

# **A Unified Classification Framework Employing Region-of-Interest Localisation for Respiratory Conditions through a Combination of Chest X-Rays and Associated Data**



**By**

**NS Asad Mansoor Khan  
(Registration No: 324955)**

**A thesis submitted to the National University of Sciences and  
Technology, Islamabad  
in partial fulfillment of the requirements for the degree of  
Doctor of Philosophy  
in  
Computer Engineering**

**Supervisor**

**Dr. Muhammad Usman Akram**

**Co-Supervisor**

**Dr. Sajid Gul Khawaja**

**DEPARTMENT OF COMPUTER and SOFTWARE ENGINEERING  
COLLEGE OF ELECTRICAL and MECHANICAL ENGINEERING  
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY (NUST)  
ISLAMABAD, PAKISTAN**

**(2023)**


## THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of PhD Thesis written by NS Asad Mansoor Khan (Registration No 00000324955) of College of Electrical and Mechanical Engineering has been vetted by undersigned, found complete in all respects as per NUST Statutes/Regulations/PhD Policy, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of PhD degree. It is further certified that necessary amendments as pointed by GEC members and foreign/local evaluators of the scholar have also been incorporated in the said thesis.

Signature of Supervisor: 

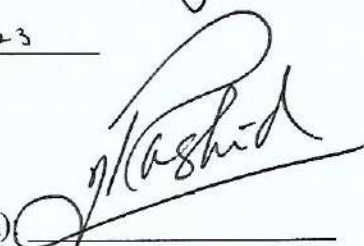
(Dr. Muhammad Usman Akram)

Date: 27-12-23

Signature (HOD): 

(Dr. Usman Qamar)

Date: 27-12-23

Signature (Dean) 

(Brig Dr. Nasir Rashid)

Date: 27-12-23

## **Dedication**

This pursuit is devoid of significance worthy of dedicating it to anyone.

Asad Mansoor Khan

## **Acknowledgement**

Gratitude is owed to those who extended their hands in assistance; they require no introduction, for they know themselves.

Asad Mansoor Khan

## Abstract

Chronic Respiratory Diseases and Chronic Obstructive Pulmonary Disorders affect millions of people around the globe, especially the population of low-middle-income countries such as Pakistan, and are the cause of millions of years lived with disability. Chest X-rays (CXR) are the most commonly used imaging methodology in radiology to diagnose these pulmonary diseases with close to 2 billion CXRs taken every year. Although CXRs are often used, their sheer volume can be a strain on the healthcare system and take a lot of radiologists' time and resources. Therefore, the need for an automated system utilizing this modality is imperative. Furthermore, merely providing an image-level diagnosis for a CXR is insufficient, as the disease affects multiple lung regions. This detailed information is crucial in assessing the severity and progression of the condition. The framework, proposed in this research, offers a unified framework capable of disease classification, providing a severity score for a subset of lung diseases by segmenting the lungs into six regions, and producing chest X-ray reports while taking these challenges into consideration. The classification sub-module proposes a modified progressive learning technique in which the amount of augmentations at each step is capped. Additionally, an ensemble of 4 EfficientNet B0 is used to improve this sub-module's performance and generalizability by taking advantage of a number of augmentation techniques. Furthermore, the segmentation task makes use of an attention map generated within and by the network itself. This attention mechanism allows to achieve segmentation results that are on par with networks having an order of magnitude or more parameters. Severity scoring is introduced for 4 thoracic diseases that can provide a single-digit score corresponding to the spread of opacity in different lung segments with the help of radiologists. The report generation sub-module of the proposed framework generates a CXR report that provides the findings from a single CXR taken either from the Anterior-Posterior (AP) or Posterior-Anterior (PA) viewing position. An encoder and a decoder are employed in the report-generation module; the former splits the image into patches to create hidden states, while the latter uses the hidden encoded states to generate word probabilities, which are then used to build the final report. A foundation model is first fine-tuned in an unsupervised manner which is then used as the Teacher for knowledge transfer to a smaller Student model via Knowledge Distillation (KD). Kullback–Leibler (KL) divergence loss is employed for KD. The distilled student model is then used as the encoder in conjunction with a decoder for report generation. The evaluation and training is done using 9 different CXR datasets, both publicly available and collected locally including BRAX, Indiana, MIMIC, JSRT, Shenzhen, SIIM and others utilising nearly 400,000 CXR images from diverse demographic and geographical locations. On the BRAX validation data set for segmentation, we achieve F1 scores of 0.924 and 0.939 without and with fine-tuning, respectively and a mean matching score of 80.8% is obtained for severity score grading. An average area under receiver operating characteristic curve of 0.88 is achieved for classification using the proposed modified progressive learning which is an improvement of almost 9% in comparison to literature. The incorporation of KD in report generation framework by first fine-tuning a foundation model and then training a student model results in an increase of BLEU-1 score for Indiana dataset by 4% and BERTScore by 7.5%. Similarly, pre-training on larger datasets for report generation, when used in combination with KD, further increases BLEU-1 score for Indiana dataset by 7.2% and BERTScore

by 3%. For MIMIC dataset, comparable performance is achieved for Findings and the Impression sections of the report while the proposed framework outperforms other techniques when both of these sections are combined. For MIMIC-PRO dataset, an  $s_{emb}$  score of 0.4069 while a RadGraph F1 score of 0.1165 is achieved outperforming other techniques in literature. With the highest BERTScore of 0.2245 on the same dataset, the difference between SOTA is just 1.06%. Finally, the proposed framework is also evaluated on locally gathered dataset BRAX subset without any re-training or fine-tuning resulting in BLEU-1 score of 0.3827 and a BERTScore of 0.4392 for local dataset and BLEU-1 score 0.1671 of and a BERTScore of 0.2186 for BRAX dataset showing generalisation ability. The results indicate that the proposed framework performs comparably to existing techniques for some sub-modules and outperforms state-of-the-art techniques for other sub-modules while using a simple architecture with a relatively small parameter size. By obtaining many insights from a single chest X-ray, the approaches used in the proposed framework have the potential to improve the precision of lung disease diagnosis and offer the medical community a comprehensive solution to expedite the chest examination procedure. Subsequent improvements in the performance of the proposed framework will increase its utility even further.

## List of Publications

The following is a list of articles that this research endeavor has produced:

- Khan, Asad, Muhammad Usman Akram, and Sajid Nazir. "Automated grading of chest x-ray images for viral pneumonia with convolutional neural networks ensemble and region of interest localization." PLoS One 18.1 (2023): e0280352.
- Khan, Asad Mansoor, Muhammad Usman Akram, Sajid Nazir, Taimur Hassan, Sajid Gul Khawaja, and Tatheer Fatima. "Multi-head deep learning framework for pulmonary disease detection and severity scoring with modified progressive learning." Biomedical Signal Processing and Control 85 (2023): 104855.
- Khan, Asad Mansoor, Taimur Hassan, Muhammad Usman Akram, Norah Saleh Alghamdi, and Naoufel Werghi. "Continual learning objective for analyzing complex knowledge representations." Sensors 22, no. 4 (2022): 1667.

# Contents

<b>List of Figures</b>	<b>ix</b>
<b>List of Tables</b>	<b>xiii</b>
<b>Chapter 1: Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	2
1.2 Problem Statement . . . . .	4
1.3 Scope and Objective . . . . .	5
1.4 Contributions . . . . .	6
1.5 Structure of Thesis . . . . .	7
<b>Chapter 2: Thoracic Anatomy and Conditions</b>	<b>9</b>
2.1 Thoracic Cavity . . . . .	9
2.2 Respiratory System Diseases . . . . .	10
2.3 Imaging Modalities . . . . .	14
2.3.1 Computed Tomography (CT) . . . . .	14
2.3.2 Magnetic Resonance Imaging (MRI) . . . . .	15
2.3.3 Ultrasonography . . . . .	15
2.3.4 Positron Emission Tomography (PET) . . . . .	16
2.3.5 X Radiography (X-ray) . . . . .	16
2.4 Summary . . . . .	21
<b>Chapter 3: Literature Review</b>	<b>23</b>
3.1 Classification of Chest Diseases . . . . .	25
3.2 Segmentation and Opacity Localisation . . . . .	34
3.3 Severity Scoring and Quantification . . . . .	40
3.4 Automated Report Generation Through Natural Language Processing . . . . .	47
3.5 Research Gaps . . . . .	57
3.6 Summary . . . . .	57
<b>Chapter 4: Materials</b>	<b>58</b>
4.1 Classification Datasets . . . . .	58
4.1.1 Chest Expert (CheXpert) . . . . .	59
4.1.2 Brazilian Labeled Chest X-ray (BRAX) . . . . .	60
4.2 Segmentation Datasets . . . . .	61
4.2.1 Montgomery County Dataset . . . . .	62
4.2.2 Shenzhen Dataset . . . . .	63
4.2.3 Japanese Society of Radiological Technology Dataset . . . . .	64
4.3 Opacity Localisation Datasets . . . . .	64
4.3.1 SIIM-FISABIO-RSNA (SIIM) . . . . .	64
4.4 Report Generation Datasets . . . . .	65



4.4.1	Indiana University Chest X-ray (IU)	66
4.4.2	Medical Information Mart for Intensive Care - Chest X-rays (MIMIC-CXR) and MIMIC Previous References Omitted (MIMIC-PRO)	67
4.4.3	Local Dataset	69
4.5	Summary	72
<b>Chapter 5: Proposed Multi-Head Deep Learning Framework for Pulmonary Disease Detection and Severity Scoring with Modified Progressive Learning</b>		<b>73</b>
5.1	Multi-Head Deep Learning Framework	74
5.1.1	Classification Head	75
5.1.2	Segmentation Head	80
5.1.3	Localisation Head	84
5.1.4	Severity Quantification through A Combination of Segmentation and Localisation	86
5.1.5	Training Parameters	90
5.2	Results	90
5.2.1	Classification	91
5.2.2	Segmentation	97
5.2.3	Opacity Localisation	99
5.2.4	Severity Score	100
5.3	Discussion	101
5.4	Summary	104
<b>Chapter 6: Proposed Framework for Radiology Report Generation from a Singular Perspective using Transformers with Knowledge Distillation</b>		<b>106</b>
6.1	Report Generation Framework	107
6.2	Encoder: Visual Feature Extractor	108
6.2.1	Image Input Representation	108
6.2.2	Self Attention Mechanism	109
6.2.3	Positional Encoding	111
6.3	Decoder	111
6.3.1	Text Input Representation	112
6.4	Foundation Model: BioBERT	113
6.5	Knowledge Distillation Module via Teacher-Student Model	114
6.6	Training the Report Generation Framework	116
6.6.1	Pre-Training of Foundation Model	116
6.6.2	Knowledge Distillation	117
6.6.3	Training of the Downstream Task: Report Generation	118
6.7	Results	119
6.7.1	Evaluation Metrics	119
6.7.2	MIMIC Dataset Evaluation	122
6.7.3	Indiana University CXR Dataset Evaluation	125
6.7.4	MIMIC-PRO Dataset Evaluation	132
6.7.5	Local Dataset Evaluation	136
6.7.6	BRAX Dataset Evaluation	137
6.7.7	Ablation Study	138
6.8	Discussion	139
6.9	Summary	142

<b>Chapter 7: Conclusion and Future Work</b>	<b>143</b>
7.1 Conclusion . . . . .	143
7.2 Future Work . . . . .	148
<b>Bibliography</b>	<b>151</b>

# List of Figures

Figure 1.1	Organ involvement in children under the age of 14 presented to the pediatric emergency department at the National Institute of Child Health in Karachi, Pakistan . . . . .	3
Figure 1.2	Physician density per 10,000 people around the globe . . . . .	4
Figure 2.1	Anatomical structure of human lungs . . . . .	10
Figure 2.2	Gaseous Exchange via Alveoli . . . . .	11
Figure 2.3	Slices from a Chest Cavity CT Scan . . . . .	15
Figure 2.4	Pulmonary MRI of a patient with COVID-19 . . . . .	16
Figure 2.5	PET scan of lung with a tumor . . . . .	17
Figure 2.6	X-ray of lungs . . . . .	18
Figure 2.7	Examples of CXR with COVID-19 given a severity score under the scheme described in . . . . .	20
Figure 3.1	Literature analysis for insights from chest X-ray . . . . .	24
Figure 4.1	The frontal (a) and lateral (b) view of the thoracic cavity of a patient captured using CXR. Image taken from [1] . . . . .	59
Figure 4.2	Example images from (a) to (e) for the various pulmonary diseases in the BRAX [2] data set . . . . .	62
Figure 4.3	Example images from (a) to (c) showcasing the segmentation masks for JSRT, Montgomery, and Shenzhen datasets respectively [3–7] respectively . . . . .	63
Figure 4.4	Presentation of lung opacity in SIIM [8] dataset . . . . .	65
Figure 4.5	A CXR study from Indiana [9] containing both a frontal and lateral scan along with the radiological report. The report contains information about indication, availability of a prior CXR, findings, and impressions among others. . . . .	66
Figure 4.6	Samples in each class in MIMIC dataset [10] . . . . .	67
Figure 4.7	A CXR study from MIMIC [10] containing both a frontal and lateral scan along with the radiological report. The radiological report has been divided into three sections: history, findings, and impressions. [10] . . . . .	68
Figure 4.8	Two distinct CXR studies from the local dataset; the one of the left is a normal CXR whereas the one on the right is abnormal. The normal reports in this dataset are shorter in length. The radiological report does not contain any sections. . . . .	69

Figure 5.1	Proposed Framework Architecture. The classification head (Bottom Left) outputs the probability of a disease using an ensemble. Opacities are first localised using the localisation head (Top Left) and are then combined with the different lung regions obtained using the segmentation head to obtain the final severity score (Bottom Right).	74
Figure 5.2	The effects of scaling along various dimensions such as width, depth, and resolution (b,c,d). These scaling factors are arbitrary, whereas compound scaling (e) scales the aforementioned factors in accordance with one another. Image taken from [11]	76
Figure 5.3	Mobile inverted convolutional layers form the bulk of the Efficient-Net B0. Image taken from [11]	77
Figure 5.4	Progressive learning vs amended progressive learning. The difference in the methodologies lies in the random augmentation factor which is kept constant for all sizes. Several augmentations can be applied to the same image. From left to right, the images show an increasing random augmentation factor with increasing input image size.	78
Figure 5.5	Classification head used in the sub-framework. The probability vector from each of the backbone is averaged to obtain the final $P(Disease)$	79
Figure 5.6	The number of layers in the encoder and decoder in the U-Net architecture is set to three. Another architectural change shown here is the replacement of transposed convolution with up sampling followed by convolution. Such modifications are possible due to the simple architecture. Because U-Net is fully convolutional, the final layer is also convolutional. Activation after convolutional layers are not depicted here.	81
Figure 5.7	Segmentation head in the framework. The output of the first U-Net is used as an attention map and concatenated with the input image for the input of the second U-Net. The concatenation of the attention map improves the performance of the model while requiring fewer parameters across both U-Nets.	83
Figure 5.8	BiFPN block structure. The flow of the features is in both directions from top to bottom and from bottom to top. Image taken from [12].	84
Figure 5.9	EfficientDet architecture with BiFPN blocks that can be repeated several times with EfficientNet [11] backbone for feature extraction. The networks at the end provide the positional coordinates along with the class label and confidence. Image taken from [12].	85
Figure 5.10	After being divided into six lung regions, the output from the segmentation head is multiplied with the output from the opacity localization head to provide the severity score for each region, which is then added together to produce the final severity score.	88
Figure 5.11	AUROC curves for the ensembles across five folds ((a) to (e)). Ensembling the models together increases the performance for each fold on BRAX [2] dataset.	94
Figure 5.12	Box plots of AUROC for all six classes for the [2] dataset at different input image sizes. The ensemble method outperforms any individual model.	95

Figure 5.13 Average AUROC for each class across different input image sizes shows the utility of progressive learning. . . . .	96
Figure 5.14 F1 score increases with an increase in the number of fine-tuning images from the new (BRAX [2]) data set showing that this technique can be used for continual learning of new data sets. . . . .	99
Figure 5.15 Distribution of BRAX [2] validation set according to the matching score. The majority of the images differ by 1 lung region at most from the markings of the radiologist. . . . .	102
Figure 6.1 Proposed report generation architecture. In the first step, the teacher BioBERT model (Bottom Left) is fine-tuned on the target dataset. Knowledge distillation (Bottom Right) utilises the trained teacher model to reduce the loss in the ViT student model. Finally, this trained ViT is used as the Encoder in the report generation module (Top Centre) in combination with the decoder. . . . .	107
Figure 6.2 Vision Transformer is based on the original encoder model and has been modified for an image. The images are divided into 196 patches of 16x16 for the input size of 224x224. The encoder is used to extract the visual features through the use of Self-Attention which is the matrix product of linear projections of the input that is scaled and then passed through the Softmax activation. . . . .	109
Figure 6.3 Query-Keys and Values-Values scaled dot product is calculated for both the teacher and the student model the similarity between them is increased using the KL Divergence. Image taken from [13] . . . . .	113
Figure 6.4 For a CXR-report pair, the latent space features are generated by both the ViT and the BioBERT which are used to as features to increase the similarity between the two by reducing the loss using KL Divergence. . . . .	115
Figure 6.5 Part of the CXR report is masked to train the foundation model in an unsupervised manner. The trained foundation model is then used as the teacher in the Knowledge Distillation step. . . . .	117
Figure 6.6 Multiple layers are used in both the encoder and the decoder with each layer containing multiple attention heads. Each successive layer in the encoder receives the input from the previous layer while the last layer passes the output to each decoder layer. The figure illustrates the last encoder and decoder layer. . . . .	119
Figure 6.7 The text highlighted in red represents portions of the text that are absent in the generated report for the Findings section of the MIMIC [10] data set. Conversely, the other highlighted sentences demonstrate the similarities between the ground truth and the generated report . . . . .	126
Figure 6.8 The impression section is briefer than the findings section. The differences have been marked in red for the MIMIC [10] data set. The model is unable to correctly identify the distance of different tubes in the image. . . . .	127

Figure 6.9	The combination of Findings and Impression sections not only improves the number of unique words in vocabulary but also improves the quality of the generated reports as well. The results shown above are for the MIMIC [10] data set. . . . .	128
Figure 6.10	In the majority of cases for IU [9] dataset, the generated report closely resembles the ground truth. The discrepancies, marked in red, involve the use of different words, but they still convey the same meaning.	130
Figure 6.11	The text highlighted in red represents portions of the text that are absent in the generated report for MIMIC-PRO [14] data set. Conversely, the other highlighted sentences demonstrate the similarities between the ground truth and the generated report. . . . .	134
Figure 6.12	The reports for BARX [2] subset have been generated without any retraining or fine-tuning. Red highlights the differences between the generated report and the report by the radiologist. Similarities are shown in different colors. It can be observed that different verbiage has been used for the same statement in the actual and the predicted reports. . . . .	140

# List of Tables

Table 3.1	Summary of recent pulmonary disease classification techniques . . .	31
Table 3.2	Summary of recent lung segmentation and opacity localisation techniques . . . . .	38
Table 3.3	Summary of recent severity scoring and quantification techniques . .	45
Table 3.4	Summary of recent CXR report generation techniques . . . . .	53
Table 4.1	Distribution of samples according to the 14 classes in the CheXpert data set [1] . . . . .	60
Table 4.2	Distribution of samples according to the 14 classes in the BRAX data set [2] . . . . .	61
Table 4.3	Number of samples for different segmentation datasets . . . . .	64
Table 4.4	Difference between original impressions from the radiological reports of the MIMIC [10] dataset and the rewritten reports from MIMIC-PRO [14] . . . . .	70
Table 4.5	A summary of datasets that were used to train different sub-modules of the proposed framework . . . . .	72
Table 5.1	Hyperparameters of different output heads in the proposed framework	89
Table 5.2	AUROC scores across 5 fold cross validation using amended progressive learning on BRAX [2] data set where the validation set contains a limited number of No Finding class . . . . .	92
Table 5.3	AUROC scores across 5 fold cross validation using amended progressive learning on BRAX [2] data set where there is no cap on the number of images of No Finding class . . . . .	93
Table 5.5	Comparison of AUROC scores for 6 classes using modified progressive learning with other techniques in the literature on BRAX [2] dataset. Equal and Unequal represents the validation sets where the samples are capped and uncapped respectively. . . . .	96
Table 5.4	AUROC scores across 5 fold cross validation using modified progressive learning with pre-training on CheXpert [1] data set where the validation set contains a limited number of No Finding class . . . . .	98
Table 5.6	F1 score for validation data sets for different segmentation model architectures on JSRT, Montgomery and Shenzhen [3–7] data sets . . . . .	98
Table 5.7	mAP score for SIIM [8] validation dataset for different models at IoU threshold of 0.5 . . . . .	100

Table 5.8	Mean matching score opacities localised by different architectures at IoU value of 0.3 on a subset of BRAX [2] data set that was manually annotated by a radiologist . . . . .	101
Table 6.1	Hyperparameters of different stages of the report generation framework . . . . .	121
Table 6.2	Performance of proposed on MIMIC [10] dataset. The framework is trained on a combination of Findings and Impressions sections while the results are computed on different sections of the report. F represents <i>Findings</i> , F+I represents <i>Findings &amp; Impressions</i> while I represents <i>Impressions</i> . The absence of (AP & PA) denotes that all samples were used for testing. . . . .	122
Table 6.3	Performance of the proposed framework on the Findings section against different techniques in literature for MIMIC [10] data set. The absence of (AP & PA) denotes that all samples were used for testing. . . .	123
Table 6.4	Performance of the proposed framework on the Impressions section against different techniques in literature for MIMIC [10] data set. The absence of (AP & PA) denotes that all samples were used for testing. . . .	124
Table 6.5	Performance of the proposed framework on the combined Findings and Impressions section against different techniques in literature for MIMIC [10] data set. The absence of (AP & PA) denotes that all samples were used for testing. . . . .	125
Table 6.6	Performance of proposed framework on Indiana University [9] dataset. The framework is trained on just the Findings sections. . . . .	129
Table 6.7	Performance of proposed framework on Indiana University [9] data set with pre-training on MIMIC [10] data set for different sub-modules. . .	131
Table 6.8	Performance of the proposed framework on the Findings section against different techniques in literature for Indiana University [9] data set.	131
Table 6.9	Performance of proposed on MIMIC-PRO [14] dataset. The framework is trained on Impressions from which prior references have been removed. I represents <i>Impressions</i> and the absence of (AP & PA) denotes that all samples were used for testing. . . . .	132
Table 6.10	Performance of proposed framework on MIMIC-PRO [14] data set with pre-training on MIMIC [10] data set for different sub-modules. . . .	133
Table 6.11	Performance of the proposed framework on the against different techniques in literature for MIMIC-PRO [14] data set. For the different techniques used in the proposed methodology, the absence of (AP & PA) denotes that all samples were used for testing. * denotes the results without pre-training on the MIMIC dataset while <sup>+</sup> and <sup>#</sup> refer to <i>Exp 1</i> and <i>Exp 2</i> from 6.10 respectively. . . . .	135
Table 6.12	Performance of proposed framework on Local data set. The framework trained on the Indiana University [9] data set was used. . . . .	136
Table 6.13	Performance of proposed framework on BRAX [2] subset. Models trained on three different datasets (Indiana University [9], MIMIC [10] and MIMIC-PRO [14]) were used. . . . .	137



Table 6.14 Effects of using Knowledge Distillation in training by using the foundation model as teacher and ViT as the student model on the Indiana University [9] data set. . . . .	138
--	-----

# Chapter 1

## Introduction

Both Upper Respiratory Infections (URI) and Chronic Respiratory Diseases (CRDs), of which Chronic Obstructive Pulmonary Disease (COPD) is a constituent can have a deleterious impact on the respiratory system affecting different parts from the pharynx to the alveoli including the larynx and sinuses as well. These ailments, which can be brought on by a variety of factors including environmental, viral, bacterial, and smoking are marked by coughing, sore throat, difficulty in breathing, and fatigue among other symptoms. Respiratory tract diseases can greatly increase the Global Burden of Disease, resulting in a significant amount of Years of Life lived with Disability (YLD) and Disability-Adjusted Life Years (DALY). YLD due to COPD amounts to 118 million years while 84.4 age-standardised DALY per hundred thousand people result from URI across all ages and sexes [15–19]. These pulmonary disorders can have a significant economic burden as well [20].

