different views of chest radiographs, AP and Lateral views performed comparably better than PA. Table 4.6 shows the detail result of each experiment for each projection. The AUC is averaged over all conditions. Figure 4.7 shows the

Exp	Anteroposterior(AP)	Posteroanterior(PA)	Lateral
1	0.87	0.72	0.78
2	0.65	0.73	**0.85**
3	0.87	0.73	0.84
4	0.89	**0.75**	0.82
5	**0.91**	0.73	0.77
6	0.9	0.75	0.82
7	0.85	0.72	0.83
8	0.88	0.72	0.81
9	0.89	0.73	0.83
10	0.88	0.73	0.83
Avg	0.86	0.73	0.82
SD	0.08	0.01	0.03

Table 4.6: All experiments result for AP, PA and Lateral views. The highest AUC is in bold for each projection.

results plot. This is an interesting fact, keeping in view that PA radiographs are more commonly performed in the outpatient setting, while AP radiographs are performed in patients who are ill or unstable and therefore unable to cooperate for a PA view.

Given the violation of normality assumptions in the data, as evidenced by the Kolmogorov-Smirnov and Shapiro-Wilk tests ($p < .001$ for AP and PA models, $p < .01$ for La model), a Kruskal-Wallis H test was employed to compare AUC values across the three projection models (AP, PA, and La). This non-parametric test revealed a statistically significant difference in AUC values among the models ($\chi^2(2) = 18.892$, $p < .001$). This finding suggests a difference in the distribution of AUC values between at least two of the projection models. Subsequent post-hoc pairwise comparisons, conducted using Dunn's test with Bonferroni correction, revealed that the PA model differed significantly from both the La model (adjusted $p = .031$) and the AP model (adjusted $p = .000$). However, no significant difference was detected between the La and AP models (adjusted $p = .236$).

These findings highlight the potential benefit of identifying a condition by a view-specialised model because it can be more reliable than a general model trained on all types of views. As the frontal and lateral views of the chest look different,important features of a frontal image can manifest differently on a lateral view resulting in uncertainty for interpretation both by the radiologist and the machine learning algorithm.

4.4 Summary

This Chapter addresses the challenge of uncertain labels and the impact of different radiograph projections in chest radiograph classification performance. The problem starts from the inherent uncertainty in labels extracted from radiograph reports using natural language processing. Two key techniques are explored and compared their performance, first, excluding uncertain samples and second, relabeling them using a Gaussian Mixture Model (GMM). The results indicate that incorporating relabeled uncertain samples significantly improves model performance across various architectures. Therefore, the hypothesis **Hyp 01**, that the GMM-based uncertain label handling on CheXpert dataset will significantly improve the classification performance is accepted. Moreover, this Chapter also explored the variability introduced by different radiograph views by comparing the performance of deep learning models trained on distinct radiograph projections (PA, AP, and lateral). The results showed, models trained on AP and lateral views outperform those trained on PA view, suggesting that certain projections offer better diagnostic

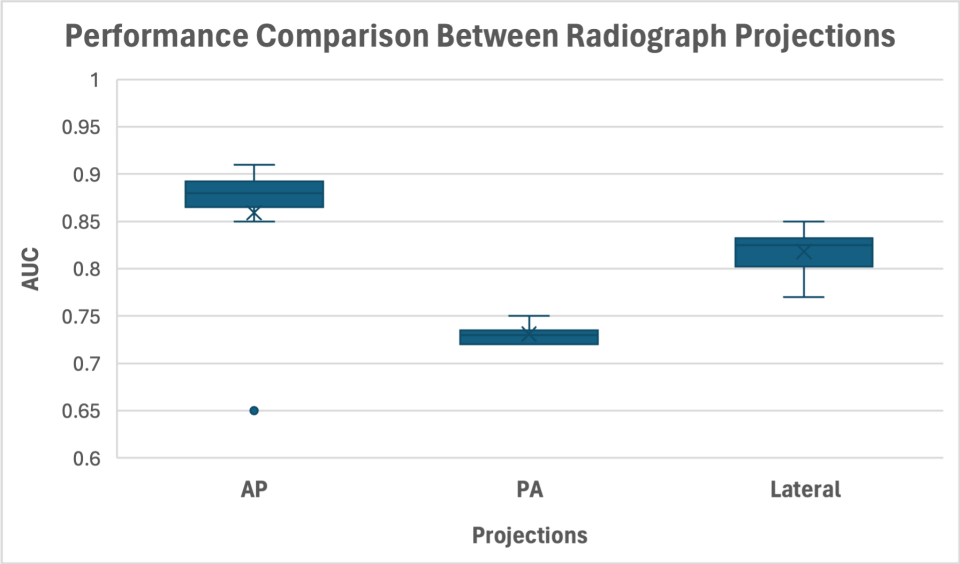


Figure 4.7: Performance comparison of models trained on different radiograph projections.

capability.

Chapter 5

Efficacy of Custom Pooling Layer

5.1 Introduction

In Section 3.4.3 of this thesis, a novel custom pooling layer was introduced that merges the strengths of both max pooling and min pooling. It explained the inherent shortcomings of the max pooling and its opposite min pooling. This Chapter delves into the experimental evaluation of this custom pooling layer, specifically assessing its efficacy in chest radiograph classification tasks in a multi label setting. The experiments involve training convolutional neural networks (CNNs). This is because pooling layers are more prevalent in CNN based networks.

The core objective lies in comparing the performance of the custom pooling layer against the standard pooling operations. The hypothesis is that the custom pooling layer's ability to capture a broader spectrum of features which encompass both high and low intensity regions, will lead to improved model performance in chest radiograph classification. To achieve this, a series of experiments were conducted using both standard pooling and custom pooling separately. To accurately measure the performance difference between the two approaches, all other parameters such as training, validation and test datasets, model architecture (except pooling layers), number of epochs, learning rate etc, were kept the same.

Specifically, this Chapter explores the performance of two convolutional neural network models on distinct datasets. Experiment 1 leverages the standard CheXpert dataset for training, validation, and testing. In contrast, experiment 2 trains on a multi-scale template-matched version of the CheXpert dataset, detailed in Section 3.3.4. Within each experiment, two models are compared: a baseline DenseNet121 and its improvised version using custom pooling.

5.1.1 Hypothesis: Hyp 02

The hypothesis for this Chapter is as follows:

The integration of a custom pooling layer in CNNs significantly enhances the performance of chest radiograph classification by effectively capturing both high and low-intensity features, compared to standard pooling methods.

5.2 Experiment 1: With Standard CheXpert Dataset

5.2.1 Data Preprocessing

This experiment investigated the performance of two distinct models, a baseline and an improvised one, trained on the standard CheXpert dataset. To avoid conflict in anatomical features, only the frontal view radiographs were utilised. Extracting these frontal views involved filtering the CheXpert label CSV files

based on the "AP/PA" column, to prepare training, validation, and test sets. The filter was set to choose only those samples with value either 'AP' or 'PA' leaving out any other value. As the CheXpert dataset contains images with varying dimensions, a pre-processing step was important. Therefore, all images were resized to a uniform size of 224 x 224 pixels for a consistent dataset.

Moreover, details on additional preprocessing techniques employed in this experiment can be found in Section 3.3.1 of this thesis. For consistency, all other hyperparameters (learning rate, optimizer, etc.) and the software/hardware specifications used in this experiment remained identical to those described in Section 3.2. This ensures that the observed performance differences can be attributed solely to the usage of custom pooling layers.

5.2.2 Baseline Model with Max Pooling

To compare the performance of the models with and without the custom pooling layer, DenseNet121 was chosen as the baseline due to its superior performance [27]. The chosen baseline model, DenseNet121, is a convolutional neural network architecture known for its efficiency and strong performance in image classification tasks. It utilizes a unique approach where feature maps are reused and connected across layers with in each dense block, which promotes feature propagation and reduce the effect of vanishing gradient compared to traditional CNNs. To adopt the model for this task, the final fully connected layer was replaced with a six-node dense layer followed by a sigmoid activation function. This modification was made to adapt the network for multi-label classification, where each node corresponds to one of the five conditions: Edema, Cardiomegaly, Atelectasis, Consolidation, Pleural Effusion, and No Finding. The sigmoid activation function is applied to each of these nodes to produce an output probability between 0 and 1 for each condition which indicates the likelihood of the presence of that condition.

First, experiments were conducted with the standard DenseNet121 architecture pre-trained on ImageNet. It is made up of four dense blocks, each containing a series of 1x1 and 3x3 convolutional layers with a growth rate of k (where each layer adds k new feature maps). Batch normalisation and rectified linear unit (ReLU) activations are employed throughout the network to improve training stability and introduce nonlinearity. Together, there are 121 layers in DenseNet121. Figure 5.1(a) shows the diagram of the standard DenseNet121 architecture.

5.2.3 Model with Custom Pooling

Following the preparation of the baseline model. In order to evaluate the impact of the custom pooling layer. A separate Imagenet pre-trained DenseNet121 model was used. In a careful process, all the pooling layers within this pre-trained model were replaced with the custom pooling layer (as detailed in Section 3.4.3 of this thesis). In this way, only the pooling layers were modified, and the integrity of the remaining architecture was preserved. Similar to the base model, the last fully connected layer was replaced with the 6 node dense layer followed by a sigmoid layer.Consequently, this careful modification allowed for a direct comparison of the influence of the custom-pooling layer on the model's performance. A visual representation of the modified DenseNet121 architecture, which incorporates the custom pooling layer, is presented in Figure 5.1(b). The python code for integrating the custom pool layer in DenseNet121 is provided in Appendix C (Custom Pooling).

5.2.4 Training and Evaluation

The standard CheXpert dataset served as the basis for these experiments. The radiographs in this dataset contain irrelevant thoracic regions. This can disturb the focus and hinder the model's performance. The details of the implications of the presence of non-thoracic areas in the radiographs are explained in Section 3.3.4 of this document. The dataset was divided into training, validation, and test sets according

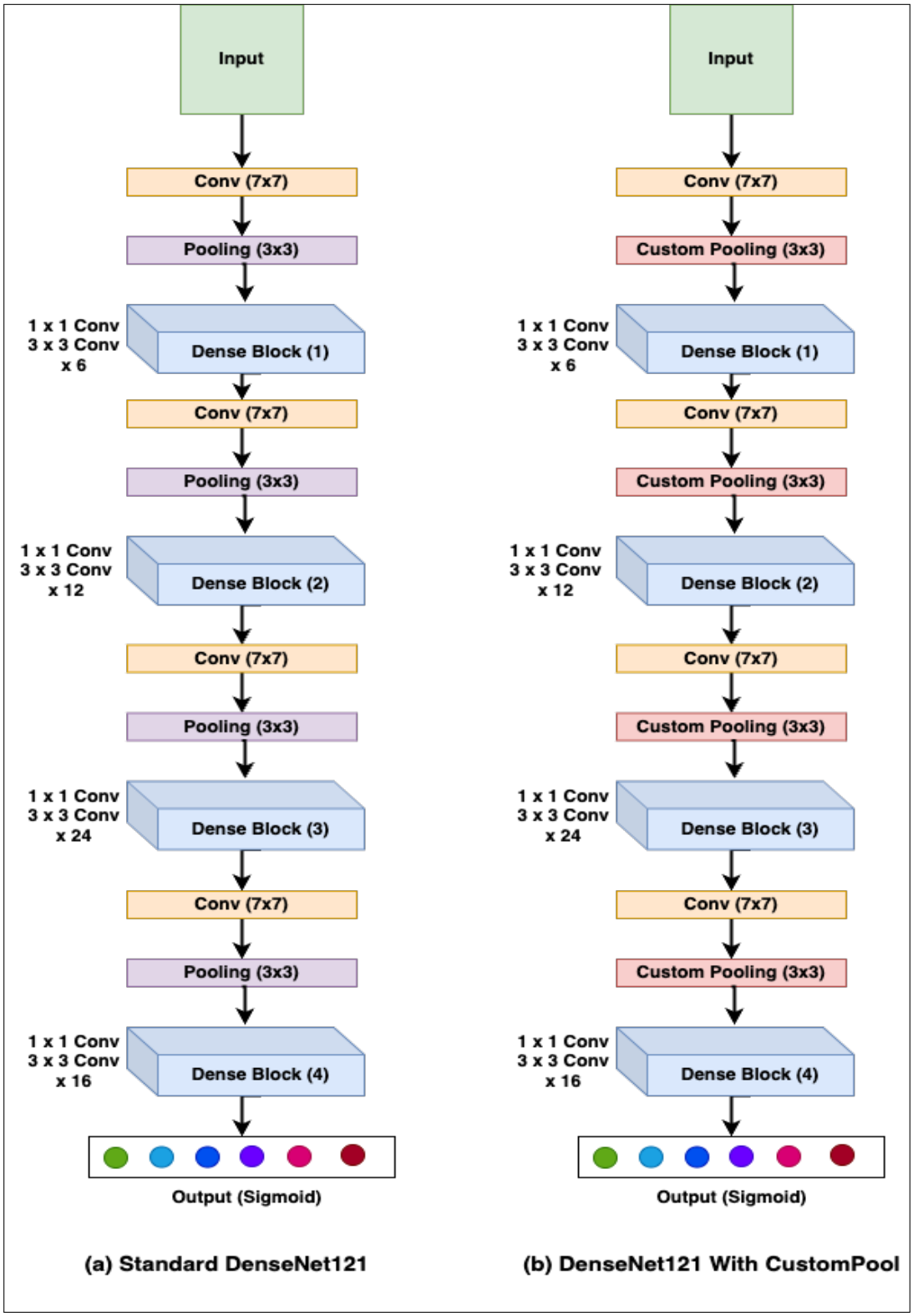


Figure 5.1: DenseNet-121 architectures before and after replacing the pooling layers. The custom pooling layers use both max-pooling and min-pooling by concatenating their resulting feature maps, which doubles the channel dimensions of the custom pooling layer output compared to the standard DenseNet-121 architecture.

Exp	Baseline Model	Custom Pool Model
1	0.812	0.868
2	**0.813**	**0.874**
3	0.811	0.871
4	0.812	0.870
5	0.813	0.871
Avg	0.812	0.870

Table 5.1: Overall AUC for all five conditions on DenseNet121 with max pool and with custom pool.

to the CheXpert guideline. The validation set was used to check models performance during the training after each epoch, while the test set was used to test the final trained model. The uncertain samples re-labelled, described in Section 3.3.2 were incorporated into the training set. This increased the number of samples and helped the model to expand its understanding. Furthermore, data augmentation techniques, detailed in Section 3.3.5 was applied to the training set. It helps to artificially expand the dataset and improve the model's ability to generalise to unseen data. Both the base model and the model with custom pooling integration explained above were trained for 100 epochs with early stopping. Each of the model was trained and tested five times to obtain a generalisable result.

5.2.5 Results and Discussion

As explained above, this experiment is performed using the standard CheXpert dataset. After training the baseline model, it was tested on the CheXpert test set, which contains more than 500 frontal chest radiographs carefully annotated by the eight board certified radiologists. The final label was determined by the majority vote of the five radiologists. This creates a strong and reliable benchmark for assessing the performance of classification models. Both models were trained and tested five times for better genralizability of the results. Table 5.1 shows the average AUC across all conditions of all experiments.

The result shows a significant improvement in the overall AUC of the model with custom pooling layers. Data normality was assessed using both the Kolmogorov-Smirnov and Shapiro-Wilk tests [129, 130]. The Kolmogorov-Smirnov test yielded p-values of 0.200 for both the Baseline and Custompool models which indicates that there is no significant deviation from normality. Similarly, the Shapiro-Wilk test showed p-values of 0.084 for the Baseline model and 0.168 for the Custompool model, further confirming the normality assumption. With normality confirmed, an independent samples t-test was conducted to compare the AUC between the two groups. Levene's test for equality of variances indicated no significant difference in variances ($F(1, 58) = 0.826$, $p = 0.367$), allowing for the assumption of equal variances. The t-test results showed a significant difference between the group means ($t(58) = -2.918$, $p = 0.005$), with a mean difference of -0.0597 (SE = 0.0204). The 95% confidence interval for the difference ranged from -0.1006 to -0.0187. These results demonstrate a statistically significant difference in AUC between the Baseline and Custompool models, supporting the hypothesis that the groups differ in this measure. Figure 5.2 shows the average AUC of all conditions in five experiments for both models.

Upon closer examination of the results plot, the baseline model demonstrates the lowest AUC of 0.69 for 'Atelectasis' and the highest AUC of 0.96 for 'No Finding'. This translates to a difference of 0.27 between the lowest and highest AUCs. Conversely, the custom pool model exhibits a lower AUC of 0.75 for 'Consolidation' and a higher AUC of 0.99 for 'No Finding'. The difference here is reduced to 0.24 which indicates a positively narrower range of AUCs across conditions. In addition to that, it is important to note that the all conditions got improvement. Considering the highest AUC's in each condition for both models, Edema and Cardiomegaly 7%, Atelectasis 10%, Consolidation 9%, Pleural Effusion 6% and No Finding got 3% AUC improvemnet on custom pooling model. Table 5.2 gives the condition wise detailed results of all experiments.

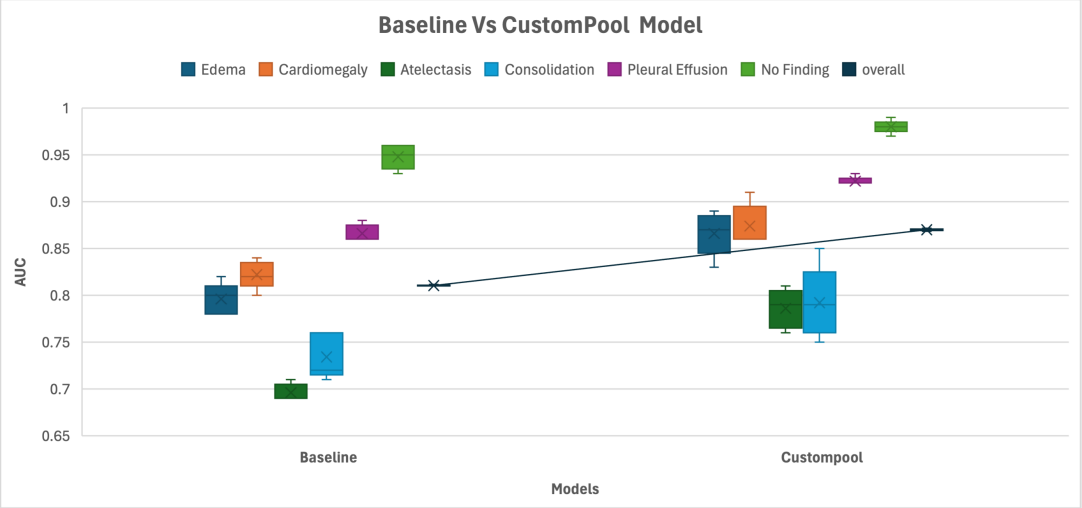


Figure 5.2: Condition wise performance comparison of baseline and custom pool models.

Furthermore, a close inspection reveals that 'Edema' and 'Cardiomegaly' share similar AUCs, as for 'Atelectasis' and 'Consolidation'. This observation might be attributed to two factors. Firstly, edema can be a consequence of cardiomegaly. Secondly, the manifestation of features for atelectasis and consolidation can appear visually similar on radiographs. This similarity could potentially contribute to lower AUCs, particularly for atelectasis and consolidation, due to the high degree of confusion between the two conditions.

To better visualise the results of this experiment. The ROC curve and the confusion matrix were plotted. Figure 5.3 and Figure 5.4 shows the ROC curves and confusion matrix for all conditions for the baseline and custom pool models respectively.

Furthermore, the confusion matrix was also plotted for both models. The confusion matrix shows the number of True positives, True Negatives, False positives and False negatives for each condition. Figure 5.5 and Figure 5.6 shows the confusion matrix for the baseline and custompool models respectively.

5.3 Experiment 2: With Multi-Scale Template Matched Dataset

This experiment is almost the replication of Experiment 1 with one exception. That is employing a specifically tailored dataset. The dataset described in detail in Section 3.3.4. This choice aimed to demonstrate two key points. Firstly, it showcases the potential for the custom pooling model to achieve even higher performance when trained with a well-preprocessed dataset. Secondly, it shows the effectiveness of the multi-scale template matching technique in data preparation.

5.3.1 Data Preprocessing

Building upon the preprocessing steps outlined in Experiment 1 and Section 3.3.1, Experiment 2 incorporates an additional preprocessing stage utilizing multi-scaled template matching. This technique aims to remove non-thoracic regions from the radiograph images within the CheXpert dataset (as detailed in Section 3.3.4). By eliminating these extraneous areas, the model focuses solely on the thorax and potentially improve the feature extraction specific to thoracic pathologies.

The multi-scaled template matching process involves applying templates of the thoracic region at various scales to the radiographs. Regions with high similarity to the templates were identified as

	Baseline Model	**Custom Pool Model**
Edema	0.8	0.86
	0.78	0.88
	0.8	0.83
	0.82	**0.89**
	0.78	0.87
Cardeomegaly	0.82	**0.91**
	0.83	0.86
	0.84	0.88
	0.8	0.86
	0.82	0.86
Atelectasis	**0.71**	0.79
	0.7	0.76
	0.69	0.77
	0.69	0.8
	0.69	**0.81**
Consolidation	0.71	0.75
	0.76	0.8
	0.72	**0.85**
	0.72	0.77
	0.76	0.79
Pleural Effusion	0.88	0.92
	0.86	**0.93**
	0.86	0.92
	0.87	0.92
	0.86	0.92
No Finding	0.94	0.98
	0.93	**0.99**
	0.95	0.97
	0.96	0.98
	0.96	0.98

Table 5.2: AUC achieved by each condition in each experiment for models. The highest achieved AUC's for both baseline and custom pool models are in bold.

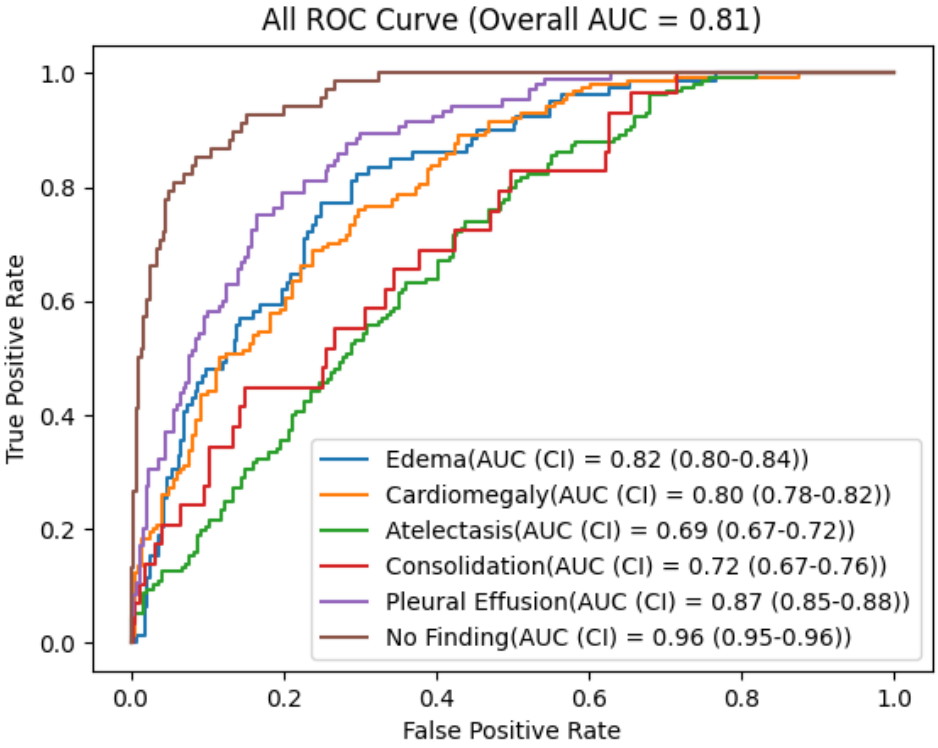


Figure 5.3: ROC curve for all conditions on the test data for baseline model. AUC and confidence intervals are mentioned for each condition.

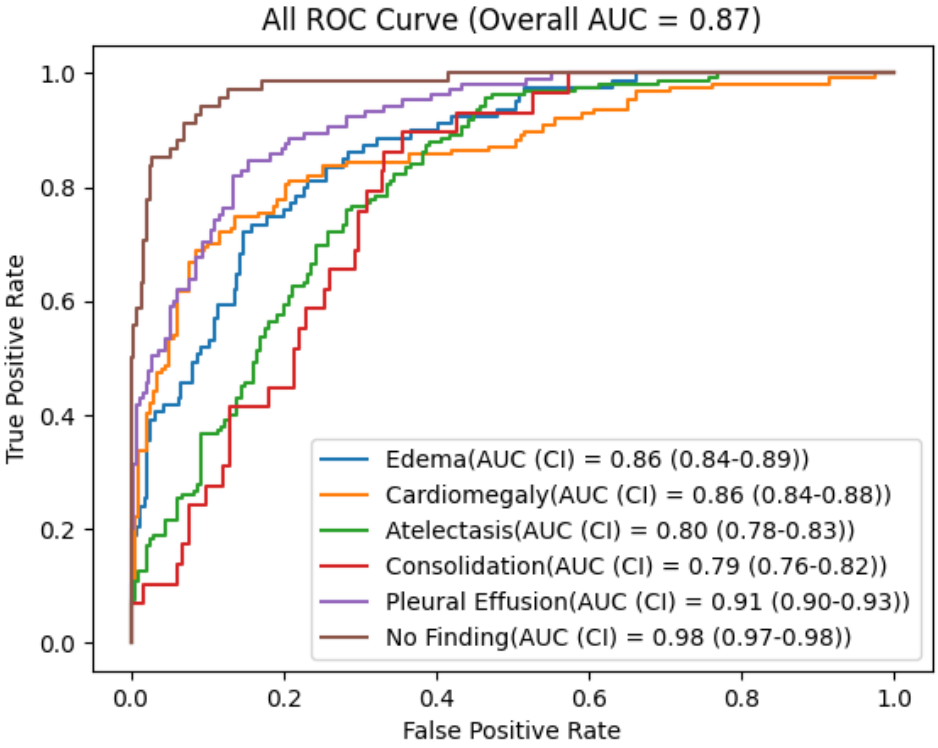


Figure 5.4: ROC curve for all conditions on the test data for custompool model. AUC and confidence intervals are mentioned for each condition.

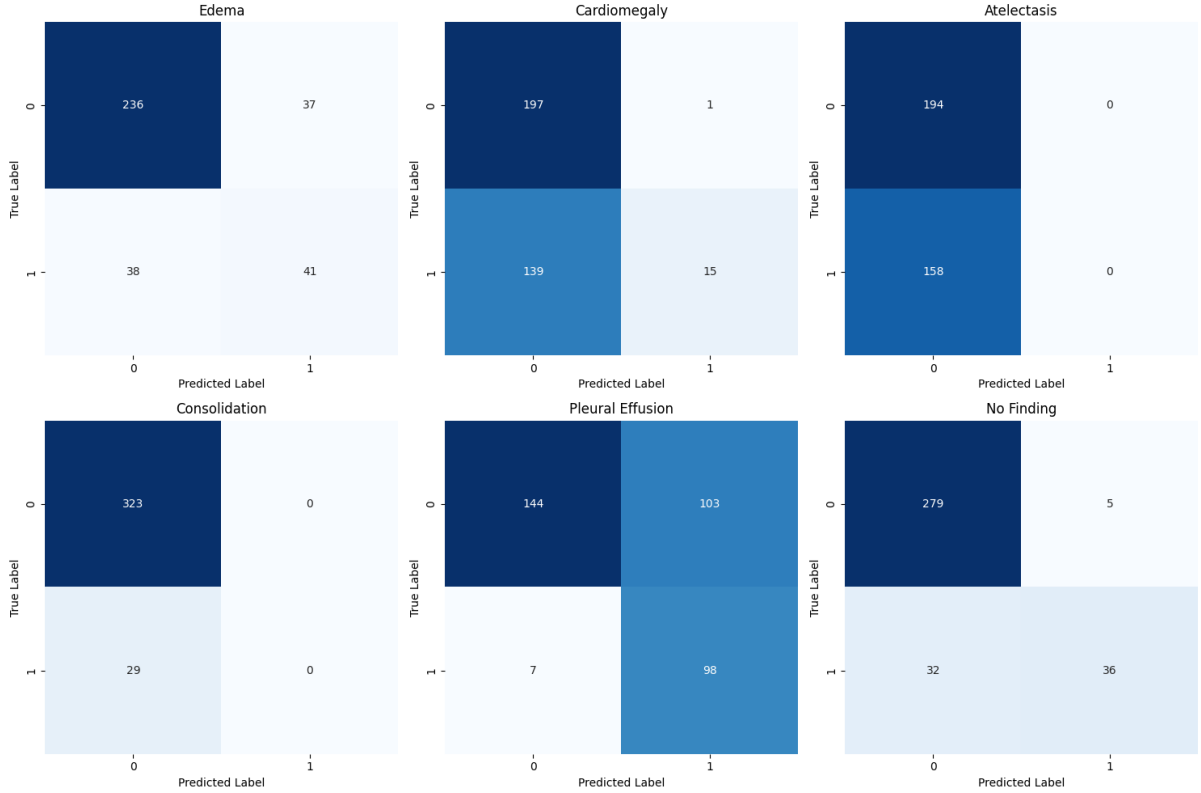


Figure 5.5: Individual confusion matrices for all conditions classified by baseline model.