World Health Organisation (WHO) notes that COPD accounts for the third leading cause of death worldwide with close to 3.29 million deaths in 2019 where Low-to Middle-Income Countries are affected disproportionately. While chronic disorders may not have a cure, different treatment strategies can help mitigate their adverse effects to a certain extent [21]. Along with dealing with these chronic illnesses, the recent COVID-19 pan-

demographic has exposed the healthcare system's flaws. Patients who have COVID-19 infections that are severe and concomitant conditions like COPD are more likely to develop respiratory distress syndrome, require a mechanical ventilator, and pass away as a result of these complications [22].

As effective treatment stems from correct diagnosis, therefore, various techniques are used for imaging the chest cavity in a non-invasive manner. These include various imaging techniques such as Computed Tomography (CT) [23], Magnetic Resonance Imaging (MRI) [24], ultrasonography [25], and Positron Emission Tomography (PET) [26]. While this range of techniques allows for accurate assessment, the cost of these imaging methods can be prohibitive in many cases. Therefore, the use of X-rays for imaging provides an alternative that is not only cost-effective but has widespread availability as well. X-rays provide a black-and-white image of the lungs where abnormalities can be identified as opacities with other markers as well.

Any imaging modality that is used to image the thoracic region requires a lot of time and expertise to infer insights that can then be used to treat the patients. In the case of Chest X-rays (CXRs), radiologists require a lot of time to analyse the CXR and produce the corresponding report.

## **1.1 Motivation**

Chronic Respiratory Diseases like asthma, bronchiectasis, and COPD are among the Non-Communicable Diseases that have a significant impact on the population of Low- and Middle-Income Countries (LMICs) starting as early as childhood [27] with the mortality rate increasing 20% for every 1% increase per capita in the Gross Domestic Product [28]. Globally, over half a billion people contracted a CRD in 2017 showing a significant increase over the past years [29]. COPD alone impacted 212.3 million individuals worldwide in 2019; 71% of these persons were from LMICs, and 84% of COPD fatalities were attributed to the disease. In the same year, asthma impacted 262.4 million people

worldwide, with 96% of deaths occurring in LMICs [30]. Just in Asia, approximately 21.4 million succumbed to respiratory diseases over a seven year period from 2010 to 2017 [28]. With 77.79 lung disease-related fatalities per 100,000 inhabitants, Pakistan is ranked 8 globally [31]. For children under the age of 14 in Pakistan, respiratory system involvement is the most prevalent illness [32] as shown in Figure 1.1.

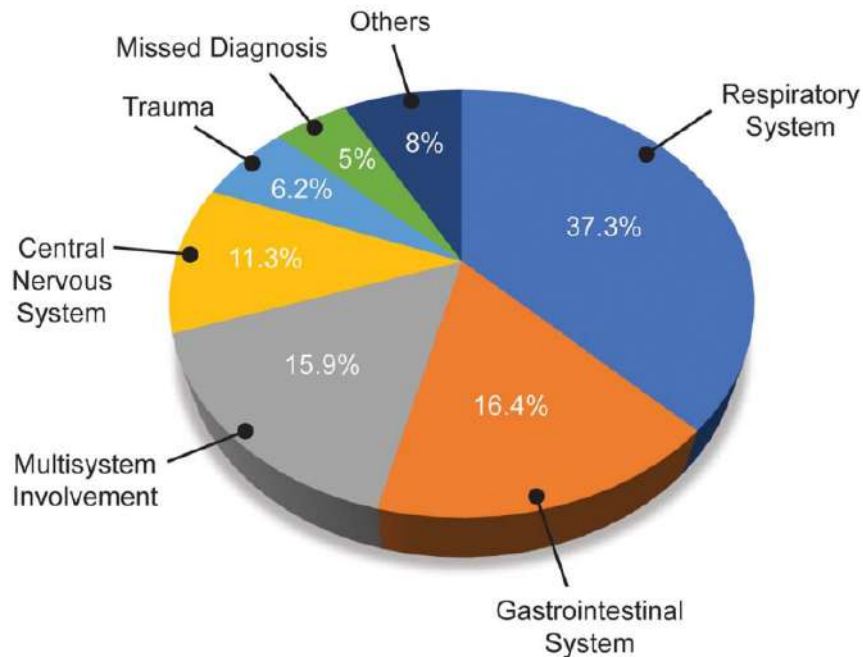


Figure 1.1: Organ involvement in children under the age of 14 presented to the pediatric emergency department at the National Institute of Child Health in Karachi, Pakistan. Taken from [32]

The recent pandemic caused by Coronavirus (COVID-19) with a fatality rate of around 2% [33] has disrupted every aspect of life. There have been 695,537,592 Coronavirus cases and 6,918,075 deaths worldwide by September 2023 [34]. Due to the wide variations in physician density worldwide, the effects of lung disorders are made much more severe [35] as demonstrated by Figure 1.2. Only 11 radiologists provided care for Rwanda's 12 million residents. With a population of 4 million, Liberia only has 2 radiologists who were actively practising [10]. Similarly, in Pakistan, only 14.5 healthcare professionals per 10,000 citizens are available, which falls significantly below the recommended 25, reducing the care that citizens need.

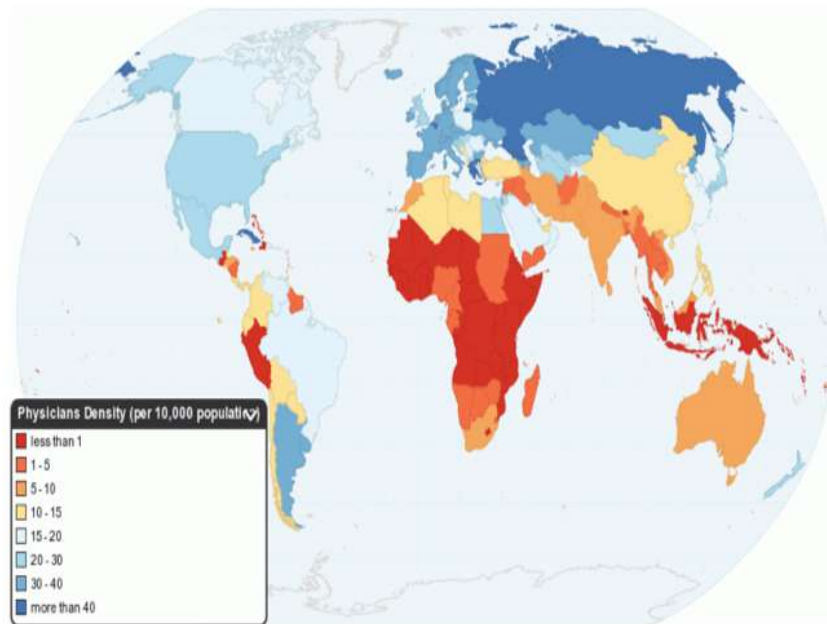


Figure 1.2: Physician density per 10,000 people around the globe. Taken from [35]

Keeping these challenges in view, the need for an automated system to classify pulmonary diseases and assess their severity is paramount. Additionally, if such an automated system can provide CXR reports that are similar in content to those generated by the radiologist, it will be of great assistance to radiologists and lessen their workload.

## 1.2 Problem Statement

Computer-Aided Diagnostic (CAD) systems that can provide a timely and accurate diagnosis of various pulmonary diseases can help to some extent with a number of problems including the prevalence of Chronic Obstructive Pulmonary Disease, particularly in LMICs, the poor to non-existent healthcare infrastructure in some parts of the world, and the strain that a sudden pandemic can exert on the system. To enable fast adoption of suitable treatment and stop the spread of infectious diseases, rapid and precise diagnosis is crucial. With over 2 billion CXRs taken each year [36], chest X-rays are the most widely used modality in the world, making a diagnostic system designed to deal with X-rays extremely important. It is difficult to draw conclusions from a CXR since it relies not only

on image-level diagnosis but also on how the disease affects various lung regions, which can reveal information about the severity and course of the disease. The time-consuming task of assessing each CXR and creating the corresponding report is further complicated by the sheer volume of X-ray examinations. The lack of a unified framework addressing all these problems poses a disadvantage. Therefore, this research aims to provide a robust and reliable solution for image-level diagnosis, organ-level segmentation, severity scoring with quantification, and automated report generation effectively targeting all the steps in the diagnostic pipeline that follow the acquisition of a CXR. Furthermore, this research work tackles the problem of analysing CXRs for image-level and local insights along with report generation using state-of-the-art deep learning and natural language processing techniques.

### **1.3 Scope and Objective**

Automated medical diagnostic systems based on deep learning methodologies can provide assistance to medical care providers in the diagnosis of pulmonary diseases in a timely manner allowing them to provide the care needed by the patients. In addition, the availability of these systems also allows for use in remote areas where medical professionals may be absent. The main objective of this research endeavor is to provide a reliable and robust method that can not only perform disease classification but can also provide an automated radiology report that can assist physicians. Not only can early disease detection benefit the patient, but it can also benefit those in proximity, decreasing the spread of infection and potentially saving lives. The sub-objectives to achieve this primary goal are:

1. To develop a single framework that not only classifies a CXR image in a particular disease class but also segments the lungs' opacity regions if the lungs are diseased.
2. To develop a system capable of performing severity classification for different pathologies and not just COVID-19 which has been the focus of recent severity classifica-

tion attempts.

3. To develop a generic CXR analysis framework that can generate radiology reports that are closer in content to a radiologist is required.
4. To validate the report generation framework using a locally collected CXR dataset with sample images from Pakistan.

## 1.4 Contributions

To address the identified gaps, the main research contribution of this work is a novel, single framework consisting of two sub-frameworks i.e. CXR manifestation analysis and radiology report generation [37, 38]. The sub-frameworks of this methodology can be categorised as follows:

1. CXR Manifestation Analysis via a Multi-Head Framework
  - (a) We propose a single (sub-)framework consisting of disease classification using modified progressive learning and severity grading [38] for different pulmonary disorders using opacity localisation [37].
  - (b) We provide segmentation masks with severity grades for a validation data set that has been validated by a radiologist [38].
  - (c) A segmentation network sub-module is utilized that despite its relatively small size is able to perform relatively close to large architectures such as U-Net [38].
  - (d) We experimentally show that while good performance can be achieved in segmentation using publicly available data sets, fine-tuning on just a small number of samples from the target data set can actually improve the segmentation performance even further [38].
2. CXR Report Generation using a Singular Perspective

- (a) We propose a report generation (sub-)framework employing foundation model fine-tuning on CXR reports for use as a Teacher model to train a smaller Student model
- (b) We propose employing Knowledge Distillation so that a smaller Student model can learn better CXR representation
- (c) Pre-training on larger CXR datasets with reports is shown to improve performance when used in conjunction with Knowledge Distillation providing reasonable performance at a relatively small size and simple architecture
- (d) Gathered local CXR images dataset with findings (reports) from a local hospital.
- (e) We also provide radiology reports generated by a radiologist for a validation data set from the BRAX [2] data set.

## 1.5 Structure of Thesis

The thesis is organised as follows: The structure of the thoracic cavity and various pulmonary diseases, as well as their presentation and symptoms, are briefly covered in Chapter 2. Additionally, various imaging techniques that enable non-invasive imaging of the chest cavity are also discussed.

A thorough review of most recent methodologies in the literature, is provided in Chapter 3, covering classification of pulmonary diseases, lung segmentation and sub-segmentation of diseased areas, severity scoring and severity quantification, and radiological report generation using CXR. Chapter 4 discusses details of the datasets utilised in this research work, specifically focusing on their different attributes.

Chapter 5 details the proposed multi-head framework for chest disease classification, segmentation, opacity localisation, and severity scoring while Chapter 6 focuses on CXR



report generation using a transformer-based network.

This is concluded in Chapter 7, which summarises the key contributions followed by future directions which can be undertaken to improve the proposed framework.

# Chapter 2

## Thoracic Anatomy and Conditions

The thoracic cavity enclosed by the ribs, spine, and sternum houses the heart, the lungs, and the trachea among other important organs. The health of these organs can be investigated using numerous imaging techniques and based on these examinations a diagnosis can then be formed. In this chapter, some aspects of lung anatomy are explored along with a look at imaging techniques and different diseases that can affect the lungs.

### 2.1 Thoracic Cavity

Forming the second largest cavity in the body [39], the chest cavity protects and supports the major portion of the two very important systems: the circulatory and the respiratory system. The heart in the circulatory system is responsible for pumping blood through the body while the lungs which form the backbone of the respiratory system are tasked with the gaseous exchange. Figure 2.1 shows the position of the lungs in relation to the rib cage and the diaphragm in the chest cavity.

Each of the lungs is encased in pleura - a thin membrane. Each lung can be further broken down anatomically into lobes; the right lung is slightly bigger than the left and consequently has three major lobes as opposed to two in the left lung. In each lobe,

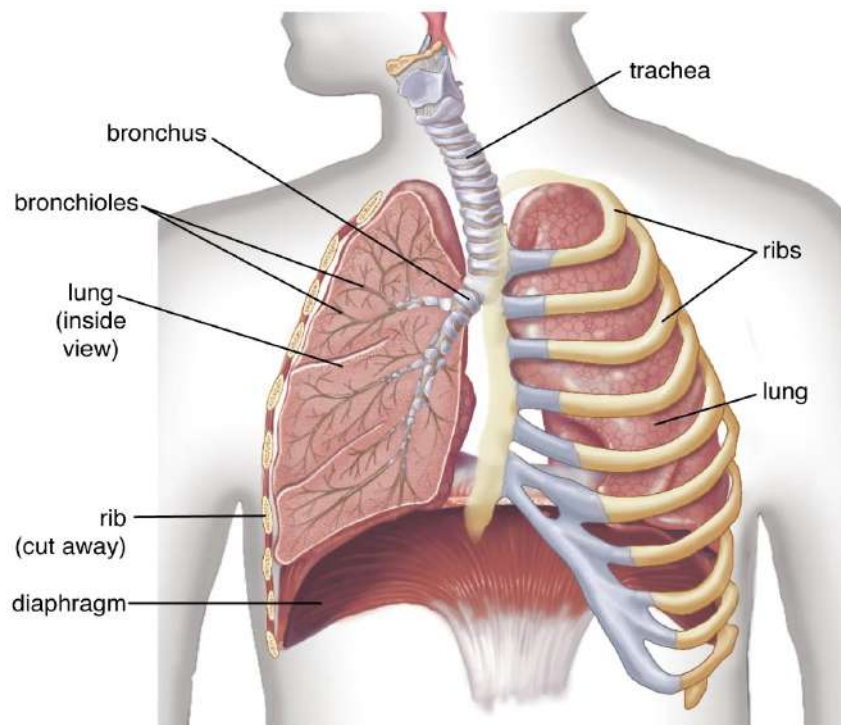


Figure 2.1: Anatomical structure of human lungs and its different parts. Taken from [39]

lobules are present that further subdivide into bronchiole which terminates in a collection of alveoli [40]. Alveoli are where the gaseous exchange happens during breathing. Figure 2.2 shows the functioning of alveoli as a part of the respiratory system.

The normal functioning of the lungs is not just limited to gaseous exchange but also includes the absorption and excretion of water vapor and pharmacological substances [40]. Whenever the functioning of the lungs is disrupted at the cellular level due to apoptosis of the cells, Chronic Obstructive Pulmonary Diseases are developed [42]

## 2.2 Respiratory System Diseases

The alveoli, among other structures found in the lungs, must function properly for the lungs to be able to exchange gases. Various components of the lungs may be impacted by environmental, biological, and genetic variables, which may cause damage or abnormal functioning and impair the lung's ability to execute gaseous exchange properly and efficiently. The following section discusses a few of the disorders that can affect the lungs.

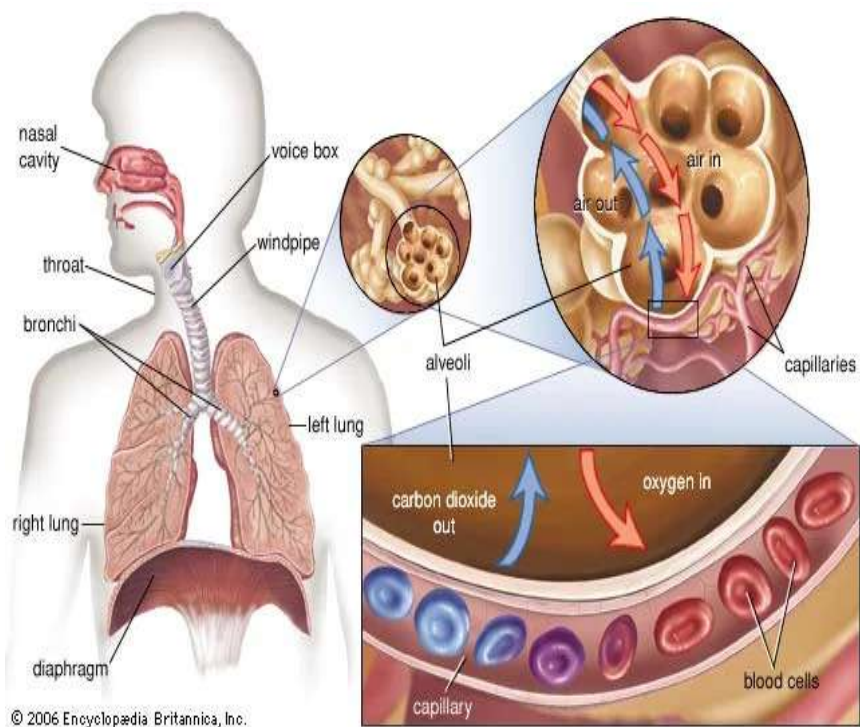


Figure 2.2: The process of gaseous exchange via Alveoli present in the lung. Taken from [41]

- **Atelectasis:** Atelectasis can be brought on by the alveoli collapsing. As they are primarily in charge of exchanging gases, their collapse significantly lowers the lung's capacity for carrying out its functions. This condition may result from a number of factors, such as an increase in lung permeability brought on by inflammation and negative chest cavity pressure [43]. When a patient is given artificial ventilation while sedated, their lung function can decline and they may develop progressive atelectasis, which can be fatal [44, 45].
- **Consolidation:** Another way that the alveoli can lose their ability to perform gaseous exchange is when they become filled with fluid which can be a result of infection, inflammation, or an IgG4-related disease [46]. This also results in difficulty in breathing. The severity of COVID-19 is also directly linked to the severity of consolidation in the lungs with the severity peaking at around 10 days after the onset of the disease. [47, 48].

- **Edema:** Similar to consolidation, pulmonary edema is also an accumulation of fluid in the lungs which results as a complication from a variety of diseases including acute respiratory distress syndrome (ARDS) and severe acute respiratory syndrome (SARS). Cardiogenic pulmonary edema is a type of pulmonary edema that leads to an increase in pulmonary capillary pressure and is caused by the failure of the left ventricular [49]. Patients suffering from lung edema can experience hypoxemia caused by the symptoms of edema which range from difficulty breathing to chest pain and rapid heartbeat [50]. Hypoxemia can lead to further complications.
- **Lung Lesion:** Any abnormal area or aberrant growth in the lung tissue can be classified as a lesion. Some lesions may be benign while others are caused by inflammation or even cancer among other causes with most of the lesions arising in the peripheral lung bands [51] with the right lung being more predisposed to lesion formation than the left [52,53]. Consolidation can sometimes be confused with lung lesion however, an accurate diagnosis can be made by using ultrasound [54]. If a lesion causes a reduction in the amount of exhaled air, it is known as an obstructive lesion in contrast to a restrictive lesion which causes a decrease in the amount of air that can be inhaled [55].
- **Lung Opacity:** Opacity is a blanket term used to specify any region of lungs with decreased transparency or increased density that shows up on different imaging modalities such as X-rays and Computed Tomography scans which can highlight the presence of consolidation and ground-glass opacities [56]. The edges of arteries and airway walls are obscured, and there is hazy enhanced lung attenuation with preserved bronchial and vascular margins. Some of the causes of lung opacity are coronavirus-associated pneumonia [57], fluid accumulation, inflammation, scarring of tissue, and infection.
- **Effusion:** Although, a normal amount of fluid between the pleura - a membrane that encases the lungs - and the chest cavity is required for normal lung operation,

an abnormal buildup from a variety of underlying causes can be detrimental to this process [58] as the excess fluid can exert pressure on the lungs and result in difficulty in breathing. There are two types of pleural effusion: exudative and transudative. With the latter, a high protein concentration is anticipated while in the former, no proteins are present in the fluid that is producing the buildup [59]. Effusion usually presents with shortness of breath, cough, pain in the chest, and fever and can be caused by lung cancer, tuberculosis, Mycoplasma pneumonia, and COVID-19 [60–62]. As an initial treatment step, fluid is drained along with treating the underlying cause.