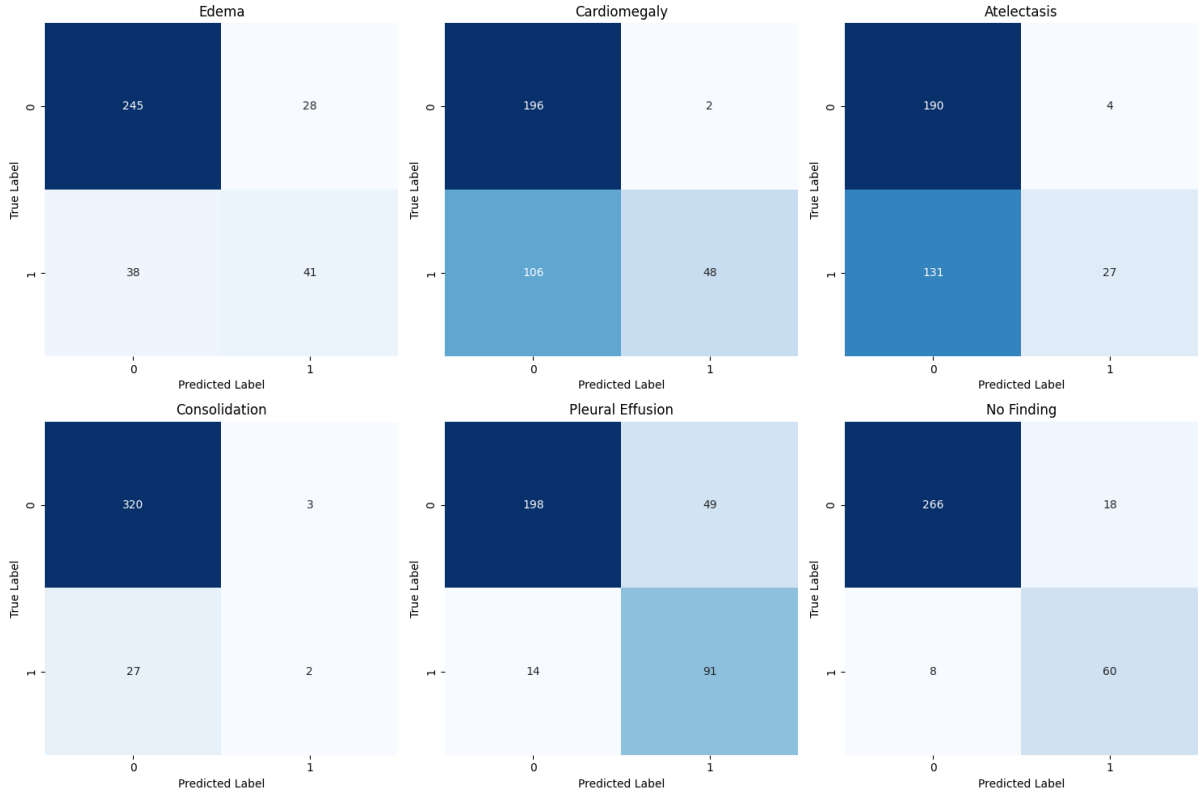


Figure 5.6: Individual confusion matrices for all conditions classified by custom pool model.

thoracic, while those with low similarity were classified as non-thoracic. These non-thoracic regions were then cropped out, resulting in images containing only the thorax. This refined dataset was then fed into the model for training, validation, and testing. By focusing on the anatomical region of interest, the model could potentially learn more discriminative features relevant to the targeted pathologies within the CheXpert dataset.

5.3.2 Baseline Model

Similar to experiment 1, DenseNet121 was chosen as the baseline model for training on the TM dataset due to its well-established success in image classification tasks. Unlike traditional Convolutional Neural Networks (CNNs) that process information in a sequential manner, DenseNet121 incorporates a unique dense block architecture that promotes efficient feature utilization. Within these dense blocks, each layer is directly connected to all subsequent layers, which increases the chances of better feature preservation throughout the architecture. This dense connectivity effectively addresses the vanishing gradient problem which is a common challenge in deep neural networks, where gradients from earlier layers can diminish as they propagate through the network. To tailor DenseNet121 for this specific task, the final fully connected layers were replaced with a new dense layer containing six nodes and a sigmoid activation function.

The first phase of the experiment leveraged the standard DenseNet121 architecture pre-trained on the massive ImageNet dataset. This pre-training process equips the model with a strong foundation for recognizing general image features, which can then be fine-tuned for the specific classification task at hand. The DenseNet121 architecture is comprised of four distinct dense blocks. Each dense block consists of a sequence of convolutional layers with varying filter sizes. These layers are typically configured with a 1x1 kernel for dimensionality reduction and a 3x3 kernel for capturing spatial information within the images. As the network progresses through these dense blocks, the number of feature maps increases which allows the model to extract complex features from the input data.

Moreover, batch normalization layers are placed throughout the network to address internal covariate shift which is a phenomenon that can hinder training stability. Additionally, ReLU (Rectified Linear Unit) activation functions are employed after each convolutional layer to introduce non-linearity into the network which enables the model to learn more complex relationships within the data. In total, the DenseNet121 architecture has 121 layers.

5.3.3 Model with Custom Pooling

The custom pooling model for experiment 2 is similar to the one prepared for experiment 1. The idea is to isolate the impact of the custom pooling layers. To achieve this, a separate DenseNet121, pre-trained on ImageNet, was used. The difference between the baseline and the custom pooling model lies in the pooling layers. In the baseline model, standard pooling layers were used. Here, all pooling layers were replaced with the custom pooling layers detailed in Section 3.4.3 of this thesis. While implementing the custom pooling model, only the pooling mechanism was altered, while the rest of the architecture, including the convolutional layers, remained unchanged.

Finally, mirroring experiment 1, the last fully connected layer was replaced with a new six-node dense layer followed by a sigmoid activation. This modification allows for a direct comparison of the custom pooling layer's influence on the model's performance. Figure 5.1(b) shows the custom pool architecture.

5.3.4 Training and Evaluation

This experiment leverages a template-matched version of the CheXpert dataset. This variation ensures that all radiographs only contains the thoracic region by eliminating the irrelevant anatomical areas that

Exp	Baseline Model	Custom Pool Model
1	0.838	0.883
2	0.838	0.886
3	0.841	0.888
4	**0.843**	0.890
5	0.843	**0.896**
Avg	0.840	0.888

Table 5.3: Baseline vs custom pool model's overall AUC on template matched dataset.

could potentially hinder the model's performance. The preprocessing explained in Section 3.3.4 was applied on standard CheXpert dataset. By utilizing a template-matched dataset, the model will be more focused to the region of interest (thoracic region) and potentially improve the classification performance.

Following the data split for training, validation, and testing, the training of the model was initiated. Data augmentation techniques, as described in Section 3.3.5, were applied to both the training and validation sets to artificially expand the dataset and enhance the model's ability to generalize to unseen data. The model undergo training for a 100 epochs with early stopping to prevent overfitting. After training and validation, the model was tested on the chexpert test set, which was preprocessed through multi scale template matching. This training and evaluation process was repeated five times to get generalizable results.

5.3.5 Results and Discussion

As explained above, this experiment is performed using the template matched CheXpert dataset. After training the baseline model, it was tested on the CheXpert test set, which contains more than 500 frontal chest radiographs carefully annotated by the eight board certified radiologists. The final label was determined by the majority vote of the five radiologists. This creates a strong and reliable benchmark for assessing the performance of classification models. Both models were trained and tested five times for better genralizability of the results. Table 5.3 shows the average AUC across all conditions for all experiments.

In line with experiment 1 results, experiment 2 confirms a significant improvement in the model's overall AUC. The normality of the data was assessed using the Kolmogorov-Smirnov test for both models. Non-significant p-values of 0.200 for each model suggested that the data met the assumption of normality. An independent sample t-test was then conducted to compare the means of the AUC variable between the groups. Levene's test for the equality of variances revealed no significant difference ($F(1, 58) = 1.888$, $p = 0.175$) and supported the assumption of equal variances. The t-test results indicated a statistically significant difference between the group means ($t(58) = -2.598$, $p = 0.012$), with a mean difference of -0.0497 (SE = 00191). The 95% confidence interval for the difference ranged from -0.0879 to -0.0114. These findings suggest a statistically significant difference in AUC between the two models, which supports the hypothesis that the groups differ in this measure.

These results indicate, both the baseline and custom pooling models achieved higher AUCs compared to the best performing models on the standard CheXpert dataset. The baseline model reached a maximum AUC of 0.843 (0.030 improvement), while the custom pooling model achieved a maximum of 0.896 (0.024 improvement). This indicates that the template-matched data itself benefited both models, with the custom pooling layer providing an additional boost. Figure 5.7 illustrates the average AUC for both models across all conditions in the five experiments.

The overall AUC improved. However, the benefit wasn't uniform across all conditions. In the

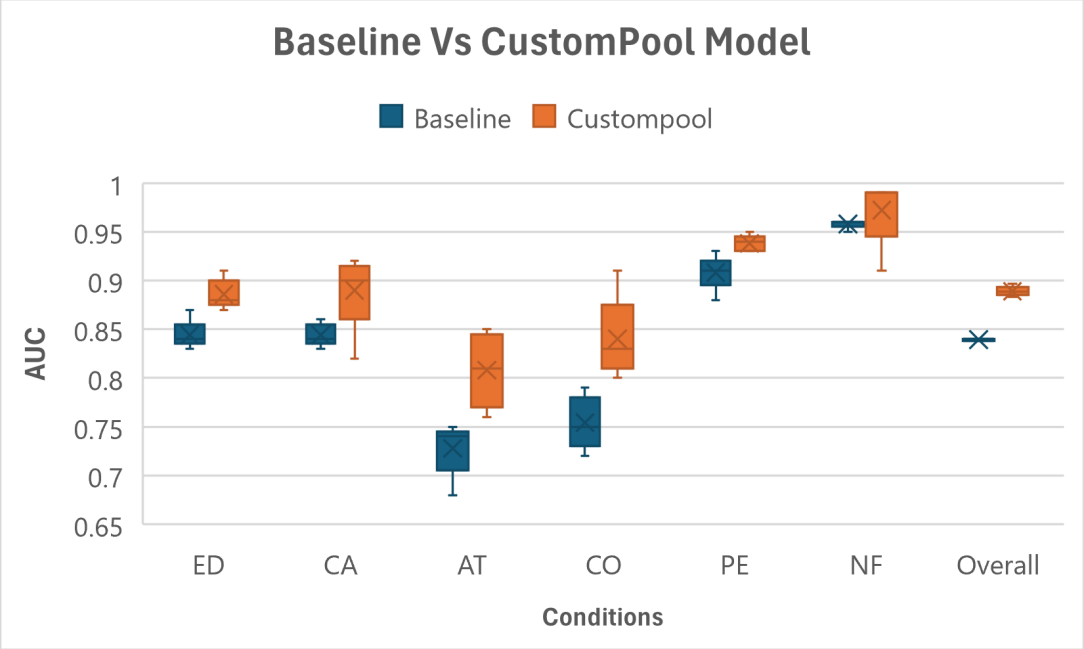


Figure 5.7: Condition wise performance comparison of baseline and custom pool models on TM dataset.

baseline model, the difference between the best and worst performing condition was 0.28, which is higher compared to the 0.27 observed in Experiment 1. This suggests that template-matched data doesn't improve all conditions equally. Which makes sense because the cropping process in multi scale template matching can alter the orientation of features.

The custom pool model presents a different picture. Here, the difference between the highest and lowest performing conditions is slightly lower (0.23) compared to Experiment 1 (0.24). Interestingly, the custom pool model's lowest AUC is associated with Atelectasis, whereas Experiment 1 found the lowest performance for Consolidation. This difference highlights that the custom pooling method might be less susceptible to feature orientation changes compared to the baseline model. Table 5.4 shows the per condition AUC achieved in each experiment of baseline and custom pool model.

Furthermore, to provide a more comprehensive visual interpretation of the experimental results, ROC curves and confusion matrices were generated. These graphical representations are employed to depict the performance of both the baseline and custom pooling models for all conditions. For ease of comparison, Figures 5.8 and 5.9 present the ROC curves and confusion matrices for each model, respectively.

To completment the ROC curve, confusion matrices were generated for both the baseline and custom pooling models. These matrices provide a detailed breakdown of classification performance for each condition. Specifically, confusion matrices illustrate the distribution of True Positives, True Negatives, False Positives, and False Negatives . Figures 5.10 and 5.11 present the confusion matrices for the baseline and custom pooling models, respectively.

5.4 Summary

In summary, this Chapter explores the limitations of standard pooling operations in CNNs for chest radiograph classification and introduces a custom pooling layer that combines max and min pooling to capture a wider range of features. The problem lies in the inability of traditional pooling methods

	Baseline Model	**Custom Pool Model**
Edema	0.84	**0.91**
	0.84	0.88
	0.83	0.89
	0.84	0.87
	0.87	0.88
Cardeomegaly	**0.86**	0.91
	0.84	0.9
	0.85	0.82
	0.84	0.9
	0.83	**0.92**
Atelectasis	0.74	0.76
	0.73	0.78
	0.74	**0.85**
	0.75	0.81
	0.68	0.84
Consolidation	0.72	0.8
	0.79	0.84
	0.75	**0.91**
	0.74	0.83
	0.77	0.82
Pleural Effusion	0.91	0.93
	0.88	0.94
	0.91	**0.95**
	0.91	0.94
	0.93	0.93
No Finding	**0.96**	**0.99**
	0.95	0.98
	0.96	0.91
	0.96	0.99
	0.96	0.99

Table 5.4: AUC achieved by baseline and custom pool model on each condition in all experiment. The bold values show the highest AUC achieved.

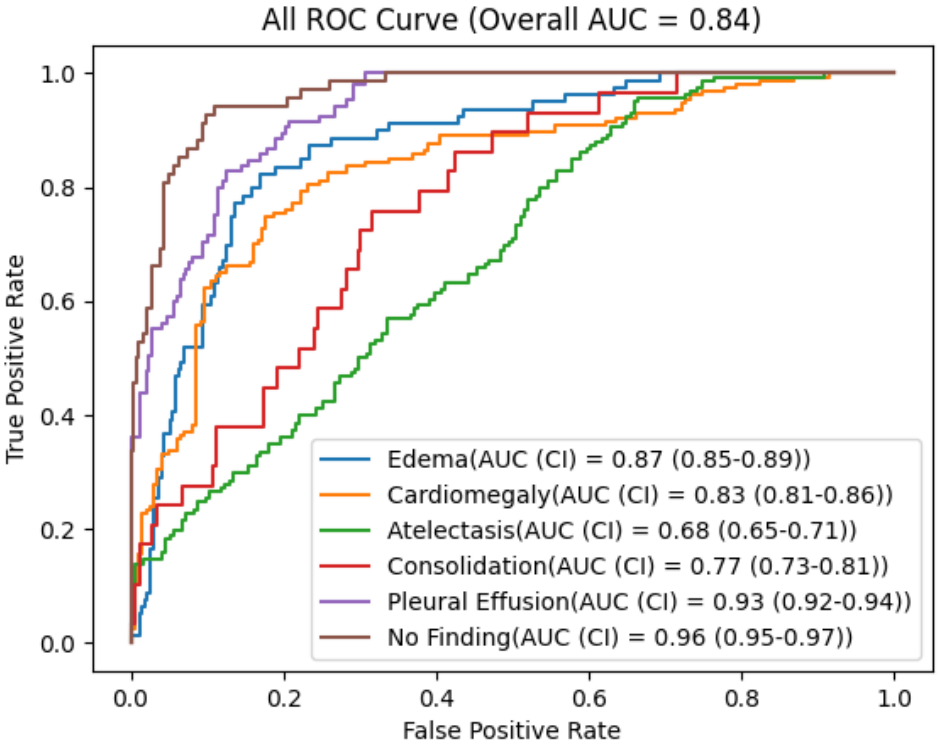


Figure 5.8: ROC curve for all conditions on the test data for baseline model. AUC and confidence intervals are mentioned for each condition.

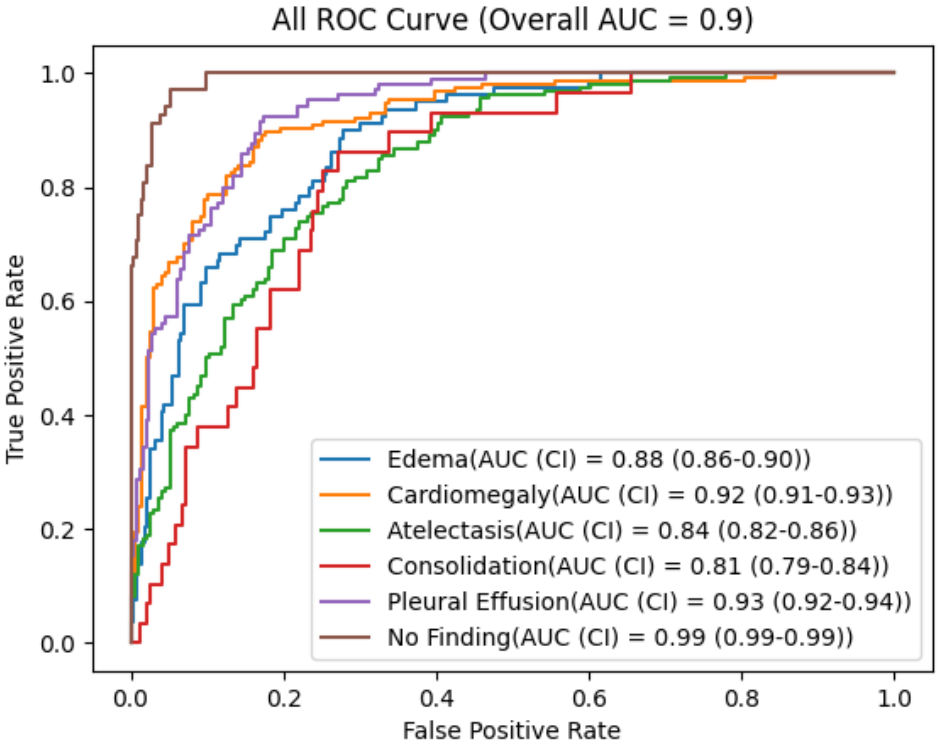


Figure 5.9: ROC curve for all conditions on the test data for custompool model. AUC and confidence intervals are mentioned for each condition.

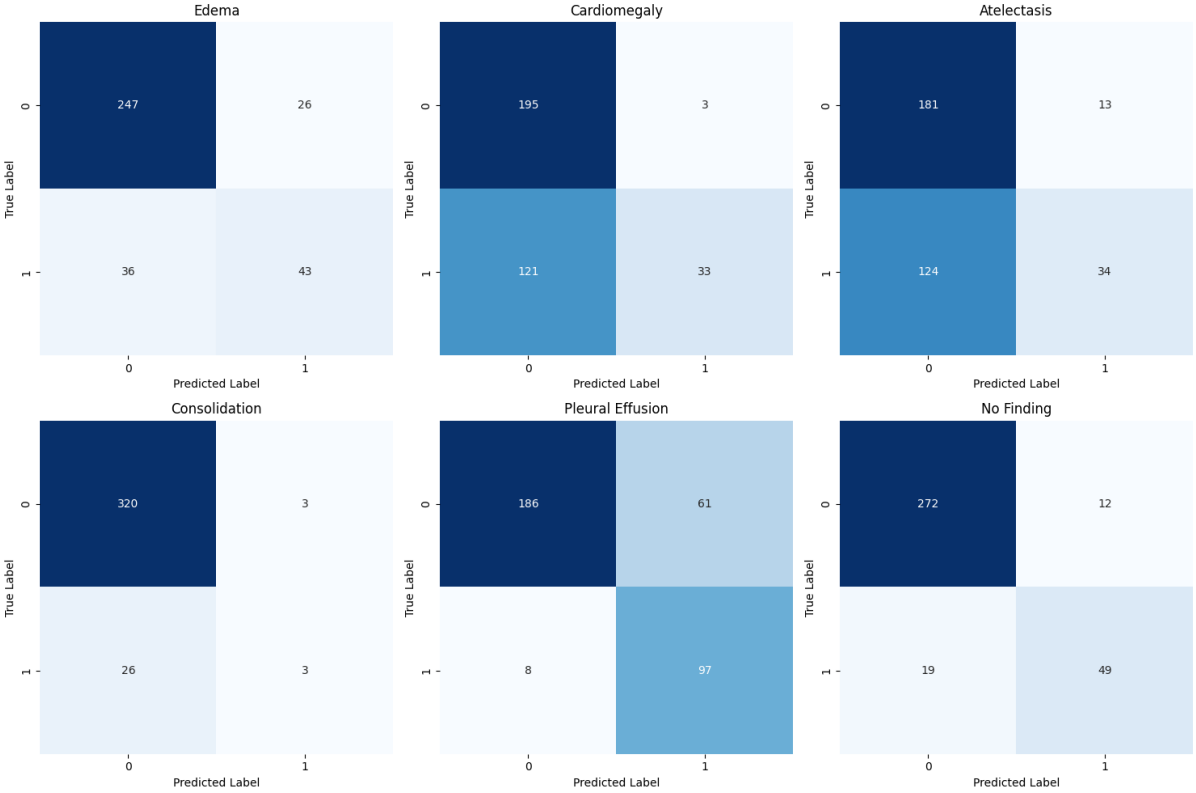


Figure 5.10: Individual confusion matrices for all conditions classified by baseline model.

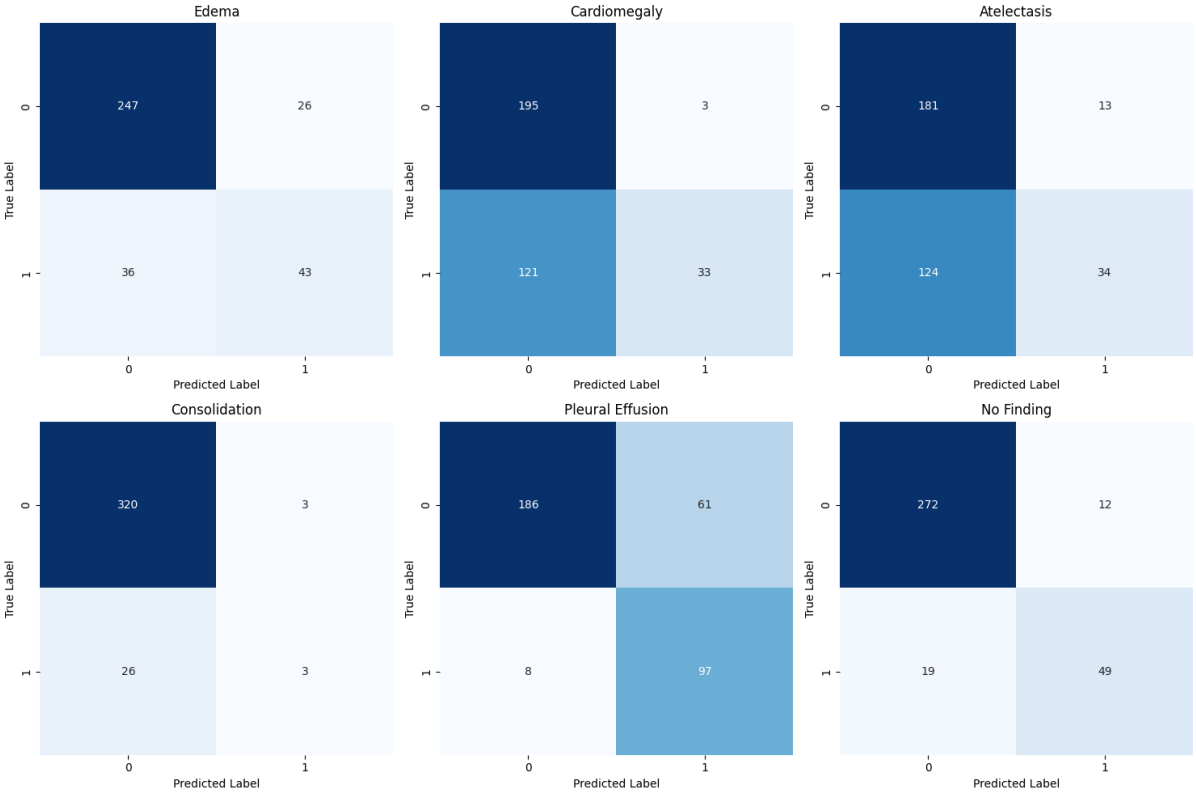


Figure 5.11: Individual confusion matrices for all conditions classified by custom pool model.

to retain both high and low-intensity features for accurate classification. Two experiments, using the standard CheXpert dataset and its multi-scale template-matched version, compare models with standard and custom pooling layers. Results show that models with custom pooling consistently achieve higher AUC scores which demonstrate the custom pooling layers effectiveness in enhancing the models feature extraction capabilities and overall performance. Therefore, the hypothesis **Hyp 02**, that the custom pooling layer integrated in CNNs will significantly enhance the performance of chest radiograph classification is accepted.

Chapter 6

Sequential Multi Label Enrichment

6.1 Introduction

Medical image classification is an important component of modern healthcare and plays a significant role in the diagnosis and management of diseases. Among various types of medical imaging, chest radiographs are particularly important to identify a wide range of thoracic conditions. Traditional approaches often focus on single class classification, where each image is classified into a binary outcome, such as the presence or absence of a specific disease. Although this method is straightforward, it does not reflect the complexity of real-world clinical scenarios, where patients are often affected by multiple coexisting conditions. In response to this limitation, the field of computer vision in medical imaging has progressed towards multiclass classification, where images are categorised into one of several possible conditions. Although this represents an improvement, it still does not fully address situations where multiple diseases may be present simultaneously. This leads to the need for multilabel multiclass classification, where each image can be labelled with multiple conditions and each independently classified. This approach aligns more closely with clinical practice but introduces significant challenges, particularly in terms of feature complexity and the ability to generalise across diverse patient populations.

Convolutional Neural Networks (CNNs) have become the cornerstone of image classification tasks, including medical imaging. These models work by automatically selecting and propagating features through multiple layers and progressively building a representation that can distinguish between different conditions. However, when it comes to multi-label classification, CNNs face the real difficulty of distinguishing overlapping features from multiple conditions. For example, on chest radiographs, the features associated with pneumonia might overlap with those associated with pulmonary edema, which confuses the model and degrades its performance. This overlapping of features presents a significant obstacle for CNN-based models, as the presence of one condition can obscure or distort the features of another. Moreover, due to the presence of a coexisting condition, there might be a completely different set of features representing the conditions. This challenge is particularly present in CheXpert data set used in this research, where the frequency of samples with multiple positive conditions decreases as the number of coexisting conditions increases. This imbalance means that models are often trained predominantly on single-condition samples, which limits their ability to accurately identify and distinguish multiple conditions within a single radiograph.

The Sequential Multi-Label Enrichment (SMLE) DenseNet model is designed to address these challenges. By employing a strategic multistage training procedure combined with a progressive expansion of the model's architecture, SMLE-DenseNet aims to enhance the model's ability to learn complex patterns associated with multiple co-existing conditions. This approach involves fine-tuning different sections of the model on subsets of the dataset. It starts with single condition samples and gradually incorporates more complex multi-condition samples sequentially in each stage. This staged training method helps the model build a robust baseline before tackling the complexities of multiple overlapping

conditions. An evaluation was conducted to assess the effectiveness of the SMLE-DenseNet model in multi-label classification compared to the baseline DenseNet121. Both models were trained and tested on the CheXpert dataset. The focus was on their ability to detect multiple concurrent conditions within a single radiograph, particularly in samples containing two, three, four, and five coexisting conditions. The number of correctly classified samples with multiple conditions served as the metric for evaluating the SMLE approach's impact. For example, ten samples contained three coexisting conditions in the test set. The baseline model accurately identified all three conditions in only three of these samples, while the SMLE model achieved accurate detection in six. This suggests the SMLE model outperforms the baseline model.

The results show that SMLE-DenseNet outperforms the baseline DenseNet121 model in detecting multiple conditions within a single radiograph. The progressive training strategy of SMLE helps the model to develop a nuanced understanding of complex feature interactions which leads to improved detection rates for samples with multiple coexisting conditions. This improvement underscores the potential of the SMLE approach to enhance the performance and reliability of multi-label classification in medical imaging to support better clinical decision-making and patient outcomes. This Chapter presents an evaluation of the SMLE-DenseNet model, along with a demonstration of its effectiveness in addressing the challenges of multi-label classification in chest radiology. A detailed comparison with the baseline DenseNet121 model is conducted. This comparison will illustrate how the SMLE-DenseNet model offers a more robust solution for detecting multiple conditions within chest radiographs. Ultimately, this contributes to advancements in the field of automated medical diagnosis.

6.1.1 Hypothesis: Hyp 03

The hypothesis for this Chapter is as follows:

The Sequential Multi-Label Enrichment (SMLE) DenseNet model will significantly improve the detection performance of multiple co-existing conditions on CheXpert test set as compared to the baseline standard DenseNet121 model.

6.2 Data Preprocessing

A data preprocessing strategy was employed within the experimental setup to address the challenges of multi-label classification tasks. The CheXpert dataset was utilized for this purpose. This dataset contains chest radiographs annotated for the presence or absence of various thoracic conditions and it provides valuable data for training and evaluating machine learning models. The availability of the predefined splits of the dataset for training, validation, and testing ensured consistency in both experiments (With Baseline and SMLE models). The intention was to participate in the CheXpert competition, which is why the predefined split of the dataset provided by the competition was used for the experiments.

Furthermore, the recognition of the inherent complexity in multi-label classification tasks necessitated the creation of distinct datasets tailored to varying numbers of coexisting conditions. To achieve this, the data set was segmented into four subsets, those containing samples with only two, three, four or five conditions. This segmentation facilitated a systematic exploration of the increasing complexity of feature interactions associated with multiple coexisting conditions. Table 6.1 gives the details about the number of samples for each subset of the training, validation and test sets. The number of samples with two positive conditions is significantly larger than any other subset. The 'Five Conditions' subset has only 275, 4 and 3 samples in training, validation and test sets respectively.

Although the large disparity in sample numbers across subsets creates the initial impression of significant data imbalance, it is important to consider that each subset contains all possible conditions, just

Dataset	Subsets	Subset Samples	% in Total Samples
Training	Two Conditions	46,716	30.63
	Three Conditions	17,278	11.33
	Four Conditions	3,327	2.18
	Five Conditions	275	0.18
Validation	Two Conditions	39	25.32
	Three Conditions	21	13.63
	Four Conditions	18	11.68
	Five Conditions	4	2.59
Test	Two Conditions	91	61.07
	Three Conditions	48	32.21
	Four Conditions	14	9.40
	Five Conditions	3	2.01

Table 6.1: Number of samples in each subset of the training, validation and test sets with their percentages in the total number of samples.

with variations in the number of coexisting ones. This mitigates the concern about the imbalance to some extent. The rationale behind preparing separate datasets for each number of coexisting conditions stemmed from two key considerations. Firstly, it allowed for the progressive training of the model on increasingly complex feature combinations. This gradual transition from simpler to more complex features aimed to enhance the generalisability of the model and the ability to accurately classify radiographs with multiple conditions. Secondly, grouping the samples by the number of conditions helped to manage the data. There were comparatively fewer samples with multiple conditions and this grouping made better use of the available data. Algorithm 1 shows the pseudo-code used to create these subsets of the CheXpert dataset. Finally, all images were resized to 224 x 224 pixels, while other preprocessing techniques used in this experiment can be found in Section 3.3.1 of this thesis.

Algorithm 1 Filtering and Saving Data Subsets Based on Number of Positive Conditions

```
BEGIN
LOAD DataFrame from 'CheXpert_training.csv'
INITIALIZE empty list dflist
for num in range(1, 6) do
    FILTER DataFrameto rows where the sum of
    ['Edema', 'Cardiomegaly', 'Atelectasis', 'Consolidation', 'Pleural Effusion'] equals num
    ADD a copy of the filtered DataFrame to dflist
end for
for i, df in ENUMERATE(dflist) do
    GENERATE filename as '{i+1}_Condition.csv'
    SAVE DataFrame df to CSV file with the generated filename
end for
END
```

6.3 Models

This experiment utilises two models: a baseline model and the SMLE model. The baseline model is the standard DenseNet121, trained on the CheXpert dataset without any modifications. It serves as a key point of reference for evaluating the performance of the SMLE model. The SMLE model, which stands for Sequential Multi-Label Enrichment, incorporates additional convolutional blocks and employs a multistage training process to handle the complexity of multi-label classification. The comparison of these two models allowed for an assessment of the effectiveness of the SMLE approach in improving the performance of detecting multiple coexisting conditions on chest radiographs.

6.3.1 Baseline Model

The selection of DenseNet121 as the baseline model for this experiment was based on its proven success in image classification tasks. DenseNet121 architectures have demonstrated superior performance in various image recognition challenges due to their efficient use of parameters and feature reuse mechanisms [1]. In this experiment, the pre-trained DenseNet121 model, originally trained on the ImageNet dataset, was adopted as the baseline. Leveraging pre-trained weights from ImageNet enables the model to initialize with learnt features that capture general image characteristics, which facilitates faster convergence and potentially improving performance on the chest radiographs.

To adapt the pre-trained DenseNet121 for the multi-label classification task of detecting thoracic conditions in chest radiographs, the last fully connected layers were replaced with a fully connected six-node layer. This layer is followed by sigmoid activation function, which gives the model the ability to output probabilities for each condition independently. By replacing the final layers of the network, the network architecture was tailored to suit the specific requirements of multi-label classification, where each image may be associated with multiple conditions simultaneously. This modification makes the model to learn to predict the presence or absence of each condition independently. The algorithm 2 explains the pseudocode used to build the baseline model.

Algorithm 2 Baseline Model Creation

```
function BASELINE_MODEL(IMG_SIZE, cls)
    LOAD DenseNet121 as base_model with:
      include_top set to False,
      weights set to 'imagenet',
      input_shape set to (IMG_SIZE[0], IMG_SIZE[1], 3)
    for all layer in base_model.layers do
       layer.trainable = True
    end for
    ADD custom classification head:
      x = base_model.output
      x = MaxPooling2D(pool_size=(2, 2))(x)
      x = GlobalAveragePooling2D()(x)
      x = Dense(1024, activation='relu')(x)
      x = Dropout(0.5)(x)
      predictions = Dense(cls, activation='sigmoid')(x)
    model = Model(inputs=base_model.input, outputs=predictions)
    print(model.summary())
    return model
end function
```

6.3.2 SMLE Densenet Model

Training a model with the entire dataset at once poses significant challenges, particularly in the context of multi-label classification. One major drawback is the dominance of features corresponding to single conditions, which can suppress the representation of features specific to multiple coexisting conditions. This imbalance in feature representation hamper the model's ability to learn complex patterns associated with multi-label scenarios which leads to suboptimal performance. To address this limitation, the Sequential Multi-Label Enrichment (SMLE) approach combines both model improvements and a novel training procedure.

The SMLE represents an extension of the DenseNet121 architecture, specifically designed to handle the complexities of multi-label classification in chest radiographs. As described in Section 3.4.4 of this thesis, SMLE incorporates additional convolutional blocks and adopts a multi-stage training process to progressively enrich the model's capability to learn complex feature interactions. In particular, the training procedure involves progressively fine-tuning different sections of the model with subsets of the dataset containing an increasing numbers of coexisting conditions at each stage. This strategic approach

leads to the model receiving sufficient exposure to diverse feature combinations, thereby enhancing its ability to generalise across a range of clinical scenarios.

A key advantage of the SMLE DenseNet model lies in its ability to retain specifically the multi label features more effectively in the later parts of the architecture compared to the earlier layers. This retention of features is important for accurately capturing the nuances of complex multi-label patterns present in chest radiographs. By allowing the later layers of the model to adapt more gradually to the increasing complexity of the input data, SMLE helps the development of more robust and discriminative representations which ultimately leads to improved classification performance. The algorithm 3 describe the pseudocode used to build and train the SMLE Densenet model in five stages. The python code defining all stages of the SMLE is given in Appendix D (SMLE) of this thesis.

Algorithm 3 SMLE Model Creation and Training

```
function NEXT_STAGE_MODEL(old_model, cls, layers_to_train)
    x ← old_model.layers[−6].output
    x ← Conv2D(256, (3, 3), activation='relu', padding='same')(x)
    x ← Conv2D(512, (3, 3), activation='relu', padding='same')(x)
    x ← Conv2D(512, (3, 3), activation='relu', padding='same')(x)
    x ← Conv2D(512, (3, 3), activation='relu', padding='same')(x)
    x ← MaxPooling2D(pool_size=(2, 2))(x)
    x ← GlobalAveragePooling2D()(x)
    x ← Dense(1024, activation='relu')(x)
    x ← Dropout(0.5)(x)
    predictions ← Dense(cls, activation='sigmoid')(x)
    model ← Model(inputs=old_model.input, outputs=predictions)
    for each layer in model.layers[:-layers_to_train] do
        layer.trainable ← False
    end for
    print(model.summary())
    return model
end function
stage_2 ← NEXT_STAGE_MODEL(baseline_model, NUM_CLASSES, 25)
TRAIN_MODEL(stage_2, 2)
stage_2 ← MODELS.LOAD_MODEL('stage_2_model.h5')
stage_3 ← NEXT_STAGE_MODEL(stage_2, NUM_CLASSES, 20)
TRAIN_MODEL(stage_3, 3)
stage_3 ← MODELS.LOAD_MODEL('stage_3_model.h5')
stage_4 ← NEXT_STAGE_MODEL(stage_3, NUM_CLASSES, 15)
TRAIN_MODEL(stage_4, 4)
stage_4 ← MODELS.LOAD_MODEL('stage_4_model.h5')
SMLE ← NEXT_STAGE_MODEL(stage_4, NUM_CLASSES, 10)
TRAIN_MODEL(SMLE, 5)
SMLE ← MODELS.LOAD_MODEL('SMLE_model.h5')
```

6.4 Training and Evaluation

The training of the baseline model began with the acquisition of the CheXpert dataset which is detailed in Section 3.3.1 of this thesis. Following the data pre-processing explained in the data pre-processing section of this Chapter, the model underwent training by utilising a predefined set of hyperparameters. These parameters were selected on the basis of previous experiments and include the learning rate and batch size. The training process involves a maximum of 100 epochs, during which the model iteratively adjusted its weights to minimise the loss function. Augmentation techniques, as discussed in Section 3.3.5, was applied to diversify the training data and enhance the robustness of the model.

Furthermore, the transfer learning technique, which is explained in Section 3.4.1 applied in the training of the baseline model by using a pre-trained DenseNet121 model to help the extraction of general image features and subsequently fine-tuned to align with the CheXpert dataset. To prevent overfitting and model performance optimisation, an early stopping mechanism was used. Additionally, a weighted loss function, detailed in Section 3.4.2, was used to address class imbalances present in the

	Two conditions (% of 91)	Three condition s(% of 48)	Four conditions (% of 14)	Five conditions
Baseline	10 (10.9)	2 (4.1)	0	0
	9 (9.8)	2 (4.1)	0	0
	9 (9.8)	2 (4.1)	0	0
	11 (12)	1 (2)	0	0
	11 (12)	2 (4.1)	0	0
SMLE	63 (69.2)	20 (41.6)	2 (0.5)	0
	62 (68.1)	20 (41.6)	2 (0.5)	0
	63 (69.2)	19 (39.5)	2 (0.5)	0
	62 (68.1)	20 (41.6)	2 (0.5)	0
	61 (67)	19 (39.5)	2 (0.5)	0

Table 6.2: The number and percentage of correctly classified samples for each test subset in all five experiments of baseline and SMLE Densenet models.

CheXpert multilabel classification task. The baseline model was trained and validated on the CheXpert's training and validation datasets respectively. Once trained, the model was tested on a separate portion of the dataset (CheXpert test set) to assess its ability to generalise to unseen data. The evaluation process involved measuring the total number of correctly classified samples per subset compared to the actual total number of samples per subset in the test set. This training, validation, and testing process was repeated five times to get better generalisable results.

On the other hand, the training procedure for the SMLE-DenseNet model followed a five stage approach, which is detailed in Section 3.4.4 of this thesis. At each stage, the model is progressively trained on the subsets of the dataset with increasing numbers of positive conditions per sample. The SMLE-DenseNet model underwent training with the same hyperparameters as the baseline model. The data augmentation, transfer learning technique, early stopping mechanisms and weighted loss function were used similarly to the baseline model to provide consistency in the training process of the two models. Finally, the SMLE-DenseNet model was evaluated using the same procedures as the baseline model and tested its performance on the CheXpert test set. Again, this training and evaluation process was repeated five times for the SMLE densenet model.

6.5 Results and Discussion

The purpose of the experiments conducted in this Chapter was to compare the performance of the baseline model with the SMLE DenseNet approach in classifying samples with multiple positive conditions. Both models were tested on the CheXpert test set. The number of samples corresponding to each test data subset is outlined in Table 6.1. The hypothesis of this Chapter mentioned in 6.1.1, says the SMLE approach will exceed in detecting multiple co-existing conditions in a single radiograph compared to the baseline model. This is because the latter part of the SMLE Denesenet model is trained on the data subsets with an increasing number of coexisting conditions, giving a better chance to the model to strengthen its understanding. The comparison between the baseline and the SMLE DenseNet model's performance was conducted by assessing the number of correctly classified samples for each subset by each model relative to the total number of samples in that subset. Table 6.2 presents the results detail of all experiments for both models. This includes the total samples in the test set, the number of correctly classified samples by the baseline model, and the number of correctly classified samples by the SMLE DenseNet model in each subset.

The results show that the SMLE approach is significantly effective compared to the baseline model. The Figure 6.1 shows the performance comparison of baseline and the SMLE model on the percentage of samples correctly classified for each subset. For two coexisting conditions, out of 91 total samples, the baseline model was only able to correctly classify 11 samples, while the SMLE Densenet model was

able to correctly classify 63 samples. Moving forward, out of total 48 samples in the three coexisting conditions subset, baseline model correctly classifies 2 and SMLE Densenet was able to classify 20. Similarly, in the four and five conditions subset, the baseline model was unable to classify even a single sample, while SMLE correctly classifies two and none in the four and five conditions subset, respectively.

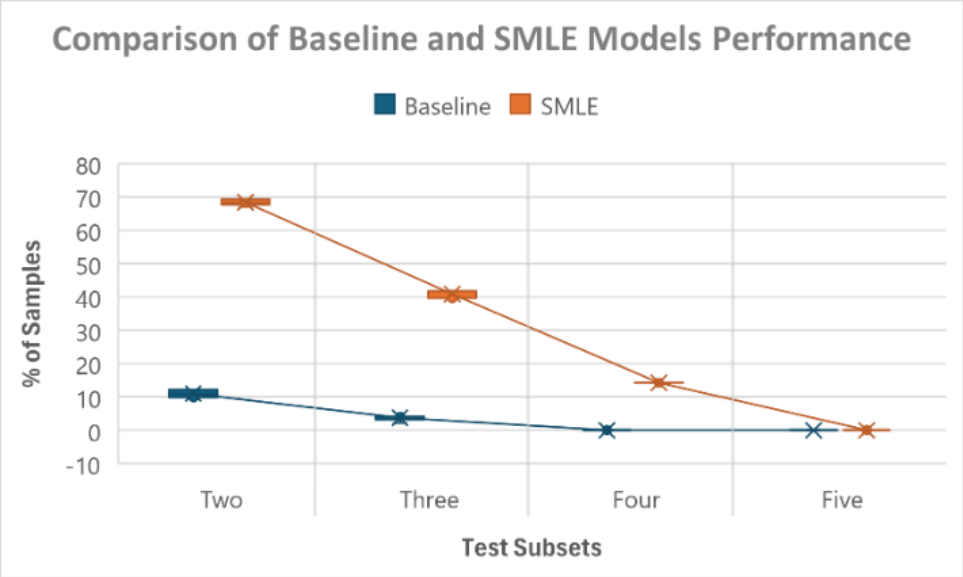


Figure 6.1: The percentage of correctly classified samples for each test subset (separated by the vertical lines) in all five experiments of baseline and SMLE Densenet models. The trend line shows the decreasing performance as the number co existing conditions increase.

To further confirm the performance superiority of the SMLE approach, a non-parametric Mann-Whitney U test was performed to check if there was a significant difference between the results of the baseline and the SMLE approach. This was done because the result dataset failed the normality test by giving the significance value for both the Kolmogorov-Smirnov and the Shapiro-Wilk tests less than 0.05. The Mann-Whitney U-test was conducted to compare the number of correct classifications between the Baseline model and the SMLE model. The independent variable was the model type (Baseline vs. SMLE), and the dependent variable was the number of correct classifications for all test subsets over five experiments as shown in Table 6.2. The results of the Mann-Whitney U-test indicated that there was a significant difference in the number of correct classifications between the Baseline model (n = 20) and the SMLE model (n = 20) with U = 117.5 and p = 0.021. Specifically, the SMLE model had higher ranks on average (Mean Rank = 24.63) compared to the Baseline model (Mean Rank = 16.38). This suggests that the SMLE model performed significantly better in terms of correct classifications than the Baseline model. The negative Z-score (-2.303) indicates that the median for correct classifications in SMLE is higher than that of the baseline model.

In addition, the results were deeply investigated, especially with regard to the remaining samples that were not correctly classified.Upon closer look, in the baseline model.

- **For two coexisting conditions subset:** out of the remaining 80 samples, one condition was correctly classified in 53 samples while no condition was detected in 27 samples.
- **For three coexisting conditions subset:** The remaining 46 samples, two conditions were correctly classified in 10 samples, one condition in 31 samples, and no condition was detected in 5 samples.
- **For four coexisting conditions subset:** While no sample was correctly classified for the four

coexisting conditions subset. Only 1 sample was correctly classified for three conditions, 7 samples for two conditions, and 6 samples for one condition.

- **For five coexisting conditions subset:** Lastly, of the 3 remaining samples, three conditions were correctly classified in 1 sample and two conditions were correctly classified in 2 samples.

In contrast, the SMLE approach shows different results on the remaining samples.

- **For two coexisting conditions subset:** out of the remaining 28 samples, one condition was correctly classified in 26 samples and none in the rest of 2 samples.
- **For three coexisting conditions subset:** 25 of 28 remaining samples were correctly classified for two conditions while one condition was correctly classified in 3 samples.
- **For four coexisting conditions subset:** out of the remaining 12 samples, 6 were correctly classified for three conditions, 5 for two conditions, and 1 for one condition.
- **For five coexisting conditions subset:** Lastly, for the five conditions subset, all 3 samples were correctly classified for three conditions.

Figure 6.2 shows a grouped bar plot to compare the performance of the baseline and SMLE models on the remaining samples, that were not hundred percent correctly classified for all co existing conditions. Figure 6.3 shows the confusion matrix of the best result of baseline and SMLE models, providing a comprehensive birds eye view of the performance of both models.

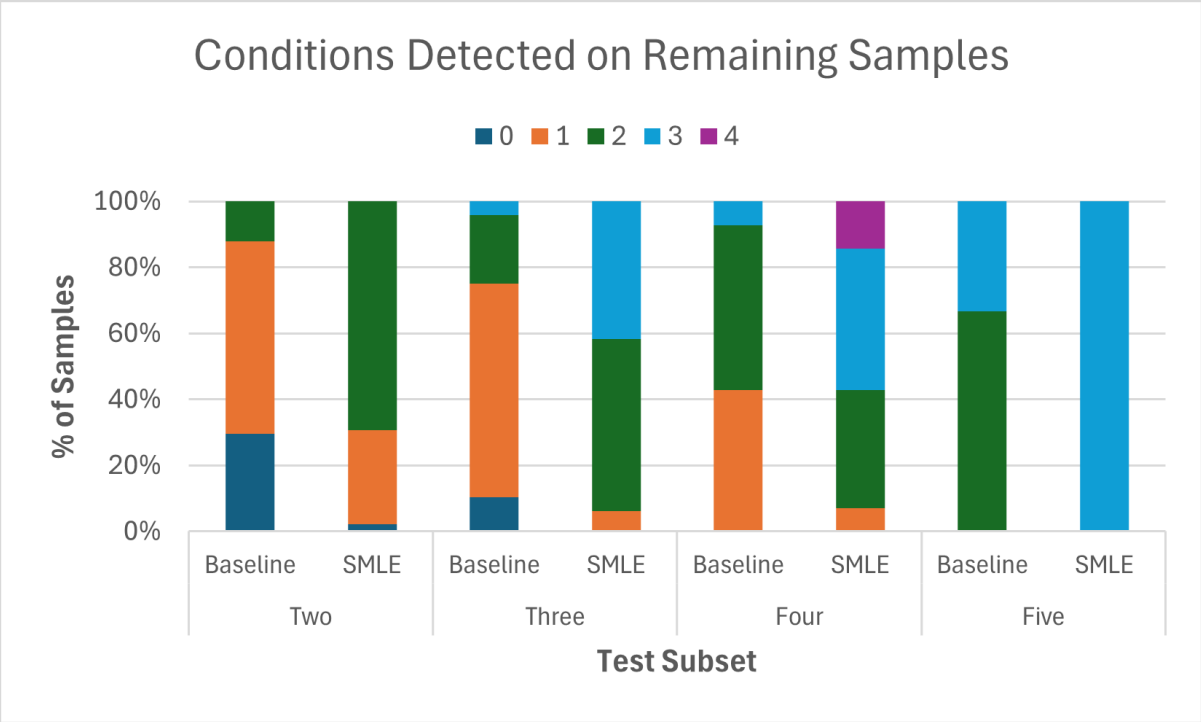


Figure 6.2: Comparison of baseline and SMLE model performance on remaining samples.

In conclusion, it is clear from the above discussion that the SMLE approach has a significantly superior performance over the baseline model in correctly identifying multiple conditions on a single radiograph. It is also observed that even if the SMLE Densenet model failed to detect all conditions on a radiograph, it still performed well in identifying a portion of the conditions correctly. For example, in the case of the five coexisting conditions subset, both the baseline and the SMLE were unable to detect

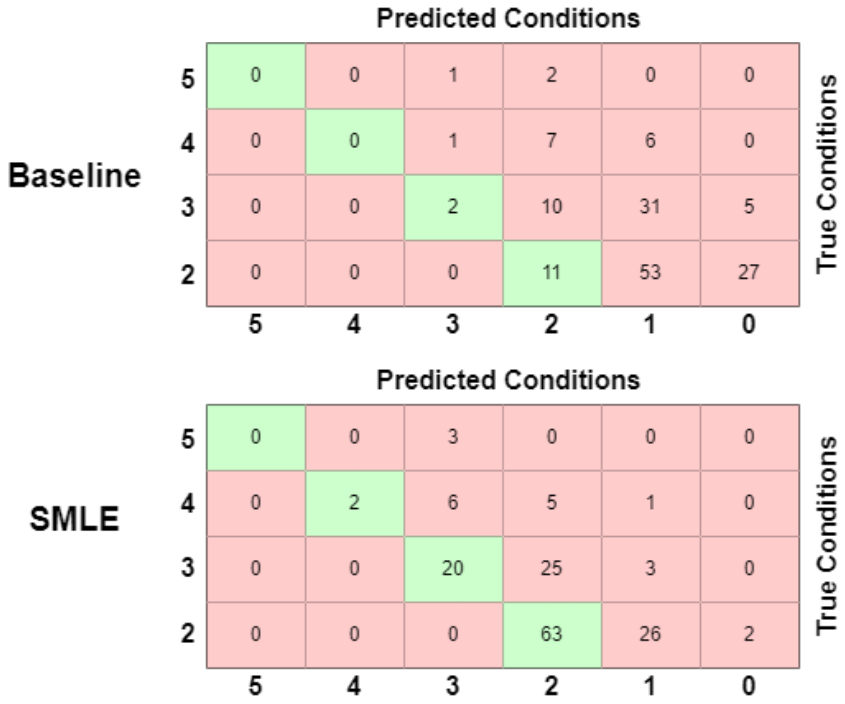


Figure 6.3: Confusion matrices for the best results of the baseline and SMLE models. The diagonal cells (Green), where the row and column indices (Co existing conditions) are equal, represent the number of perfectly classified samples.

all five conditions. However, comparing the results reveals that the baseline model was able to identify two conditions in two samples and three conditions in one sample, while the SMLE Densenet was able to correctly identify three out of five conditions in all three samples.

6.6 Summary

Identifying multiple objects within a single image is a well-studied problem in the field of computer vision. There are several approaches and techniques that have been developed to classify multiple co existing objects in a single image [131–133]. This Chapter tackles the complexity of classifying multiple co existing conditions in a chest radiographs, where overlapping features from multiple conditions pose significant challenges. The Sequential Multi-Label Enrichment (SMLE) DenseNet model is introduced, which is a multi-stage training process that progressively incorporates samples with increasing numbers of coexisting conditions. This technique helps the model build a strong understanding of complex feature interactions. The results show that the SMLE-DenseNet model significantly outperforms the baseline DenseNet121 in detecting multiple conditions within a single radiograph. This highlight its potential to enhance multi-label classification and support more accurate clinical decision-making. Therefore, the hypothesis **Hyp 03**, that the SMLE-DenseNet model will significantly improve the detection performance of multiple co-existing conditions on CheXpert test set is accepted. Moreover, the potential for broader applicability of SMLE-DenseNet extends beyond medical imaging. For example, in the domain of general image classification, the technique can be used to identify multiple co-existing objects within a single image.

Chapter 7

Radiologist's Perception of AI in Chest Radiology

7.1 Introduction

Chest radiographs are one of the most frequently performed imaging modalities for medical diagnosis in clinical setting [134]. It plays an extremely important role in identifying thoracic diseases and abnormalities, such as heart and lungs conditions. However, the increasing demand of trained and skilled radiologists every year is not being met by a proportional increase in qualified professionals. According to a report by The Royal College of Radiologists in 2022, the shortfall of radiologists in the UK is 29% which is expected to reach 40% and an additional 3,365 clinical radiologists will be required by 2027. Moreover, the diagnostic activity demand is increasing every year by 5% [135]. This leads to an additional pressure on the already stressed existing staff in radiology departments.

Due to this increasing demand of radiologists, the medical imaging field is desperately asking for help. Fortunately, the recent advancements in the field of computer vision for medical imaging is here to help in interpreting chest radiographs. However it is an active area of research and requires human radiologists involvement to build and strengthen the capabilities of AI based radiograph interpretation systems. As the primary validators, radiologist's trust is crucial to ensure the reliability of these systems before they are deployed in the clinical setting. Additionally, their overall perception of integrating these AI models, along with any fears and concerns they might have needs to be understood for a successful AI based system implementation.

Therefore, a study is conducted to get the perception of radiology professionals about the use and integration of AI models in their day to day clinical workflow. This Chapter delves into the details of this study. The study comprises of three stages. Stage one includes a questionnaire called pre-presentation questionnaire aimed to capture the demographics and initial perception of radiologists about the use of AI models in chest radiograph interpretation. At stage two, the participants were shown a presentation about the capabilities of AI models in detecting conditions on chest radaiographs. Not only the overall results, the actual radiographs including those with multiple coexisting conditions from the CheXpert test set were included. This way they were able to analyse the radiographs and compare their interpretation with the one done by the AI model. Finally, a post-presentation questionnaire was presented to them, which contains most of the same questions asked in pre-presentation questionnaire to measure the shift in perception. The post-presentation questionnaire also includes some additional questions along with open ended ones.

This Chapter discuss the perception of radiologists towards the use of AI models in radiology. Following the introduction, the methodology section will explain the participants demographics, ethical approval for the study and data acquisition through the pre and post presentation questionnaires. The data

analysis section details about the analysis setup and the study goals. Finally, the results and discussion section will present and discuss the results for each of the specific goals set in the data analysis section of this Chapter.

7.1.1 Hypothesis: Hyp 04

The hypothesis for this Chapter is as follows:

The perception and trust of radiologists in AI-based chest radiograph interpretation systems significantly improve after being exposed to the AI model's diagnostic capabilities.

7.2 Methodology

This study employed a pre-post questionnaire, designed to investigate the perceptions of radiology professionals especially radiologists, towards AI based chest radiograph interpretation. The design involved administering a questionnaire to them before and after they were shown a presentation on the performance of AI based model detecting conditions on chest radiographs. The pre-presentation questionnaire assessed baseline trust, confidence, and views on AI integration. While, the post presentation questionnaire captured any changes in their perception after viewing and analysing the radiographs in the presentation. This includes the comparison of responses given to the shared questions in pre and and post questionnaires and the analysis of a set of direct questions asking about their shift in perception in post presentation questionnaire. Finally open ended questions were also included to give them a chance to express their specific views about AI in radiology to capture a broader understanding of their perception.

7.2.1 Participants and Ethical Considerations

To run the study, the aim was to target only the radiologist participants working in various clinical settings to gain a perspective on their perceptions of AI in chest radiology. This selection was done purposely because of the fact that radiologists are the clinicians who work with chest radiographs in their day to day clinical work as their primary job. This inclusion criteria ensured participants have experience in interpreting chest radiographs. Participants were requested from Princess Alexandra and Luton & Dunstable Hospitals. However, when the survey was shared with these two hospitals, 4 participants, one from each of the other clinical speciality (Acute Medicine (AM), Intensive Care (IC), Digital transformation (DT) and GP Trainee (GPT)) with experience working on chest radiographs, also responded to the survey. This further diversify the sample set having participants with different levels of experience, age groups and specialities. At the time of writing this Chapter, a total of 15 valid responses have been received, out of which 11 are from the radiologists. To describe the characteristics of the participants, Table 7.1 shows the demographic information collected in the pre presentation questionnaire.

Before starting the study, ethical approval was taken from the University of Hertfordshire Health, Science, Engineering and Technology, Ethics Committee with Delegated Authority (HSET ECDA). The Ethics Protocol number for this study is SPECS/PGR/UH/05596. Getting the ethical approval process involved reviewing the study design, informed consent procedures, and data confidentiality measures. Moreover, participants were provided with an informed consent form which outlines the purpose of the study, their voluntary participation rights, disposal of the data after the completion of the study and assurance of complete anonymity of their responses.