- **Pneumonia:** Pneumonia being a form of lung infection affects the alveoli in the lungs and causes them to fill up with fluid affecting their normal functioning of gaseous exchange. Pneumonia causes can either be bacterial, viral, or other microorganisms such as hookworms and Ascaris [63,64]. In order to differentiate one type of pneumonia from the other, imaging such as X-rays and CT Scans can play an important role as viral pneumonia may present with nodules and crazy-paving appearance while bacterial pneumonia is characterized by effusion and centrilobular nodules [65,66]. Symptoms of pneumonia include but are not limited to cough, fever, and shortness of breath. Pneumonia can be lethal depending on its severity, especially in persons with compromised immune systems [67] but can be treated with antibiotics and antiviral medications.
- **Pneumothorax:** When the space between the chest cavity and the pleura of the lungs is filled up with gas or air and exerts pressure on the lungs resulting in collapse, this condition is called pneumothorax. As this condition's causes can be numerous, it can be ruled out if lung sliding is present in a lung ultrasound in contrast to a chest X-ray. [68]. Pneumothorax also presents with symptoms such as shortness of breath, fatigue and fever.
- **COVID-19** First discovered in late 2019, the novel SARS-Cov-2 coronavirus is re-

sponsible for causing COVID-19 [69]. The symptoms of COVID-19 can range in severity from going away on their own to contracting COVID-19-associated pneumonia necessitating hospitalisation. Individuals are more likely to experience severe symptoms if they have underlying illnesses like hypertension [70]. The most common symptoms include cough, fever, and breathing problems [71] which can even last months after the initial infection resulting in a phenomenon known as Long COVID [72]. Owing to the fact that COVID-19 emerged recently, the long-term effects and risk factors are not understood completely [73].

## **2.3 Imaging Modalities**

Modern doctors' arsenal is incomplete without non-invasive imaging. These imaging techniques give them the ability to rapidly and precisely evaluate the state of the body's interior structures, on the basis of which they can develop a diagnosis and a treatment strategy. There are two main categories of imaging: structural and functional. The structures of the organs, such as the connections between muscles, can be seen through structural imaging, whereas the functionality of an organ, such as the brain, can be understood through functional imaging [74]. Some of the common imaging modalities used to investigate are Magnetic Resonance Imaging, Computed Tomography, Positron Emission Tomography, ultrasonography, nuclear scanning, Chest X rays and Pulmonary Functional Test (PFT) among others. Some of these modalities are discussed in this section.

### **2.3.1 Computed Tomography (CT)**

Using a series of x-ray beams that are localised to a certain region at multiple angles, an image of that region can be formed once the exiting x-rays are measured by a computer [23]. In some cases, in order to obtain a better image, a contrast dye can be injected before a CT scan is performed. While a CT scan can provide more detail as the beams can be made to pass in multiple axis, the resultant dose of radiation from a CT scan is also higher

as compared to that of a simple Chest X-ray. Owing to this fact, CT scans are used less commonly than x-rays. Figure 2.3 shows the different slices obtained from a CT scan of the chest cavity.



Figure 2.3: Slices from a Chest Cavity CT Scan in which different organs can be seen. Taken from [75]

### 2.3.2 Magnetic Resonance Imaging (MRI)

MRI, originally known as Nuclear Magnetic Resonance Imaging (NMRI) [24], is a non-invasive imaging technique that employs powerful magnetic fields and radio waves to image the internal organs of the body [76]. One of the limitations of MRI is that the patient has to remain still during the procedure in order to ensure that a usable image is obtained. MRI, unlike CT and X-rays, does not use ionising radiation, which can be harmful to live tissues [77]. Functional MRI (fMRI) is a more advanced type of MRI that can be used to study real-time changes in different organs of the human body, such as visual cortical activity in humans during spatial attention [78]. Figure 2.4 shows the cross-sectional MRI of a patient with COVID-19.

### 2.3.3 Ultrasonography

Pulmonary ultrasonography is a non-invasive technique that uses a range of sound frequencies (from 3MHz to 13MHz) to image the pulmonary and surrounding region in order to diagnose different lung pathologies [25]. The imaging is done using a specialised



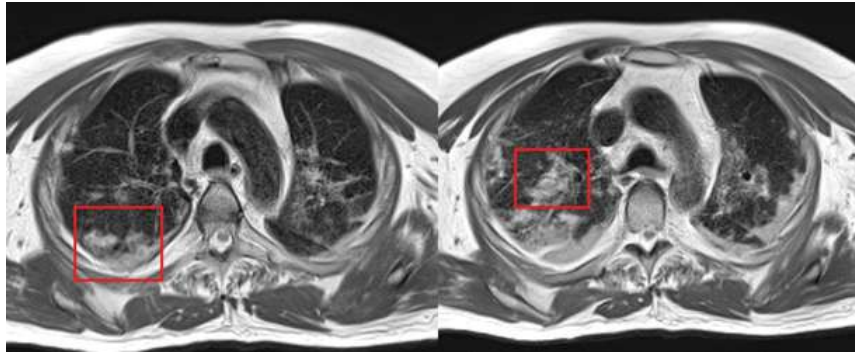


Figure 2.4: Pulmonary MRI of a patient with COVID-19. Taken from [79]

probe that can vary based on the structure that is to be examined [80]. As ultrasound relies only on sound waves, therefore it does not have any negative effects that are usually associated with high-energy radiation such as Computed Tomography scans, Magnetic Resonance Imaging, or X-rays. Ultrasound has been shown to have a high sensitivity for pulmonary pathologies such as pneumothorax, acute respiratory distress syndrome, and even pneumonia [25].

### **2.3.4 Positron Emission Tomography (PET)**

Positron Emission Tomography [26] is a nuclear imaging technique that uses a radioactive substance to map internal organs such as the brain or the lungs. The radioactive substance is injected into the body, and when it decays, it emits a positron, which when it comes into contact with an electron emits gamma rays [81]. These gamma rays are then used for imaging by the appropriate detectors. While a PET scan can provide valuable information, it is also costly to set up and exposes the patient to radioactive traces, which can be harmful in some cases [82]. Figure 2.5 shows a tumor in a PET scan.

### **2.3.5 X Radiography (X-ray)**

This subsection discusses the imaging modality, X-ray, which has been used to train and evaluate the proposed framework in this work.

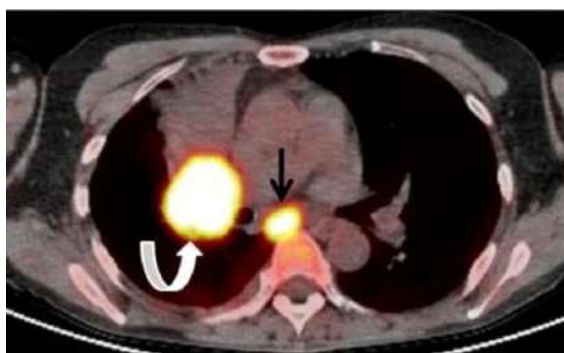


Figure 2.5: PET scan of lungs with a tumor. The arrows mark the position of the tumor. Taken from [83]

### 2.3.5.1 Acquisition

X-rays - another form of ionising radiation - can also be used to image the human body non-invasively. Chest X-Rays are used by radiologists because they are a simple, low-cost method that can be brought to the patient in certain situations and thus can be useful in pulmonary disease diagnosis. [84, 85]. As X-rays can easily pass through the low-density area such as lungs filled with air or fluid, therefore these regions appear as darker regions. In contrast, X-rays are absorbed by high-density areas such as bones and therefore, these regions appear as white. The density of other objects is depicted as a scale of different shades of grey on the X-ray image. Figure 2.6 shows major landmarks in the a chest X-ray.

As X-rays can only provide a two-dimensional view of the body, therefore different projections or views of the chest can be obtained as a result of the relative position of the imaging plate and the X-ray beam. Posteroanterior (PA), anteroposterior (AP), and lateral views are the most common.

When the X-ray beam is positioned such that it enters through the back of the patients and exits from the front to hit the imaging plate, such a view is referred to as posteroanterior. On the other hand, the anteroposterior view is formed when the patient faces the X-ray beam such that the X-rays exit from the back and hit the imaging plate. Generally, PA CXRs are preferred to AP CXRs as they are easier to read. The third type of project is the

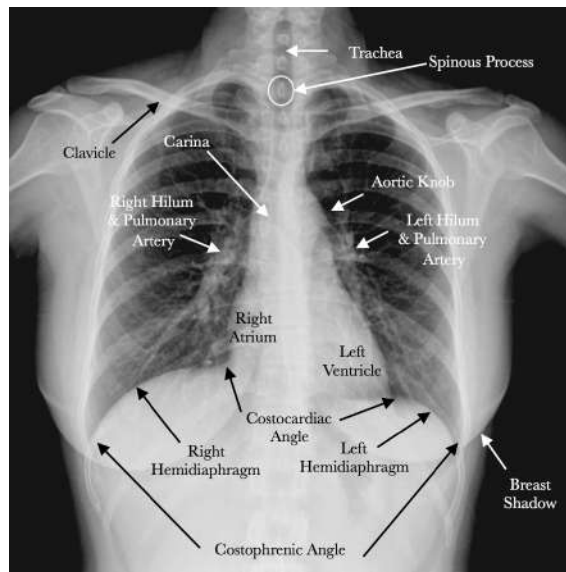


Figure 2.6: Chest X-ray with different landmarks. Taken from [86]

lateral view which is obtained when the X-rays exit the left side of the patient. To obtain the lateral view, both arms need to be out of the way. A PA or AP view is paired with a lateral view in order to better understand the organs that have been photographed because X-rays are unable to provide a three-dimensional view.

### 2.3.5.2 Clinical Manifestations of Pulmonary Diseases

Although X radiography does not produce images with the same level of detail as an MRI or a CT scan, its accessibility, and relatively easy setup make it an ideal method for initial screening and diagnosis. Many pulmonary diseases can not only be diagnosed with a chest X-ray but the progression of these diseases can also be monitored.

Typically, atelectasis is identified by the displacement of the interlobar fissure and the opacification of the collapsed lobe. Moreover, there may be some indirect symptoms such as elevated diaphragm, and a change in the position of the trachea, mediastinum, and heart [87]. Examining the X-ray can also be used to diagnose consolidation. It causes the damaged area of the lung to become uniformly opaque, making the surrounding vessels more noticeable in a tree-like pattern. The opacity also conceals the lungs' typical characteristics [88].

The condition known as pulmonary edema causes the alveoli in the lungs to swell up with fluid. On an x-ray, this fluid is visible as ground glass opacity (GGO). When opacity develops in pulmonary edema, it is characterised by being symmetric in both lungs and covering a sizable portion of the lungs. While ultrasounds provide higher sensitivity when it comes to detection of pleural effusion [89], CXR are still useful if the accumulated fluid is over a certain volume [90]. When the fluid between the pleura and the lungs is present and has a higher density than air, it manifests on the CXR as opacity [91] with the fluid building up towards the lower parts of the lungs. A horizontal air-fluid level is reached in the pleural cavity if fluid and air are both present, as opposed to buildup at the bottom [92]. Similar to this, different lung disorders also show up on the chest X-ray with a variety of other abnormalities that can be used to correctly diagnose them.

### **2.3.5.3 Severity Quantification of Pulmonary Diseases**

Chest x-rays are an efficient way to monitor a disease's severity and progression. A one-digit severity score can be utilised to obtain the most details on the condition of the lungs from a single x-ray and can significantly affect how the patient is managed. Some of the thoracic diseases such as acute respiratory infections [93,94] require physical examination or questionnaires to assess the severity of the disease while the severity of others can be determined by the opacification of the lungs.

Recent severity scoring approaches have focused on COVID-19 which require an X-ray [95–97]. These techniques either assign a score to each lung or further divide the lung into different zones, with each zone receiving a separate score, and the overall score being the sum of the individual zone scores. The division of the each lung can either be into two or three zones resulting in either four quadrants or six zones. Similarly, the scores allotted to each lung or the zone can either be binary or can vary on a spectrum [95,98–100]. Figure 2.7 depicts the severity score that has been assigned to different zones of the lungs for a patient with COVID-19. As opacification is a major manifestation in many of pulmonary

diseases, therefore a similar scoring system based of the scores of individual zones can be beneficial for those diseases as well.

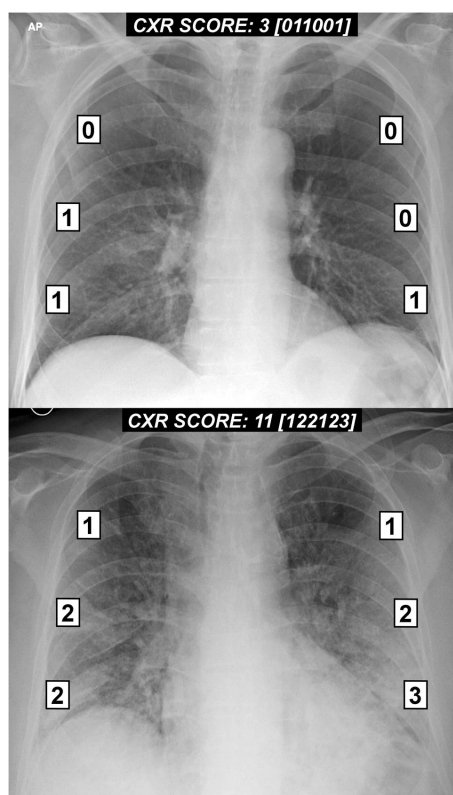


Figure 2.7: Examples of CXR exhibiting COVID-19 that have been a severity score under the scheme described in [101]. Image taken from [101]

#### 2.3.5.4 Reading and Reporting

A chest X-ray provides a plethora of information that can be used to accurately identify multiple thoracic ailments. This information can be described in the form of a report after careful examination of the x-ray. The X-ray can be examined and the problems summarised using the ABCDEF [102] method which entails the following:

- **Airways:** The status of the airways, including the trachea and hilar points, is one aspect of a chest X-ray examination. Hilar point enlargement may be a sign of an underlying disease like sarcoidosis or malignancy. [103].
- **Breast Shadows/Bones:** Rib fractures and other bones in the chest cavity can be confirmed using the CXR.

- **Cardic Silhouette/Costophrenic Angles:** Cardiomegaly is the enlargement of the heart that can point to an underlying problem and is easily seen on the CXR. Similarly, blunted or obscured costophrenic angles also indicate the presence of pleural effusion or similar disease.
- **Diaphragm:** The diaphragm can be displaced when the region between pleura and the lungs is filled up with air indicating an abnormal finding.
- **Edges:** Some disorders can affect the apices, or the edges of the lungs, and they can either cause the tissues to thicken or result in the creation of scar tissues there [104, 105].
- **Fields/Failure:** In the event of some disorders, the lung parenchyma (fields) of the lungs may be flooded. Similar to this, alveoli can fill with fluid or pus, causing them to seem opaque and signalling the presence of a disease.

To generate a thorough report, all of the CXR's aforementioned factors must be taken into account. A skilled radiologist can describe all the findings in a specific CXR in 10 to 15 minutes on average. Moreover, even skilled radiologists occasionally fail to detect some results, which might have grave consequences [106].

## 2.4 Summary

The anatomy of the thoracic cavity and the lungs, imaging methods, and diseases that can affect respiratory health are all covered in this chapter. The circulatory and respiratory systems are shielded and supported by the thoracic cavity, which is made up of the ribs, spine, and sternum. The pleura that covers each lung can be further broken down into lobes and lobules, which in turn can be further broken down into bronchioles and alveoli. Gaseous exchange occurs in the alveoli during breathing, and cellular abnormalities can cause chronic obstructive pulmonary disease. The ability of the lungs to undergo gaseous exchange can be impacted by a variety of respiratory disorders when the alveoli are af-

ected in some manner, whether it be by fluid buildup or abnormal lung tissue growth.

The respiratory system can be examined and a diagnosis made using imaging techniques like MRI, CT scan, and PET scan. The degree of opacification in the lungs is correlated with the severity of respiratory disorders, including COVID-19 and others. The successful management of respiratory illnesses depends on an accurate diagnosis.

X-rays are a helpful diagnostic and monitoring tool for pulmonary illnesses because they provide non-invasive images of the human body. High-density structures, such as bones, absorb X-rays and appear white, whereas low-density structures, such as lungs packed with air or fluid, appear as darker regions. A chest projection or view can be obtained by moving the imaging plate and X-ray beam relative to one another. X-rays only give a two-dimensional perspective of the body. Several lung disorders can be diagnosed and their development and severity tracked using chest X-rays. The important findings from the CXR can also be summarised in the form of an accompanying report.

# Chapter 3

## Literature Review

Chest X-rays remain one of the most used imaging methodologies used to diagnose and track the progression of several chest ailments due to their low cost, accessibility, and portability with over 2 billion X-rays taken annually [36, 84, 85]. The recent COVID-19 pandemic has also overwhelmed radiologists because of an influx in the number of patients as, now, they have to deal with the unprecedented challenges of diagnosing a significantly large number of CXRs [107]. While it is true that the COVID-19 pandemic has put significant strain on an already stretched medical system, even before the pandemic, the CXR evaluation and interpretation demands far exceeded the number that could be handled by the radiologists [108, 109]. A robust diagnosis system able to work on CXRs can help alleviate some of these problems such as reducing the exposure of healthcare staff to the disease [110]. In addition, automated detection and diagnosis systems can aid the healthcare workforce in making better decisions regarding the level of care needed by a patient. Although, Computed Tomography scans are more sensitive than CXR and better for diagnosis [56] and have been used for the classification of COVID-19 as well [111, 112], the ubiquity of the CXR makes it far more practical as a diagnosis tool along with the added benefit of less exposure to radiation.

Various techniques have been proposed for the detection of pulmonary diseases to the



segmentation of lungs, the segmentation of affected regions to localisation of opacities, and from scoring the state of the disease to report generation from the chest X-ray. This section provides an overview of such techniques that have come to rely on the use of deep learning in its various forms. This chapter is mainly divided into 4 sections: classification, segmentation and opacity localisation, severity scoring and quantification, and report generation where each section provides an overview of the techniques used. For all these tasks, methods like artificial neural networks, convolutional and fully convolutional neural networks, vision transformers, recurrent neural networks, and attention-based models among others have been adopted over recent years. The literature analysed in this chapter is structured as shown in Figure 3.1.

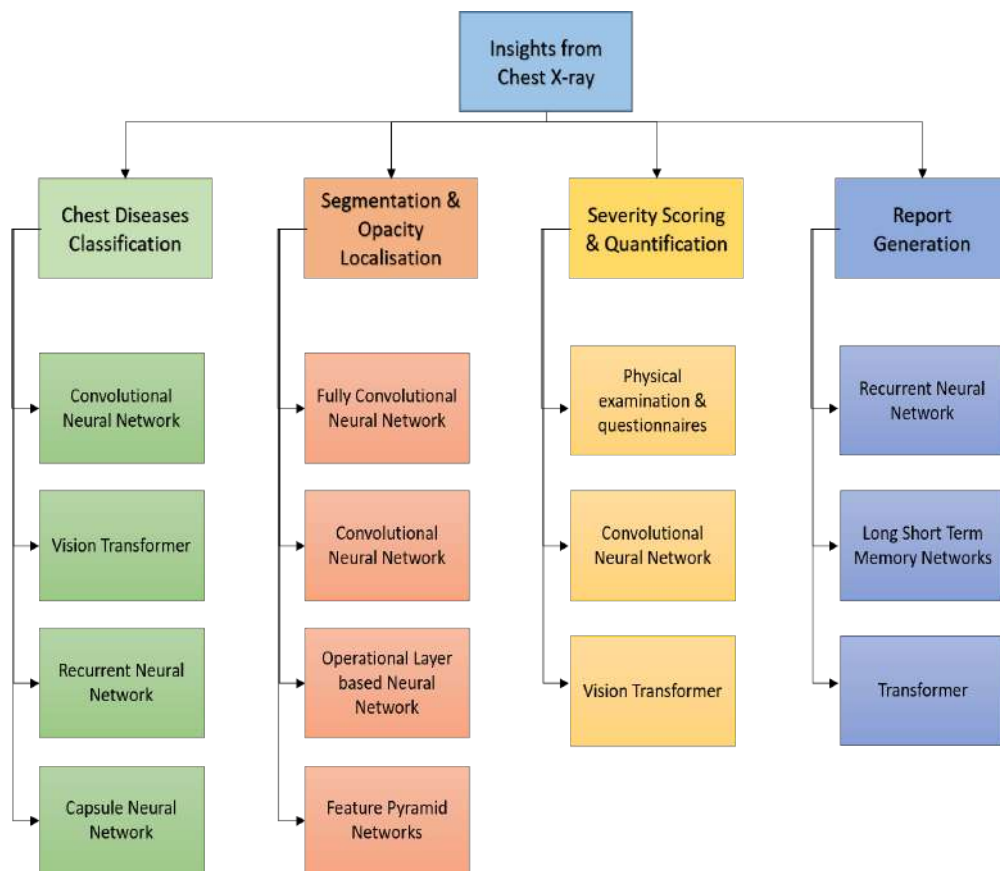


Figure 3.1: Literature analysis for insights from chest X-ray. The techniques discussed in this chapter can be broadly divided into different sub-categories as shown here.