7.2.2 Data Acquisition

Two customised questionnaires were designed to collect data from the participants. An extensive literature review was conducted to identify an existing questionnaire that suited the design of this study. All reviewed surveys employed their own set of questions [34–50] and no existing questionnaire was

Speciality	Count	Experience (Years)	Age
Radiologist (RAD)	11	4 (6-10) 2 (>15) 1 (< 1) 3 (11-15) 1 (1-5)	3 (25 - 34) 5 (35 - 44) 1 (45 - 54) 2 (55 - 64)
Acute Medicine (AM)	1	<1	25-34
Intensive Care (IC)	1	11-15	35-44
Digital transformation (DT)	1	<1	35-44
GP Trainee (GPT)	1	<1	25-34
Total	15		

Table 7.1: Demographics of the studys participants

Likert Scale in Questionnaires									Score
Strongly Disagree	Extremely not confident	Not at all likely	Not at all	Definitely not involved	No trust at all	Very negatively	Much less confident	Hinder	1
Disagree	Somewhat not confident	Slightly likely	Slightly	Probably not involved	Low trust	Negatively	Less confident	Slightly hinder	2
Neutral	Neutral	Moderately likely	Moderately	Neutral	Moderate trust	Neutral	No change	No change	3
Agree	Somewhat confident	Very likely	Very	Probably involved	High trust	Positively	More confident	Enhance	4
Strongly Agree	Extremely confident	Extremely likely	Extremely	Definitely involved	Complete trust	Very positively	Much more confident	Significantly enhance	5

Table 7.2: The Likert scales used in pre post questionnaires and their corresponding score.

found that fit for the purpose. However, it was observed that some questions with rephrased wording, appeared repeatedly across multiple studies. To ensure consistency with the previous studies, some of these frequently repeated questions (such as, "Do you believe AI/ML will devalue chest radiology as a profession?"), relevant to the purpose of this study, were incorporated into the questionnaires. To further ensure the quality of the questions and the design of the study, two senior radiologists from the Princess Alxendra hospital were consulted and their feedback was incorporated. With that, pre and post presentations questionnaires were finalised and shared to the participants via Microsoft Forms. Following sections explain the pre and post presentation questionnaires and the results presentation in detail.

7.2.2.1 Pre-presentation Questionnaire

The pre-presentations questionnaire consists of two sections. The first section captures the demographics of the participants such as gender, age group, clinical speciality, experience and their level of computer knowledge. The second section has fifteen questions to gauge the participants overall level of trust and confidence on the AI models, the level of optimism, the concern of being replaced by AI, impact on their profession and openness to use AI models in their day to day clinical work. The questions utilized various Likert scales [136] with five response options (e.g., strongly disagree - strongly agree, extremely not confident - extremely confident and not at all likely - extremely likely). For instance, one question used a five-point Likert scale (strongly disagree, disagree, neutral, agree, strongly agree) to ask participants about their level of agreement with the statement: 'Do you believe AI/ML will devalue chest radiology as a profession?'. Table 7.2 shows all Likert scales used in the pre and post presentation questionnaires with the corresponding score for each Likert scale value. The full details of the general and pre presentation questions can be found in the Appendix E (Pre and Post Questionnaires) of this thesis.

7.2.2.2 Presentation

Following the pre presentation questionnaire, the participants were provided with a slide presentation which they could access through a link given right after the pre questionnaire. The presentation starts with explaining the role of machine learning in chest radiology, followed by the details of the CheXpert dataset used to train the models, explaining evaluation metrics used to guage the models performance and the results in the form of AUC plots for each of the five conditions (Edema, Cardiomegaly, Atelectasis, Consolidation and Pleural Effusion). Furthermore, during this presentation, participants viewed fifteen radiographs from the CheXpert test set. All of them were labelled by the best model trained during this research. The radiographs presented in the presentation had single as well as multiple co existing conditions. The collection comprises of radiographs having different medical conditions. Specifically, two radiographs exhibit signs of cardiomegaly given by an enlarged heart. Another two radiographs show evidence of atelectasis, which indicates a partial or complete collapse of the lung. Additionally, two radiographs show the presence of edema, which is a condition marked by excess fluid in the lungs. Furthermore, two radiographs shows evidence of pleural effusion, where fluid accumulates in the pleural space around the lungs. Lastly, two radiographs are categorized as having no findings which indicate the absence of any abnormalities.

In addition to these, three radiographs present two co existing conditions each, which shows the complexity of multiple simultaneous conditions. Moreover, two radiographs are even more complex because of showing three coexisting conditions. The participants were asked to have a look at the radiographs and compare their interpretations with the ones labelled by the AI models. Finally results from other studies were also presented. The idea was to show them the capabilities of AI models in detecting single as well as multiple co exising conditions on radiographs. The actual presentation used in this study can be found in Appendix E (Results Presentation).

7.2.2.3 Post-presentation Questionnaire

At the end of the presentation. The last slide instructed the participants to go to the post presentation questionnaire to rerecord their responses for a comparison. The post presentation questionnaire consists of 25 total questions. Among these, 13 questions were deliberately repeated from the pre-presentation questionnaire. This intentional repetition was aimed at capturing any shifts in the participant's perceptions after they had viewed the presentation results. The list of these 13 pre- and post- questions can be found in Table 7.5. Moreover, Four questions were added to this questionnaire, to directly ask participants on the impact of the result presentation on their perception about AI in radiology. This addition of questions was made due to the anticipated low participation rate and the possibility that participant's prior exposure to AI might lead to a minimal difference in responses when pre and post are compared. Table 7.6 lists all of the four direct questions from post presentation questionnaire.

Furthermore, to gain a deeper understanding of the participant's viewpoints on the integration of AI in chest radiology, the final five questions of the questionnaire were designed in an open-ended format. These questions were intended to delve deeply into various aspects, including the potential impacts on clinical workflow, levels of trust and usability, the perceived strengths and limitations of the AI models, and any specific fears or concerns the participants might have. This open-ended section was important for capturing detailed insights and comprehensive feedback from the participants.

7.3 Analysis of Responses

Analyzing the responses that utilize Likert scales, such as the ones used in pre- and post- questionnaires, involves a systematic examination of the data to derive meaningful insights about the respondent's attitudes, perceptions, and behaviors. These Likert scales are carefully designed to measure a variety of dimensions, including agreement, confidence, likelihood, involvement, trust, impact, and others. Each

response option is associated with a numerical score that ranges from 1 to 5 to facilitate the quantitative analysis. Table 7.2 shows the detail of the Likert scales and the score associated to each response option. For instance, the Likert scale ranges from "Strongly Disagree" to "Strongly Agree" also have intermediate options such as "Disagree", "Neutral" and "Agree." The scores assigned to these responses (from 1 to 5) allow for straightforward computation of descriptive statistics such as mean, median, and mode, as well as more advanced statistical analyses.

7.3.1 Analysis Setup

The analysis of responses in this study began with a thorough setup, so that the data collected from the pre- and post-presentation questionnaires is systematically examined for meaningful insights. The first step involved data cleaning and preparation. This includes verifying the completeness of the responses, the handling of any missing data appropriately, and ensuring that all responses are coded accurately according to the Likert scale scores outlined in Table 7.2. Once the data was prepared, the next step was to summarize the responses. Descriptive statistics such as mean, median, mode, and standard deviation was computed for each question to provide an overview of the general trends and central tendencies within the data. Frequency distributions was created to visualize how often each response option was selected by the participants. To have a clear visual summary of the data, these distributions were visualized through bar charts and boxplots.

Following the descriptive analysis, inferential statistics were employed to explore relationships and differences within the data. Mann-Whitney U test [137] was used to compare pre- and post-presentation responses to identifying any statistically significant shifts in perceptions due to the results presentation. Analysis was performed to capture the qualitative insights from the open-ended questions. This involves coding the text responses to identify recurring themes and patterns. The qualitative data was used to provide context and depth to the quantitative findings which helps in building more comprehensive understanding of the radiologist's perceptions.

7.3.2 Study Goals

This study was designed to investigate the perceptions of radiology professionals regarding the interpretation of chest radiographs using AI models. The primary objectives are listed as follows:

- **Initial Overall Perception:** To assess the overall trust and confidence of radiology professionals in AI models for radiology at the start of the study. This was achieved through the analysis of a set of 10 questions in pre-questionnaire, which aimed to capture their baseline attitudes towards AI integration in chest radiograph interpretation.

- **Initial Role-Specific Perception:** To determine if there are differences in the initial opinions of radiology professionals based on their specific roles, such as radiologists, intensive care and others outilned in Table 7.1. The understanding these role-specific initial perceptions can provide insights into the varying levels of acceptance and trust in AI across different specialties.

- **Pre-Post Opinion Shift:** To measure any shift in beliefs and attitudes towards AI after the participants were exposed to a detailed presentation on the capabilities and performance of AI in interpreting chest radiographs. This pre-post survey comparison aimed to capture shifts in perception that might occur due to increased knowledge and familiarity with AI's diagnostic potential.

- **Pre-Post Role-Specific Shift:** To investigate if the shift in opinions, as measured in the Pre-Post Opinion Shift, varies across different roles. This analysis helps to understands if the impact of the presentation differs between radiologists and other professionals having experience in chest radiology.

- **Concerns and Suggestions:** To analyze responses to open-ended questions and identify specific concerns and suggestions from participants regarding the implementation of AI models in clinical practice. This qualitative analysis aimed to gather detailed feedback that can inform strategies for better integration of AI systems to address potential barriers and enhance their acceptance among radiology professionals.

By addressing the above mentioned goals, this study aimed to provide a comprehensive understanding of the current perceptions of AI in chest radiology, identify areas for improvement, and offer actionable insights for the successful implementation of AI technologies in clinical settings.

7.4 Results and Discussion

This section contains the results based on the analysis done for each study goal explained in section 7.3.2 and discusses them in detail.

7.4.1 Initial Overall Trust and Confidence

To assess the general trust and confidence of radiology professionals in AI for radiograph interpretation, ten questions were identified from the pre-presentation questionnaire. These questions aimed to gauge participants' general beliefs in AI in radiology, specifically addressing their anxieties about AI potentially replacing radiologists, confidence in using AI suggestions when uncertain, the associated potential benefits and risks, the potential of AI to expedite diagnosis, concerns about devaluing chest radiology as a profession, AI's role in addressing the current shortage of radiologists, the possibility of AI surpassing radiologists in disease detection, AI's capability to flag abnormal findings, and enhancing work efficiency in clinical radiology. Figure 7.1 shows the response frequency on all questions. Table 7.3 lists all ten

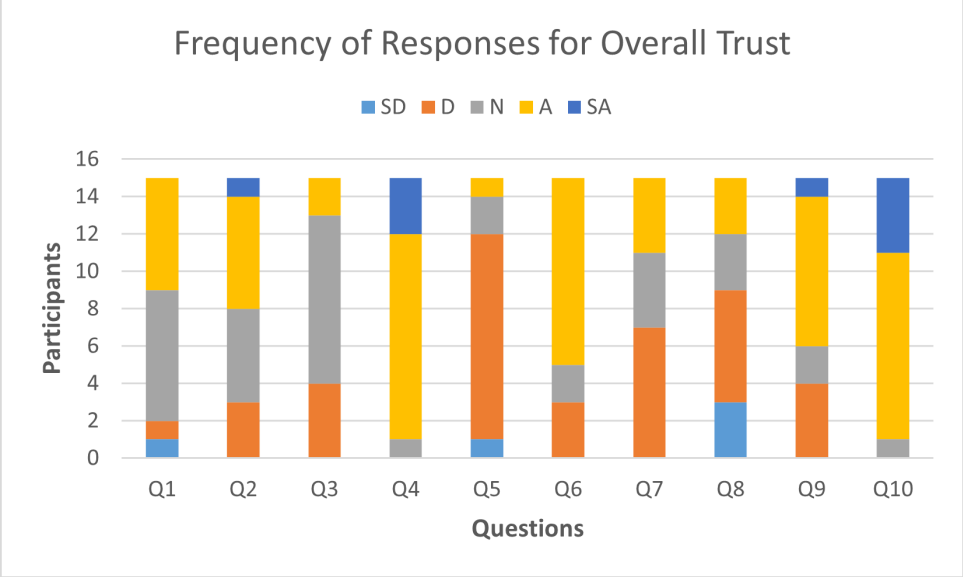


Figure 7.1: Frequency of participant responses for overall trust across the 10 questions, categorized by strongly disagree (SD), disagree (D), neutral (N), agree (A), and strongly agree (SA)

questions used to analyze the overall trust, the median response scores associated with each question, and their corresponding interquartile ranges (IQR). The Table provides a snapshot of the initial perceptions of radiology professionals regarding the integration of AI in their field. The median score offers a measure of central tendency, while the IQR provides insights into the variability of responses. The boxplot in Figure 7.2 visually represents the distribution of responses for each of these ten questions. The median

No	Questions	Median	IQR (25-75%)
1	To what extent do you believe that AI/ML models could replace the role of a chest radiologist in interpreting X-rays?	3	1
2	How confident are you in AI/ML making diagnostic suggestions when you are unsure about chest X-ray findings?	4	1
3	What is your level of trust in AI/ML tools for chest radiology in terms of the balance of risks versus benefits?	3	1
4	Do you agree AI/ML has a great potential to speed up diagnosis and can be a great tool at the hand of the radiologist?	4	0
5	Do you believe AI/ML will devalue chest radiology as a profession?	2	0
6	Do you believe AI/ML have the potential to address the shortage of radiologists in the future?	4	1
7	How confident are you in the prediction that AI/ML will reduce the number of chest radiologists in each imaging department?	3	2
8	In your opinion, will AI/ML outperform chest radiologists for disease detection in chest X-rays within a decade?	2	3
9	To what extent do you think AI/ML could alert chest radiologists to abnormal findings in chest X-rays?	4	1
10	Do you believe AI/ML could increase work efficiency in chest imaging, particularly in X-ray interpretation?	4	0

Table 7.3: Questions used to gauge the overall belief and confidence of the participants on the use of AI in chest radiology.

values are indicated by the line within each box, while the boxes themselves represent the interquartile range (IQR), encompassing the middle 50% of the data. The whiskers extend to the minimum and maximum values within 1.5 times the IQR, and any points outside this range are considered outliers.

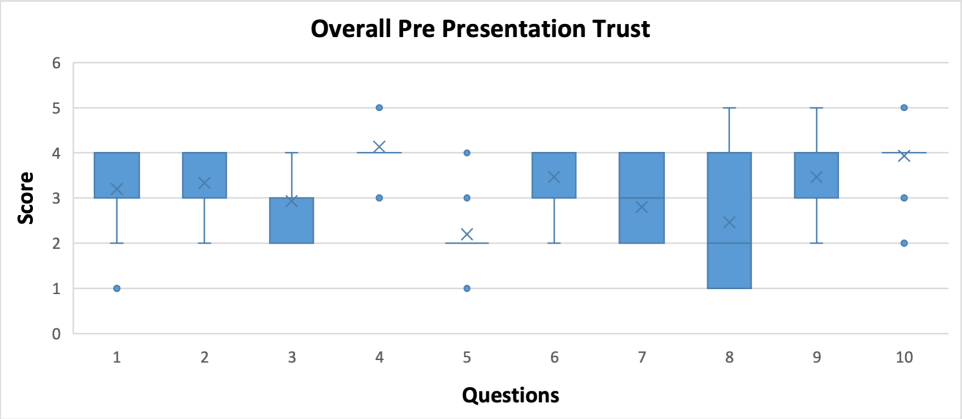


Figure 7.2: Overall Trust pre presentation

The initial perceptions of radiology professionals as summarized in Table 7.3 and visualized in Figure 7.2 provides several key insights. On the question about potential replacement by AI (Question 1), The median score of 3 with an IQR of 1 indicates a neutral stance on whether AI could replace chest radiologists. This suggests uncertainty and mixed feelings among participants about the potential for AI to fully take over their roles. On confidence in AI suggestions (Question 2), With a median score of 4 and an IQR of 1, there is a relatively high level of confidence in AI making diagnostic suggestions when radiologists are unsure about findings in chest radiographs. This highlights a positive attitude towards AI assistance in challenging cases. On trust in AI tools (Question 3), the median score of 3 and an IQR of 1 reflect a balanced view of the risks and benefits associated with AI tools in chest radiology. Participants seem to recognize both the potential and the limitations of these models. Moreover, on AI's potential to speed up diagnosis (Question 4), the high median score of 4 and an IQR of 0 demonstrate strong agreement that AI has significant potential to expedite the diagnostic process which makes it a valuable tool for radiologists. On the devaluation of the radiologist profession (Question 5), A lower median score of 2 with an IQR of 0 suggests that most participants do not believe AI will devalue chest radiology as a profession which indicates confidence in the strong value of human expertise.

No	Questions	Median	IQR (25-75%)
1	To what extent do you believe that AI/ML models could replace the role of a chest radiologist in interpreting X-rays?	3	1
2	How confident are you in AI/ML making diagnostic suggestions when you are unsure about chest X-ray findings?	4	2
3	What is your level of trust in AI/ML tools for chest radiology in terms of the balance of risks versus benefits?	3	1
4	Do you agree AI/ML has a great potential to speed up diagnosis and can be a great tool at the hand of the radiologist?	4	1
5	Do you believe AI/ML will devalue chest radiology as a profession?	2	1
6	Do you believe AI/ML have the potential to address the shortage of radiologists in the future?	4	1
7	How confident are you in the prediction that AI/ML will reduce the number of chest radiologists in each imaging department?	2	2
8	In your opinion, will AI/ML outperform chest radiologists for disease detection in chest X-rays within a decade?	2	2
9	To what extent do you think AI/ML could alert chest radiologists to abnormal findings in chest X-rays?	3	1
10	Do you believe AI/ML could increase work efficiency in chest imaging, particularly in X-ray interpretation?	4	0

Table 7.4: Questions used to gauge the overall belief and confidence of the radiologists on the use of AI in chest radiology.

Furthermore, on the question about addressing the radiologist shortage (Question 6), the median score of 4 and an IQR of 1 indicate optimism that AI can help address the shortage of radiologists which highlights the perceived practical benefits of AI integration. On question about reduction in number of radiologists (Question 7), the median score of 3 with a wider IQR of 2 suggests mixed views on whether AI will reduce the number of chest radiologists required in imaging departments.Outperformance by AI (Question 8): The median score of 2 and a wide IQR of 3 indicate skepticism about AI outperforming chest radiologists in disease detection within the next decade, reflecting caution about overestimating AI capabilities. On capabilities of AI being able to alert for abnormal findings (Question 9), the high median score of 4 and an IQR of 1 show strong confidence in AI's ability to alert radiologists to abnormal findings shows AI as a supportive tool. Lastly, on the question about the increase in work efficiency (Question 10), the high median score of 4 with an IQR of 0 suggests that participants believe AI can significantly enhance work efficiency in chest radiograph interpretation.

Overall, this initial assessment indicates that while radiology professionals recognize the potential benefits of AI, there remains a cautious optimism, particularly regarding AI's ability to fully replace human radiologists or completely outperform them in the near future. These responses reflect a balance of trust in AI's supportive capabilities and recognition of its current limitations.

7.4.2 Role-Specific Trust Level

To delve deeper into the perceptions of different roles within radiology, the initial trust and confidence in AI for radiograph interpretation was analysed while focusing on specific roles such as radiologists (RAD), acute medicine specialists (AM) , intensive care specialists (IC), digital transformation professionals (DT), and the GP trainees (GPT). This section aims to explore how these different roles perceive AI's integration into chest radiology, with a particular focus on the responses from radiologists due their primary affiliation with radiology. Table 7.4 lists the same ten questions used to analyze the overall trust and confidence with the median and IQR specific to radiologists. These responses are also compared to the overall initial trust and confidence provided in Table 7.3 to identify any radiologist specific nuances. Furthermore, the boxplot in Figure 7.3 provides a visual representation of the overall trust scores segmented by role. The median values are indicated by the line within each box, while the boxes themselves represent the interquartile range (IQR), which shows the middle 50% of the data. The whiskers extend to the minimum and maximum values within 1.5 times the IQR, and any other points outside this range are considered outliers.

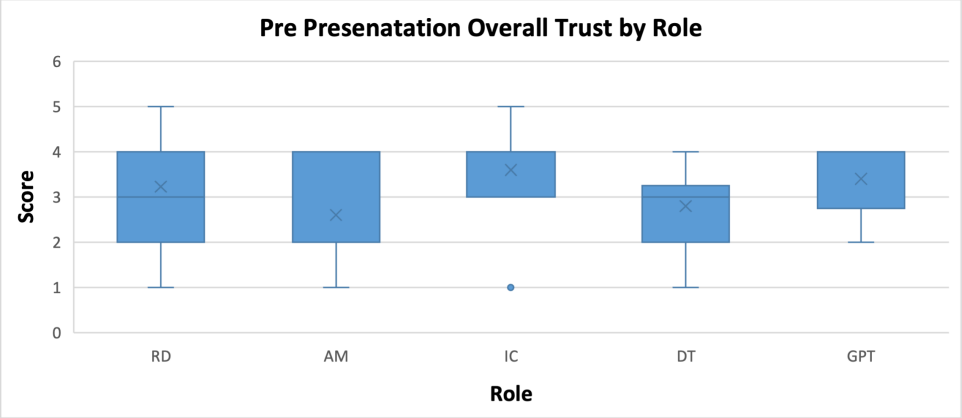


Figure 7.3: Overall Trust by role

The initial perceptions of radiologists, as summarized in Table 7.4 and visualized in Figure 7.3, provide several key insights. On the question about potential replacement by AI (Question 1), the median score of 3 with an IQR of 1 among radiologists indicates a neutral stance on whether AI could replace chest radiologists. On confidence in AI suggestions (Question 2), the radiologists reported a median score of 4 with an IQR of 2, which shows a relatively high level of confidence in AI making diagnostic suggestions when they are unsure about chest radiograph findings. Compared to the overall initial trust and confidence (median of 4 and IQR of 1), radiologists exhibit a similar level of confidence but with more variability in their responses which indicates a wider range of opinions among radiologists. Moreover, on the trust in AI tools (Question 3), the median score of 3 and an IQR of 1 among radiologists reflects a balanced view of the risks and benefits associated with AI tools in chest radiology. This aligns with the overall initial trust and confidence, which also had a median of 3 and IQR of 1.

On the question about AIs potential to speed up diagnosis (Question 4), radiologists showed a high median score of 4 with an IQR of 1, This is consistent with the overall initial trust and confidence, which had a median of 4 and an IQR of 0. The slight difference in IQR (1 for radiologists vs. 0 overall) suggests radiologist exhibit slightly more variability in their responses. On the devaluation of the radiologist profession (Question 5), A lower median score of 2 with an IQR of 1, compared to the overall initial trust and confidence, which also had a median of 2 but with a slightly narrower IQR of 0 also shows more variability in their responses. Furthermore, on the question about addressing the radiologist shortage (Question 6), the median score of 4 and an IQR of 1 among radiologists also aligns with the overall perception. On question about reduction in number of radiologists (Question 7), the median score of 2 with an IQR of 2 among radiologists, compared to the overall initial trust and confidence (median of 3 and IQR of 2), radiologists are slightly more skeptical about AI reducing the number of chest radiologists which indicates more variability and uncertainty in their responses.

On the outperformance by AI (Question 8), The median score of 2 and an IQR of 2, compared to the overall initial trust and confidence, which had a median of 2 and a wider IQR of 3, radiologists show similar skepticism but with slightly less variability in their responses which suggests a more consolidated views among radiologists. On capabilities of AI being able to alert for abnormal findings (Question 9), the median score of 3 and an IQR of 1 among radiologists, compared to the overall initial trust and confidence, which had a median of 4 and IQR of 1, radiologists seem slightly less confident in this capability which indicates a slightly more cautious stance regarding AI's reliability in identifying abnormalities. Lastly, on increasing work efficiency (Question 10), the high median score of 4 with an IQR of 0 among radiologists is consistent with the overall perception which shows AI's potential to improve workflow efficiency.

Looking at the boxplot in Figure 7.3 shows the overall trust scores segmented by role which indicate

differences in perceptions across various specialties. The box representing radiologists, display a balanced view with a median score of around 3-4, which shows a cautious optimism towards AI integration. The acute medicine shows slightly higher trust with a broader range of scores and reflect diverse opinions within this group. Moreover, intensive care presents a more concentrated range of scores, with high median values which indicate strong trust in AI's potential. The digital transformation exhibit variability in trust with scores ranging widely showing mixed feelings about AI in radiology. Finally the GP trainee displays a more optimistic view with higher median scores. The analysis shows that while there are variations in trust and confidence levels across different roles, there is a general consensus on the potential benefits and current limitations of AI in chest radiology. Radiologists, in particular, share similar views with the overall perception with a balanced approach to AI integration.

7.4.3 Evaluation of Shift in Perception

To evaluate the shift in perception of radiology professionals regarding the use of AI in chest radiology, a comparison was made between the responses of 13 shared questions in pre- and post-presentation questionnaires. This comparison aimed to identify any changes in trust, confidence, and overall perception after participants were exposed to detailed presentation about AI capabilities in chest radiograph interpretation. Table 7.5 presents the median scores for each of the 13 questions, both before and after the presentation, along with the P-Value, U-Value, and Z-Value obtained from the Mann-Whitney U test. The Table provides a comprehensive view of the statistical analysis performed to determine the significance of the observed shifts. The Figure 7.5 and Figure 7.6 visually represent the distribution of pre- and post-presentation responses across the 13 questions for all roles and specifically for radiologists, respectively. The boxplots display the median scores, interquartile ranges (IQR), and the outliers which offers a visual comparison of the shifts in perception. Figure 7.4 shows the comparison of the response frequencies between pre- and and post- across the 13 questions.

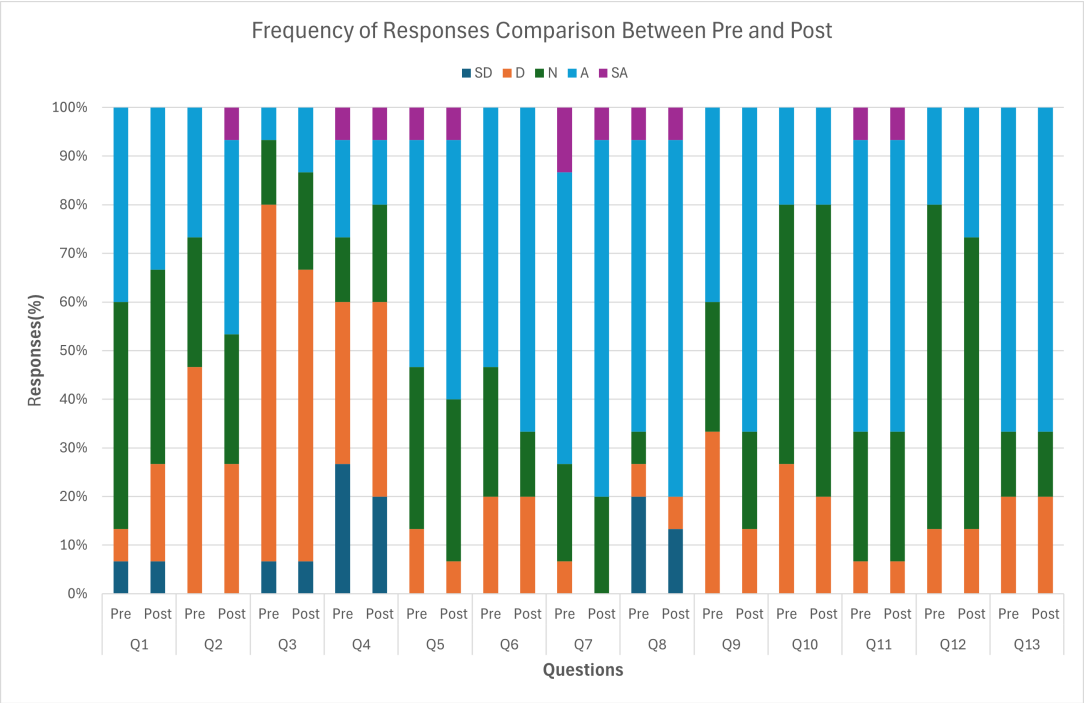


Figure 7.4: Comparison of Pre and Post questionnaire responses across 13 questions, displaying the frequency of responses on different scales (Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree).