### 3.1 Classification of Chest Diseases

Similar to its use in other domains, Convolutional Neural Networks (CNNs) have been used for automated medical image analysis [113] signaling a departure from classical machine learning techniques and have become state-of-the-art methods for the aforementioned task along with other domains as well. This performance can be attributed to the number of large-scale image datasets that have become available over time, the improved architectures of the deep learning models, and greater computational resources. However, if the data is not representative of the problem domain then the results can be underwhelming. An advantage that deep learning models have over earlier machine learning techniques is that these can automatically infer significant features. This is why the recent focus on diagnosis using CXRs has shifted to CNN as well. For chest X-rays, the diagnosis using images with deep learning models can provide performance that is at par with expert radiologists [114]. As a result, many researchers have employed a variety of CNN architectures for the classification of different pathologies [36, 115, 116], a process that involves the detection of lesions that are classified [117]. The more generalizable such systems are, the more widespread they can be.

In [118] evaluated the performance of six different neural network architectures used for binary classification of CXR images for pneumonia. The researchers documented the effect of transfer learning using CNN architectures of various depths and complexity for a binary classification problem. Their work showed that the number of trainable parameters is not directly related to the performance of the network. A total of six neural network models, with four pre-trained models (VGG16, VGG19, ResNet50, Inception-v3), and two models consisting of two and three convolutional layers, were used for binary classification of CXR images for pneumonia. The custom models designed by the researchers only consisted of convolutional layers followed by activation and max pooling with the final layer being a fully connected layer. The custom networks employed Adam optimiser and cross-entropy loss. The researchers found out that the custom model with three con-

volutional layers and the VGG network had the best performance among all six models with a recall of 98% and 95%, and F1 scores of 94% and 91% respectively.

In [119] introduced a novel training methodology termed curriculum learning in which a ResNet-50 model with multi-label output was trained using a patch-based strategy that focused on anomaly detection. The authors used a varying range of proportions such as 1%, 5%, 20%, 50%, and 100% for the patches extracted from the image around the lesion. After the training of the patch-based model, transfer learning was then used to train the same architecture on entire X-ray images. The results showed that increasing the proportion of the patch extracted around the lesion generally improved the results with the use of 50% data producing the best results for most classes with 95.04% AUC and 91.92% accuracy for the external dataset and 92.60% AUC and 94.54% accuracy for the test split. Gradient Class Activation maps were used to localise abnormalities as well.

In [120], the researchers introduced an attention mechanism by combining two CNN architectures with the ResNet-50 [121] and DenseNet-121 [122] backbone with a neural network. The global CNN branch examined the complete CXR and produced a lesion localisation heatmap which was used to produce a lesion mask. Using the generated mask, the sub-region of the CXR was cropped and the local CNN branch inspected this region. Both the networks were trained in tandem and their feature space was combined using the fusion branch which consisted of only fully-connected layers. This not only allowed the framework to have a shared loss but also allowed the framework to concatenate both local and global features. The researchers validated their methodology on Chest X-ray 14 [123] dataset and achieved a mean AUC of 0.868 for ResNet-50 and 0.871 for DenseNet-121. ResNet-50 was able to outperform DenseNet-121 for three classes.

In order to make use of the free-text reports available with the CXR to localise the region of interest to a certain extent, [124] leveraged a Recurrent Neural Network in combination with a CNN in a novel architecture called TieNet using a combination of text embeddings and image features. Using the text reports as a priori knowledge to generate

attention maps, the authors combined multi-level attention models into a single framework. This not only improved the baseline Area Under the Curve scores but also allowed the researchers to re-purpose TieNet purely for report generation as well. The proposed framework was evaluated on Chest X-ray 14 [123] and Indiana University Chest X-ray datasets [9] and achieved a weighted AUC of 0.748 and 0.798 respectively. In addition, around 900 reports were hand-labeled by radiologists from [123] and the framework achieved a weighted AUC score of 0.719.

Moreover, the recent COVID-19 pandemic was addressed by utilising convolutional neural networks to develop clinical decision support systems. A transfer learning approach was used for avoiding over and underfitting [125]. A VGG16 model pre-trained on ImageNet Large Scale Visual Recognition Challenge (ILSVRC) weights was used. VGG16 has over 138 million trainable parameters with six blocks of 13 convolutions, five max pooling, and three fully connected layers. The model was fine-tuned with CXR images after the diaphragm has been removed using pre-processing steps. The image dataset had 8474 CXR images acquired using the [126–130] repositories, and the model classified the images into normal, pneumonia, and COVID-19 classes. The results without data augmentation were significantly lower compared to the results with data augmentation highlighting the importance of data augmentation for CXR datasets. This model achieved a COVID-19 detection sensitivity of 98.4%, and a three-class accuracy of 94.5%.

Similar to [125], researchers in [131] also worked on data augmentation and hyperparameter optimisation for improving the results of multi-class classification for three classes: normal, pneumonia, and COVID-19. The proposed optimisations increased the VGG-19 and ResNet-50 accuracy by 11.93% and 4.97% respectively. The dataset used for the evaluation of the proposed optimisations consisted of the combination of the datasets from [123, 130] and was termed as COVIDcxr. EfficientNet-B0 [11] was found to achieve the best results based on accuracy, precision, recall and F1-scores compared to other network architectures achieving 95.69%, 96.24%, 94.76% and 95.48% in the aforementioned

metrics respectively. Data augmentation used translation ( $\pm 10\%$ ), intensity shift ( $\pm 10\%$ ), zoom ( $\pm 15\%$ ), horizontal flip ( $\pm 10\%$ ), and rotation ( $\pm 10\%$ ) [132].

Using CXR, a classification network called DFFCNet was proposed by [94] for COVID-19 diagnosis that combined the deep features obtained from different networks using a Deep Feature Fusion Module (DFFM). The model utilised the EfficientNetV2 [133] backbone network for feature extraction along with ResNet which were combined using DFFM and passed onto Support Vector Machine (SVM) for final classification. In order to improve feature information, Spatial Attention (SA) and Channel Attention (CA) modules were also incorporated; the former generates a SA map containing the information location using the spatial relationship of features while the latter produces a CA graph using a sequence of average and maximum pooling along with feature relationship. The suggested framework outperformed the other selected models in experiments achieving 99.9% accuracy for COVID-19 along with F1 score of 99.6% on a dataset that was modified from [134].

Instead of relying on a single CNN classifier for the final output, methods that rely on an ensemble of several classifiers have also been proposed. In [135] the team came up with an ensemble approach comprising Inceptionv3, DenseNet121, Xception, InceptionResNetv2 for the classification of COVID-19, Pneumonia, and normal CXR images. They were able to achieve 98.33% and 92.36% accuracy for binary and multi-class classification respectively on a dataset that was constructed using a combination of different online repositories and consisted of 10,046 images. Similarly, a study compared 16 classifiers for COVID-19 in CXR images for classification for three classes (COVID-19, normal, viral Pneumonia) and different ensemble classification techniques, determining that the majority voting technique yields an accuracy of 99.314% [136].

Vision Transformer (ViT) [137] was more recently repurposed by [96] for the classification of COVID-19. To enable the framework to include the low-level corpus findings, the researchers first trained the transformer architecture using a sizable, publicly accessible

data set [1]. This backbone was then utilised to extract patch embeddings required for input to the vision transformer. Then input from the encoder was then passed through a classification head containing fully connected layers to obtain the final class probabilities. Performance significantly improved because of the combination of vision transformer and the large-scale data set that was used for pre-training. The researchers validated their results on both publicly available [1, 95, 123, 138] and local datasets that were collected in Asan Medical Center, Chonnam National University Hospital, Yeungnam University Hospital, and Kyungpook National University Hospital and were marked by radiologists. The highest AUC achieved by the framework was 0.973 and 0.935 for normal and others class on Yeungnam University Hospital dataset.

Similarly, [139] proposed a novel vision transformer architecture called Input Enhanced ViT (IEViT) for chest X-ray image classification. Inspired by ResNet, the proposed framework builds CNN in parallel to the ViT whose output of the entire image is concatenated to the corresponding transformer encoder layer. The researchers evaluated their framework on four datasets [127, 128, 132, 140] totaling a little over 56000 images and containing four classes: normal, pneumonia, tuberculosis, and COVID-19. An F1 score between 96.39% to 100% was achieved by the framework across all four datasets.

In addition to CNNs, Capsule Networks were used for identifying COVID-19 in CXR images [141]. Their models achieved an accuracy of 98.02% on 1019 images on a dataset constructed from four repositories [128, 130, 142, 143] containing images as normal, COVID-19, and Pneumonia and constructed. In addition, the researchers also worked on a cloud-based application for faster computation.

Some studies have used the combination of CXR and CT images for improving the classification performance [144, 145]. Pre-trained models like Xception, InceptionV3, and EfficientNetV2 were used to identify COVID-19 in CXR and CT images. For the CXR dataset, EfficientNetV2 with fine-tuning performed the best, but the LightEfficientNetV2 model performed the best for the CT data set [146]. The dataset used for training and

evaluation was curated using public datasets for both CXR and CT [147–149]. In a similar vein, in [150], a multi-classification model was proposed for four classes (normal, COVID-19, Pneumonia, and lung cancer) by combining CXR and CT images. The study used VGG19+CNN, ResNet152, ResNet152V2+Gated Recurrent Unit (GRU), and ResNet152V2 + Bidirectional GRU and achieved the best scores with VGG19+CNN model with a 98.05% accuracy. A number of different datasets [3, 126, 148, 151, 152] were combined to create the dataset that the researchers utilised to assess the proposed methodology.

In [153], the researchers used a hierarchical approach for the classification of lung and heart diseases. In the first step, using an ensemble of 3 different models, multi-class classification was performed for three classes: normal, heart disease, and lung disease. The second step consisted of using the same ensembling technique but performed binary classification between normal and seven other diseases; three of which were for heart disease and the other four were pulmonary abnormalities. The researchers used both CNNs in the ensemble and the transformer-based architectures. Architectures based on a modified version of Swin Transformer [154] outperformed the CNN architectures. This methodology was validated on a combined dataset created from [1, 155]. Their proposed methodology achieved an AUC of 95.13% for the multi-class classification and an average AUC of 99.26% for heart diseases and 99.57% for lung pathologies for binary classification.

In [156], presented a novel approach named PaulDi-COVID that ensembled eight different CNN architectures (VGG16, VGG19, ResNet50, ResNet152V2, DenseNet169, DenseNet201, MobileNetV2, and NASNetMobile) for classification of 9 different lung diseases including COVID-19. Transfer learning was used to fine-tune the aforementioned CNNs on an amalgamation of a subset of three different datasets [132, 157, 158]. Different iterations of the models were tested and the best-performing models were used in the ensemble. Using this approach, PaulDi-COVID achieved 99.70% accuracy and 99.24% AUCROC for the classes under consideration. Table 3.1 provides a comprehen-

sive summary of the majority of the methods discussed in this section.

Table 3.1: Summary of recent pulmonary disease classification techniques

<b>Author, Year</b>	<b>Dataset</b>	<b>Methodology</b>	<b>Results</b>
Ibrahim et al., [150], 2021	Public datasets [151, 152], [3, 126, 148]	<ul style="list-style-type: none"> <li>• VGG19 backbone with a custom CNN architecture for classification of COVID-19 using CXR and CT scans</li> </ul>	98.05% accuracy
Cho et al. [119], 2021	Local Dataset (AMCenter & SNU Bundang Hospital) consisting of 9534 images	<ul style="list-style-type: none"> <li>• ResNet-50 for multi-label classification trained on patches extracted from normal and diseased images</li> <li>• Fine-tuning of the ResNet-50 model trained on patches using whole CXRs</li> </ul>	.95 AUC and 91.92% accuracy for the external dataset and .926 AUC and 94.54% for test split
Liu et al. [94], 2022	Curated dataset for COVID-19 [134]	<ul style="list-style-type: none"> <li>• Deep feature fusion using EfficientNetv2 and ResNet101 backbone</li> <li>• Spatial Attention and Channel Attention modules to generate SA and CA maps for better feature representation</li> <li>• SVM for classification using fused features</li> </ul>	99.9% accuracy with an F1 score of 99.6%



**Table 3.1 continued from previous page**

<b>Author, Year</b>	<b>Dataset</b>	<b>Methodology</b>	<b>Results</b>
Guan et al. [120], 2019	ChestX-ray14 [123]	<ul style="list-style-type: none"> <li>• ResNet-50 and DenseNet121 backbone for global and local CNN for entire CXR and lesion region</li> <li>• Fusion deep neural network to combine local and global features from both CNNs for the final classification</li> </ul>	0.868 mean AUC for ResNet50 & 0.871 mean AUC for DenseNet121
Park et al., [96], 2022	Public [95, 123, 138] [1] and local datasets	<ul style="list-style-type: none"> <li>• Pre-train of CNN backbone using [1] for feature extraction</li> <li>• Features from CNN backbone used as embeddings for Transformer architecture</li> <li>• Encoder outputs passed through deep neural network for final classification probabilities</li> </ul>	0.973 and 0.935 AUC for normal and others class on local dataset
Monshi et al., [131], 2021	Combined dataset from [123, 130]	<ul style="list-style-type: none"> <li>• CovidXrayNet based on EfficientNet-B0 with aggressive data augmentation and hyperparameter optimisation</li> </ul>	95.69%, 96.24%, 94.76% & 95.48% accuracy, precision, recall & F1 score respectively

**Table 3.1 continued from previous page**

<b>Author, Year</b>	<b>Dataset</b>	<b>Methodology</b>	<b>Results</b>
Wang et al., [124], 2018	ChestX-ray14 & Indiana University datasets	<ul style="list-style-type: none"> <li>• A ResNet-50 backbone based CNN combined with LSTM for combining the feature maps from the image and word embeddings from the text reports for joint learning of classification labels and producing the text report.</li> </ul>	Weighted AUC of 0.748 & 0.719 on [123] & 0.798 on [9]
Bhardawaj et al., [135], 2021	Combined dataset consisting of 10,046 images	<ul style="list-style-type: none"> <li>• Ensemble consisting of Inceptionv3, DenseNet121, Xception and Inception-ResNetv2</li> </ul>	98.33% & 92.36% accuracy for binary and multi-class respectively
Jain et al. [118], 2020	Kaggle dataset consisting of 5840 images	<ul style="list-style-type: none"> <li>• Custom CNN model with 3 convolutional layers</li> <li>• Pre-trained VGG16, VGG19, ResNet50 and Inception-V3</li> </ul>	98% Recall with 94% F1 score
Heidari et al., [125], 2021	Combined datasets from [126–130]	<ul style="list-style-type: none"> <li>• Fine-tuning of VGG16 pre-trained on ImageNet after the removal of the diaphragm and data augmentation</li> </ul>	98.4% sensitivity & 94.5% accuracy.

## 3.2 Segmentation and Opacity Localisation

Segmentation tasks have benefited from fully convolutional neural networks that are capable of constructing a pixel-level mask for the region of interest, similar to the performance benefits for classification specifically in medical imaging [71, 159–162]. The construction of these masks allows to examine the structural irregularities and sizes of different organs in the chest cavity. While many methods have been proposed for medical image segmentation, most of the segmentation methodologies rely on a U-Net-like architecture with an Encoder-Decoder configuration. This configuration is capable of taking an entire image as input, reducing it to a latent space using successive convolutions and max pooling operations, and then reconstructing a mask with the same size as the input image.

In [163], the team used a pair of CNNs to generate the lung mask. The initial network based on AlexNet [164] performs a two-class classification into lung and no lung region. This network is trained using a 32x32 patch from the original image where a patch is considered to be lung if 20% or more of its region is the lung. This technique allows for better performance on anomalous images. The reconstruction network which was based on ResNet18 [121] and modified such that the last layer performs two-dimensional max-pooling on the output of the preceding fully connected layer, takes the output of the classification CNN, and produces a mask. Due to computational constraints, the input size for the former CNN is 512x512 while for the latter, it is reduced to 128x128. The final mask was a combination of the two masks. The researchers evaluated their methodology on [4] and achieved a dice coefficient of 0.94, and accuracy, sensitivity, and specificity of 96.79%, 97.54%, and 96.79% respectively.

In [84], researchers proposed a hybrid method for generating lung masks for a large dataset [165] using a human-machine collaborative approach. Using [4], as the initial dataset, U-Net [121], U-Net++ [166], and Feature Pyramid Networks (FPN) [167] with a number of backbones were trained. Using the best of these trained models, the team was

then able to generate labels for another subset of the complete dataset which was then vetted by a team of radiologists. This batch methodology was repeated until the masks for the dataset have been generated. In addition to training the segmentation models for complete lungs, the researchers also trained similar models but lung regions that have been affected by COVID-19. This set of segmentation models created two masks for each picture, one for the full lung and the other for just the COVID-19-affected area. The proposed approach achieved sensitivity and specificity values above 99% for COVID-19 detection. In order to aid the classification of pneumonia, [168] also used U-Net segmentation to isolate the area of the lungs for further processing.

In order to aid the classification of COVID-19 through a patch-based classifier, [169] employed a segmentation model based on DenseNet103 to generate the lung segmentation mask. In order to retain just the region containing the lungs, the generated segmentation mask was then used. The classification algorithm produced a class for each of the  $K$  patches from the segmented lungs and leveraged majority voting as the final result. The researchers observed that the network tended to yield under-segmentation when applied to abnormal CXRs due to the considerable disparity between the available normal and abnormal images. Using [3, 4] datasets, the segmentation framework achieved a Jaccard similarity coefficient of 0.955 and 0.932 on [3] and [4] respectively.

To tackle the problem of lack of segmentation masks for medical images, a few-shot semantic segmentation approach with sparse labeled images was proposed by [170]. The proposed method used dense labels in the meta-test and used sparse labels in meta-learning to predict dense labels from sparse ones. In order to train for meta-learning, the dataset not only consisted of CXRs [3–7, 171, 172] but also included CT scans [173] and mammograms [174, 175] and leave-one-out strategy was employed. For [3, 171], the researchers showed that increasing the number of points that are used for segmentation and shots used for training improved the performance of the framework.

Degerli et al. [176] proposed a substantial change to the typical U-Net-like architecture

of segmentation networks by substituting the convolutional layers in the decoder segment with operational layers [177] whose architecture of generative neurons permits a heterogenous network which can be used to model any non-linear transformation. Using DenseNet121 and Inceptionv3 for the encoder, the authors used operational 2D transpose - repeated multiple times - to upsample from the latent space and operational layer with sigmoid activation to obtain the final output. This new architecture was used for the segmentation of COVID-19-affected lung tissue in the [178] dataset which was further expanded in this study as well. The proposed model achieved an accuracy of 99.65% and a precision of 98.09% on the aforementioned dataset.

When the lungs lose their ability to effectively exchange oxygen, this presents as increased density on chest X-rays and is known as opacity. Opacity can be indicative of a number of pulmonary disorders [179–183] therefore its localisation can be of vital importance and even help in the classification of different diseases. In [184], the researchers modeled the classification problem such that the abnormality class could be found using an ensemble of different object detection architectures. Using an ensemble of Yolov5 [185], EfficientDet [12], and FasterRCNN [186], the team trained and tested their model on [155] dataset. To combine the predictions from all three models, Weighted Box Fusion [187] combines the bounding boxes from all the models in a weighted manner. The proposed approach achieved a mean Average Precision (mAP) score of 0.292.

Along the same lines as [184], the researchers in [188], also employed an ensemble of object detectors for opacity localisation for [189] dataset. The outputs from RetinaNet [190] with Mask R-CNN [191] were combined using a custom weighted average in which the individual outputs from each architecture were trimmed using Non-Maxima Suppression (NMS). RetinaNet performed better than Mask R-CNN for validation, hence the final bounding box was altered with a weight ratio of 3:1 (RetinaNet: Mask R-CNN) for each predicted bounding box that overlapped with a predicted bounding box from Mask R-CNN with an IoU threshold larger than 0.5. For the bounding boxes from RetinaNet with

no corresponding bounding box from Mask R-CNN, they were kept as is. This proposed methodology achieved an mAP score of 0.204 on the test set.

In [192], in order to localise opacity for pneumonia detection on [189] dataset, the team used U-Net for pixel-level segmentation. Once the segmentation mask was produced, the bounding box for opacity was generated using the outermost coordinates. In order to improve the performance of their methodology, the researchers also implemented a novel batch control method (BCM) in which the ratio of the samples of each class in each training cycle was varied. Using different numbers of randomly selected positive and negative samples in the batch with size six, going all the way from 100% positive samples to just 17% positive samples, the proposed methodology was able to achieve an F1 score and accuracy of 0.78 when the ratio of positive samples was kept at 83%. U-Net was also substituted with SegNet [193] and PSPNet [194] however, there was not much variation between the results of all these architectures.

In [195], the researchers devised a method to divide the lung into six anatomically correct regions by using a registration method to warp an image to a pre-defined normal CXR template using an affine transformation. The generated bounding boxes were further refined by using heuristics obtained using medical insights. Using a subset of [123] containing 13911 images, each of the six regions in each image was then assigned one of two labels: opacity and normal. These labels were automatically generated using the associated text reports with the CXRs and all the diseases were clumped together as opacity. Using these opacity bounding boxes, RetinaNet [190] with ResNet50 backbone was trained for 15 epochs with an input size of 1024x1024. The team was able to achieve an mAP of 0.29 and an image-level classification accuracy of 77% on the test set. Table 3.2 provides a comprehensive summary of the majority of the methods discussed in this section.

Table 3.2: Summary of recent lung segmentation and opacity localisation techniques

<b>Author, Year</b>	<b>Dataset</b>	<b>Methodology</b>	<b>Results</b>
Oh et al., [169], 2020	JSRT, [3], Montgomery [4]	<ul style="list-style-type: none"> <li>• Patch-based classifier trained on lung region segments segmented using DenseNet103</li> </ul>	Dice coefficient of 0.955 and 0.932 on [3] and [4] respectively
Souza et al., [163], 2019	Montgomery [4]	<ul style="list-style-type: none"> <li>• Generation of lung masks by patch-wise classification using AlexNet</li> <li>• Mask generated by the classification network is used as input for ResNet18 based segmentation network</li> <li>• Combination of both the masks generated for the final mask</li> </ul>	Dice coefficient of 0.94, accuracy, sensitivity, and specificity of 96.79%, 97.54%, and 96.79% respectively
Dergerli et al., [176], 2022	Combined dataset [178]	<ul style="list-style-type: none"> <li>• DenseNet121 and Inceptionv3 used for encoder</li> <li>• Substitution of convolutional layers in the decoder by operational layers</li> <li>• Segmentation masks obtained for only the COVID-19 affected region</li> </ul>	Accuracy of 99.65% and a precision of 98.09%

**Table 3.2 continued from previous page**

<b>Author, Year</b>	<b>Dataset</b>	<b>Methodology</b>	<b>Results</b>
Wu et al., [195], 2020	Chest X-ray 14 [123]	<ul style="list-style-type: none"> <li>• Registration of CXR using a template for dividing the lungs in anatomically correct six regions refined using medical heuristics</li> <li>• RetinaNet architecture with ResNet50 backbone</li> </ul>	0.29 mean Average Precision and 77% image level classification accuracy
Pham et al., [184], 2021	Vindr-cxr, [155]	<ul style="list-style-type: none"> <li>• Ensemble of YOLOv5, EfficientDet, and Faster R-CNN using Weighted Box Fusion</li> </ul>	0.292 mean Average Precision (mAP)
Sirazitdinov et al., [188], 2019	RSNA pneumonia dataset [189]	<ul style="list-style-type: none"> <li>• Ensemble of RetinaNet and Mask R-CNN</li> <li>• Combination of predicted bounding boxes using a ratio of 3:1 (RetinaNet:Mask R-CNN) after the application of Non-maxima suppression on each network's predictions</li> <li>• RetinaNet predicted bounding box kept if no corresponding bounding box found for Mask R-CNN</li> </ul>	0.204 mean Average Precision (mAP) on [189]



**Table 3.2 continued from previous page**

<b>Author, Year</b>	<b>Dataset</b>	<b>Methodology</b>	<b>Results</b>
Chang et al., [192], 2022	RSNA pneumonia dataset [189]	<ul style="list-style-type: none"> <li>• Bounding box generation after pixel-level segmentation using U-Net</li> <li>• Novel batch control method with an 83% ratio of positive to negative samples</li> </ul>	F1 score and accuracy of 0.78
Tahir et al., [84], 2021	Covid- qu [165]	<ul style="list-style-type: none"> <li>• Generation of lung segmentation masks for a large dataset [165] using a human-machine collaborative approach</li> <li>• U-Net, U-Net++, and FPN used for mask segmentation</li> <li>• In addition to complete lung masks, masks for the COVID-19 affected regions also generated using the same approach</li> </ul>	Sensitivity & specificity above 99% on [165]

### **3.3 Severity Scoring and Quantification**

Knowing the severity of the diseases can aid medical care providers in better clinical management of a variety of thoracic diseases because they can vary widely in severity. While in the past, methods like CXR findings, questionnaires, and physical examinations have been used in order to gauge the severity of disorders such as Chronic Lung Dis-

ease (CLD), Acute Respiratory Infections (ARIs), and Severe Acute Respiratory Infection (SARI) [93, 196–198], the recent focus has been on the use of deep learning methods such as CNNs and transformers to provide an object score using just a CXR. The ability to estimate the severity score from a single CXR allows for a quick assessment of the patient’s level of care and can significantly affect how the patient is managed. The recent COVID-19 pandemic has also received attention for severity quantification.

In [179], Warren et al. proposed a non-invasive method of determining the severity of pulmonary edema in Acute Respiratory Distress Syndrome by the assessment of chest X-ray. The Radiological Assessment of Lung Oedema (RALE) score proposed by the team took into account the density of opacification and the extent of consolidation. The lungs were divided into four quadrants with each lung having two portions and each of these four quadrants was given a score for the extent of consolidation and opacification density with the former having a range of 0 to 4 while the latter was graded on a scale of 1 to 3 where 1 denotes hazy, 2 denotes moderate and 3 denotes dense. For each quadrant, the product of both these scores was taken and for the RALE score for the complete CXR, all the quadrant scores were summed up, thus, giving a range of 0 to 48. This scoring protocol proved to be an excellent measure for assessing the severity of edema. [199] came up with a modified RALE (mRALE) score, an alteration of score in [179] such that each lung was not divided into quadrants, instead the score was applied on each lung. The researchers also presented a novel Siamese network based on DenseNet121 which was capable of producing a Pulmonary X-Ray Severity (PSX) score by comparing a COVID-19 positive image with a group of normal images taken from [1] which was also used during the pre-training phase. Their results showed that the severity score from a CXR was a predictor of need for intubation or death within 3 days of when the CXR was taken with an Area Under Receiver Operating Characteristic Curve (AUROC) of 0.8.

[200] devised a COVID-19 severity scoring system by dividing each lung anatomically into three regions for a total of six regions where each region was given a score of 1 if

opacity was present and 0 if it was absent resulting in a maximum score of 6 if opacity was prevalent in all six regions. The upper region covered the region from the apices to the superior hilar markings, the middle region was between the superior hilar and inferior hilar markings while the lower region was between the inferior hilar markings to costophrenic sulcus. To validate this scoring system, the study included a cohort of 338 patients with a median age of 39 years and the results showed that a severity score of 2 or more predicted that the patient was admitted to the hospital while a severity score of 3 or more was a predictor of whether the patient would be intubated or not.

This scoring methodology in [179] was adapted by [99] after modification such that the extent of ground-glass opacities in each lung instead of each quadrant was graded on a scale of 0 to 4 with 0 showing no involvement, 1 showing less than 25% involvement, 2 showing between 25% and 50% involvement, 3 showing between 50% and 75% involvement and 4 showing greater than 75% involvement. The final score was the aggregate of scores of both lungs and this method yielded a sensitivity of 69% for a cohort of 64 patients.

[98] trained a pair of regression models based on DenseNet to predict the extent of lung involvement and the degree of opacity, a scoring system which was inspired by both [99, 179]. To each lung, two scores were assigned: the extent of consolidation which was scored in the exact manner as [99], and the degree of opacity which could have a value of 0 to 3 where 0 meant no opacity, 1 meant ground glass opacity, 2 meant consolidation and 3 indicated white-out due to opacity. The pre-training used 7 publicly available datasets [1, 10, 123, 201–204] with COVID-19 samples while the testing was on [130] which included COVID-19 samples. The DenseNet architecture was trained with varying numbers of output classes including a single output for lung opacity, four outputs for lung opacity, infiltration, consolidation, and pneumonia, and 18 outputs containing almost all the classes from the non-COVID datasets. The proposed mechanism was able to achieve a Mean Absolute Error (MAE) of 1.14 for the extent and 0.78 MAE

for the degree of opacity. VGG16 with frozen convolutional layers and transfer learning was used by [97] for the [130] dataset with an 80/20 split while still keeping the same scoring protocol as [98]. Zhu et al's [97] approach achieved a Mean Absolute Error of 0.93 for the extent and 0.91 MAE for the degree of opacity.

Irmak et al [100] suggested using CNN to categorise the severity of COVID-19 CXR images into four classes: mild, moderate, severe, and critical. The study used a total of 3260 images from nine publicly available CXR datasets [128, 130, 205–209]. A two radiologists' opacity score served as the foundation for the disease severity score. The custom CNN model employed by the team had 16 layers with an input size of 227x227 pixels and softmax activation was used in the final layer. The results of the suggested proposed architecture were superior to ResNet-101, AlexNet, VGG-16, and other networks achieving an accuracy of 95.52%. The hyperparameters for the proposed architecture were optimised using grid search. A similar approach to classify COVID-19 cases into severe and non-severe classes was used by researchers [210–215] as well. However, their methodologies relied upon CT scans instead of CXRs and included varying numbers of samples in the input data with varying performance.

By dividing the lungs into three equal segments, [95] employed the Brixia score [101], which assigned a value from 0 to 3 to each of the six lung regions. A score was assigned to each segment to represent the degree of lung damage in COVID-19 cases, with 0 denoting no abnormalities and 3 denoting substantial aberrations. In addition, the researchers proposed BS-Net, which utilised a pyramid technique to combine features gathered at various scales. Before the CXR could be used as the input for BS-Net, automatic alignment of the segmentation masks generated using U-Net++ occurred using a multi-feature region aligner and a multi-feature area aligner. The aligned features were passed through a series of convolutional blocks which outputted the Brixia score in a 3x2 grid. In order to improve the explainability of the framework, the team also came up with novel super-pixel explainability maps which first grouped together the pixels based on their similarity

and then using the predictions from the network, produced an Explainability map  $E$  as an aggregate of the differences of the probabilities for each class. The methodology was validated on a local dataset as well as other public datasets and achieved a Mean Absolute Error of 0.424 compared to the markings of radiologists containing 450 images on the former.

The framework based on a vision transformer for COVID-19 diagnosis by [96] also had a severity quantification component to it. The deep features from the backbone model that were used for the prediction of disease class were also employed for severity maps after being processed by a lightweight network. The saliency maps produced by this network were combined with the segmentation masks of the lungs. The researchers utilised a scoring method similar to that used by [95]. This scoring system was devised such that each of the six lung areas could only receive a maximum score of 1, with a total score of 6 being assigned to the overall CXR. The greatest value in each of the six regions was obtained using max pooling, and it was then thresholded to make it either a zero or a one. In order to validate the proposed approach, local datasets along with [95] dataset was used. This methodology was able to achieve a minimum Mean Square Error (MSE) of 1.441 and an MAE of 0.843 on one of the external datasets. The same approach achieved an MSE of 1.683 on the consensus test subset from [95].

In order to quantify the severity of COVID-19 [84] made use of both the lung segmentation mask and the segmentation mask indicating the area affected by COVID-19. The severity was expressed as a percentage that was determined by dividing the total lung pixels by the affected lung pixels. The value allowed for an independent assessment of each lung and showed that the infection had severely harmed a significant percentage of the lungs if the value was high. This technique achieved a dice coefficient of 0.882. Table 3.3 provides a comprehensive summary of the majority of the methods discussed in this section.

Table 3.3: Summary of recent severity scoring and quantification techniques

Author, Year	Dataset	Methodology	Results
Park et al., [96], 2022	Local dataset & test set from [95]	<ul style="list-style-type: none"> <li>• Pre-trained CNN backbone using [1] for feature extraction</li> <li>• Features from CNN backbone used as embeddings for Transformer architecture</li> <li>• Encoder outputs passed through a lightweight CNN for generation of pixel-wise severity scoring</li> <li>• Max pooling along with thresholding applied to each region to obtain a final severity score for each region</li> </ul>	<p>Mean Square Error (MSE) of 1.441 and an MAE of 0.843 on one of the external datasets along with an MSE of 1.683 on the consensus test subset from [95]</p>
Warren et al., [179], 2018	Local Dataset	<ul style="list-style-type: none"> <li>• Division of lungs into 4 quadrants with each quadrant covering half a lung</li> <li>• Scoring based on the extent of consolidation (0 to 4) and opacification density (1 to 3)</li> <li>• A single-digit score between 0 to 48 obtained by the summation of the product of both scores in each quadrant</li> </ul>	<p>An excellent indicator of severity of ARDS by just using the CXR</p>

**Table 3.3 continued from previous page**

<b>Author, Year</b>	<b>Dataset</b>	<b>Methodology</b>	<b>Results</b>
Wong et al., [99], 2020	Local dataset	<ul style="list-style-type: none"> <li>• Similar to the scoring protocol used by [179] with each lung being scored instead of each quadrant</li> </ul>	69% sensitivity for 64 patients
Singo- roni et al., [95], 2022	Local dataset	<ul style="list-style-type: none"> <li>• Brixia [101] scoring system used to assign a score (0 to 3) to each of the six lung zones</li> <li>• Automatic alignment of the segmentation masks generated using U-Net++ using a multi-feature region aligner and a multi-feature area aligner</li> <li>• BS-Net with a pyramid technique to combine features at various scales</li> <li>• Explainability maps using a super-pixel approach</li> </ul>	Mean Absolute Error (MAE) of 0.424 on a test set of 450 images
Irmak et al., [100], 2021	Combined dataset from [128, 130, 205], [206–209], with 3260 images	<ul style="list-style-type: none"> <li>• Severity quantification using four classes</li> <li>• Custom CNN with 16 layers</li> <li>• Hyperparameter optimization using grid search</li> </ul>	95.52% accuracy

**Table 3.3 continued from previous page**

<b>Author, Year</b>	<b>Dataset</b>	<b>Methodology</b>	<b>Results</b>
Cohen et al., [98], 2020	Combined dataset from [1, 123, 201], [10, 202, 203], [130, 204],	<ul style="list-style-type: none"> <li>• Scoring system inspired by both [99, 179]</li> <li>• Scoring based on the extent of consolidation (0 to 4) and opacification density (0 to 3) with the higher number being more severe</li> <li>• DenseNet-based pair of regression models to predict each score individually</li> <li>• Different number of output classes for 1 to 18</li> </ul>	Mean Absolute Error (MAE) of 1.14 for the extent of consolidation and 0.78 MAE for the opacification degree on [130]

### **3.4 Automated Report Generation Through Natural Language Processing**

Medical reports provide additional information for different imaging modalities that may not be apparent from the imaging itself making such a report quite useful. Automated report generation from a single or a pair of CXRs can allow for rapid evaluation of a patient’s condition as not only it can be generated quickly but can be made less error-prone as compared to a non-experienced radiologist [216]. The use of diverse deep learning-based techniques has become more prevalent where the problem can either be modeled as a retrieval problem due to a radiology report following a template [217] or a generation problem where a free-text report is generated from scratch.



[218] employed the retrieval-based approach and proposed Knowledge-Driven Encode, Retrieve, Paraphrase (KERP) framework. At each stage of the architecture, a graph transformer is employed using either the latent space output or the graph output from the previous module. Using a Graph Transformer (GTR), the features in the latent space generated by the CNN are flattened such that each element represents a node in a graph and then turned into an abnormality graph. The abnormality graph is again transformed using GTR into a template sequence, where each template sequence's length is denoted by  $N_s$ . The generated sequences of templates are transformed in the *Paraphrase* module where the template sequences are enriched with case-specific information along with improving the text itself to make it more dynamic. Additionally, using the abnormality graph, GTR is once more used to transform it into a disease classification graph with multi-labels. In order to validate their approach, the researchers used two datasets [9], one of which was private. This methodology was able to achieve a Bilingual Evaluation Understudy (BLEU) 1 score [219] of 0.482 and a BLEU-4 score of 0.162 on [9] while on the private dataset, a score of 0.673 and 0.473 was achieved for BLEU-1 and BLEU-4 respectively.

Li et al. [220] proposed a Hybrid retrieval-generation reinforced (HRGR) agent - a similar approach to [218] of encoding visual features extracted through a CNN [122] as a context vector which was used to generate topic states by the use of stacked RNN layers with attention. A retrieval policy module was then used to determine whether a template should be retrieved or a new sentence generated depending on the probability and this module was trained using hierarchical reinforcement learning. On the [9] dataset, this research methodology achieved a BLEU-1 score of 0.438 and a BLEU-4 score of 0.151. On the CH-CXR private dataset, scores of 0.673 and 0.486 for BLEU-1 and BLEU-4 were achieved.

Researchers in [221] introduced two additional modules to a traditional sequence-to-sequence architecture where latent space features extracted from an image are regarded as input sequences, putting forth a novel framework for radiology report generation. A rela-

tional memory module was added that is utilised to store the pattern information in order to take advantage of similarities between the images. The matrix from timestep  $t - 1$  is used as the query, along with the other two components which are simply the concatenation of the same matrix with the preceding output to produce the key and value pairs for each attention head at timestep  $t$ . ResNet101 [121] is used as the feature extractor and the feature length is capped at 2048. Using a Multilayer Perceptron to predict the  $\Delta\gamma$  and  $\Delta\beta$  from the matrix from the memory module, the second module called Memory-driven Conditional Layer Normalization (MCLN) is responsible for improving the generalisation capabilities of the decoding process. The performance of the framework is tested on two datasets: Indiana [9] and MIMIC-CXR [10]. The researchers were able to achieve a BLEU-1 and BLEU-4 score of 0.47 and 0.165 respectively on the former dataset and a BLEU-1 and BLEU-4 score of 0.353 and 0.103 on the latter.

[222] employed Long Short-Term Memory (LSTM) [223] in conjunction with CNN and attention module to generate X-ray reports. The VGG19 architecture is modified by adding fully connected layers at the end which output a feature vector of length 256. Using an embedding layer, the latent space vector is converted to 256x22 embedding where 22 represents the maximum size of the findings in [9] dataset. An attention mechanism generates a context vector prior to the embeddings being used by the LSTM for the generation of the output words, which is then combined with the hidden state to forecast the following word. The inclusion of a context vector allows the framework to focus on only the information that is most relevant to the region of interest. Similar to [221], two datasets ([9] and [10]) were used for gauging the performance of the proposed methodology which achieved a BLEU-1 score of 0.58 and BLEU-4 score of 0.155 on Indiana dataset and BLEU-4 score of 0.153 on MIMIC-CXR.

Liu et al. [224] used the Knowledge Graph Auto-Encoder (KGAE), an unsupervised method that relies on a pre-built knowledge graph, to reduce the reliance on datasets that contain image-report pairs. The knowledge graph, which is utilised for the encoder and

decoder, is built using [10] reports instead of the images, with the abnormalities acting as nodes and the normalised co-occurrence of these abnormalities in the reports serving as edge weights. ResNet50 [121] and Transformer [137] are used to extract the embeddings which are then used as queries to the pre-built knowledge graph making the encoder Knowledge-driven (KE). In a similar fashion, a knowledge-driven decoder is used to generate textual report from the graph representations of the image. The researchers also modified their proposed approach to make it semi-supervised and supervised by incorporating some of the image-report pair in the generation of the knowledge graph. The supervised approach outperformed both the semi-supervised and the unsupervised approach achieving a BLEU-1 score of 0.512 and 0.369 for [9] and [10] respectively and a BLEU-4 score of 0.179 and 0.118 for the same datasets.

Srinivasan et al. [225] suggested a report-generation methodology that utilised Image Level Chest Features (ICLF) extracted using a custom CNN and Tag Level Chest Features (TCLF) extracted using Multi-Head Attention (MHA) by using only the lung area clipped using the Single Shot Detector (SSD) [226]. The CNN used for ICLF was modified such that it could classify the image as normal or abnormal based on 16 overlapping patches of 128x128 each in order to retrieve the tags for only diseased images. Triplet loss was used for training this sub-module. Using a concatenated version of ICLF passed through MHA, the top 16 tags are selected from a total of 237. It employs two [137]-like encoders, one of which uses TCLF as input to produce embeddings and the other of which uses ICLF. Similarly, two decoders are used as well; one for generating *Findings* output based on both embeddings while the other one generates *Impressions* based on output features from the first decoder. Both *Impressions* and *Findings* are concatenated together to produce the final report. On [9] dataset, the proposed methodology was able to achieve a BLEU-1 score of 0.464 and a BLEU-4 score of 0.158.

In order to mimic the process of report writing by the radiologists, [227] proposed a Posterior and Prior Knowledge Exploring and Distilling (PPKED) approach. Three

primary components formed this approach: Multi-domain Knowledge Distiller (MKD), Prior Knowledge Explorer (PrKE), and Posterior Knowledge Explorer (PoKE). PoKE being the first component of the framework was used to extract the abnormality information from the image embeddings from ResNet152 [121] feature extractor and the embeddings from a bag of Tags  $T$  containing 20 of the most common abnormality tags. Using multi-head attention and Feed Forward Network (FFN), the output from both the image and tag embeddings is normalised and added to create the final output of PoKE. The next component, PrKE, uses the results of PoKE to generate two more outputs: Prior Working Experience ( $W_{Pr}$ ), which is created by generating embeddings from the text reports of the 100 closest images using cosine similarity from the training corpus, and Prior Medical Knowledge ( $G_{Pr}$ ), which is created by combining the results of PoKE with embeddings from a knowledge graph created using the same abnormality tags. MKD performs the task of the decoder by using the output of PoKE and PrKE. The final probability of the words is calculated using Softmax activation with Cross-Entropy loss for training the model. [9] and [10] were used for gauging the performance of the proposed framework which achieved a BLEU-1 score of 0.483 and BLEU-4 score of 0.168 on the Indiana dataset and a BLEU-1 score of 0.36 and BLEU-4 score of 0.149 on MIMIC-CXR.

[228] proposed including two different forms of knowledge — general domain knowledge and image-specific knowledge — in the report generation task, the former of which is independent of the input and the latter depends on the input. The manually constructed RadGraph [229] was utilised for generic domain knowledge. Latent-space features of the input image extracted via a CNN are used to retrieve similar reports from a report pool in order to obtain image-specific information. In order to obtain triplets (source entity, relation, and target entity), related reports are fed via the [229] relation extractor. The novel knowledge-enhanced attention module aggregates the embeddings generated from the generic domain knowledge with the embeddings extracted from the visual feature extractor. The triplets derived from similar reports are concatenated, passed through Clinical BERT [230], and aggregated with the visual features in the same way as generic

domain knowledge for image-specific domain knowledge. For report generation, the decoder from [137] is used with the generic knowledge, specific knowledge, and the visual features concatenated as input. Two datasets ([9] and [10]) were used for gauging the performance of the proposed framework which achieved a BLEU-1 score of 0.496 and BLEU-4 score of 0.178 on the Indiana dataset and a BLEU-1 score of 0.363 and BLEU-4 score of 0.115 on MIMIC-CXR. In addition, the framework also achieved a CIDEr [231] score of 0.382 and 0.203 on the two aforementioned datasets respectively.

In order to mimic the practice of consolidating the opinion of multiple experts for difficult cases, researchers in [232], presented a novel transformer-based framework termed ME-Transformer. Similar to the Mixture of Experts technique, the team modified the Transformer Encoder and Decoder to incorporate a Multi-expert Bilinear Attention Encoder and Decoder. Special learnable *expert* tokens are added in both the Encoder and Decoder. Using these *expert* tokens embeddings, the orthogonal loss is computed to maximise the similarity between the tokens in the encoder and decoder. This approach allows the framework proposed in [232] to have an ensemble-like behavior without having the disadvantage of having a lot of parameters. The performance was gauged using [9] and [10] datasets where a BLEU-1 score of 0.483 and 0.386, and a BLEU-4 score of 0.172 and 0.124 was achieved respectively.