No	Questions	Median (Pre/Post)	P-Value	U-Value	Z-Value
1	To what extent do you believe that AI/ML models could replace the role of a chest radiologist in interpreting X-rays?	3/3	0.51	98	-0.68
2	How confident are you in the prediction that AI/ML will reduce the number of chest radiologists in each imaging department?	3/3	0.18	82	-1.33
3	Do you believe AI/ML will devalue chest radiology as a profession?	2/2	0.47	98	-0.72
4	In your opinion, will AI/ML outperform chest radiologists for disease detection in chest X-rays within a decade?	2/2	0.93	110	-0.08
5	To what extent do you think AI/ML could alert chest radiologists to abnormal findings in chest X-rays?	4/4	0.66	103	-0.43
6	How confident are you in AI/ML making diagnostic suggestions when you are unsure about chest X-ray findings?	4/4	0.57	100	-0.56
7	Should there be a dedicated AI task force or subcommittee for chest radiology in relevant medical conferences or associations?	4/4	0.90	110	-0.12
8	If there was an initial AI task force meeting at the next conference, would you attend or like to be involved, specifically focusing on chest radiology?	4/4	0.51	99.5	-0.64
9	Would you be more willing to learn about AI/ML if educational tools were specifically tailored for chest radiology rather than general radiology?	3/4	0.12	79	-1.53
10	What is your level of trust in AI/ML tools for chest radiology in terms of the balance of risks versus benefits?	3/3	0.78	106.5	-0.27
11	How do you anticipate the job of the average chest radiologist will be impacted in the next 5 years from AI?	4/4	1	112.5	0
12	Do you think jobs in chest radiology will be more or less impacted than other radiology subspecialties by AI/ML?	3/3	0.75	106	-0.31
13	Do you believe AI/ML have the potential to address the shortage of radiologists in the future?	4/4	1	112.5	0

Table 7.5: Detail of questions used to evaluate the pre and post shift in perception. The median values of pre and post scores for each questions and the P-Values for Mann-Whitney U test are also given.

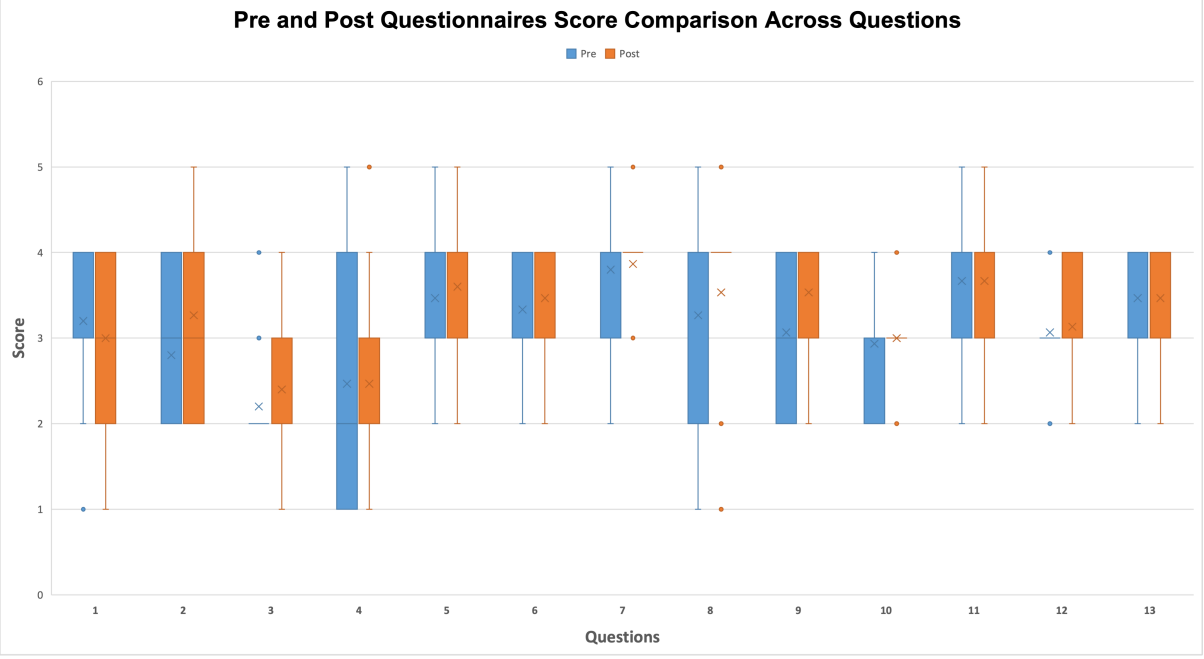


Figure 7.5: Pre and Post questionnaire shift in perception. The blue box represents the pre-presentation scores and the orange box represents the post-presentation scores for each of the 13 questions to measure the shift in perception.

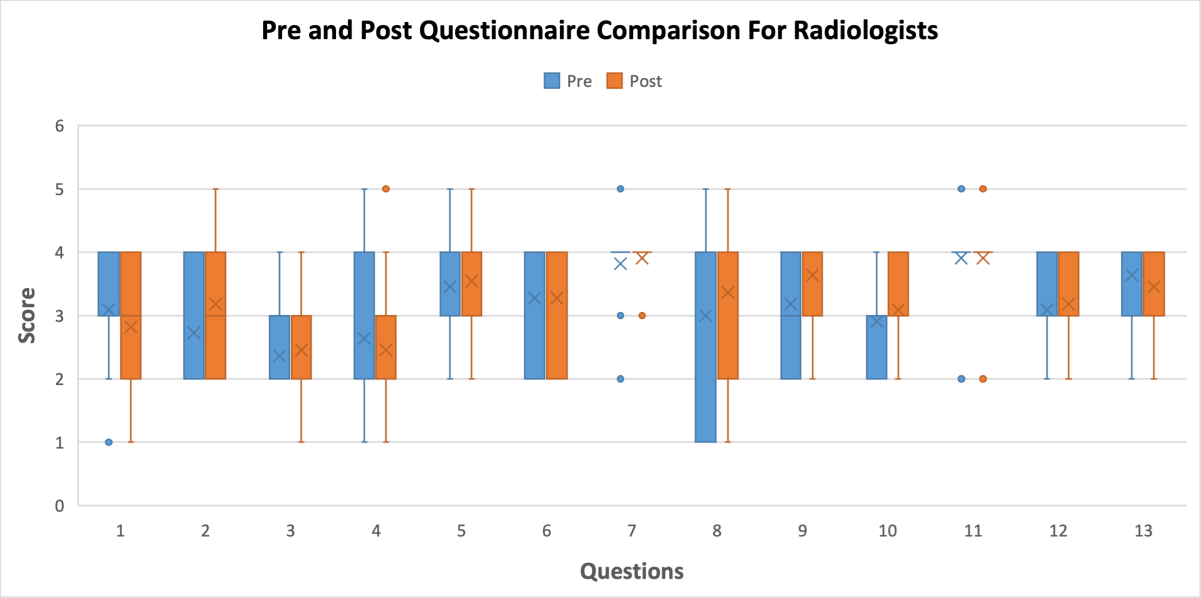


Figure 7.6: Comparison of Pre-Post responses given only by the radiologists.

Moreover, it was important to check if there is a significant change in perception. To check that, first a test of normality was performed to see if the data set is normally distributed, which is a prerequisite to a t test. Both the pre and post presentation dataset failed to pass the tests of normality because the p-value for both the Kolmogorov-Smirnov and Shapiro-Wilk tests was less than 0.05. Therefore, a non parametric, Mann-Whitney U test was conducted to see if their is a significant difference between the pre and post responses. The test gives an overall p-value of 0.171, which was greater than the commonly used threshold of 0.05 to reject the null hypothesis. This means, there was not enough statistical evidence to conclude a significant difference between the radiology professional's perception of AI in chest radiology before and after the presentation.

Looking at the evaluation of question wise shift in perception, for Question 1, concerning AI replacing radiologists, the median score remained at 3 pre- and post-presentation which indicates a neutral stance and supported by a p-value of 0.51, which suggests no significant shift. Similarly, Question 2, about the reduction of radiologists due to AI, also showed no change with a median score of 3 and a p-value of 0.18. The perception of AI devaluing the profession (Question 3) stayed stable with a median of 2 and a p-value of 0.47. In terms of AI outperforming radiologists (Question 4), the median remained at 2, and the p-value of 0.93 indicated no significant change. For AI's capability to alert abnormal findings (Question 5), the median score consistently stayed at 4, with a p-value of 0.66. Confidence in AI diagnostic suggestions (Question 6) also maintained a median of 4, with a p-value of 0.57, while the need for a dedicated AI task force (Question 7) had a stable median of 4 and a p-value of 0.90.

Moreover, attendance at AI task force meetings (Question 8) showed no shift, with a median of 4 and a p-value of 0.51. However, Question 9 showed a positive trend, as the median shifted from 3 to 4 which indicates increased interest in AI education tailored for chest radiology, though this was not statistically significant with a p-value of 0.12. Trust in AI tools (Question 10) remained at a median of 3, with a p-value of 0.78. The impact on jobs in the next 5 years (Question 11) and the impact on chest radiology jobs compared to other subspecialties (Question 12) both remained stable with median scores of 4 and 3 respectively, and p-values of 1 and 0.75. Finally, the potential of AI to address the radiologist shortage (Question 13) also showed no shift with a median of 4 and a p-value of 1. These results suggest a consistent perception among radiology professionals pre- and post-presentation with a notable positive trend in the interest for tailored educational tools.

No	Questions	Median Score	IQR (25-75%)
1	After reviewing the AI/ML-detected results in the slide presentation, do you feel more confident in the capabilities of AI/ML for chest X-ray interpretation?	3	1
2	To what extent do you believe the AI/ML-detected results align with your own interpretations of the chest X-rays shown in the presentation?	3	1
3	Did the presentation of AI/ML-detected results change your perception of the potential impact of AI/ML on the field of chest radiology?	2	2
4	Would you be more or less likely to collaborate with AI/ML systems in your daily practice after seeing the presentation results?	4	1

Table 7.6: Detail of the questions used to explicitly inquire the participants about the change in their perception after looking at the results presentation. The individual score of for each question is also provided.

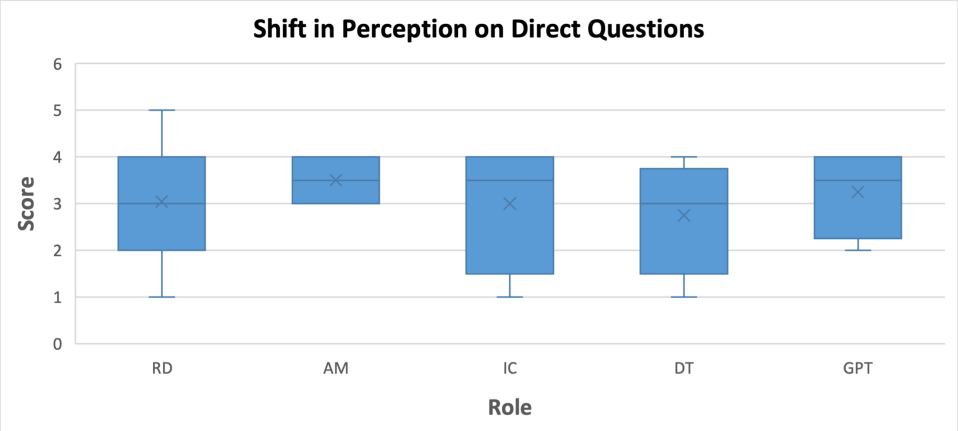


Figure 7.7: Change in perception of the radiology professionals after viewing the results presentation, evaluated using the four explicit questions.

Despite the lack of significance shift observed in the comparison of the 13 repeated questions, the analysis of four additional direct questions in the post-presentation questionnaire provides further insight into the shift in perception. The detail of those questions is given in the Table 7.6. For confidence in AI capabilities (Question 1), the median score was 3 with an IQR of 1 which suggests a neutral to positive shift in confidence after viewing the presentation. Regarding alignment with own interpretations (Question 2), the median score was 3 with an IQR of 1 which indicates moderate agreement that AI-detected results align with their own interpretations. The change in perception of AI impact (Question 3) had a median score of 2 with an IQR of 2, which reflects a slight change in perception regarding AI's potential impact on chest radiology. Lastly, for willingness to collaborate with AI (Question 4), the median score was 4 with an IQR of 1 which shows a positive shift in willingness to collaborate with AI systems in daily practice.

In conclusion, while the overall statistical analysis of the 13 repeated questions indicates no significant shift in perception, the responses to the direct questions highlight a more positive trend. Participants, particularly radiologists, showed increased confidence in the capabilities of AI, alignment of AI results with their own interpretations, and a greater willingness to collaborate with AI systems. These findings suggest that, despite initial familiarity with AI, exposure to AI performance can foster a more favorable perception which indicate a potential for increased acceptance and integration of AI tools in chest radiology. This positive trend, though not statistically significant in the broader analysis shows a shift that have been influenced by the results presentation.

No	Questions
1	How do you think the integration of AI/ML into chest radiology practice might affect your day-to-day work after witnessing the AI/ML results in the presentation?
2	In your opinion, what are the main strengths of the AI/ML model presented in the slides for detecting conditions in chest X-rays?
3	What limitations or challenges do you perceive in the AI/ML model's performance based on the results shown in the presentation?
4	Can you share any specific concerns or fears you have regarding the integration of AI/ML in chest radiology?
5	In your opinion, what steps could be taken to enhance the trust and usability of AI/ML models in chest radiology practice?

Table 7.7: Open ended questions from the post presentation questionnaire.

No	Responses	Identified Codes
1	Improve accuracy. Using model for triage could help to prioritize clinically important cases	Improve accuracy, Triage, Prioritization
2	Would alert me to the abnormality and help me tailor my report accordingly.	Alert abnormalities, Tailor report
3	Increase my speed and accuracy in Diagnosis.	Increase speed, Improve accuracy
4	Good patient care	Patient care
5	Will be useful to see if AI picks up things I do not notice like subtleties.	Detect subtleties
6	It might be of value when non-radiology doctors are analysing X-rays while waiting for the radiologist report. Also AI can highlight potential positive findings so these X-rays can be reported sooner.	Assist non-radiologists, Highlight findings, Prioritization
7	Will increase efficiency.	Increase efficiency
8	See interesting cases	Interesting cases
9	increase efficiency.	Increase efficiency
10	It would help clear the long list of unreported chest radiographs and highlight ones that need further imaging or investigations.	Clear backlog, Highlight findings

Table 7.8: The responses to first open ended question and the corresponding codes identified.

7.4.4 Analysis of Open Ended Questions

The post presentation questionnaire also included five open ended questions to give participants a chance to provide their inner thoughts and reactions towards the impact on workflow, strengths of the AI models, concerns and fears, enhancement in trust and usability, and limitations of the AI models. Table 7.7 shows all five open ended questions.

The analysis of the open-ended questions involved a systematic approach to make sure that the insights taken, were accurate and representative of the participant's views. This process started with the organization of all responses. Each response was read carefully and broken down into meaningful segments (codes), that captured the feeling of the feedback. These codes were then grouped into broader themes that reflected common ideas and sentiments across multiple responses. This method allowed the identification of patterns within text. By applying this technique, It was aimed to uncover the underlying attitudes, concerns, and suggestions of radiology professionals regarding the integration of AI/ML into chest radiology practice. The coding process was first done with one specific question to show how individual responses were categorized, followed by a comprehensive Table that shows the themes identified across all open-ended questions.

Table 7.8 shows the coding process of question 1 which is **'How do you think the integration of AI/ML into chest radiology practice might affect your day-to-day work after witnessing the AI/ML results in the presentation?'**. Moreover, Table 7.9 summarise the themes identified from the codes of question 1. Following the same process of coding, the themes for the rest of the questions were also identified and detailed in Table 7.10.

The responses to the open-ended questions as outlined in Table 7.11, reflect a cautious optimism

Codes	Themes
Improve accuracy, Detect subtleties	Accuracy
Increase speed, Increase efficiency, Clear backlog	Efficiency
Triage, Prioritization, Highlight findings	Prioritization
Patient care	Patient Care
Assist non-radiologists	Assistance
Tailor report	Reporting
Interesting cases	Engagement

Table 7.9: Codes and the corresponding themes for the responses of question 1.

Question	Themes
1	Accuracy, Efficiency, Prioritization, Patient Care, Assistance, Reporting, Engagement
2	Diagnostic accuracy, Speed, Assistance in decision-making, Learning, Sensitivity, Triage
3	Accuracy, Reproducibility, Data sufficiency, Clinical context, Trust, Deskilling, Overcalling
4	False positives, Deskilling, Income, Clinical context, No concerns, Bias, Role of radiologists
5	Research, Diversity in training sets, Licensing, Integration, Validation, Feedback, Focus

Table 7.10: Themes identified for all five open ended questions.

towards the integration of AI in radiological practices. For Question 1, participants mentioned the importance of improving accuracy, efficiency, and prioritization in diagnostics, along with aspects of patient care, assistance, reporting customization, and engagement. Responses to question 2 highlighted the need for AI to enhance identification, providing guidance, deliver quick results, facilitate learning, ensuring accurate diagnoses, and supporting triage. Question 3 showed concerns about the need for further improvement in AI models, accuracy issues, integration challenges, over calling, and differentiation of conditions. Moreover, question 4 mentioned the false positives, potential deskilling of radiologists, financial impacts, integration concerns, and the evolving role of radiologists, with some participants expressing no concerns. Finally, Question 5 responses emphasized the need for reliable research, practical integration, user feedback, accuracy confirmation, and a specific focus on early cancer detection and second reader AI reports. Overall, participants recognize the potential benefits of AI but stress the importance of addressing accuracy, integration, and practical implementation concerns.

7.5 Conclusion

The field of radiology is undergoing a transformation driven by advancements in artificial intelligence (AI). Chest radiographs are among the most frequently performed imaging modalities in clinical settings which plays an important role in diagnosing thoracic diseases and abnormalities. However, the increasing demand for radiologists has not been matched by a proportional increase in the number of trained professionals. This shortage places immense pressure on the existing workforce which leads to delays in diagnosis and increased workloads. AI offers a promising solution to this challenge by aiding in the interpretation of chest radiographs. However, the successful integration of AI into clinical practice requires not only technological advancements but also the acceptance and trust of radiologists who are the primary validators of these systems. Understanding their perceptions, concerns, and openness to AI is extremely important for developing AI tools that are both effective and trusted by healthcare professionals. This study was conducted to capture the current attitudes of radiology professionals towards AI, there shift in perception, and insights into how AI can be effectively integrated into the radiology workflow. By investigating them, the study aims to contribute to the broader effort of leveraging AI to improve healthcare delivery and patient outcomes.

This study delved into the perceptions of radiology professionals, particularly radiologists, regarding

Question	Theme	Coded Responses	Total
Q1	Accuracy	- Improve accuracy (Response 1) - Increase accuracy in Diagnosis (Response 3) - Detect subtleties (Response 5)	3
	Efficiency	- Increase speed (Response 3) - Increase efficiency (Responses 7, 9) - Clear backlog (Response 10)	4
	Prioritization	- Triage (Response 1) - Prioritization (Responses 1, 6) - Highlight findings (Responses 6, 10)	5
	Patient Care	- Good patient care (Response 4)	1
	Assistance	- Assist non-radiologists (Response 6)	1
	Reporting	- Tailor report (Response 2)	1
	Engagement	- Interesting cases (Response 8)	1
Q2	Identification	- Identify cases without pathology (Response 1) - Detect obvious conditions (Response 1) - High sensitivity (Response 8) - Identifying abnormal radiographs (Response 9)	4
	Guidance	- Point in the right direction (Response 2) - Helping radiologists to be more accurate and faster (Response 3)	2
	Speed	- Quick results (Response 4)	1
	Learning	- Ability to learn from each X-ray (Response 5)	1
	Accuracy	- Diagnosing major findings with good accuracy (Response 6)	1
	Triage	- Triage (Response 7)	1
Q3	Improvement Needed	- Further improvement of model metrics (Response 1) - Reproducibility in a clinical setting (Response 2)	2
	Accuracy Concerns	- Accuracy and sufficiency of training data (Response 3) - Poor specificity (Response 8)	2
	Integration Challenges	- Relating XR findings to patient symptoms (Response 4) - Deskilling clinicians (Response 5) - Biasing reports (Response 7)	3
	Overcalling	- Overcalling and unnecessary investigations (Response 6)	1
	Differentiation Issues	- Differentiating similar conditions (Response 9)	1
Q4	False Positives	- Increasing FP due to model suggestions (Response 1) - AI bias in reports (Response 7)	2
	Deskilling	- De-skilling of junior/less experienced radiologists (Response 2)	1
	Financial Impact	- Decreasing radiologist's income (Response 3)	1
	Integration Concerns	- AI not correlating with patient symptoms (Response 4) - Hinderance rather than help (Response 1)	2
	Role of Radiologists	- Radiologists reviewing AI reports (Response 8)	1
	No Concerns	- No concerns (Responses 5, 6)	2
Q5	Research and Validation	- More reliable research (Response 1) - Diverse training sets (Response 2)	2
	Practical Integration	- Integration in daily practice (Response 4) - Free licensing for evaluation (Response 3)	2
	User Feedback	- Talk to end users and address concerns (Response 6)	1
	Accuracy Confirmation	- Initial phase of radiologist reports (Response 5)	1
	Specific Focus	- Focus on early cancer detection (Response 8) - Second reader AI reports (Response 7)	2

Table 7.11: The number of responses corresponding to each of the themes identified for all five open ended questions.

the integration of AI models into chest radiograph interpretation. The initial overall perception of radiologists towards AI in chest radiology was generally positive. Participants recognized AI's potential to enhance diagnostic accuracy, increase efficiency, and address the shortage of radiologists. The median scores for questions related to trust and confidence in AI capabilities indicated a balanced view, with a cautious optimism about AI's role in the future of radiology. Despite this optimism, there was a consistent theme of skepticism about AI's ability to fully replace human radiologists or outperform them within the next decade. This skepticism shows the importance of human expertise in interpreting complex medical images and making complex clinical decisions. Moreover, while examining the role-specific trust levels, it was evident that radiologists shared similar views with the overall perception but exhibited slightly more variability in their responses. This suggests a range of opinions within the radiology community and reflects both hope and caution. Radiologists, given their extensive experience and direct interaction with chest radiographs, provide an important perspective on the potential integration of AI. Their insights are invaluable in understanding the practical implications of AI tools and their readiness to adopt such technologies in their daily workflow.

Furthermore, the analysis of the shift in perception using pre- and post- questionnaires revealed that the results presentation on AI capabilities did not result in a statistically significant shift in attitudes. This could be attributed to the participant's pre-existing familiarity with AI technologies and their well-formed opinions on the subject. Therefore, the hypothesis **Hyp 04**, that the perception of radiologists towards AI in radiology will significantly improve after being exposed to the AI models diagnostic capabilities is rejected. However, a positive trend was noted both in the pre-post and the direct questions in post-questionnaire, which indicates and increased confidence in AI capabilities and a greater willingness to collaborate with AI systems in daily practice. This suggests that while the presentation may not have dramatically shifted perceptions, it reinforced and enhanced the existing positive views towards AI.

In addition to that, the open-ended responses provided deeper insights into the participant's views. Key themes identified including the potential for AI to improve accuracy and efficiency, aid in the prioritization of cases, and assist non-radiologists in preliminary interpretations. However, concerns were raised about the clinical context, potential for deskilling of new radiologists, risk of misclassification, and ethical implications of AI integration. These responses shows the importance of addressing these concerns to enhance the trust and usability of AI models in clinical practice. This feedback shows a blend of optimism and caution and emphasize the need for AI to complement rather than replace human expertise. Despite these valuable insights, this study had several limitations that should be acknowledged. Firstly, the sample size was relatively small, with only 15 participants in total, of which 11 were radiologists. This limited sample size may not fully capture the diversity of opinions within the broader radiology community.

Additionally, the participants already exhibited a reasonably good attitude towards AI, which may have contributed to the lack of significant shift in perception. The study's findings, while informative, should be interpreted with caution and may not be generalizable to all radiology professionals. Future studies should aim to include a larger and more diverse sample to capture a broader spectrum of views. Moreover, To build a more comprehensive understanding of radiologist's perceptions of AI, engaging radiologists from different regions, with varying levels of experience and exposure to AI technologies, would provide a more representative views. Additionally, longitudinal studies that track changes in perception over time as AI technologies evolve and become more integrated into clinical practice would be extremely valuable. Such studies could provide deeper understanding into how ongoing advancements and increased familiarity with AI influence perceptions and acceptance among radiologists. Another key takeaway from this study could be, the specific concerns raised by radiologists, to put efforts on addressing the challenge of transparency and explainability of AI models. Collaborative efforts between AI developers, radiologists, and regulatory bodies could be crucial in achieving these goals. By addressing these concerns through focused research and development can mitigate fears and enhance the practical benefits of AI..

Chapter 8

Discussion and Conclusions

8.1 Discussion

This thesis investigates new strategies to enhance the performance of deep learning models for multi-label classification of chest radiographs, with a particular focus on the CheXpert dataset. The research addresses several important challenges in medical imaging, which includes, the handling of uncertain labels, the impact of different radiograph projections, and the efficacy of custom pooling layer. Additionally, this research explores the Sequential Multi-Label Enrichment (SMLE) DenseNet model for multiple coexisting conditions detection and examine the perception of radiology professionals towards the integration of AI models in clinical practice. The scope of this thesis covers both technical advancements in AI methodologies and the human factors influencing the acceptance and trust of AI in healthcare aiming to bridge the gap between technological potential and practical application.

Based on the findings of previous studies as well as this research, it is evident that there is a substantial potential for AI in medical imaging, particularly in chest radiology [27, 70, 86]. The deep learning research community has long been dedicated to improving the efficiency of these models. Efforts have included increasing the depth of the models, reducing the number of trainable parameters for quicker training and inference by employing various feature extraction techniques, and developing reliable loss functions, among other advancements. While this focus on improving model architecture has yielded significant success in many applications, such as detecting conditions on radiographs, recent years have not seen a major breakthrough in deep learning model architectures, despite the availability of extreme computational power. This stagnation has highlighted the need for a paradigm shift from model-centric methods to data-centric methods in machine learning. Data-centric methods emphasize the importance of preprocessing data effectively to improve data quality, rather than solely relying on new deeper and wider model architectures. This shift is particularly important for the medical imaging domain, where there is often a shortage of high-quality data. By focusing on data quality, the performance and reliability of AI systems in medical imaging can be enhanced. High-quality data is essential for training robust AI models that can generalize well to new, unseen data, thereby improve diagnostic accuracy and patient outcomes.

8.1.1 Model-Centric Methods

This work has employed several model-centric methods to enhance the performance of deep learning models for chest radiograph interpretation. These include transfer learning, a weighted loss function, a custom pooling layer, a sequential multi-stage training procedure and model enhancement.

8.1.1.1 Interpretation of Findings

The use of transfer learning has allowed the models to leverage basic image features learned from other datasets, which aids in developing a more complex understanding of chest radiographs. This approach has shown improved performance in detecting thoracic diseases, though its effectiveness is limited due to the pre-trained models being trained on non medical imaging tasks. The weighted loss function has been instrumental in balancing the overall loss by penalizing classes with more samples and guide the optimization algorithm more accurately. This method has addressed class imbalance to an extent, although challenges remain with extremely imbalanced classes, specially in coexisting conditions.

The custom pooling layer, which captures both low and high-intensity features using min and max pooling operations, has enhanced feature extraction capabilities on rdaiographs and results in better classification outcomes. However, this approach increases computational complexity which leads to longer training times and higher resource consumption. The SMLE-DenseNet approach which is a multi-stage model training method combined with an extended DenseNet121 architecture, has significantly improved the model's ability to detect multiple co-existing conditions and demonstrated the efficacy of multi stage training procedures.

8.1.1.2 Comparison with Existing Work

Compared to the previous studies, this research introduced two novel model-centric methods for chest radiograph interpretation. The custom pooling layer and the SMLE approach, which showed significant improvements over the baseline state of the art architecture in chest radiograph classification tasks. Previous research has primarily focused on evaluating model performance by assessing both single and co existing conditions combined. In contrast, through SMLE-DenseNet, this study addresses the complexities of co-existing conditions separately which provides a more comprehensive and realistic approach to chest radiograph interpretation. Similarly, previous studies utilising CheXpert dataset, mainly relied on the off the shelf available pooling layers which pools the radiographic features similar to any other non medical image, this study developed a task specific custom pooling operation. While other studies have employed transfer learning and weighted loss functions, their combination with custom pooling and SMLE in this work, provides a unique advantage in handling the multi-label classification challenges in chest radiographs.

8.1.1.3 Implications for Theory and Practice

The findings of this research have substantial implications for both theory and practice in medical imaging and AI:

- The novel model-centric methodologies proposed, such as the custom pooling layer and the SMLE approach, can serve as a foundation for future studies aiming to improve AI diagnostic tools for medical imaging.
- Practically, these advancements can lead to more accurate and reliable diagnostic support tools in clinical settings and improve the workflow and decision-making of radiologists.

8.1.1.4 Limitations

Despite the promising results shown, several limitations of the model-centric methods were also identified:

- While transfer learning helps the model's performance, it is limited by the nature of the pre-trained models. The features learned from non-medical datasets may not always perfectly transfer to medical imaging tasks which could leads to a suboptimal performance in some cases.

- Although the weighted loss function helps in balancing class distributions, it may not entirely eliminate the impact of class imbalance. Extremely imbalanced classes may still pose challenges and affect the model's ability to generalize. Specially, in cases with rare multi co existing conditions.
- The custom pooling layer, while effective in capturing diverse features, increases the computational complexity of the model. This can lead to longer training times and higher resource consumption, which may not be feasible in all settings. Such as running them on smart devices.
- The SMLE approach, though effective, requires a carefully curated dataset with multiple co-existing conditions. This may limit its applicability in scenarios where such datasets are not available. Such as the cases where all five conditions are simultaneously present on a single radiograph.