In order to learn multi-level visual representation and adaptively condense the data with contextual and clinical knowledge for word prediction, [233] proposed a Knowledge-injected U-Transformer (KiUT). To specifically characterize interactions between various modalities, a U-connection were established between the encoder and decoder. In order to further improve the performance of the framework, a symptom graph of the common lung pathologies was created using the correlation, location and characteristics of different symptoms of these pathologies and used as Clinical Knowledge Signal in conjunction with Visual Knowledge Signal and Contextual Knowledge Signal. The combination of these three signals was treated as injected knowledge and was fed into the Injected

Knowledge Distiller which combined this knowledge with the Decoder embeddings using Multi-Head Attention. Using this elaborate technique, the framework proposed in this paper achieved a BLEU-1 score of 0.525 and 0.393 and a BLEU-4 score of 0.185 and 0.113 on [9] and [10] datasets respectively. Table 3.4 provides a comprehensive summary of the majority of the methods discussed in this section.

Table 3.4: Summary of recent CXR report generation techniques

<b>Author, Year</b>	<b>Dataset</b>	<b>Methodology</b>	<b>Results</b>
Liu et al., [224] 2021	Indiana [9], MIMIC [10]	<ul style="list-style-type: none"> <li>• A pre-built knowledge graph using [10] reports with the abnormalities as nodes and the normalised co-occurrence of the abnormalities serving as edge weights</li> <li>• Knowledge-driven encoder uses embeddings and the knowledge graph</li> <li>• Knowledge-driven decoder uses image graph representation for generating reports</li> <li>• Proposed approach also modified to work semi-supervised and supervised by incorporating some of the image-report pair in the generation of the knowledge graph.</li> </ul>	BLEU-1: 0.512 BLEU-2: 0.327 BLEU-3: 0.240 BLEU-4: 0.179 on [9] BLEU-1: 0.369 BLEU-2: 0.231 BLEU-3: 0.156 BLEU-4: 0.118 on [10]

Table 3.4 continued from previous page

Author, Year	Dataset	Methodology	Results
			BLEU-1: 0.438
			BLEU-2: 0.298
			BLEU-3: 0.208
		• Context vector generated using image encoder and is based on	BLEU-4: 0.151
Li et al., [220], [220], 2018	Indiana [220], Private Dataset	image latent space features	CIDEr: 0.343
		• Stacked RNN layers with self-attention used to generate hidden topic states	on [9]
			BLEU-1: 0.673
			BLEU-2: 0.587
		• Sentence generation or retrieval is decided by a retrieval module	BLEU-3: 0.530
			BLEU-4: 0.486
			CIDEr: 0.2895
			on private dataset
			BLEU-1: 0.470
			BLEU-2: 0.304
		• Addition of relational memory module for pattern information between similar images	BLEU-3: 0.219
Chen et al., [221], 2020	Indiana [9], MIMIC [10]		BLEU-4: 0.165
		• Memory driven Conditional Layer Normalization for predicting $\Delta\gamma$ and $\Delta\beta$ for better generalization capabilities	METEOR [234]: 0.187
			on [9]
			BLEU-1: 0.353
			BLEU-2: 0.218
			BLEU-3: 0.145
			BLEU-4: 0.103
			METEOR: 0.142 on [10]

**Table 3.4 continued from previous page**

<b>Author, Year</b>	<b>Dataset</b>	<b>Methodology</b>	<b>Results</b>
Li et al., [218], 2019	Indiana [9], Private Dataset	<ul style="list-style-type: none"> <li>• Latent space visual features are encoded as an abnormality graph using Graph Transformer</li> </ul>	BLEU-1: 0.482 BLEU-2: 0.325 BLEU-3: 0.226 BLEU-4: 0.162
		<ul style="list-style-type: none"> <li>• Template sequence generated using Graph Transformer from the abnormality graph</li> <li>• Paraphrase Graph Transformers takes the template sequences and converts them to the final report</li> <li>• Abnormality Graph also used for the prediction of diseases class for the input image</li> </ul>	CIDEr [231]: 0.28 on [9] BLEU-1: 0.673 BLEU-2: 0.588 BLEU-3: 0.532 BLEU-4: 0.473 CIDEr: 0.285 on private dataset
Yang et al., [228] 2022	Indiana [9], MIMIC [10]	<ul style="list-style-type: none"> <li>• Integration of generic domain knowledge and image-specific knowledge obtained from manually constructed knowledge graph with visual features</li> </ul>	BLEU-1: 0.496 BLEU-2: 0.327 BLEU-3: 0.238 BLEU-4: 0.178 on [9]
		<ul style="list-style-type: none"> <li>• Novel Knowledge-enhanced attention head was used for the aggregation of embeddings</li> <li>• Standard decoder was used for report generation</li> </ul>	BLEU-1: 0.363 BLEU-2: 0.228 BLEU-3: 0.156 BLEU-4: 0.115 on [10]



**Table 3.4 continued from previous page**

<b>Author, Year</b>	<b>Dataset</b>	<b>Methodology</b>	<b>Results</b>
Sirshar et al., [235], 2022	Indiana [9], MIMIC [10]	<ul style="list-style-type: none"> <li>• Feature extraction through VGG19 which is then converted to 256x22 embeddings using an embedding layer</li> <li>• Context vector generation using attention mechanism from the embeddings</li> <li>• Word prediction using LSTM from the context vector and embeddings</li> </ul>	BLEU-1: 0.582 BLEU-2: 0.342 BLEU-3: 0.263 BLEU-4: 0.155 on [9] BLEU-4: 0.153 on [10]
Srinivasan et al., [225], 2020	Indiana [9]	<ul style="list-style-type: none"> <li>• Image Level Chest Features (ICLF) are extracted through CNN along with Tag Level Chest Features (TCLF) using Multi-Head Attention</li> <li>• Two encoder-decoder pairs are used with ILCF and TLCF for generating Findings and Impressions</li> <li>• Findings and Impressions are combined for the final report.</li> </ul>	BLEU-1: 0.464 BLEU-2: 0.301 BLEU-3: 0.212 BLEU-4: 0.158 on [9]

## 3.5 Research Gaps

Through the literature review of techniques for disease classification, lung segmentation, severity scoring, and report generation, the following research gaps have been identified:

1. Lack of a single framework that not only classifies a CXR image in a particular disease class but also segments the lungs' opacity regions if the lungs are diseased.
2. While COVID-19 has been the focus of recent severity classification attempts, a lack of severity classification for different pathologies still exists.
3. A generic CXR analysis framework that can generate radiology reports that are closer in content to a radiologist is required.
4. There is no benchmark CXR dataset available with sample images from Pakistan or any South Asian country with the labels of chest diseases.

## 3.6 Summary

This literature survey provides a brief overview of the techniques that have been used for lung disease classification, lung segmentation, severity scoring and quantification, and radiological report generation. It is quite clear that all the modern approaches for the aforementioned tasks rely on the use of deep learning methodologies such as convolutional neural networks, and more recently, vision transformers. The use of these techniques has been made possible due to the availability of large-scale CXR datasets containing both the images and the reports. Using a single framework for all the tasks requiring just one of two chest X-rays has been the main focus of recent techniques. In addition, the optimisation of already existing techniques whether it is in terms of the number of trainable parameters or reduction in inference time has also been the focus of recent studies.

# Chapter 4

## Materials

Chest X-rays are the primary tool used by radiologists to identify pulmonary disorders. The rise of the availability of several large, public CXR datasets has allowed for the training of different artificial intelligent decision support systems that can be integrated with existing infrastructure and can help combat the unprecedented challenges of diagnosing a significantly large number of CXRs by the radiologists all around the globe [107]. The two frameworks that serve as the foundation of this research endeavor are one that generates medical reports from a single chest X-ray and the other that does classification, segmentation, and severity grading. This chapter discusses the utilisation of various data sets that can be categorised according to their intended use for the different frameworks.

### 4.1 Classification Datasets

The main goal of classification is to assign one or more labels, each with a different level of confidence, to the complete contents of the provided image. The underlying condition can be swiftly identified with the aid of models trained for classification, and radiologists can then thoroughly examine it for further insights.

### 4.1.1 Chest Expert (CheXpert)

A group of subject matter experts from various Stanford University departments has generated the dataset known as CheXpert [1] that has been gathered over a period of 15 years from October 2002 to July 2017. It consists of a significantly large number of CXRs totaling 224,316 which have been gathered from 65,240 patients and were made public in 2019. The chest X-rays included in this dataset are primarily captured using either the frontal view (both AP and PA) or the lateral view. For each patient, the screening study can contain both the frontal and the lateral view. Figure 4.1 shows the chest X-ray of one such study from the dataset.

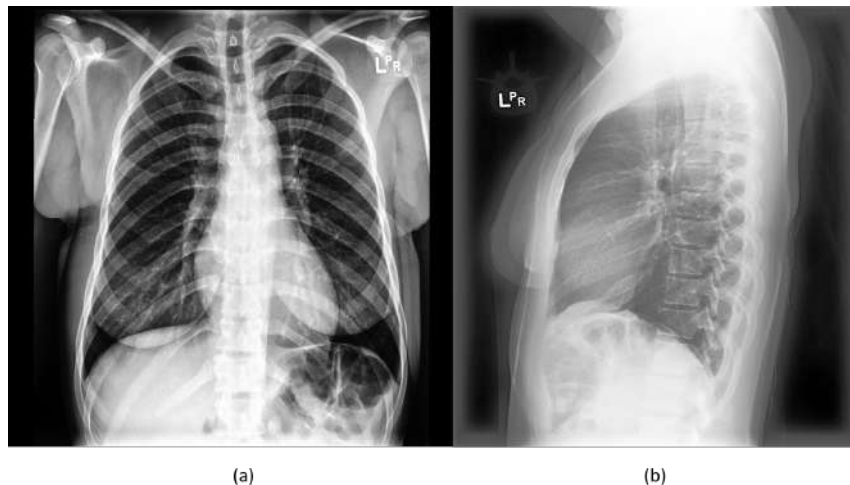


Figure 4.1: The frontal (a) and lateral (b) view of the thoracic cavity of a patient captured using CXR. Image taken from [1]

The CXRs included in this study only had associated reports which were used to generate the class labels assigning one or more labels based on the report. These reports were parsed through a rule-based labeler to extract possible mentions of the classes and then classify those mentions as positive, uncertain, or negative. The researchers opted for 14 classes covering major ailments of the chest cavity including *cardiomegaly*, *enlarged cardiomeastinum*, and even the presence of *support devices*. The labeler results were verified using a set of 1000 radiology reports [1]. Table 4.1 shows the class distribution in terms of samples and percentages by the mentions extracted by the labeler.

Table 4.1: Distribution of samples according to the 14 classes in the CheXpert data set [1]

<b>Pathology</b>	<b>Positive (%)</b>	<b>Uncertain (%)</b>	<b>Negative (%)</b>
Atelectasis	29333 (15.63)	29377 (15.66)	128931 (68.71)
Cardiomegaly	23002 (12.26)	6597 (3.52)	158042 (84.23)
Consolidation	12730 (6.78)	23976 (12.78)	150935 (80.44)
Edema	48905 (26.06)	11571 (6.17)	127165 (67.77)
Enlarged Cardiom.	9020 (4.81)	10148 (5.41)	168473 (89.78)
Fracture	7270 (3.87)	484 (0.26)	179887 (95.87)
Lung Lesion	6856 (3.65)	1071 (0.57)	179714 (95.78)
Lung Opacity	92669 (49.39)	4341 (2.31)	90631 (48.30)
No Finding	16627 (8.86)	0	171014 (91.14)
Pleural Effusion	75696 (40.34)	9419 (5.02)	102526 (54.64)
Pleural Other	2441 (1.30)	1771 (0.94)	183429 (97.76)
Pneumonia	4567 (2.43)	15658 (8.35)	167407 (89.22)
Pneumothorax	17313 (9.23)	2663 (1.42)	167665 (89.35)
Support Devices	105831 (56.40)	898 (0.48)	80912 (43.12)

Ignoring the class *support devices*, the maximum number of samples are found of *lung opacity* followed by *pleural effusion* and *edema* respectively. This dataset also contains a separate test subset containing 235 images from 200 patients that has been annotated by a team of three radiologists for the presence or absence of the 14 pathologies.

#### 4.1.2 Brazilian Labeled Chest X-ray (BRAX)

Brazilian labeled Chest X-ray [2] dataset was generated from Picture Archiving and Communication System (PACS) from the Hospital Israelita Albert Einstein (HIAE) in Sao Paulo and was made public in 2022 in Digital Imaging and Communications in Medicine (DICOM) format [236]. DICOM allows the X-rays to be stored in a higher bit format such as 12 bits per pixel which can be helpful for applying the windowing operation. Although compared to CheXpert [1], BRAX is a relatively small dataset, however, it still consists of 24,959 chest exams and 40,967 X-rays. Similar to CheXpert, X-rays in BRAX also contain both the frontal and the lateral view. Building upon the image labeler in [1], the team modified it for Portuguese to generate the class labels as the associated reports were in the aforementioned language. The rest of the methodology for the labeling was kept the same. The distribution of the number of samples for different pulmonary pathologies

Table 4.2: Distribution of samples according to the 14 classes in the BRAX data set [2]

<b>Pathology</b>	<b>Positive (%)</b>	<b>Uncertain (%)</b>	<b>Negative (%)</b>
Atelectasis	3518 (8.59)	0	41 (0.1)
Cardiomegaly	3984 (9.72)	0	28000 (68.35)
Consolidation	3157 (7.71)	0	19 (0.05)
Edema	50 (0.12)	0	0
Enlarged Cardiom.	71 (0.17)	2 (0.00)	26212 (63.98)
Fracture	624 (1.52)	0	16405 (40.04)
Lung Lesion	1290 (3.15)	19 (0.05)	46 (0.11)
Lung Opacity	4065 (9.92)	17 (0.04)	52 (0.13)
No Finding	29009 (71)	0	11958 (29)
Pleural Effusion	1822 (4.45)	0	31422 (76.7)
Pleural Other	117 (0.29)	0	1 (0.00)
Pneumonia	774 (1.89)	0	46 (0.11)
Pneumothorax	214 (0.52)	0	189 (0.46)
Support Devices	8791 (21.46)	0	21 (0.05)

is given in table 4.2.

When utilising this dataset to train classification models, it can be challenging because the number of images in BRAX that are classified as *no finding* is significantly higher than the number of samples in all other classes combined excluding *support devices*. The next class to have maximum samples is *lung opacity* followed by *cardiomegaly* containing 4065 and 3984 positive samples respectively. Figure 4.2 illustrates the presentation of a subset of diseases from the BRAX [2] dataset. Due to the similarity of symptoms, several lung diseases may be hard to distinguish from one another.

## 4.2 Segmentation Datasets

While an image-level classification is helpful, examining the specific lung regions can provide additional information from the X-ray. This necessitates that the lungs are segmented from the overall CXR and used for additional processing, such as severity quantification or the creation of pixel-level masks for the affected areas.

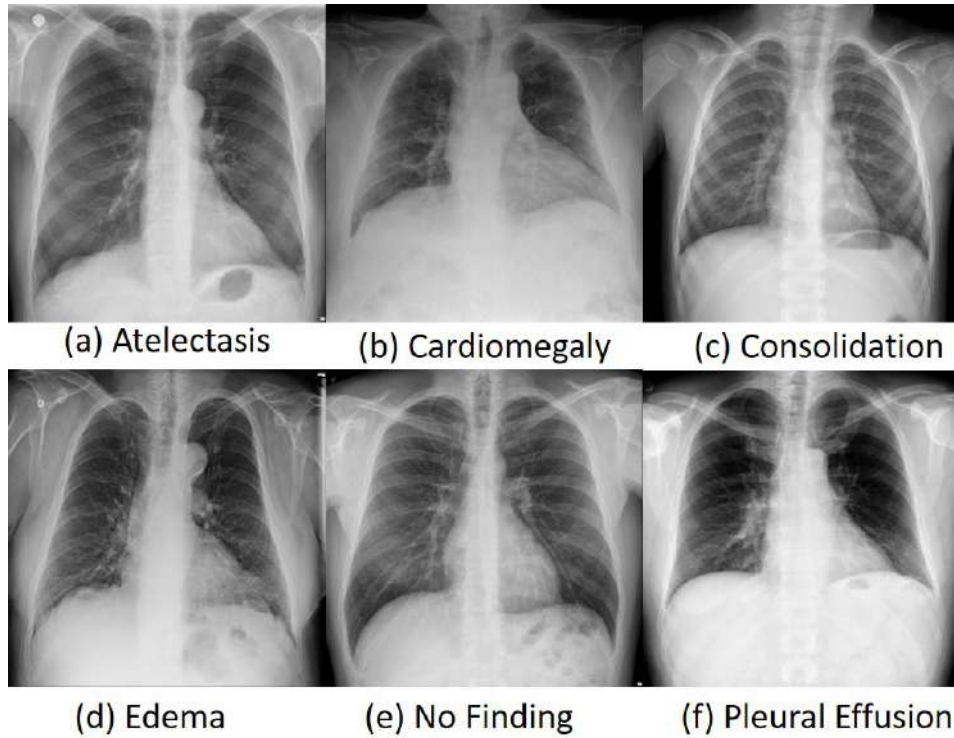


Figure 4.2: Example images from (a) to (e) for the various pulmonary diseases in the BRAX [2] data set

### 4.2.1 Montgomery County Dataset

138 frontal chest X-ray images from the Montgomery County Tuberculosis Screening Program are included in the Montgomery County dataset [4], which was compiled by the Department of Health and Human Services in collaboration with Montgomery County, Maryland in the United States. 80 of the images show normal functioning and 58 show symptoms of tuberculosis. These chest X-ray images range in resolution from 4020 x 4892 to 4892 x 4020 pixels providing excellent detail. Along with the images, the dataset also contains segmentation masks for the CXRs which have been annotated by a team of radiologists. Figure 4.3 shows an image and its corresponding segmentation mask from all three segmentation datasets used in this work.

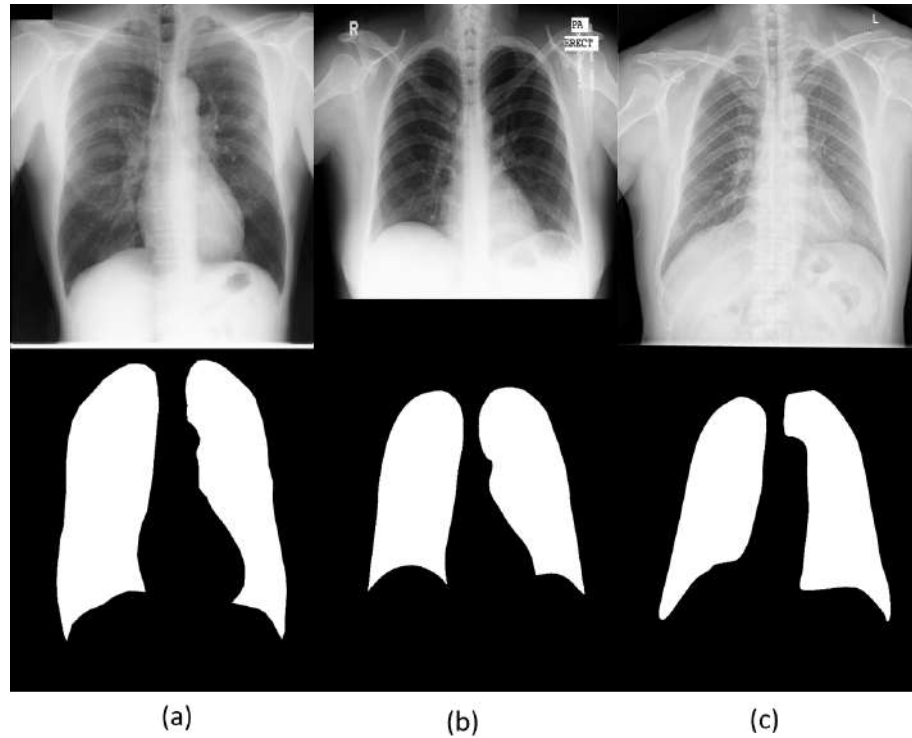


Figure 4.3: Example images from (a) to (c) showcasing the segmentation masks for JSRT, Montgomery, and Shenzhen datasets respectively [3–7] respectively

#### 4.2.2 Shenzhen Dataset

This dataset contains X-ray images that were gathered by the Shenzhen No. 3 Hospital in Shenzhen, Guangdong Province, China [5, 6] The Shenzhen Hospital acquired these X-rays as part of its routine care in JPEG format. The dataset was originally acquired for image-level classification as the dataset contains 326 normal and 336 abnormal x-rays where the abnormality is tuberculosis.

The segmentation masks for a major portion of this dataset were made available in another study [7] where the masks were prepared by a team at the Computer Engineering Department, Faculty of Informatics and Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine.



### 4.2.3 Japanese Society of Radiological Technology Dataset

For lung segmentation, the Japanese Society of Radiological Technology (JSRT) dataset is frequently utilised [3]. The JSRT collection offers masks for the heart and both clavicles that have been annotated by radiologists, even though it is best known for lung segmentation. Out of a total of 247 CXRs, 154 images also have image-level classification labels in JSRT. 100 of the 154 images with nodules are cancerous, while 54 are benign. The dataset also shows where each nodule is located.

All CXR scans within the JSRT dataset have a resolution of  $2048 \times 2048$  pixels with a 12-bit pixel depth. Table 4.3 summarizes segmentation datasets that are used for training the segmentation head in one of the frameworks in this research work.

Table 4.3: Number of samples for different segmentation datasets

Data set	Samples
JSRT [3]	247
Montgomery [4]	138
Shenzhen [5–7]	566

## 4.3 Opacity Localisation Datasets

Numerous lung conditions cause the lungs to become opaque, which impairs the lungs’ ability to perform the gaseous exchange. These appear as areas of greater density on the X-ray. The extent of the disease can also be determined by the opacification of the lung tissue, which is particularly helpful for monitoring the disease’s progression over time. Therefore, determining where the lung opacities are located is crucial.