8.1.1.5 Future Work

Future work can address these limitations and further enhance the model-centric methods:

- Future research could explore more advanced transfer learning techniques, such as domain adaptation, to better tailor pre-trained models to medical imaging tasks. This could involve using larger and more diverse medical datasets for pre-training. Targeted transfer learning can be employed by pre-training the model on a specialised dataset to specifically learn targeted features (e.g., bluntness of the costophrenic angle).
- Developing more sophisticated loss functions that dynamically adjust weights during training could further mitigate the impact of class imbalance. Additionally, incorporating focal loss or other advanced techniques might help in dealing with extremely imbalanced datasets.
- To reduce computational complexity, future work could focus on optimizing custom pooling layer. This could involve developing intelligent pooling strategies that balance feature capture with computational efficiency. A possible approach could be a multi stage pooling, that creates a single feature map out of the two created by the custom pooling layer proposed in this work. This will keep the computational requirement similar to the standard architecture.
- Making the SMLE approach more broadly applicable. Future research could investigate new methods for feature wise staged training, based on the hierarchies present among the conditions. Additionally, exploring the integration of SMLE with other model architectures could further enhance its robustness and versatility.
- Validating these models in real-world clinical settings is crucial. Collaborating with medical institutions to test and refine the models. Finally, working on the explainability of the models to make sure that they are not giving the right answer for the wrong reason.

8.1.2 Data-Centric Methods

Data-centric methods emphasize the importance of high-quality data in enhancing the performance and reliability of deep learning models. In the context of chest radiograph interpretation, data-centric strategies involve image data preprocessing, careful handling of uncertain labels, and techniques to augment and standardize the dataset. This research highlighted the crucial role of data preparation by employing methods such as GMMs for relabeling uncertain labels and multi-scale template matching to focus on the relevant thoracic regions. These techniques aim to improve the quality of the training data to ensure that the models are trained on the most accurate version of the data. This is important for developing robust and reliable AI systems, especially in medical imaging where patient lives are at risk. Moreover, this study explored various data augmentation techniques to address limitations posed by data variability. By generating transformed versions of training samples that mimic real-world variations, these techniques help create a more diverse dataset which could enhance the model's generalization

ability. Additionally, this research investigated the impact of different radiograph projections on model performance which shows the importance of data-centric methods and the need for ongoing innovation in data preparation and management for chest radiograph classification.

8.1.2.1 Interpretation of Findings

The findings demonstrate that data-centric strategies significantly enhance model performance. For example, relabeling uncertain labels using GMM improved the overall model performance and highlight the importance of addressing label uncertainty in CheXpert dataset. Moreover, the Multi-scale template matching approach crops and resize the CheXpert radiographs and center it on the thoracic regions which improves feature extraction and ultimately classification performance. The promising results indicate that high-quality, well-prepared data is essential for training robust and effective DL models for chest radiograph classification tasks.

8.1.2.2 Comparison with Existing Work

Compared to existing work, this research introduced new data-centric techniques that showed improvements in chest radiograph classification. Previous studies often overlooked the impact of uncertain labels or handle them in a very simple way by assigning them positive or negative labels. This study has paid special attention to the uncertain labels in the CheXpert dataset and thoroughly investigated the possible ways to relabel and effectively use them for training DL models. This study has uniquely leveraged a GMM based semi supervised approach to relabel the uncertain labels which proved effective. Moreover, while very few of the previous studies explored the significance of precise region-focused CheXpert training data. By employing an improvised multi scale template matching mechanism to preprocess the CheXpert dataset, this work has demonstrated classification performance improvement. While other studies have explored data augmentation, the combination of GMM relabeling and multi-scale template matching presented in this work provided a unique advantage in handling the complexities of chest radiograph classification.

8.1.2.3 Implications for Theory and Practice

The findings from this research have important implications for both theory and practice:

- The new data-centric methodologies proposed can serve as a foundation for future research on uncertain label handling and precise cropping of chest radiographs.
- Practically, these advancements can lead to the development of more accurate and reliable data preprocessing tools that improve the data quality for medical imaging applications.

8.1.2.4 Limitations

The data-centric methods employed in this work showed effectiveness but also presented several limitations:

- The CheXpert dataset utilized in this research comprises radiographs from a single hospital which limits the generalizability of the models to broader populations.
- Handling uncertain labels remains a significant challenge. Despite employing GMM for relabeling, there is no absolute guarantee that the relabeling perfectly reflects the true condition. Mislabeling can still propagate errors and affect model performance.
- The data imbalance in a multi-label setting remains a problem. It is extremely difficult to perfectly balance all co-existing conditions.

- The employed multi-scale template matching technique sometimes crops the image too tightly by either cutting off the upper or side regions of the lungs, which could adversely affect model performance.

8.1.2.5 Future Work

Future research could aim to address these limitations to enhance the effectiveness of data-centric methods:

- Collaborating with multiple medical institutions to collect a more diverse set of chest radiographs. This would involve creating a multi-institutional dataset that includes radiographs from various hospitals, regions, and demographic groups. Incorporating data from different sources would enhance the generalizability and robustness of the models. Additionally, cross-institutional studies could validate the models across different clinical settings.
- Exploring more sophisticated labeling techniques to improve the accuracy and reliability of uncertain label handling. Employing the latest NLP tools to extract more accurate labels from radiology reports could enhance label quality. Involving radiologists in the relabeling process by providing them a random set of relabeled samples to measure the effectiveness of the labeller could further improve robustness and overall model performance.
- To mitigate data imbalance for co-existing conditions, generative models like GANs could be used to create realistic radiograph images for rare combinations of co-existing conditions.
- Developing adaptive and context-aware cropping techniques to ensure that the entire relevant region of the lungs is preserved while preprocessing CheXpert radiographs. One approach could be using machine learning models to identify and delineate the boundaries of the lungs more accurately before cropping. Training segmentation models specifically to recognize the anatomical structure of the lungs and other thoracic regions could achieve this.

8.1.3 Perception Study

This research also examines the social implications of using deep learning models for chest radiograph interpretation by focusing on how radiology professionals perceive the integration of these models into their clinical workflow. This study aims to bridge the gap between technological advancements and their practical acceptance in healthcare settings.

8.1.3.1 Interpretation of Findings

The findings indicate a general positive trend in radiologist's perceptions towards AI integration after being exposed to the capabilities of deep learning models in identifying conditions on chest radiographs. The study involved a three-stage approach: an initial questionnaire to gauge general perceptions of AI in radiology, a presentation which shows the model results, and a follow-up questionnaire to capture any shifts in perception. Although the shift in perception was not significant, this may be attributed to the limited number of responses at the time of writing this thesis. Nevertheless, the trends suggest a growing readiness among radiologists to embrace AI as a valuable tool in clinical practice.

8.1.3.2 Comparison with Existing Work

Compared to existing literature, which generally captures a broad view of radiologist's attitudes towards AI, this study specifically addressed the change in perception after direct interaction with AI model outputs. Previous studies have largely focused on theoretical acceptance, whereas this research provides practical insights by involving radiologists in hands-on evaluation of AI capabilities by comparing their interpretation with AI model. This approach provides a more deep understanding of the factors influencing acceptance and trust on AI in radiology.

8.1.3.3 Implications for Theory and Practice

The study's findings have important implications for both theory and practice:

- The positive trend in perception towards AI suggests that continued exposure and education about AI capabilities can enhance acceptance among radiologists.
- Practically, the integration of AI tools in clinical settings would require addressing transparency and explainability to build trust among users.
- The study shows the need for involving end-users in the development and evaluation of AI tools to ensure they meet clinical needs and expectations.

8.1.3.4 Limitations

Several limitations were identified in the perception study conducted in this research:

- First of all, the limited number of responses may not provide a comprehensive view of the broader radiology community's perceptions.
- As there is no standard approach to quantify the perception of medical imaging community. The study's design may not have fully captured the nuances of radiologist's concerns and expectations regarding AI integration.
- There was no longitudinal follow-up to assess how perceptions might evolve over a more extended period.

8.1.3.5 Future Work

Future research could address these limitations and further explore the following areas:

- Expanding the study to include a larger and more diverse group of radiologists to provide a more comprehensive understanding of perceptions across different demographics and regions.
- Developing an AI tool and give access to radiologists to use it and conducting a system usability scale SUS survey after some time.
- Conducting longitudinal studies to track changes in perceptions over time and after extended use of AI tools in clinical practice.
- Exploring more in-depth qualitative methods, such as interviews or focus groups, to capture the detailed views and specific concerns of radiologists regarding AI integration.
- Investigating the impact of educational interventions and hands-on training sessions on radiologist's acceptance and trust in AI tools.
- Developing strategies to improve the transparency and explainability of AI models to address potential concerns and build trust among radiologists.

8.2 Conclusions

This research has made contributions to the field of medical imaging and artificial intelligence, particularly in the context of chest radiograph interpretation. The growing need of trained radiologists and the high number of diagnostic errors serves as the main motivation of this work. By addressing the challenge of uncertain labels in the CheXpert dataset, the study introduced advanced techniques such as relabeling using GMM and demonstrated their positive impact on model performance. Additionally, the exploration

of the effects of different radiograph projections (AP, PA, and lateral) provided valuable insights into the strengths and limitations of each view to guide the development of more robust and accurate diagnostic models. The introduction of a custom pooling layer which combines the max and min pooling showcased a new approach to enhance feature extraction for chest radiographs in CNN based networks, which results in improved classification performance.

The Sequential Multi-Label Enrichment (SMLE) approach was another key contribution which offers a structured method to progressively train model on increasingly complex multi-label datasets. This technique improved the model's ability to detect multiple coexisting conditions in chest radiographs. Furthermore, the assessment of radiologist's perceptions of AI integration provided essential feedback on the practical applicability and acceptance of AI tools in clinical settings. While the primary focus of this research is on enhancing AI-based interpretation of chest radiographs, the techniques developed have broader applications. The methods are designed with general principles that can be adapted to other similar problems in computer vision, which extends the impact of this research to a wider range of applications.

8.2.1 Summary of Key Findings

Section 3.3.2 provides a detailed explanation of the significant number of uncertain labels present in the CheXpert dataset. This issue arises because these labels were extracted from radiograph reports using an NLP system [9]. Previous studies have employed various strategies to address these uncertain labels for model training. This thesis proposes a novel method of relabeling these samples using a semi-supervised learning approach with Gaussian Mixture Models (GMM). Using this approach, samples with definite (1/0) labels were first plotted which revealed multiple clusters per condition. This clustering occurs because GMM assumes that the data is generated from a mixture of several Gaussian distributions and each distribution represents a cluster. The samples with uncertain labels were then relabeled by mapping them to these clusters. When these relabeled samples were included in the training set, the model's performance showed significant improvement compared to a model that excluded these samples. Chapter 4 details the experiments, and these results were also published in a conference paper [86]. This addresses the research question **RQ1**.

Furthermore, Section 3.4.3 of this thesis discusses the limitations of the max pooling and min pooling layers in the context of radiograph classification. Previous studies on chest radiograph classification have predominantly utilized the max pooling layer. This thesis proposes a novel pooling layer for chest radiograph classification, which combines max pooling and min pooling to capture both low and high-intensity important features in a radiograph. The effectiveness of this custom pooling layer was evaluated by comparing the performance of a standard DenseNet121 model with that of a custom pooled DenseNet121 model. The results demonstrate that the custom pooled DenseNet121 outperforms the standard DenseNet121 model. To further validate these findings, both the standard and custom pooled models were trained on the standard dataset as well as the multiscale template-matched CheXpert dataset. Chapter 5 provides a detailed explanation of these results which highlight the efficacy of the custom pooling layer in effectively detecting significant features in chest rdaiograph classification. This addresses the research question **RQ2**.

In addition to that, Section 3.4.4 of this thesis addresses a gap in previous studies on chest radiograph interpretation using the CheXpert dataset, specifically the lack of evaluations on the performance of deep learning models in accurately identifying multiple coexisting conditions in a single radiograph. This research not only fills this gap by evaluating the model's performance in detecting two, three, four, and five coexisting conditions in a single radiograph from the CheXpert dataset but also proposes a new model architecture and training procedure named SMLE (Sequential Multi Label Enrichment) to improve the performance. The SMLE approach integrates a DenseNet121 model with extended layers and employs

a staged training procedure, where each stage is trained on a subset of the data. Chapter 6 provides detailed explanations of the experiments and compare the performance of the standard DenseNet121 and the SMLE models in detecting multiple coexisting conditions. The results indicate that the SMLE approach shows a significant improvement in the model's ability to detect multiple coexisting conditions. This addresses the research question **RQ3**.

Lastly, this thesis also examines the social implications of using deep learning models for chest radiograph interpretation by focusing on how radiology professionals feel about integrating these models into their clinical workflow. While existing studies capture the overall perception of radiology professionals towards AI in radiology, none have specifically addressed the shift in perception after exposure to the capabilities of deep learning models in identifying conditions on chest radiographs. This research involved a three-stage study, an initial questionnaire to gauge participant's general perceptions of AI in radiology, a presentation to showcase the model's results and allowing participants to compare their interpretations with those of the model, and a follow-up post presentation questionnaire to capture any shifts in perception. Chapter 7 provides a detailed explanation of this study. The results indicate a shift in perception, however, it was not significant, potentially due to the limited number of responses at the time of writing this thesis. This addresses the research question **RQ4**.

8.2.2 Contributions to the Field

In terms of contributions, this research has provided several key advancements. the development of a novel method for handling uncertain labels using GMM, significantly improved models performance. The introduction of a custom pooling layer combining max and min pooling to enhance feature extraction in CNN-based networks. The proposal of the SMLE approach for better detection of multiple coexisting conditions in chest radiographs, demonstrates improved model performance. And finally insights into radiologist's perceptions of AI integration, which highlights the practical acceptance and readiness for AI tools in clinical practice. These contributions not only advance the technical aspects of AI in medical imaging but also address the human factors influencing the adoption of these technologies.

8.2.3 Final Thoughts

Reflecting on the journey of this research, it is evident that the integration of AI in medical imaging have immense potential as well as desperate need for transforming healthcare. The positive shift trend in radiologist's perceptions highlights the growing acceptance and readiness within the medical community to embrace AI as a valuable assistant in clinical practice. This research not only addresses current challenges but also paves the way for continuous improvements and innovations in chest radiology and potentially medical imaging as a whole. As AI technologies continue to evolve, they will increasingly become indispensable tools in the medical field to augment the capabilities of healthcare professionals and take the quality of patient care to next level. Embracing these technologies with a collaborative and open mindset will be key to realizing its full potential and ensuring that future advancements in AI continue to benefit both clinicians and patients alike.

Bibliography

[1] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[2] RK.MD, "Reading chest x-rays." `https://rk.md/2017/reading-chest-x-rays/`. Accessed on 2021-07-05.

[3] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[5] Medium.com, "Everything you need to know about vgg16." `https://medium.com/@mygreatlearning/everything-you-need-to-know-about-vgg16-7315defb5918`. Accessed on 2024-01-15.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[8] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

[9] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 590–597, 2019.

[10] R. Gupta, D. Srivastava, M. Sahu, S. Tiwari, R. K. Ambasta, and P. Kumar, "Artificial intelligence to deep learning: machine intelligence approach for drug discovery," *Molecular diversity*, vol. 25, pp. 1315–1360, 2021.

[11] A. İ. Tekkeşin *et al.*, "Artificial intelligence in healthcare: past, present and future," *Anatol J Cardiol*, vol. 22, no. Suppl 2, pp. 8–9, 2019.

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[13] NHS, "Diagnostic imaging dataset annual statistical release 2019/20." `https://www.england.nhs.uk/statistics/wp-content/uploads/sites/2/2020/10/Annual-Statistical-Release-2019-20-PDF-1.4MB.pdf`. Accessed on 2022-05-03.

[14] NHS, "Respiratory disease." https://www.england.nhs.uk/ourwork/clinical-policy/respiratory-disease/#:~:text=About%20respiratory%20disease,the%20biggest%20causes%20of%20death. Accessed on 2022-05-03.

[15] Statista, "Respiratory disease in the united kingdom (uk) - statistics & facts." https://www.statista.com/topics/5908/respiratory-disease-in-the-uk/#topicOverview. Accessed on 2023-12-02.

[16] T. R. C. of Radiologists, "Clinical radiology census reports." https://www.rcr.ac.uk/media/3gjdr23o/clinical_radiology_census_report_2020.pdf. Accessed on 2022-01-15.

[17] M. A. Bruno, E. A. Walker, and H. H. Abujudeh, "Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction," *Radiographics*, vol. 35, no. 6, pp. 1668–1676, 2015.

[18] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[19] P. Lakhani and B. Sundaram, "Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology*, vol. 284, no. 2, pp. 574–582, 2017.

[20] S. Bharati, P. Podder, and M. R. H. Mondal, "Hybrid deep learning for detecting lung diseases from x-ray images," *Informatics in Medicine Unlocked*, vol. 20, p. 100391, 2020.

[21] T. Rahman, M. E. Chowdhury, A. Khandakar, K. R. Islam, K. F. Islam, Z. B. Mahbub, M. A. Kadir, and S. Kashem, "Transfer learning with deep convolutional neural network (cnn) for pneumonia detection using chest x-ray," *Applied Sciences*, vol. 10, no. 9, p. 3233, 2020.

[22] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[23] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106, 2017.

[24] Y. LeCun *et al.*, "Generalization and network design strategies," *Connectionism in perspective*, vol. 19, no. 143-155, p. 18, 1989.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[27] H. H. Pham, T. T. Le, D. Q. Tran, D. T. Ngo, and H. Q. Nguyen, "Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels," *Neurocomputing*, vol. 437, pp. 186–194, 2021.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[29] G. Kumar, N. Sharma, and A. Paul, "An extremely lightweight cnn model for the diagnosis of chest radiographs in resource-constrained environments," *Medical Physics*, vol. 50, no. 12, pp. 7568–7578, 2023.

[30] Q. Li, Y. Lai, M. J. Adamu, L. Qu, J. Nie, and W. Nie, "Multi-level residual feature fusion network for thoracic disease classification in chest x-ray images," *IEEE Access*, vol. 11, pp. 40988–41002, 2023.

[31] H. N. Saleem, U. U. Sheikh, and S. A. Khalid, "Classification of chest diseases from x-ray images on the chexpert dataset," in *Innovations in Electrical and Electronic Engineering: Proceedings of ICEEE 2021*, pp. 837–850, Springer, 2021.

[32] G.-H. Huang, Q.-J. Fu, M.-Z. Gu, N.-H. Lu, K.-Y. Liu, and T.-B. Chen, "Deep transfer learning for the multilabel classification of chest x-ray images," *Diagnostics*, vol. 12, no. 6, p. 1457, 2022.

[33] K. K. Bressem, L. C. Adams, C. Erxleben, B. Hamm, S. M. Niehues, and J. L. Vahldiek, "Comparing different deep learning architectures for classification of chest radiographs," *Scientific reports*, vol. 10, no. 1, p. 13590, 2020.

[34] J. van Hoek, A. Huber, A. Leichtle, K. Härmä, D. Hilt, H. von Tengg-Kobligk, J. Heverhagen, and A. Poellinger, "A survey on the future of radiology among radiologists, medical students and surgeons: students and surgeons tend to be more skeptical about artificial intelligence and radiologists may fear that other disciplines take over," *European journal of radiology*, vol. 121, p. 108742, 2019.

[35] M. Chen, B. Zhang, Z. Cai, S. Seery, M. J. Gonzalez, N. M. Ali, R. Ren, Y. Qiao, P. Xue, and Y. Jiang, "Acceptance of clinical artificial intelligence among physicians and medical students: a systematic review with cross-sectional survey," *Frontiers in medicine*, vol. 9, p. 990604, 2022.

[36] Q. Waymel, S. Badr, X. Demondion, A. Cotten, and T. Jacques, "Impact of the rise of artificial intelligence in radiology: what do radiologists think?," *Diagnostic and interventional imaging*, vol. 100, no. 6, pp. 327–336, 2019.

[37] P. Kasetti and R. Botchu, "The impact of artificial intelligence in radiology: as perceived by medical students," , vol. 10, no. 4, pp. 179–185, 2020.

[38] G. I. G. Brandes, G. DIppolito, A. G. Azzolini, and G. Meirelles, "Impact of artificial intelligence on the choice of radiology as a specialty by medical students from the city of são paulo," *Radiologia brasileira*, vol. 53, pp. 167–170, 2020.

[39] A. Bin Dahmash, M. Alabdulkareem, A. Alfutais, A. M. Kamel, F. Alkholaiwi, S. Alshehri, Y. Al Zahrani, and M. Almoaiqel, "Artificial intelligence in radiology: does it impact medical students preference for radiology as their future career?," *BJR| Open*, vol. 2, no. 1, p. 20200037, 2020.

[40] A. E. Eltorai, A. K. Bratt, and H. H. Guo, "Thoracic radiologists versus computer scientists perspectives on the future of artificial intelligence in radiology," *Journal of Thoracic Imaging*, vol. 35, no. 4, pp. 255–259, 2020.

[41] F. Coppola, L. Faggioni, D. Regge, A. Giovagnoni, R. Golfieri, C. Bibbolino, V. Miele, E. Neri, and R. Grassi, "Artificial intelligence: radiologists expectations and opinions gleaned from a nationwide online survey," *La radiologia medica*, vol. 126, pp. 63–71, 2021.

[42] A. A. Qurashi, R. K. Alanazi, Y. M. Alhazmi, A. S. Almohammadi, W. M. Alsharif, and K. M. Alshamrani, "Saudi radiology personnels perceptions of artificial intelligence implementation: a cross-sectional study," *Journal of Multidisciplinary Healthcare*, pp. 3225–3231, 2021.

[43] I. Yurdaisik and S. Aksoy, "Evaluation of knowledge and attitudes of radiology department workers about artificial intelligence," *Ann Clin Anal Med*, vol. 12, pp. 186–90, 2021.

[44] Y. Chen, C. Stavropoulou, R. Narasinkan, A. Baker, and H. Scarbrough, "Professionals responses to the introduction of ai innovations in radiology and their implications for future adoption: a qualitative study," *BMC Health Services Research*, vol. 21, pp. 1–9, 2021.

[45] M. Huisman, E. Ranschaert, W. Parker, D. Mastrodicasa, M. Koci, D. Pinto de Santos, F. Coppola, S. Morozov, M. Zins, C. Bohyn, *et al.*, "An international survey on ai in radiology in 1,041 radiologists and radiology residents part 1: fear of replacement, knowledge, and attitude," *European radiology*, vol. 31, pp. 7058–7066, 2021.

[46] M. Huisman, E. Ranschaert, W. Parker, D. Mastrodicasa, M. Koci, D. Pinto de Santos, F. Coppola, S. Morozov, M. Zins, C. Bohyn, *et al.*, "An international survey on ai in radiology in 1041 radiologists and radiology residents part 2: expectations, hurdles to implementation, and education," *European Radiology*, vol. 31, no. 11, pp. 8797–8806, 2021.

[47] S. S. Lim, T. D. Phan, M. Law, G. S. Goh, H. K. Moriarty, M. W. Lukies, T. Joseph, and W. Clements, "Non-radiologist perception of the use of artificial intelligence (ai) in diagnostic medical imaging reports," *Journal of Medical Imaging and Radiation Oncology*, vol. 66, no. 8, pp. 1029–1034, 2022.

[48] M. A. Khafaji, M. A. Safhi, R. H. Albadawi, S. O. Al-Amoudi, S. S. Shehata, and F. Toonsi, "Artificial intelligence in radiology: Are saudi residents ready, prepared, and knowledgeable?," *Saudi Medical Journal*, vol. 43, no. 1, p. 53, 2022.

[49] M. M. Abuzaid, W. Elshami, H. Tekin, and B. Issa, "Assessment of the willingness of radiologists and radiographers to accept the integration of artificial intelligence into radiology practice," *Academic Radiology*, vol. 29, no. 1, pp. 87–94, 2022.

[50] S. C. Shelmerdine, K. Rosendahl, and O. J. Arthurs, "Artificial intelligence in paediatric radiology: international survey of health care professionals opinions," *Pediatric Radiology*, pp. 1–12, 2022.

[51] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[52] N. Howard, N. Chouikhi, A. Adeel, K. Dial, A. Howard, and A. Hussain, "Brainos: a novel artificial brain-alike automatic machine learning framework," *Frontiers in computational neuroscience*, vol. 14, p. 16, 2020.

[53] S. Schmidgall, R. Ziaei, J. Achterberg, L. Kirsch, S. Hajiseyedrazi, and J. Eshraghian, "Brain-inspired learning in artificial neural networks: a review," *APL Machine Learning*, vol. 2, no. 2, 2024.

[54] A. Nayebi, R. Rajalingham, M. Jazayeri, and G. R. Yang, "Neural foundations of mental simulation: Future prediction of latent representations on dynamic scenes," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[55] F. Messaoudi and M. Loukili, "E-commerce personalized recommendations: a deep neural collaborative filtering approach," in *Operations Research Forum*, vol. 5, pp. 1–25, Springer, 2024.

[56] Y. Liu, J. Lu, J. Yang, and F. Mao, "Sentiment analysis for e-commerce product reviews by deep learning model of bert-bigru-softmax," *Mathematical Biosciences and Engineering*, vol. 17, no. 6, pp. 7819–7837, 2020.

[57] C. Deng and Y. Liu, "A deep learning-based inventory management and demand prediction optimization method for anomaly detection," *Wireless Communications and Mobile Computing*, vol. 2021, pp. 1–14, 2021.

[58] L. Dai, L. Wu, H. Li, C. Cai, Q. Wu, H. Kong, R. Liu, X. Wang, X. Hou, Y. Liu, *et al.*, "A deep learning system for detecting diabetic retinopathy across the disease spectrum," *Nature communications*, vol. 12, no. 1, p. 3242, 2021.

[59] M. Wen, Z. Zhang, S. Niu, H. Sha, R. Yang, Y. Yun, and H. Lu, "Deep-learning-based drug–target interaction prediction," *Journal of proteome research*, vol. 16, no. 4, pp. 1401–1409, 2017.

[60] W. L. Alyoubi, W. M. Shalash, and M. F. Abulkhair, "Diabetic retinopathy detection through deep learning techniques: A review," *Informatics in Medicine Unlocked*, vol. 20, p. 100377, 2020.

[61] K. B. Nielsen, M. L. Lautrup, J. K. Andersen, T. R. Savarimuthu, and J. Grauslund, "Deep learning–based algorithms in screening of diabetic retinopathy: a systematic review of diagnostic performance," *Ophthalmology Retina*, vol. 3, no. 4, pp. 294–304, 2019.

[62] M. D. Abràmoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. C. Folk, and M. Niemeijer, "Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning," *Investigative ophthalmology & visual science*, vol. 57, no. 13, pp. 5200–5206, 2016.

[63] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *jama*, vol. 316, no. 22, pp. 2402–2410, 2016.

[64] Q. H. Nguyen, R. Muthuraman, L. Singh, G. Sen, A. C. Tran, B. P. Nguyen, and M. Chua, "Diabetic retinopathy detection using deep learning," in *Proceedings of the 4th international conference on machine learning and soft computing*, pp. 103–107, 2020.

[65] F. Tang, P. Luenam, A. R. Ran, A. A. Quadeer, R. Raman, P. Sen, R. Khan, A. Giridhar, S. Haridas, M. Iglicki, *et al.*, "Detection of diabetic retinopathy from ultra-widefield scanning laser ophthalmoscope images: a multicenter deep learning analysis," *Ophthalmology Retina*, vol. 5, no. 11, pp. 1097–1106, 2021.

[66] T. Saba, "Computer vision for microscopic skin cancer diagnosis using handcrafted and non-handcrafted features," *Microscopy Research and Technique*, vol. 84, no. 6, pp. 1272–1283, 2021.

[67] G. Yunchao and Y. Jiayao, "Application of computer vision and deep learning in breast cancer assisted diagnosis," in *Proceedings of the 3rd International Conference on Machine Learning and Soft Computing*, pp. 186–191, 2019.

[68] C. R. Pereira, D. R. Pereira, F. A. Silva, J. P. Masieiro, S. A. Weber, C. Hook, and J. P. Papa, "A new computer vision-based approach to aid the diagnosis of parkinson's disease," *Computer Methods and Programs in Biomedicine*, vol. 136, pp. 79–88, 2016.

[69] W. Yin, L. Li, and F.-X. Wu, "Deep learning for brain disorder diagnosis based on fmri images," *Neurocomputing*, vol. 469, pp. 332–345, 2022.

[70] Y. Feng, H. S. Teh, and Y. Cai, "Deep learning for chest radiology: a review," *Current Radiology Reports*, vol. 7, pp. 1–9, 2019.

[71] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, "Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs," *arXiv preprint arXiv:1901.07042*, 2019.

[72] RadiologyMasterclass, "Understanding chest x-ray abnormalities: A radiology masterclass." `https://www.radiologymasterclass.co.uk/tutorials/chest/chest_pathology/chest_pathology_start`. Accessed on 2021-09-15.