### 4.3.1 SIIM-FISABIO-RSNA (SIIM)

SIIM-FISABIO-RSNA COVID-19 detection dataset was made available in the form of a public challenge at Kaggle [8]. The purpose of this dataset is the detection of COVID-19 and associated pneumonia types with subsequent localisation of lung opacity regions in

the CXR images. The training dataset has a total of 6336 images of varying resolution ranging from 846x1353 to 4891x4020. The competition organizers provided the labels against the training dataset with four distinct classes; *negative for pneumonia*, *typical*, *indeterminate* and *atypical* appearance of COVID-19 associated pneumonia. The number of samples for each class is 1737, 3007, 1108, and 484 respectively. Out of these, Only 4224 of the total number of images have opacity annotations which were used in one of our frameworks. Figure 4.4 highlights the opacity in different CXRs represented here by red rectangles.

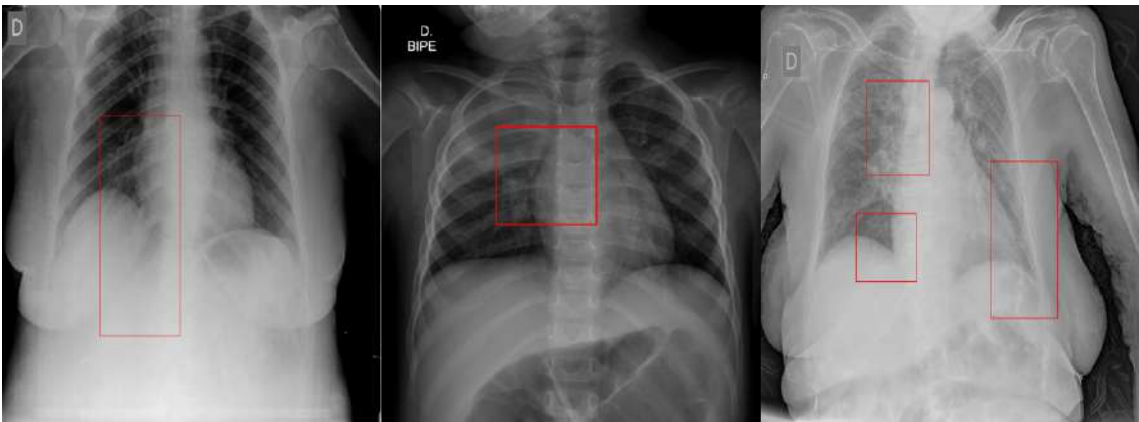


Figure 4.4: Presentation of lung opacity in SIIM [8] dataset

The test dataset is divided into two portions: the public test dataset, and the private dataset, which was not been made public. The public test dataset consists of 1214 images while the complete test dataset is around the same size as the training dataset.

## 4.4 Report Generation Datasets

Every chest X-ray examination is accompanied by a report that summarises the key findings of the CXR. These results offer a thorough analysis of the state of the systems in the chest cavity. One or more labels for the full CXR are also obtained using these reports [1, 2]. In recent years, datasets with a focus on report generation have also become available. These datasets serve as an important training resource for transformer-based models that generate reports from a single CXR. The following datasets are used by one

of the research work's frameworks.

#### 4.4.1 Indiana University Chest X-ray (IU)

Gathered from different hospitals affiliated with the Indiana University School of Medicine, this dataset is a collection of 7784 frontal and lateral chest radiographs [9] available in the DICOM format. Accompanying these CXRs are 3927 unique reports that contain the key findings of the scan under different sections. The overall number of words is 1,22,096 and there are 3257 unique terms in the lexicon with over 75% of the paragraphs being unique. In order to anonymise the dataset, the Private Health Information (PHI) has been replaced with xxxx keyword. Other identifiable parameters have been removed from the scans as well. Figure 4.5 shows the report and the accompanying chest X-ray of one such study from the dataset.

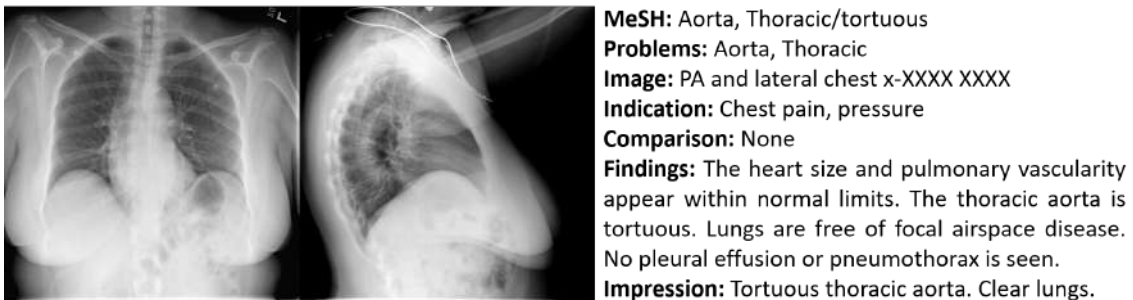


Figure 4.5: A CXR study from Indiana [9] containing both a frontal and lateral scan along with the radiological report. The report contains information about indication, availability of a prior CXR, findings, and impressions among others.

Along with the free-text reports, image-level labels for classes such as cardiomegaly, edema, pleural effusion, pneumothorax, etc. are also present with the data that can be used for training a classification model as well. However, IU has been used primarily for report-generation tasks.

## 4.4.2 Medical Information Mart for Intensive Care - Chest X-rays (MIMIC-CXR) and MIMIC Previous References Omitted (MIMIC-PRO)

MIMIC [10] dataset was collected by a team of researchers at Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA over a period of five years from 2011 to 2016. Currently, this dataset which includes 377,110 images acquired from 227,835 radiography studies of 65,379 patients—is the largest collection of chest X-rays that is publicly accessible. The CXRs in this dataset were predominantly obtained in frontal (PA and AP) and lateral projections utilising a variety of different equipment, both portable and fixed. The screening may include one or more images with one or more projections for a single patient. Although the free text reports in this dataset primarily serve the purpose of training the AI models for report generation, these associated reports are also used to generate the class labels assigning one or more labels based on the report. These class-based labels can be used for training classification models. The rule-based labeler used by the team is similar to the one used by [1] as the images in this dataset have also been divided into 14 classes. Figure 4.6 shows the class distribution for all the classes.

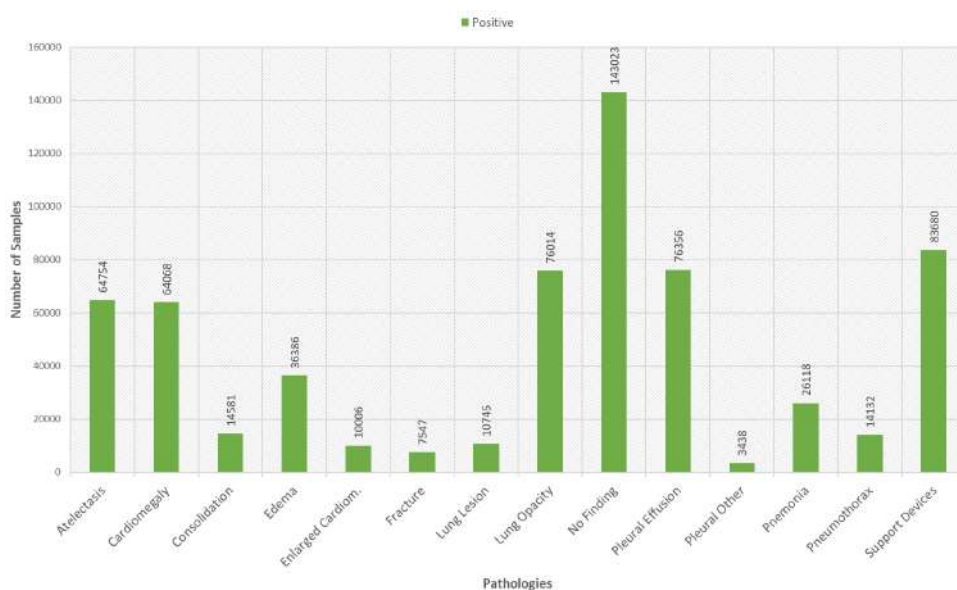


Figure 4.6: Samples in each class in MIMIC dataset [10]

*No finding* has the maximum number of samples around a hundred and forty thousand followed by *support devices*. *Pleural effusion* and *lung opacity* have the second and third most samples respectively when *support devices* are ignored.

Each of the 227,835 CXR study reports has been provided in the form of a text file with Private Health Information replaced with three underscores (" \_ \_ \_ "). There are 324,641 words total in the dataset, with 145 words and 642 characters on average in each report. Each provided report has three main sections: findings, impressions, and free-form text without any heading. The findings provide an elaborate description of the key findings in the CXR while the impressions are far more concise. A report may contain both the findings, impressions, and free-form text but one or more may also be absent from a report. The number of reports containing both the findings and the impressions is close to 117,000. Figure 4.7 shows the report and the accompanying chest X-ray of one such study from the dataset.

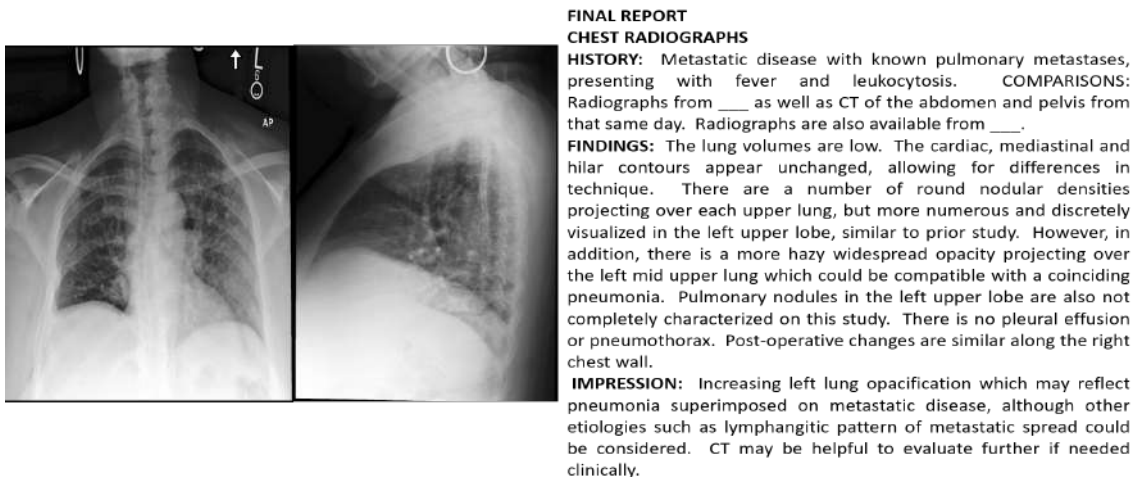


Figure 4.7: A CXR study from MIMIC [10] containing both a frontal and lateral scan along with the radiological report. The radiological report has been divided into three sections: history, findings, and impressions. [10]

The references to earlier CXR studies that may or may not be contained in the dataset are one issue that such a huge dataset poses. These prior references may limit the deep learning models' capacity to learn the representation, leading to reports that make references to erroneous data. Previous data might be ignored because the major goal of a

radiological report is to highlight the findings in the most recent scan. This issue has been resolved by [14] by automatically rewriting the reports and eliminating past references. This dataset which can be considered a modified version of [10] contains 371,951 reports for an equivalent number of images in the train portion and 2188 reports in the test set. Table 4.4 shows the difference between impressions from MIMIC and MIMIC-PRO for the same samples.

### 4.4.3 Local Dataset

A local dataset has also been acquired from the Health Ways Laboratories and Hospital located in Rawalpindi, Pakistan during the duration of January 2020 to November 2020. This dataset contains 1054 frontal chest X-rays which is the same number of patients. The resolution of the scans is 3072x3072 pixels and they have been captured using 'FXRD-1717NB' machine. Figure 4.8 shows sample images and reports from the dataset.

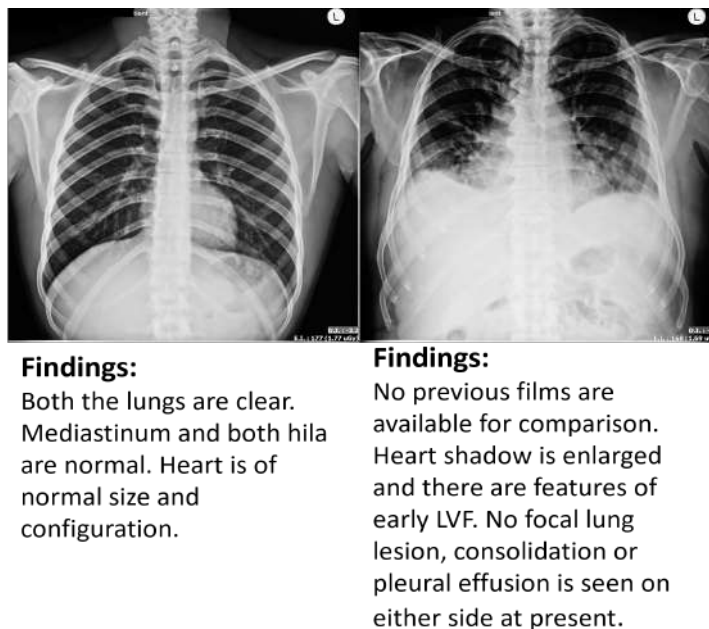


Figure 4.8: Two distinct CXR studies from the local dataset; the one of the left is a normal CXR whereas the one on the right is abnormal. The normal reports in this dataset are shorter in length. The radiological report does not contain any sections.

Table 4.4: Difference between original impressions from the radiological reports of the MIMIC [10] dataset and the rewritten reports from MIMIC-PRO [14]

MIMIC [10]	MIMIC PRO [14]
<p>Increasing left lung opacification which may reflect pneumonia superimposed on metastatic disease, although other etiologies such as lymphangitic pattern of metastatic spread could be considered. CT may be helpful to evaluate further if needed clinically.</p>	<p>Left lung opacification which may reflect pneumonia superimposed on metastatic disease, although other etiologies such as lymphangitic pattern of metastatic spread could be considered. Ct may be helpful to evaluate further if needed clinically.</p>
<p>New nodular opacities within both upper lobes, left greater than right. Findings are compatible with metastases, as was noted in the lung bases on the subsequent CT of the abdomen and pelvis performed later the same day.</p>	<p>Nodular opacities within both upper lobes, left greater than right. Findings metastases, the lung bases the abdomen and pelvis day.</p>
<p>Compared to chest radiographs since ---, most recently one ----. Previous mild pulmonary edema and possible concurrent pneumonia has all cleared. Heart is top-normal size, improved, and pleural effusions have resolved. Right hilar vessels are still enlarged, perhaps due to pulmonary arterial hypertension. Lateral view shows atherosclerotic coronary calcification in the left circumflex.</p>	<p>, one - - -. Mild pulmonary edema and possible concurrent pneumonia cleared. Heart is top - normal size, , and pleural effusions resolved. Right hilar vessels are enlarged, perhaps due to pulmonary arterial hypertension. Lateral view shows atherosclerotic coronary calcification in the left circumflex.</p>
<p>In comparison to previous radiograph of 1 day earlier, support and monitoring devices are unchanged in position. Pulmonary vascular congestion has improved. Airspace opacity at the left lung base has worsened, and additional patchy opacities have developed at the right lung base. Findings could potentially be due to aspiration or evolving aspiration pneumonia in the appropriate clinical setting. Exam is otherwise remarkable for probable small bilateral pleural effusions.</p>	<p>, support and monitoring devices position. Pulmonary vascular congestion improved. Airspace opacity at the left lung base additional patchy opacities at the right lung base. Findings could potentially be due to aspiration or evolving aspiration pneumonia in the appropriate clinical setting. Exam for probable small bilateral pleural effusions.</p>

The CXRs were present in the JPEG (.jpeg) format in addition to the DICOM format that the system used to capture the images produced. Instead of using the original DICOM files, these images in this format were used. Every patient in the dataset was given a unique 6-digit ID in the pattern I00001, with the final digit denoting the total number of X-rays performed for that patient. With the youngest patient being 3 months old and the oldest being 92 years old, the patients in the dataset have a median age of 38 years. Of the patients in this dataset, about 55% are men and the remaining patients are women.

The X-rays are accompanied by radiological reports written by the physicians from Health Ways Laboratories describing the findings in the scan as a Word (.doc) file. Every image in the dataset has been assigned a label from one of two classes: Normal or Abnormal, in addition to the CXR report that is included with the dataset. 330 of the CXRs are abnormal, and 774 of them are normal making this dataset skewed towards normal images. The radiology reports for the normal imaging are brief and only state some variations of the finding that the lungs are clear. Approximately 11 reports also contain prior references to earlier X-ray examinations which were not removed. This dataset has roughly 570 unique words, with an average report length of 27 words. The shortest report in the dataset is only 5 words long, while the longest report has a length of 98 words. Approximately 300 reports are longer than average, with the remaining reports being shorter. In contrast to the previously listed datasets [9,10,14], the reports in this dataset are comprised entirely of free text rather than being separated into distinct sections.

The dataset has been anonymised by removing any identifiable information from the reports and the CXRs. Since the patient's information was located in the upper left corner of each image, it was obscured in each image. Every report was also thoroughly reviewed to make sure no patient information remained.

This dataset enables testing the effectiveness of the suggested models on a local population, which can aid in enhancing the potential of the suggested framework.



## 4.5 Summary

The availability of different datasets that can be utilised for classification, segmentation, opacity localisation, and report generation has made it possible to train large, deep-learning models for decision support systems. The inclusion of small-scale local datasets can also be used to test the performance of the decision support systems. Table 4.5 presents a comprehensive overview of the aforementioned publicly available datasets and the local dataset that were employed in this research. The table details the specific purpose of each dataset and its other key characteristics.

Table 4.5: A summary of datasets that were used to train different sub-modules of the proposed framework

<b>Dataset</b>	<b>Institute</b>	<b>Images</b>	<b>Patients</b>	<b>Pathologies</b>	<b>Sub-module Utilisation</b>
CheXpert [1]	Stanford	224316	65240	14	Classification
BRAX [2]	HIEA	40967	19351	14	Classification, Opacity Localisation, Segmentation, Report Generation
Montgomery [4]	DHHS Shenzhen	246	–	2	Segmentation
Shenzhen [5–7]	No. 3 Hospital	566	–	2	
JSRT [3]	JSRT	247	–	2	
SIIM [8]	RSNA	4293	–	4	Opacity Localisation
Indiana University [9]	Indiana University	7784	–	2	Report Generation
MIMIC [10]	BIMDC	377110	65379	14	
MIMIC-PRO [14]	BIMDC	371,951	65379	14	
Local Dataset	Healthways	1054	1054	2	

## **Chapter 5**

# **Proposed Multi-Head Deep Learning Framework for Pulmonary Disease Detection and Severity Scoring with Modified Progressive Learning**

A comprehensive analysis of a chest X-ray is crucial to improve diagnostic outcomes, which in turn can guide patient care effectively. Three essential components — image-level classification, opacity localisation, and severity quantification — all obtained from a single chest X-ray can be combined to accomplish this. By integrating these three components, a more precise and accurate assessment of the chest X-ray can be obtained, enabling healthcare professionals to make informed decisions. Our proposed framework targets all three components to ensure a thorough and reliable analysis.

## 5.1 Multi-Head Deep Learning Framework

Our proposed solution is a multiple output framework that allows for multiple insights from a single CXR image by means of different heads where each head is responsible for a distinct task. The classification head through an ensemble of multiple convolutional neural network backbones outputs an image-level classification probability of different pulmonary pathologies. The lung segmentation mask obtained from the segmentation head is used in tandem with opacity localisation – obtained from the localisation head – to define the severity of a subset of pathologies that the framework has been trained for. The sections below describe the different heads in detail. An overview of the entire framework is provided in Figure 5.1.

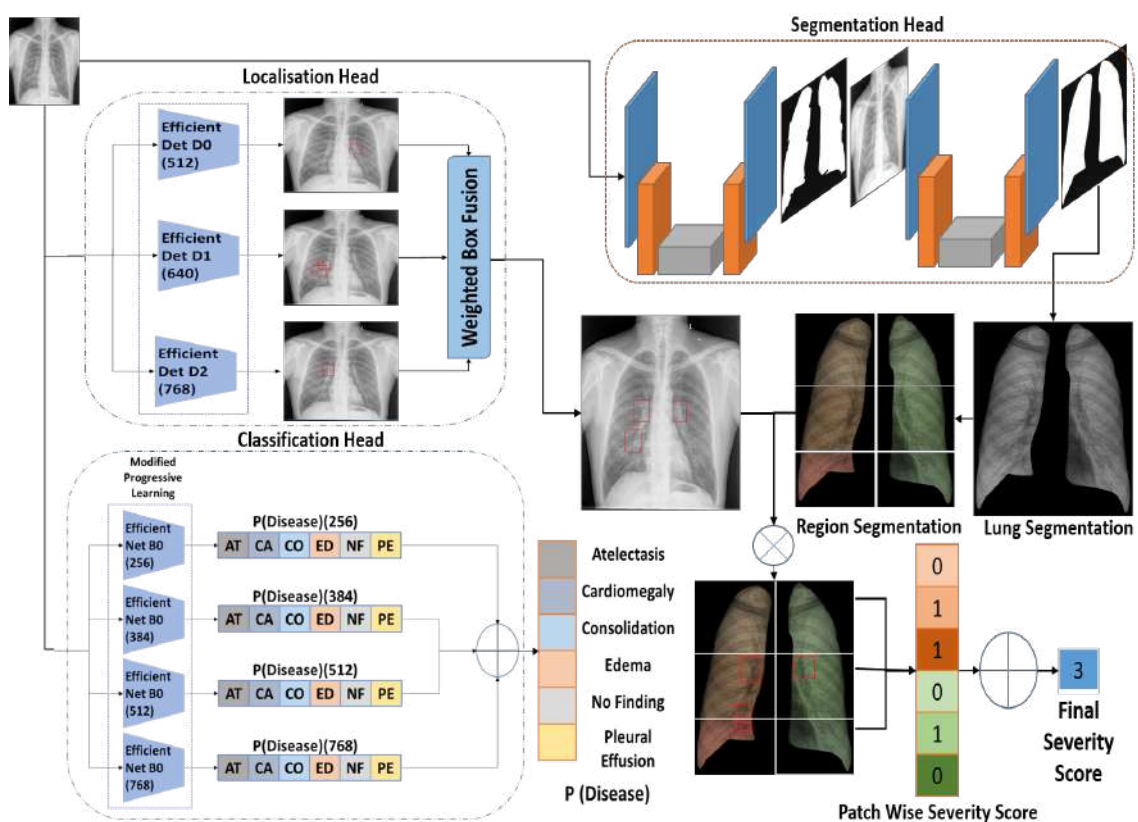


Figure 5.1: Proposed Framework Architecture. The classification head (Bottom Left) outputs the probability of a disease using an ensemble. Opacities are first localised using the localisation head (Top Left) and are then combined with the different lung regions obtained using the segmentation head to obtain the final severity score (Bottom Right).