[73] S. Horng, R. Liao, X. Wang, S. Dalal, P. Golland, and S. J. Berkowitz, "Deep learning to quantify pulmonary edema in chest radiographs," *Radiology: Artificial Intelligence*, vol. 3, no. 2, p. e190228, 2021.

[74] S.-C. Pei and C.-N. Lin, "Image normalization for pattern recognition," *Image and Vision computing*, vol. 13, no. 10, pp. 711–723, 1995.

[75] R. H. Philipsen, P. Maduskar, L. Hogeweg, and B. van Ginneken, "Normalization of chest radiographs," in *Medical Imaging 2013: Computer-Aided Diagnosis*, vol. 8670, pp. 106–111, SPIE, 2013.

[76] T. Rahman, A. Khandakar, Y. Qiblawey, A. Tahir, S. Kiranyaz, S. B. A. Kashem, M. T. Islam, S. Al Maadeed, S. M. Zughaier, M. S. Khan, *et al.*, "Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images," *Computers in biology and medicine*, vol. 132, p. 104319, 2021.

[77] Z. Xue, D. You, S. Candemir, S. Jaeger, S. Antani, L. R. Long, and G. R. Thoma, "Chest x-ray image view classification," in *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*, pp. 66–71, IEEE, 2015.

[78] C. F. Sabottke and B. M. Spieler, "The effect of image resolution on deep learning in radiography," *Radiology: Artificial Intelligence*, vol. 2, no. 1, p. e190015, 2020.

[79] Z. Xue, D. You, S. Candemir, S. Jaeger, S. Antani, L. R. Long, and G. R. Thoma, "Chest x-ray image view classification," in *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*, pp. 66–71, IEEE, 2015.

[80] I. Sirazitdinov, M. Kholiavchenko, R. Kuleev, and B. Ibragimov, "Data augmentation for chest pathologies classification," in *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, pp. 1216–1219, IEEE, 2019.

[81] R. C. Hidayatullah and S. Violina, "Convolutional neural network architecture and data augmentation for pneumonia classification from chest x-rays images," *Int J Innov Sci Res Technol*, vol. 5, pp. 158–164, 2020.

[82] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[83] S. Sundaram and N. Hulkund, "Gan-based data augmentation for chest x-ray classification," *arXiv preprint arXiv:2107.02970*, 2021.

[84] S. Albahli, "Efficient gan-based chest radiographs (cxr) augmentation to diagnose coronavirus disease pneumonia," *International journal of medical sciences*, vol. 17, no. 10, p. 1439, 2020.

[85] J. Kauffmann, L. Ruff, G. Montavon, and K.-R. Müller, "The clever hans effect in anomaly detection," *arXiv preprint arXiv:2006.10609*, 2020.

[86] M. Ahmad, K. Koay, Y. Sun, V. Jayaram, G. Arunachalam, and F. Amirabdollahian, "Deep learning for condition detection in chest radiographs: A performance comparison of different radiograph views and handling of uncertain labels," in *ACHI 2023: The Sixteenth International Conference on Advances in Computer-Human Interactions*, 2023.

[87] R. Kundu, R. Das, Z. W. Geem, G.-T. Han, and R. Sarkar, "Pneumonia detection in chest x-ray images using an ensemble of deep learning models," *PloS one*, vol. 16, no. 9, p. e0256630, 2021.

[88] L. Visuña, D. Yang, J. Garcia-Blas, and J. Carretero, "Computer-aided diagnostic for classifying chest x-ray images using deep ensemble learning," *BMC Medical Imaging*, vol. 22, no. 1, p. 178, 2022.

[89] F. Ahmad, A. Farooq, and M. U. Ghani, "Deep ensemble model for classification of novel coronavirus in chest x-ray images," *Computational intelligence and neuroscience*, vol. 2021, no. 1, p. 8890226, 2021.

[90] S. H. Khan, A. Sohail, A. Khan, M. Hassan, Y. S. Lee, J. Alam, A. Basit, and S. Zubair, "Covid-19 detection in chest x-ray images using deep boosted hybrid learning," *Computers in Biology and Medicine*, vol. 137, p. 104816, 2021.

[91] M. Adimoolam, K. Govindharaju, A. John, S. Mohan, A. Ahmadian, and T. Ciano, "A hybrid learning approach for the stage-wise classification and prediction of covid-19 x-ray images," *Expert Syst*, vol. 2021, p. e12884, 2021.

[92] K. Shaheed, P. Szczuko, Q. Abbas, A. Hussain, and M. Albathan, "Computer-aided diagnosis of covid-19 from chest x-ray images using hybrid-features and random forest classifier," in *Healthcare*, vol. 11, p. 837, MDPI, 2023.

[93] J. Zhao, M. Li, W. Shi, Y. Miao, Z. Jiang, and B. Ji, "A deep learning method for classification of chest x-ray images," in *Journal of Physics: Conference Series*, vol. 1848, p. 012030, IOP Publishing, 2021.

[94] R. Qin, K. Qiao, L. Wang, L. Zeng, J. Chen, and B. Yan, "Weighted focal loss: An effective loss function to overcome unbalance problem of chest x-ray14," in *IOP Conference Series: Materials Science and Engineering*, vol. 428, p. 012022, IOP Publishing, 2018.

[95] Z. Ge, D. Mahapatra, S. Sedai, R. Garnavi, and R. Chakravorty, "Chest x-rays classification: A multi-label and fine-grained problem," *arXiv preprint arXiv:1807.07247*, 2018.

[96] L. Li, Y. Long, B. Huang, Z. Chen, Z. Liu, and Z. Yang, "Research on chest disease recognition based on deep hierarchical learning algorithm," *Journal of Healthcare Engineering*, vol. 2022, no. 1, p. 6996444, 2022.

[97] R. Belo, J. Rocha, A. M. Mendonça, and A. Campilho, "An active learning approach for support device detection in chest radiography images," in *Fifteenth International Conference on Machine Vision (ICMV 2022)*, vol. 12701, pp. 271–278, SPIE, 2023.

[98] M. Costa, S. C. Pereira, J. Pedrosa, A. M. Mendonça, and A. Campilho, "Deep feature-based automated chest radiography compliance assessment," in *2023 IEEE 7th Portuguese Meeting on Bioengineering (ENBENG)*, pp. 64–67, IEEE, 2023.

[99] J. Fachrel, A. A. Pravitasari, I. N. Yulita, M. N. Ardhisasmita, and F. Indrayatna, "Enhancing an imbalanced lung disease x-ray image classification with the cnn-lstm model," *Applied Sciences*, vol. 13, p. 8227, 2023.

[100] A. Trivedi, C. Robinson, M. Blazes, A. Ortiz, J. Desbiens, S. Gupta, R. Dodhia, P. K. Bhatraju, W. C. Liles, J. Kalpathy-Cramer, *et al.*, "Deep learning models for covid-19 chest x-ray classification: Preventing shortcut learning using feature disentanglement," *Plos one*, vol. 17, p. e0274098, 2022.

[101] J. C. Seah, C. H. Tang, Q. D. Buchlak, X. G. Holt, J. B. Wardman, A. Aimoldin, N. Esmaili, H. Ahmad, H. Pham, J. F. Lambert, *et al.*, "Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study," *The Lancet Digital Health*, vol. 3, no. 8, pp. e496–e506, 2021.

[102] A. Majkowska, S. Mittal, D. F. Steiner, J. J. Reicher, S. M. McKinney, G. E. Duggan, K. Eswaran, P.-H. Cameron Chen, Y. Liu, S. R. Kalidindi, *et al.*, "Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation," *Radiology*, vol. 294, no. 2, pp. 421–431, 2020.

[103] P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz, *et al.*, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists," *PLoS medicine*, vol. 15, no. 11, p. e1002686, 2018.

[104] E. J. Hwang, J. G. Nam, W. H. Lim, S. J. Park, Y. S. Jeong, J. H. Kang, E. K. Hong, T. M. Kim, J. M. Goo, S. Park, *et al.*, "Deep learning for chest radiograph diagnosis in the emergency department," *Radiology*, vol. 293, no. 3, pp. 573–580, 2019.

[105] X. Wang, J. Yu, Q. Zhu, S. Li, Z. Zhao, B. Yang, and J. Pu, "Potential of deep learning in assessing pneumoconiosis depicted on digital chest radiography," *Occupational and environmental medicine*, vol. 77, no. 9, pp. 597–602, 2020.

[106] R. Singh, M. K. Kalra, C. Nitiwarangkul, J. A. Patti, F. Homayounieh, A. Padole, P. Rao, P. Putha, V. V. Muse, A. Sharma, *et al.*, "Deep learning in chest radiography: detection of findings and presence of change," *PloS one*, vol. 13, no. 10, p. e0204155, 2018.

[107] NVIDIA, *CUDA C Programming Guide: Compute Capability 8.x*, 2024. Accessed: 2024-08-01.

[108] S. M. Group, "Chexpert." `https://stanfordmlgroup.github.io/competitions/chexpert/`. Accessed on 2020-04-01.

[109] T. Gluecker, P. Capasso, P. Schnyder, F. Gudinchet, M.-D. Schaller, J.-P. Revelly, R. Chiolero, P. Vock, and S. Wicky, "Clinical and radiologic features of pulmonary edema," *Radiographics*, vol. 19, no. 6, pp. 1507–1531, 1999.

[110] E. Milne, M. Pistolesi, M. Miniati, and C. Giuntini, "The radiologic distinction of cardiogenic and noncardiogenic edema," *American Journal of Roentgenology*, vol. 144, no. 5, pp. 879–894, 1985.

[111] D. A. Reynolds *et al.*, "Gaussian mixture models.," *Encyclopedia of biometrics*, vol. 741, no. 659-663, 2009.

[112] X. Jing, X.-f. WANG, Z.-f. YANG, and C.-w. XU, "Comparison of supervised clustering methods for the analysis of dna microarray expression data," *Agricultural Sciences in China*, vol. 7, no. 2, pp. 129–139, 2008.

[113] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal processing magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[114] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[115] R. Brunelli, *Template matching techniques in computer vision: theory and practice*. John Wiley & Sons, 2009.

[116] Z. Zhou, J. Shin, L. Zhang, S. Gurudu, M. Gotway, and J. Liang, "Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7340–7351, 2017.

[117] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[118] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[119] A. Kulesza, N. Jiang, and S. Singh, "Low-rank spectral learning with weighted loss functions," in *Artificial Intelligence and Statistics*, pp. 517–525, PMLR, 2015.

[120] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[121] C. Ranjan, "Theory of pooling," *PrePrint, Research-Gate (Nov. 2020). doi*, vol. 10, 2020.

[122] J. M. Porcel and R. W. Light, "Diagnostic approach to pleural effusion in adults," *American family physician*, vol. 73, no. 7, pp. 1211–1220, 2006.

[123] D. Yu, H. Wang, P. Chen, and Z. Wei, "Mixed pooling for convolutional neural networks," in *Rough Sets and Knowledge Technology: 9th International Conference, RSKT 2014, Shanghai, China, October 24-26, 2014, Proceedings 9*, pp. 364–375, Springer, 2014.

[124] M. D. Jankowich and S. I. Rounds, "Combined pulmonary fibrosis and emphysema syndrome: a review," *Chest*, vol. 141, no. 1, pp. 222–231, 2012.

[125] S. Assaad, W. B. Kratzert, B. Shelley, M. B. Friedman, and A. Perrino Jr, "Assessment of pulmonary edema: principles and practice," *Journal of cardiothoracic and vascular anesthesia*, vol. 32, no. 2, pp. 901–914, 2018.

[126] C. M. Walker, "Subsegmentaland rounded atelectasis," *Chest Imaging*, p. 105, 2019.

[127] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[128] J. Huang and C. X. Ling, "Using auc and accuracy in evaluating learning algorithms," *IEEE Transactions on knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.

[129] H. W. Lilliefors, "On the kolmogorov-smirnov test for normality with mean and variance unknown," *Journal of the American statistical Association*, vol. 62, no. 318, pp. 399–402, 1967.

[130] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3-4, pp. 591–611, 1965.

[131] F. Kang, R. Jin, and R. Sukthankar, "Correlated label propagation with application to multi-label learning," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, pp. 1719–1726, IEEE, 2006.

[132] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "Cnn-rnn: A unified framework for multi-label image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2285–2294, 2016.

[133] F. Lyu, Q. Wu, F. Hu, Q. Wu, and M. Tan, "Attend and imagine: Multi-label image classification with visual attention and recurrent neural networks," *IEEE Transactions on Multimedia*, vol. 21, no. 8, pp. 1971–1981, 2019.

[134] M. O. Wielpütz, C. P. Heußel, F. J. Herth, and H.-U. Kauczor, "Radiological diagnosis in lung disease: factoring treatment options into the choice of diagnostic modality," *Deutsches Ärzteblatt International*, vol. 111, no. 11, p. 181, 2014.

[135] T. R. C. of Radiologists, "Clinical radiology census reports." https://www.rcr.ac.uk/news-policy/policy-reports-initiatives/clinical-radiology-census-reports/#:~:text=The%20UK%20now%20has%20a,up%20with%20demand%20for%20services. Accessed on 2024-01-15.

[136] D. Edmondson, "Likert scales: A history," in *Proceedings of the Conference on Historical Analysis and Research in Marketing*, vol. 12, pp. 127–133, 2005.

[137] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, pp. 50–60, 1947.

Appendix A

Custom Pooling Layer

```python
class CustomMaxPool2D(tf.keras.layers.Layer):
    def __init__(self, pool_size=(2, 2), strides=[2,2], padding='VALID', **kwargs):
        super(CustomMaxPool2D, self).__init__(**kwargs)
        self.pool_size = pool_size
        self.strides = strides
        self.padding = padding.upper()
    def get_config(self):
        config = super(CustomMaxPool2D, self).get_config()
        config.update({
            'pool_size': self.pool_size,
            'strides': self.strides,
            'padding': self.padding,
        })
    return config
    def call(self, inputs):
        if self.padding == 'SAME':
            # Calculate padding needed based on stride and kernel sizes
            pad_along_height = max((self.pool_size[0] - self.strides[0]) // 2, 0)
            pad_along_width = max((self.pool_size[1] - self.strides[1]) // 2, 0)
            paddings = [[0, 0], [pad_along_height, pad_along_height],
                        [pad_along_width, pad_along_width], [0, 0]]
            inputs = tf.pad(inputs, paddings)

        output_MAX = tf.nn.pool(inputs, window_shape = [2,2], pooling_type='MAX',
    strides=self.strides, padding=self.padding)

        output_MIN = -tf.nn.pool(-inputs, window_shape=[2, 2], pooling_type='MAX',
    strides=self.strides, padding=self.padding)
        concatenated_output = tf.concat([output_MAX, output_MIN], axis=3)
        return concatenated_output
```

Python code for cutom pooling operation.

Weighted Loss Function

```python
def compute_class_freqs(labels):
    N = len(labels)
    positive_frequencies = np.sum(labels, axis=0)/N
    negative_frequencies = (N - np.sum(labels, axis=0))/N
    return positive_frequencies, negative_frequencies

def get_weighted_loss(pos_weights, neg_weights, epsilon=1e-7):
    def weighted_loss(y_true, y_pred):
        loss = 0.0
        y_true = tf.cast(y_true, tf.float32)  # Cast y_true to float32

        for i in range(len(pos_weights)):
            loss += -K.mean(pos_weights[i] * y_true[:, i] * K.log(y_pred[:, i] +
    epsilon) + neg_weights[i] * (1 - y_true[:, i]) * K.log(1 - y_pred[:, i] +
    epsilon))
        return loss
freq_pos, freq_neg = compute_class_freqs(train_df[CLASSES].values)
pos_weights = freq_neg
neg_weights = freq_pos
custom_loss = get_weighted_loss(pos_weights, neg_weights)
```

Python code for defining Weighted Loss Function.

Appendix B

GMM for Relabelling

```
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
import os
import pickle
import numpy as np
from sklearn.metrics import accuracy_score
import matplotlib.pyplot as plt
from PIL import Image
import pandas as pd
from sklearn.metrics import silhouette_samples, silhouette_score
from sklearn.cluster import MiniBatchKMeans
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
from sklearn import metrics
import matplotlib as mpl
from sklearn.mixture import GaussianMixture
from sklearn.model_selection import StratifiedKFold

def retrieve_info(cluster_labels,y_train):
        reference_labels = {}
        for i in range(len(np.unique(cluster_labels))):
            index = np.where(cluster_labels == i,1,0)
            num = np.bincount(y_train[index==1]).argmax()
            reference_labels[i] = num
        return reference_labels

def run_model(certain,nf,test,view):
    print('certain samples: ', len(certain), ' nf samples: ', len(nf))
    certain['Label'] = 1
    nf['Label'] = 0
    certain = certain[['Path','Label']]
    nf = nf[['Path','Label']]
    if len(nf) > 2000:
      nf = nf.sample(2000).reset_index(drop=True)

    sample = min(len(nf),len(certain))
    certain = certain.sample(sample)
    nf = nf.sample(sample)
    certain = certain.append(nf).reset_index(drop=True)
    print('train')
    print(certain.Label.value_counts())

    x_train = certain['Path'].to_frame()
    y_train = certain['Label'].to_frame()
    x_test = test['Path'].to_frame()
    x_train = np.array([np.array(Image.open('/home/ma20aac/TM_Data/'+row.Path)) for
     i, row in x_train.iterrows()])
    y_train = np.array([np.array(row.Label) for i, row in y_train.iterrows()])
```

```
    x_test = np.array([np.array(Image.open('/home/ma20aac/TM_Data/'+row.Path)) for
i, row in x_test.iterrows()])
    x_train =x_train/255.0
    x_train = x_train.reshape(len(x_train),-1)
    x_test =x_test/255.0
    x_test = x_test.reshape(len(x_test),-1)

    n_classes = 300
    print('n_classes',n_classes)
    #GMMs using different types of covariances.
    estimators = {
        cov_type: GaussianMixture(
            n_components=n_classes, covariance_type=cov_type, max_iter=200 #,
random_state=0
        )
        for cov_type in ["spherical","diag"]
    }

    n_estimators = len(estimators)

    for index, (name, estimator) in enumerate(estimators.items()):
        print(index, name, estimator)

        # Training the other parameters using the EM algorithm.
        estimator.fit(x_train)

        y_train_pred = estimator.predict(x_train)

        reference_labels = retrieve_info(y_train_pred,y_train)
        print('reference_labels')
        print(reference_labels)

        train = []
        for j in y_train_pred:
            train.append(reference_labels[j])
        train = np.array(train)
        train_accuracy_correct = accuracy_score(y_train, train, normalize=False)
        train_accuracy = accuracy_score(y_train, train)
        print('train_accuracy',len(train),train_accuracy,train_accuracy_correct)
        y_test_pred = estimator.predict(x_test)
        test_y = []
        for j in y_test_pred:
            test_y.append(reference_labels[j])
        test_y = np.array(test_y)
        if name == 'spherical':
            GMM_AT_UN_SP = test.copy()
            GMM_AT_UN_SP['Label'] = test_y
            GMM_AT_UN_SP['covariance_type'] = name
            GMM_AT_UN_SP.to_csv('GMM_AT_'+view+'_UN_SP.csv',index=False)
```

Python code for relabelling uncertain labels of Atelectasis.

Appendix C

Custom Pooling with DenseNet121

```
def create_base_model():
    base_model = DenseNet121(weights='imagenet', include_top=False, input_shape=
    IMG_SIZE + (3,))
    for i, layer in enumerate(base_model.layers):
        if isinstance(layer, layers.MaxPooling2D):
            base_model.layers[i] = CustomPool2D(pool_size=(2, 2))
    base_model.trainable = True
    x = base_model.output
    x = layers.GlobalAveragePooling2D()(x)
    x = layers.Dense(512, activation='relu')(x)
    x = layers.Dropout(0.5)(x)
    predictions = layers.Dense(NUM_CLASSES, activation='sigmoid')(x)
    model = tf.keras.Model(inputs=base_model.input, outputs=predictions)
    return model
```

Python code for integrating the custom pooling layer within DenseNet121.

Appendix D

SMLE

```python
def initial_model(IMG_SIZE, cls):
    base_model = DenseNet121(
    include_top=False,
    weights='imagenet',
    input_shape=(IMG_SIZE[0], IMG_SIZE[1], 3)
    )
    # Freeze the base model
    for layer in base_model.layers:
        layer.trainable = True

    # Add custom classification head
    x = base_model.output
    x = layers.MaxPooling2D(pool_size=(2, 2))(x)
    x = layers.GlobalAveragePooling2D()(x)
    x = layers.Dense(1024, activation='relu')(x)
    x = layers.Dropout(0.5)(x)
    predictions = layers.Dense(cls, activation='sigmoid')(x)
    model = models.Model(inputs=base_model.input, outputs=predictions)
    print(model.summary())
    return model

def next_stage_model(old_model, cls, layers_to_train):
    x = old_model.layers[-6].output
    x = layers.Conv2D(256, (3, 3), activation='relu', padding='same')(x)
    x = layers.Conv2D(512, (3, 3), activation='relu', padding='same')(x)
    x = layers.Conv2D(512, (3, 3), activation='relu', padding='same')(x)
    x = layers.Conv2D(512, (3, 3), activation='relu', padding='same')(x)
    x = layers.MaxPooling2D(pool_size=(2, 2))(x)
    x = layers.GlobalAveragePooling2D()(x)
    x = layers.Dense(1024, activation='relu')(x)
    x = layers.Dropout(0.5)(x)
    predictions = layers.Dense(cls, activation='sigmoid')(x)
    model = models.Model(inputs=old_model.input, outputs=predictions)
    for layer in model.layers[:-layers_to_train]:
        layer.trainable = False
    print(model.summary())
    return model
```

Python functions defining each stage of the model.

Appendix E

Pre and Post Questionnaires

General Questions

1. **Your gender?**
    - Male
    - Female
    - Prefer not to say
    - Other
2. **What country are you currently residing in?** _____
3. **Your age group?**
    - Under 25
    - 25-34
    - 35-44
    - 45-54
    - 55-64
    - 65-74
    - 75 or above
4. **What type of setting do you work in (Choose multiple if apply)?**
    - Hospital
    - Private practice
    - Academic institution
    - Research institution
    - Other
5. **What is your clinical specialty?**
    - Radiologist
    - Other
6. **How many years of work experience do you have in radiology?**
    - Less than 1 year
    - 1-5 years

- 6-10 years
- 11-15 years
- More than 15 years
- Other

7. **What percentage of your job role involves interpreting chest X-rays?**
    - 0-20%
    - 21-40%
    - 41-60%
    - 61-80%
    - 81-100%
8. **Do you have any computer coding knowledge?**
    - None
    - Basic
    - Intermediate
    - Advanced
    - Expert
9. **How optimistic are you about AI/ML in radiology 10 years in the future?**
    - Not at all optimistic
    - Slightly optimistic
    - Moderately optimistic
    - Very optimistic
    - Extremely optimistic

Pre-Presentation Questionnaire

1. **To what extent do you believe that AI/ML models could replace the role of a chest radiologist in interpreting X-rays?**
    - Strongly Disagree
    - Disagree
    - Neutral
    - Agree
    - Strongly Agree
2. **Do you agree AI/ML has a great potential to speed up diagnosis and can be a great tool at the hand of the radiologist?**
    - Strongly Disagree
    - Disagree
    - Neutral
    - Agree
    - Strongly Agree

3. **How confident are you in the prediction that AI/ML will reduce the number of chest radiologists in each imaging department?**
    - Extremely confident
    - Somewhat confident
    - Neutral
    - Somewhat not confident
    - Extremely not confident
4. **Do you believe AI/ML will devalue chest radiology as a profession?**
    - Strongly Disagree
    - Disagree
    - Neutral
    - Agree
    - Strongly Agree
5. **In your opinion, will AI/ML outperform chest radiologists for disease detection in chest X-rays within a decade?**
    - Not at all likely
    - Slightly likely
    - Moderately likely
    - Very likely
    - Extremely likely
6. **To what extent do you think AI/ML could alert chest radiologists to abnormal findings in chest X-rays?**
    - Not at all
    - Slightly
    - Moderately
    - Very
    - Extremely
7. **How confident are you in AI/ML making diagnostic suggestions when you are unsure about chest X-ray findings?**
    - Extremely confident
    - Somewhat confident
    - Neutral
    - Somewhat not confident
    - Extremely not confident
8. **Should there be a dedicated AI task force or subcommittee for chest radiology in relevant medical conferences or associations?**
    - Strongly Disagree
    - Disagree

- Neutral
- Agree
- Strongly Agree

9. **If there was an initial AI task force meeting at the next conference, would you attend or like to be involved, specifically focusing on chest radiology?**
    - Definitely not attend/involved
    - Probably not attend/involved
    - Neutral
    - Probably attend/involved
    - Definitely attend/involved
10. **Would you be more willing to learn about AI/ML if educational tools were specifically tailored for chest radiology rather than general radiology?**
    - Strongly Disagree
    - Disagree
    - Neutral
    - Agree
    - Strongly Agree
11. **What is your level of trust in AI/ML tools for chest radiology in terms of the balance of risks versus benefits?**
    - No trust at all
    - Low trust
    - Moderate trust
    - High trust
    - Complete trust
12. **How do you anticipate the job of the average chest radiologist will be impacted in the next 5 years from AI?**
    - Very negatively
    - Negatively
    - Neutral
    - Positively
    - Very positively
13. **Do you think jobs in chest radiology will be more or less impacted than other radiology subspecialties by AI/ML?**
    - Much less impacted
    - Less impacted
    - Neutral
    - More impacted
    - Much more impacted

14. **Do you believe AI/ML have the potential to address the shortage of radiologists in the future?**
    - Strongly Disagree
    - Disagree
    - Neutral
    - Agree
    - Strongly Agree
15. **What do you believe should be the most important driving force behind the development of AI tools for chest radiology?**
    - Improving accuracy
    - Increasing efficiency
    - Reducing workload
    - Enhancing patient care
    - Other

Post-Presentation Questionnaire

1. **To what extent do you trust that AI/ML models could replace the role of a chest radiologist in interpreting X-rays?**
    - Strongly Disagree
    - Disagree
    - Neutral
    - Agree
    - Strongly Agree
2. **How confident are you in the prediction that AI/ML will reduce the number of chest radiologists in each imaging department?**
    - Extremely confident
    - Somewhat confident
    - Neutral
    - Somewhat not confident
    - Extremely not confident
3. **Do you believe AI/ML will devalue chest radiology as a profession?**
    - Strongly Disagree
    - Disagree
    - Neutral
    - Agree
    - Strongly Agree
4. **In your opinion, will AI/ML outperform chest radiologists for disease detection in chest X-rays within a decade?**
    - Not at all likely

- Slightly likely
- Moderately likely
- Very likely
- Extremely likely

5. **To what extent do you think AI/ML could alert chest radiologists to abnormal findings in chest X-rays?**
    - Not at all
    - Slightly
    - Moderately
    - Very
    - Extremely
6. **Do you believe AI/ML could increase work efficiency in chest imaging, particularly in X-ray interpretation?**
    - Strongly Disagree
    - Disagree
    - Neutral
    - Agree
    - Strongly Agree
7. **How confident are you in AI/ML making diagnostic suggestions when you are unsure about chest X-ray findings?**
    - Extremely confident
    - Somewhat confident
    - Neutral
    - Somewhat not confident
    - Extremely not confident
8. **Should there be a dedicated AI task force or subcommittee for chest radiology in relevant medical conferences or associations?**
    - Strongly Disagree
    - Disagree
    - Neutral
    - Agree
    - Strongly Agree
9. **If there was an initial AI task force meeting at the next conference, would you attend or like to be involved, specifically focusing on chest radiology?**
    - Definitely not attend/involved
    - Probably not attend/involved
    - Neutral
    - Probably attend/involved
    - Definitely attend/involved

10. **Would you be more willing to learn about AI if educational tools were specifically tailored for chest radiology rather than general radiology?**
    - Strongly Disagree
    - Disagree
    - Neutral
    - Agree
    - Strongly Agree
11. **What is your level of trust in AI/ML tools for chest radiology in terms of the balance of risks versus benefits?**
    - No trust at all
    - Low trust
    - Moderate trust
    - High trust
    - Complete trust
12. **How do you anticipate the job of the average chest radiologist will be impacted in the next 5 years from AI?**
    - Very negatively
    - Negatively
    - Neutral
    - Positively
    - Very positively
13. **Do you think jobs in chest radiology will be more or less impacted than other radiology subspecialties by AI/ML?**
    - Much less impacted
    - Less impacted
    - Neutral
    - More impacted
    - Much more impacted
14. **Do you believe AI/ML have the potential to address the shortage of radiologists in the future?**
    - Strongly Disagree
    - Disagree
    - Neutral
    - Agree
    - Strongly Agree
15. **What do you believe should be the most important driving force behind the development of AI tools for chest radiology?**
    - Improving accuracy
    - Increasing efficiency

- Reducing workload
- Enhancing patient care
- Other

16. **After reviewing the AI/ML-detected results in the slide presentation, do you feel more confident in the capabilities of AI/ML for chest X-ray interpretation?**
    - Much less confident
    - Less confident
    - No change
    - More confident
    - Much more confident
17. **To what extent do you believe the AI/ML-detected results align with your own interpretations of the chest X-rays shown in the presentation?**
    - Not at all
    - Slightly
    - Moderately
    - Very
    - Completely
18. **Did the presentation of AI/ML-detected results change your perception of the potential impact of AI/ML on the field of chest radiology?**
    - Not at all
    - Slightly
    - Moderately
    - Very
    - Completely
19. **Would you be more or less likely to collaborate with AI/ML systems in your daily practice after seeing the presentation results?**
    - Much less likely
    - Less likely
    - No change
    - More likely
    - Much more likely
20. **Do you believe that integrating AI/ML into chest radiology would enhance or hinder the quality of patient care?**
    - Hinder
    - Slightly hinder
    - No change
    - Enhance
    - Significantly enhance

21. **How do you think the integration of AI/ML into chest radiology practice might affect your day-to-day work after witnessing the AI/ML results in the presentation?**

    _____

22. **In your opinion, what are the main strengths of the AI/ML model presented in the slides for detecting conditions in chest X-rays?**

    _____

23. **What limitations or challenges do you perceive in the AI/ML model's performance based on the results shown in the presentation?**

    _____

24. **Can you share any specific concerns or fears you have regarding the integration of AI/ML in chest radiology?**

    _____

25. **In your opinion, what steps could be taken to enhance the trust and usability of AI/ML models in chest radiology practice?**

    _____

Results Presentation

Machine Learning for Chest Radiograph Interpretation

Machine Learning

- Machine learning is a subset of artificial intelligence (AI) that empowers computers to acquire knowledge autonomously, without the need for explicit programming.

- Machine Learning algorithms are trained on data to identify patterns and generalise them to make predictions on unseen data.

- Machine Learning is used in a wide variety of applications, including image recognition such as recognising a disease on a chest radiograph.

Machine Learning for Chest Radiology

- Machine learning models undergo training using a substantial dataset comprising a vast number of chest radiographs.
- These radiographs come with labels indicating the presence of either a single condition or multiple conditions within them.
- It is crucial for the performance of machine learning models to be trained on a diverse set of radiographs, all having the same condition. This allows the model to discern and internalize the intricate patterns unique to each specific medical condition.
- Following the training phase, the model's efficacy can be evaluated by subjecting it to radiographs that were not part of the training data. This testing phase assesses the model's generalization ability.

Data we used to train the model

In this study, a machine learning model was trained using an extensive dataset comprising 152,472 frontal chest radiographs. The dataset covers five medical conditions, including Cardiomegaly, Oedema, Consolidation, Atelectasis, and Pleural Effusion. The table shows the number of radiographs that contain one to five conditions per radiograph.

Number of Radiographs	Number of Conditions Present in Each Radiograph	Percentage
67,903	One	44.5%
46,715	Two	30.6%
17,278	Three	11.3%
3,327	Four	2.1%
275	Five	0.18%
16,974	None	11.1%

Evaluating Machine Learning Model's Performance

After training the model, it is important to evaluate the performance of the model. For that, a test set of 352 frontal radiographs is used with the following details.

Number of Radiographs	Number of Conditions Present in Each Radiograph
128	One
91	Two
48	Three
14	Four
3	Five
68	None

Model Evaluation

When presented with a radiograph, our machine learning model outputs the likelihood of five conditions. A threshold determines the predicted labels. We then compare these labels to the true test set labels, resulting in four possible scenarios.

- **True Positive (TP)**
  The model predicts a positive label (condition present), and the true label is also positive.

- **True Negative (TN)**
  The model predicts a negative label (condition not present), and the true label is also negative.

- **False Positive (FP)**
  The model predicts a positive label (condition present), but the true label is negative.

- **False Negative (FN)**
  The model predicts a negative label (condition not present), but the true label is positive.

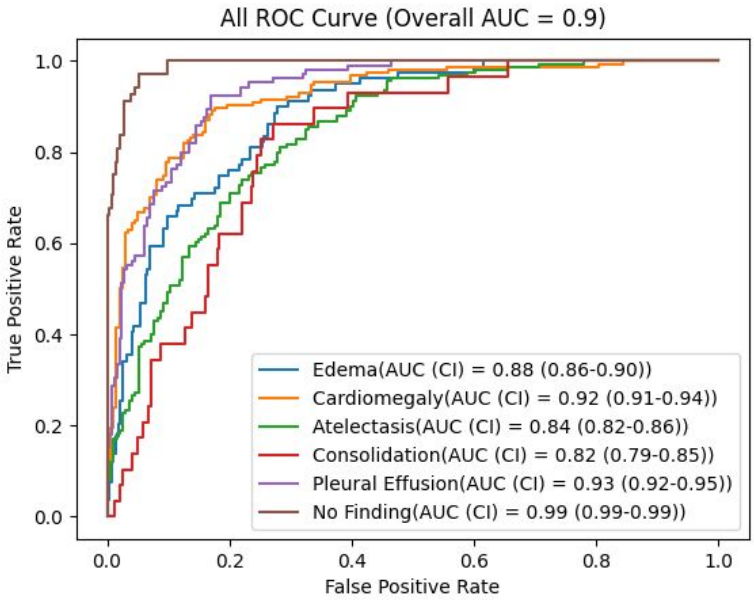
Evaluation Metrics

To assess the model's performance, we constructed an AUC (Area Under the Curve) plot for each condition. This plot captures the AUC score, mapping it onto the True Positive Rate (TPR) along the Y-axis and the False Positive Rate (FPR) along the X-axis. Higher value of AUC indicates better performance. The TPR and FPR are derived from the following expressions, utilizing True Positives, True Negatives, False Positives, and False Negatives

TPR (Sensitivity) = True Positives / (True Positives + False Negatives)
FPR (1 - Specificity) = False Positives / (False Positives + True Negatives)

This visualization provides a comprehensive view of the model's ability to discriminate between conditions, leveraging key metrics for evaluation.

Our Machine Learning Model Performance on Five Conditions



Confidence Interval (CI)
Confidence interval measures a range of AUC values within which we can be 95% confident that the true value of the AUC lies within that range.

True Positive Rate (TPR)
Also known as sensitivity or recall. Measures the proportion of actual positive cases that the model correctly identifies.

False Positive Rate (FPR)
Measures the proportion of actual negative cases that the model incorrectly classifies as positive.

Condition-wise Predictive Values

Condition	PPV	NPV
Pleural Effusion	0.62	0.89
Atelectasis	0.62	0.70
Oedema	0.73	0.85
Consolidation	0.33	0.94
Cardiomegaly	0.62	0.72

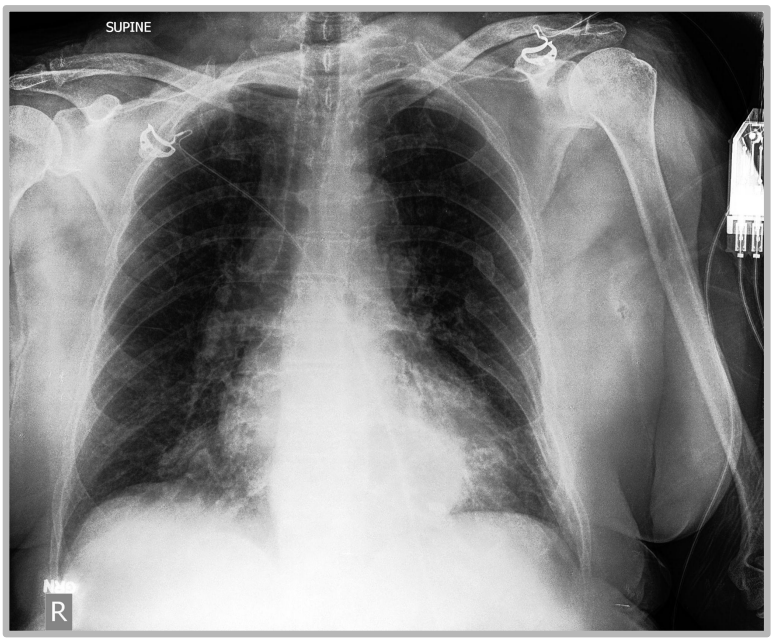
Positive Predictive Value (PPV)
Measures the likelihood that when the machine learning model identifies a condition as "positive" (e.g., having a disease), it is indeed correct.

Negative Predictive Value (NPV)
Measures the likelihood that when the machine learning model identifies a condition as "negative" (e.g., not having a disease), it is indeed correct.

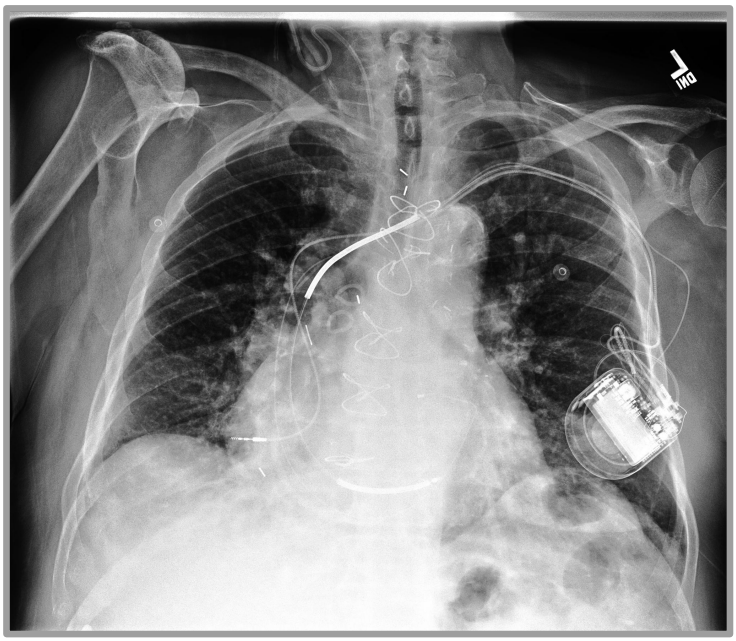
The following slides display radiographs annotated with conditions detected by our Machine Learning model.

Single Condition

Cardiomegaly (True Positive)

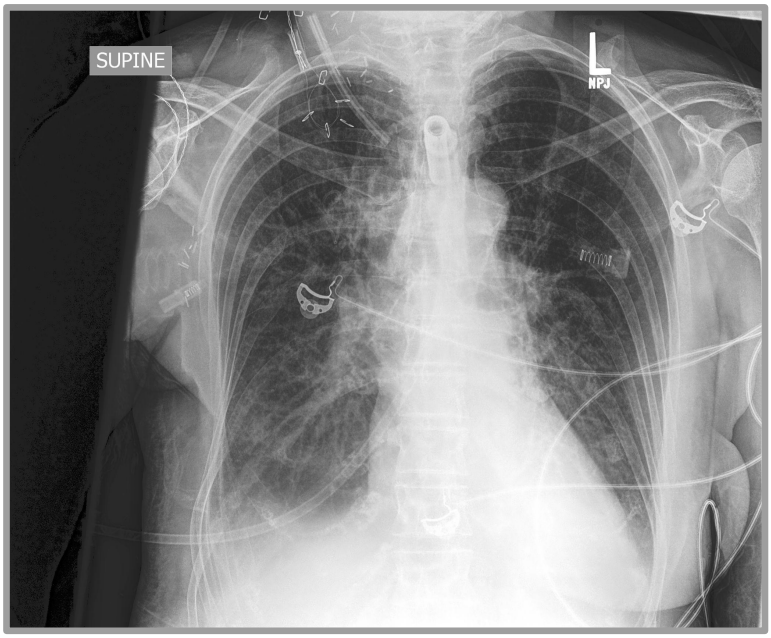


Cardiomegaly (True Positive)

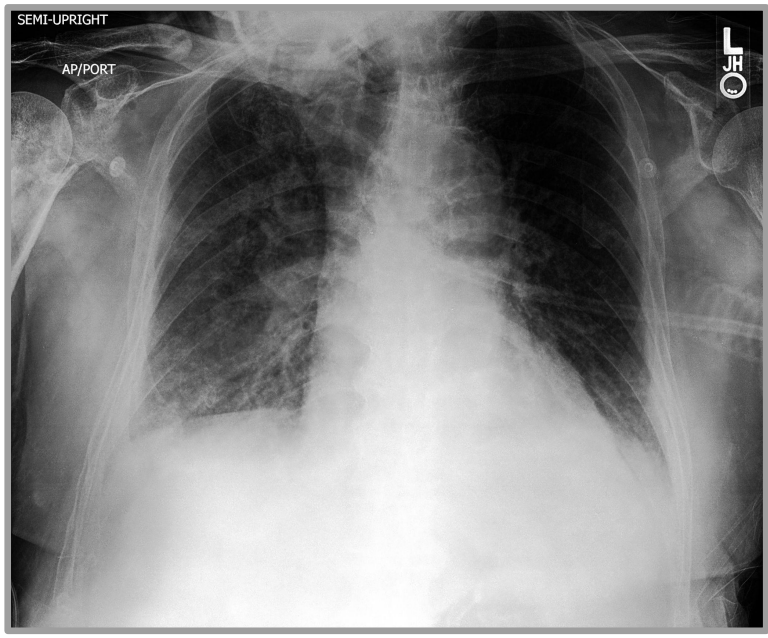


Single Condition

Oedema (True Positives)

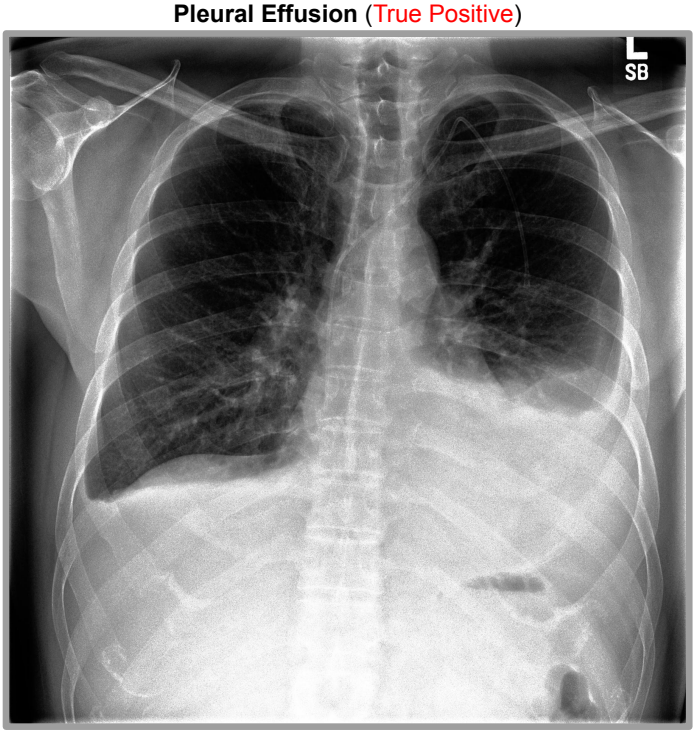


Oedema (True Positives)

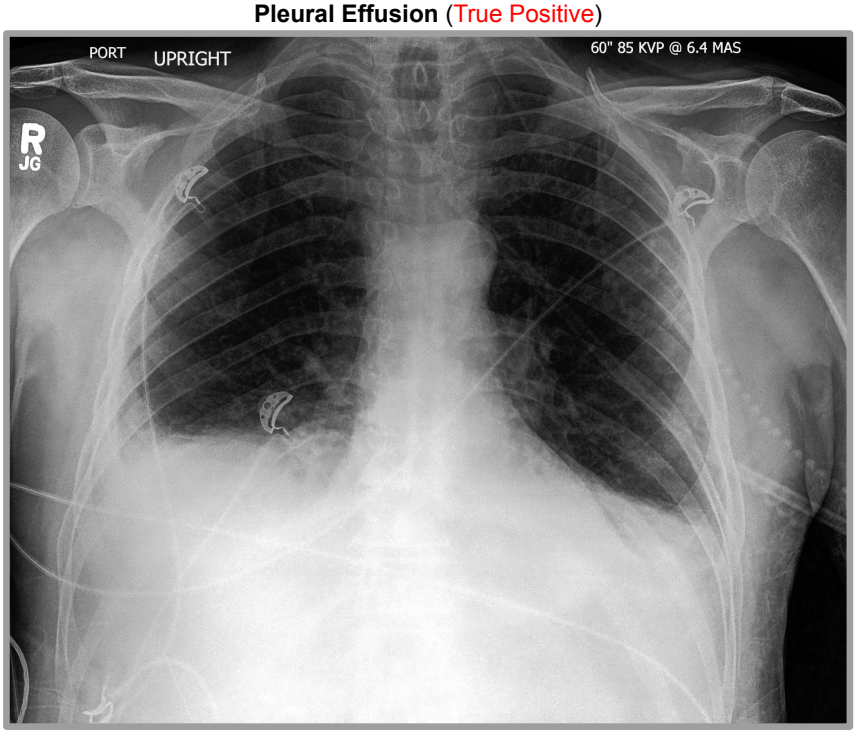


Single Condition

Pleural Effusion (True Positive)

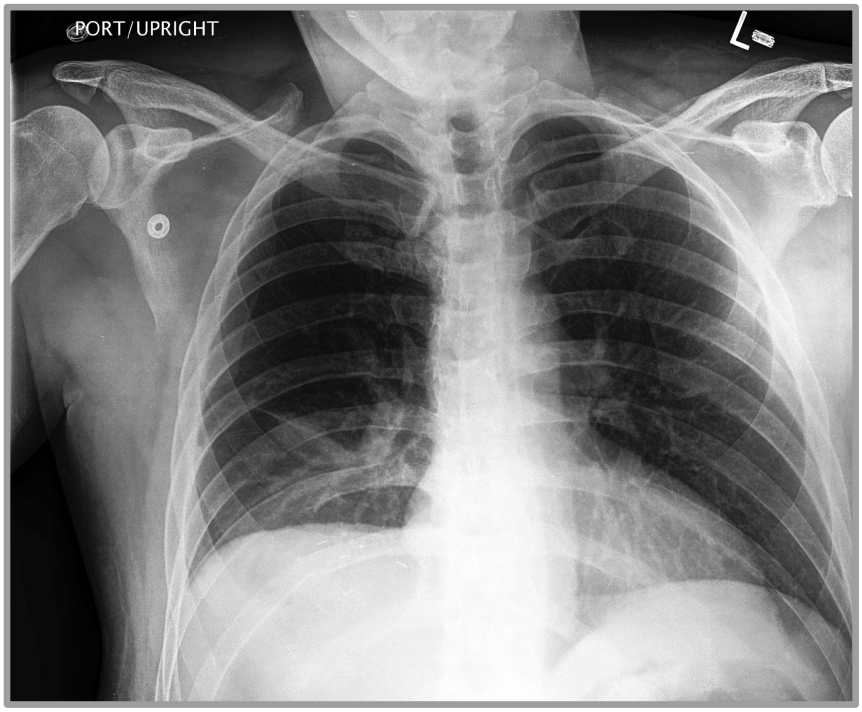


Pleural Effusion (True Positive)

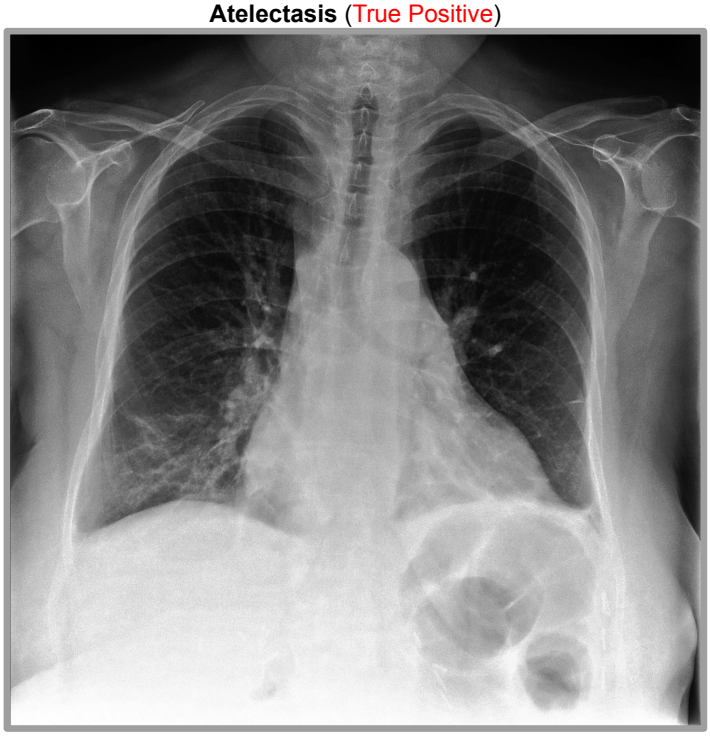


Single Condition

Atelectasis (True Positive)



Atelectasis (True Positive)



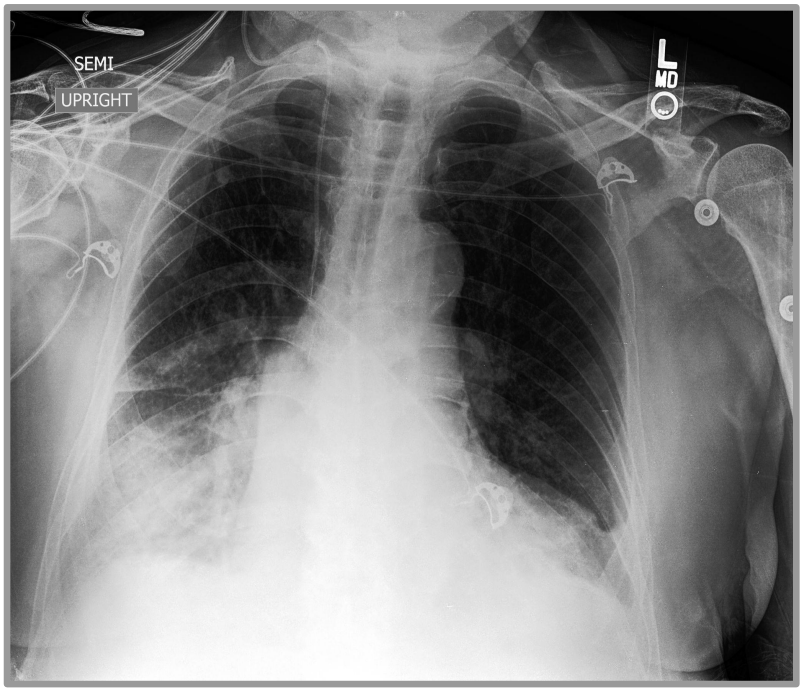
Single Condition

No Finding (True Positive)

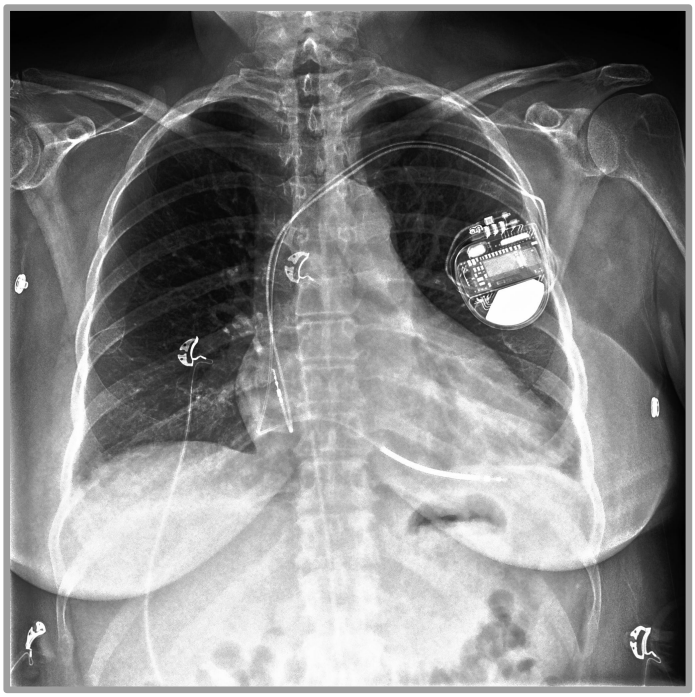
No Finding (True Positive)

Two Conditions

Pleural Effusion and Atelectasis (True Positives)

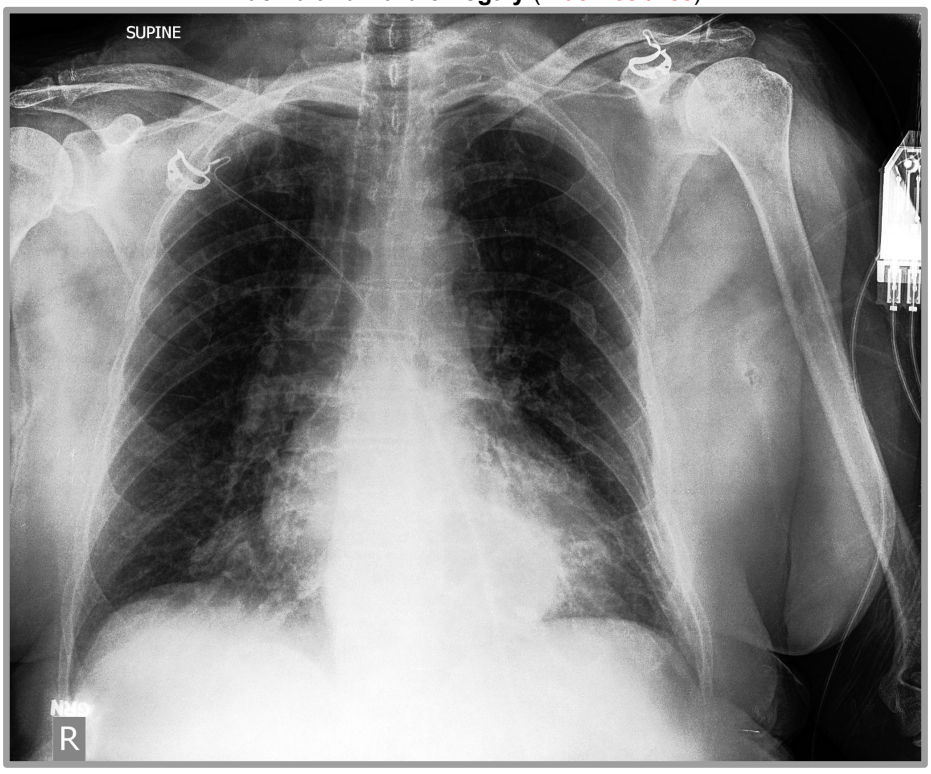


Cardiomegaly and Pleural Effusion (True Positives)



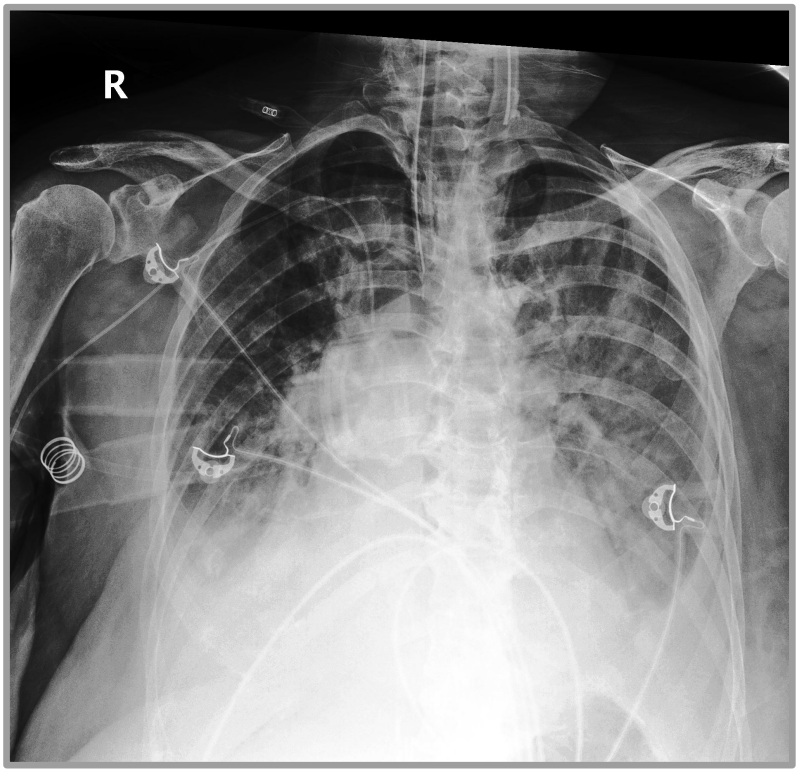
Two Conditions

Edema and Cardiomegaly (True Positives)



Three Conditions

Oedema, Consolidation and Pleural Effusion (True Positives)



Three Conditions

Atelectasis, Consolidation and Pleural Effusion (True positives)

Other Studies

Other studies have shown that machine learning based models have outperformed 2.8 out of 3 radiologists in chest radiograph interpretation.

Table 3: Averaged Testing AUC Scores on CheXpert. NBRC means the # of radiologists out of 3 are beaten by AI algorithms.

Model	AUC	NRBC	Rank
Stanford Baseline [22]	0.9065	1.8	85
YWW [40]	0.9289	2.8	5
Hierarchical Learning [31]	0.9299	2.6	2
DAM (Ours)	**0.9305**	**2.8**	**1**

https://arxiv.org/abs/2012.03173

Thanks for viewing the results presentation.

Please go back to Microsoft Forms (Previous Tab) to complete the post presentation questionnaire.