### **5.1.1 Classification Head**

The proposed framework's classification head combines the EfficientNet architecture, which offers cutting-edge performance at lower computational costs, with a modified version of progressive learning, one of the fundamental building blocks of the EfficientNetV2.

#### **5.1.1.1 EfficientNet and EfficientNetV2**

Using neural architecture search and scaling [133] created the EfficientNetV2 family of CNN architectures that resulted in improved parameter efficiency and training speed. In addition, the researchers proposed a pyramid-style strategy to train models with progressively larger input sizes by adaptively modifying regularisation. This training strategy was termed progressive learning. They showed that increasing regularisation with each increment in the input size can solve the problem of performance degradation that occurred during the training of large models. The suggested model outperformed previous models and was faster and more effective. The family of EfficientNetv2 classifiers has been leveraged for classification of COVID-19 owing to their fewer parameters resulting in comparatively smaller size. For COVID-19 diagnosis using chest X-ray data, a classification network namely DFFCNet was proposed. For feature extraction, the model utilised EfficientNetV2 as the backbone network. In comparison to the other chosen backbones, the proposed framework performed better in experiments [94].

In a study by [146], COVID-19 was detected using pre-trained models like Xception, InceptionV3, and EfficientNetV2 from CXR and CT images. EfficientNetV2 with fine-tuning produced the best performance for the CXR data set, whereas the LightEfficientNetV2 model produced the highest performance for the CT data set. Both versions of EfficientNet have proven to be successful in various other domains. Several pre-trained models including EfficientNet B3, EfficientNetV2, HrNet, and ResNet50d were used for an automatic diagnosis of Myocarditis in Cardiac Magnetic Resonance (CMR) images.

The EfficientNetV2 had the best performance compared to the other pre-trained models [237].

To improve the performance of a CNN architecture, these architectures are typically scaled by adding more layers or changing the input image dimensions [11]. One of the drawbacks of this technique is that it is random in nature, requiring empirical experimentation to find a scaled version of the baseline architecture that performs better.

To tackle this issue, [11] came up with the compound scaling approach in which the depth, width, the input resolution of the image was scaled uniformly using a single compound coefficient. Figure 5.2 highlights the effects of the scaling along the depth, width, resolution, and the combination of all three.

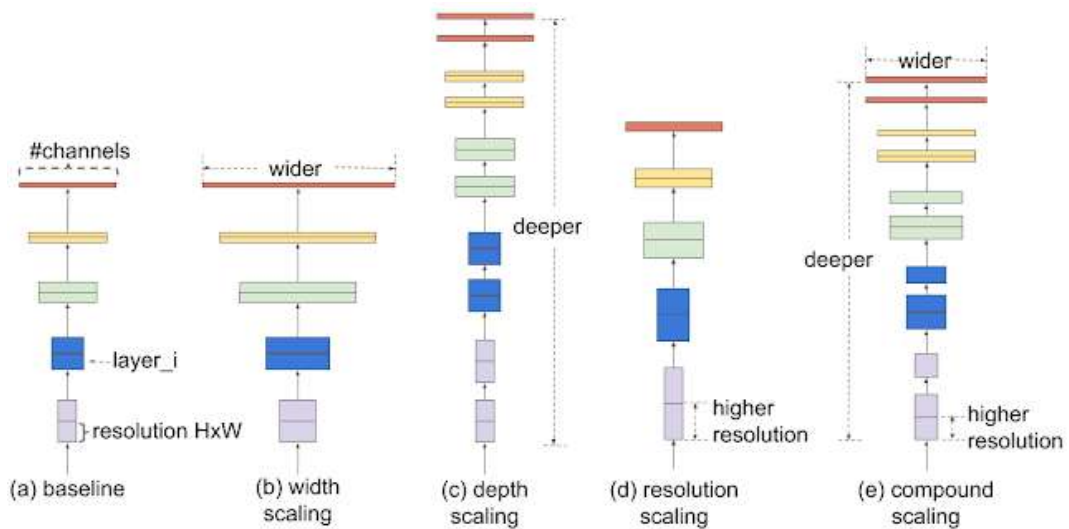


Figure 5.2: The effects of scaling along various dimensions such as width, depth, and resolution (b,c,d). These scaling factors are arbitrary, whereas compound scaling (e) scales the aforementioned factors in accordance with one another. Image taken from [11]

To arrive at this compound factor  $\phi$ , a scaling problem was formulated through which the value of depth  $d$ , width  $w$ , and resolution  $r$ . Specifically,  $d$ ,  $w$ , and  $r$  are set as follows:  $d = \alpha^\phi$ ,  $w = \beta^\phi$ , and  $r = \gamma^\phi$  subject to the constraint that they satisfy equation 5.1.  $\alpha$ ,  $\beta$ , and  $\gamma$  must always be equal or greater than 1.  $\phi$  represents the constraint such as target memory or target Floating Points Operation Per Second (FLOPS) according to which the architecture had to be scaled and the values of  $\alpha$ ,  $\beta$ , and  $\gamma$  determine how these

resources should be assigned to each individual scaling which, in essence, a grid search may determine. It is also evident from equation 5.1 that for any value of  $\phi$  other than 1, the total increase in the resources such as FLOPs is equivalent to  $2^\phi$ .

$$\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \quad (5.1)$$

EfficientNet B0 was introduced as a baseline architecture using Neural Architecture Search [11] in order to take advantage of the compound scaling mechanism. One of the convolutional blocks that lead to improvement in efficiency was the mobile inverted convolutional block (MBConv). By combining three steps—an expansion layer, a depthwise convolution, and a pointwise convolution layer—the MBConv operation is able to attain its efficiency. A  $1 \times 1$  convolutional operation is used to add more channels which are then passed through a depthwise convolution block. In order to decrease the number of input channels after this operation, a pointwise convolution with a  $1 \times 1$  filter size is applied last. [238]. Figure 5.3 illustrates the EfficientNet B0’s architecture, which mainly employs MBConv.



Figure 5.3: Mobile inverted convolutional layers form the bulk of the EfficientNet B0. Image taken from [11]

The values of  $\alpha$ ,  $\beta$ , and  $\gamma$  for EfficientNet B0 can be calculated using equation 5.1, and it turns out that for  $\phi = 1$ , these values are 1.2, 1.1, and 1.15, respectively, suggesting that depth, width, and resolution are scaled by that factor. The value of  $\phi$  can then be changed to generate scaled-up versions of the underlying design while maintaining these variables constant.

### 5.1.1.2 Modified Progressive Learning

Image level classification tends to be an important aspect of any framework as it provides an overall diagnosis of the image. In order to incorporate the progressive learning [133] strategy in the proposed framework, the classification head consists of 4 backbones that form an ensemble based on EfficientNet B0 [11] which offer similar performance to the state-of-the-art architectures at a reduced computational cost.

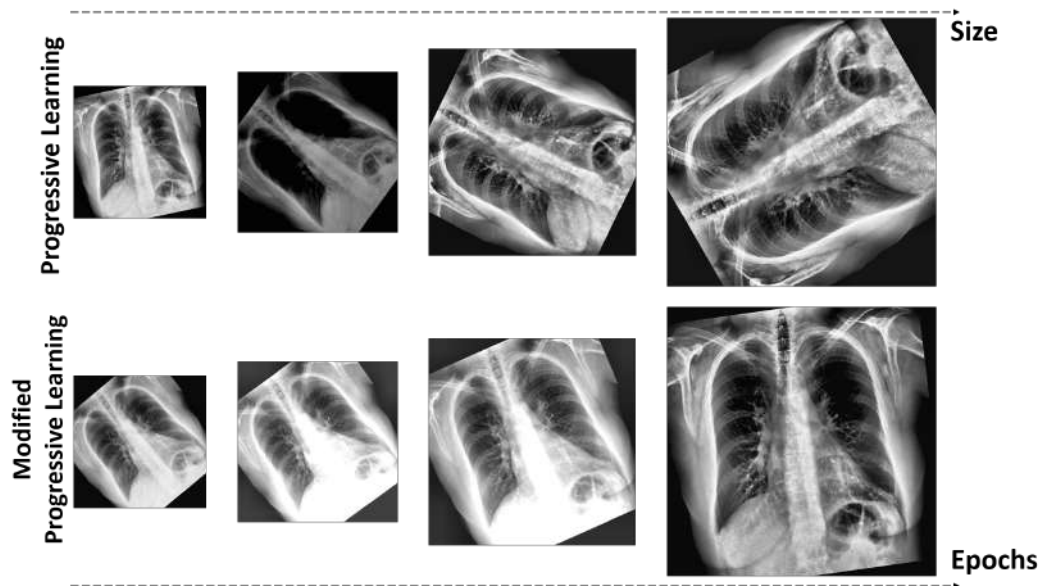


Figure 5.4: Progressive learning vs amended progressive learning. The difference in the methodologies lies in the random augmentation factor which is kept constant for all sizes. Several augmentations can be applied to the same image. From left to right, the images show an increasing random augmentation factor with increasing input image size.

The image input sizes range from 256 up to 768 to adhere to progressive learning. This training strategy proposes using the initial weights from the preceding, smaller network for the subsequent, larger network rather than random initialization. To counter the effects of using the larger network, increasingly aggressive regularization is applied at each stage as a means of making the examples harder for the network. However, in contrast to the original idea of increasingly aggressive regularization, in our training strategy, the augmentations for each progressive stage have been capped at the same level as that of the first stage. For instance, if the *rotation* augmentation is applied, the angle for this augmentation will not change at each stage but rather remain constant at every input size.

Similarly, for *blurring*, the exact same blurriness factor for each input size will be used instead of it being varied. Figure 5.14 highlights the difference between the progressive learning strategy proposed by [133] and the changes made to it.

### 5.1.1.3 Training the Classification Head

A subset of classes from the commonly used fourteen classes namely Atelectasis, Cardiomegaly, Consolidation, Edema, No Finding, and Pleural Effusion were selected from BRAX [2]. Using the EfficientNet B0 as the feature extractor, Global Average Pooling was applied to the latent vector space. Six output nodes with Softmax activation made up the model’s final output layer. The weight migration from one trained model to the next is fairly simple because the same backbone architecture is employed at all input sizes. Once all the models have been trained, they are combined in an ensemble with simple averaging to get the final disease probability  $P(Disease)$ . For each backbone, the images were re-shaped to the appropriate size but were not normalised. Figure 5.5 shows the architecture of the classification head of the framework in detail.

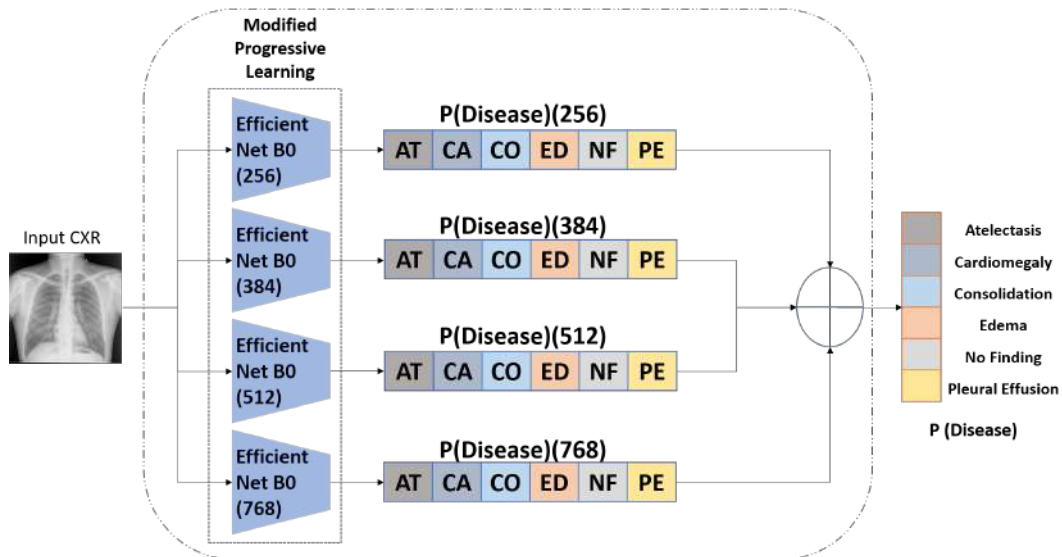


Figure 5.5: Classification head used in the sub-framework. The probability vector from each of the backbone is averaged to obtain the final  $P(Disease)$

The performance of the model was assessed using a different validation data set’s Area Under the ROC curve, and the model with the best performance was then saved.



## 5.1.2 Segmentation Head

Image segmentation is necessary for focusing on the region of interest for different problems. Fully convolutional networks - networks that do not have a fully connected layer at the end - are useful for generating masks of the desired region in an image.

### 5.1.2.1 U-Net, Related Architectural Designs and Little W-Net

The majority of segmentation methodologies use an architecture akin to a U-Net [162] which is a fully convolutional architecture with an encoder-decoder structure, where the encoder is essentially a CNN without the output layer and the decoder is the mirror of the encoder. The U-Net architecture's encoder consists of two main operations: convolution to generate feature maps and max pooling to reduce the spatial resolution of the feature maps. The number of feature maps grows with each successive convolutional layer. The kernel size is kept at 3x3 and is activated by a rectified linear unit (ReLU). The U-Net architecture allows for the customization of the number of layers, how many times they are repeated, and the kernel size at each step.

As the size of the input image has been reduced drastically after consecutive max pooling operations, therefore, in the decoder section of the architecture, this needs to be remedied so the model outputs a segmentation mask that has the same dimensions as the input image. This is achieved through 2x2 transposed convolution operation essentially doubling the spatial dimensions. In order to retain the information from the encoder section, the up-scaled feature maps are concatenated with corresponding feature maps from the decoder section. On the concatenated feature maps, a convolution operation with a 3x3 kernel size is applied with ReLU activation. The final layer of the network is again a convolution layer with 1x1 kernel size but with softmax activation resulting in a pixel-wise probability map for the input image. Applying a threshold on this probability map converts it to a segmentation mask containing two or more classes. Figure 5.6 shows one possible version of the U-Net architecture.

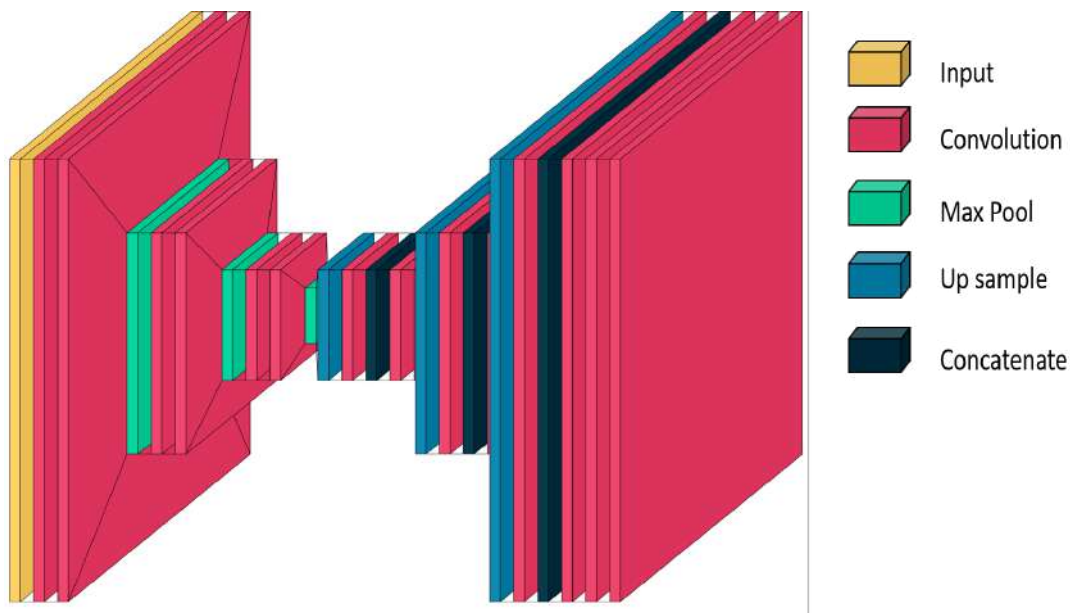


Figure 5.6: The number of layers in the encoder and decoder in the U-Net architecture is set to three. Another architectural change shown here is the replacement of transposed convolution with up sampling followed by convolution. Such modifications are possible due to the simple architecture. Because U-Net is fully convolutional, the final layer is also convolutional. Activation after convolutional layers are not depicted here.

Various enhancements have been proposed to improve the performance of the U-Net architecture in order to enhance the segmentation output [159, 166]. [166] employed a stacked U-Net architecture with the addition of auxiliary loss to each U-Net block. With the addition of auxiliary loss to different blocks, the model can learn discriminative features at each step, resulting in improved performance. [166] demonstrated the architecture's efficacy by generating segmentation masks for various medical images such as nodule segmentation, nuclei segmentation, and polyp segmentation. Similarly, segmentation of arteries and veins has been addressed using the Little W-Net architecture. Only two traditional U-Nets which are concatenated together make up this model. This method allows the model to maintain a manageable set of parameters, earning it the moniker "Little." [159]. Based on the notion in [239], the aggregate loss of each U-Net in training. The proposed model, which requires 1-3 orders of magnitude less computational resources than earlier CNNs, was tested on 10 distinct data sets and demonstrated to be superior to earlier approaches [159]. The researchers were able to show that the network's ability to

produce an attention map, which can help to improve the final segmentation mask, is the main advantage of such a design. For many medical applications, this type of network architecture has proven to have good performance. Binary vessel segmentation is challenging due to the atrophic changes and the restricted vascular architecture [161]. AutoMorph was suggested for the fundus images as a method of automating retinal morphology analysis. In order to minimise the large segmentation errors, the authors employed the Little W-Net architecture [161].

Little W-Net has also been used to address other problems as well. The reliability and accuracy of the size are crucial for polyp size measurements in colonoscopy images. The segmentation was performed with Little W-Net architecture due to its reduced number of parameters which was accompanied by increased feature representation [160].

### 5.1.2.2 Training the Segmentation Head

Isolating the lungs in a CXR is necessary whenever there is a need to provide a conclusion that relies on the specific sub-regions of the lungs. Therefore, we have utilised the aforementioned architecture in this proposed framework. As mentioned earlier, in the W-Net architecture [159], two U-Nets are strung together to form the  $W$  shape in which the output of the first U-Net is concatenated with the original image before passing it to the second one. The U-Net, generally, can be parameterised by two values: depth  $d$  and number of filters  $f$  in the first convolutional layer and can be represented as  $\phi_{f,d}$  and the output from such an architecture is denoted as  $y = \phi_{f,d}(x)$ . Here  $\phi_{f,d}^n$  denotes the architecture of a U-Net where  $n = [0, 1]$ ,  $y$  denotes the output or the *feature map* from such an architecture and  $Loss(\phi_{f,d}^n)$  denotes the Categorical Crossentropy loss for a particular model  $n$ . For every succeeding convolutional layer, the number of filters in that layer is doubled until the defined depth is reached. The final output from a WNet can be represented as equation 5.2.

$$y = \phi_{f,d}^2(x, \phi_{f,d}^1(x)) \quad (5.2)$$

The number of parameters in this type of architecture is a function of the depth and filters in the first convolutional layers. Increasing both these parameters results in a larger architecture with a greater number of trainable parameters. Figure 5.7 shows the detailed architecture of the segmentation head being used.

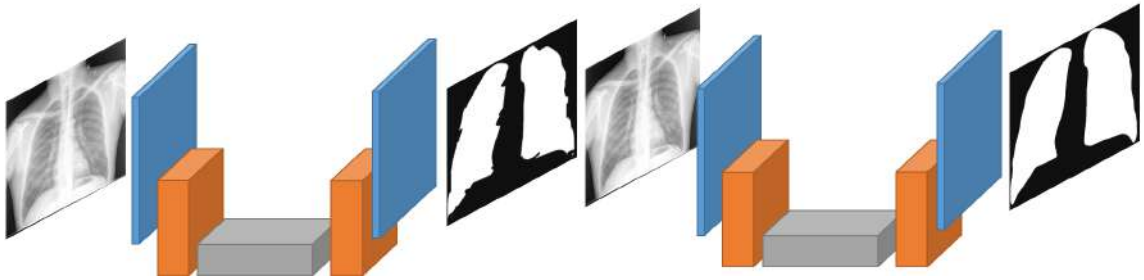


Figure 5.7: Segmentation head in the framework. The output of the first U-Net is used as an attention map and concatenated with the input image for the input of the second U-Net. The concatenation of the attention map improves the performance of the model while requiring fewer parameters across both U-Nets.

During training, the loss minimization is performed on the linear summation of the losses of both the U-Nets ( $Loss = \frac{1}{2}(Loss(\phi^1)) + \frac{1}{2}(Loss(\phi^2))$ ) which is then used for back-propagation via the Adam optimizer. The learning rate strategy is kept the same as that for the classification training employing a cyclic scheduler.

The data set used for training is a combination of JSRT and Montgomery data sets [3, 4]. This ensures that the model is machine agnostic i.e. the model performs equally well for any CXR irrespective of the capturing machine. Images are resized to  $512 \times 512$  resolution and as chest X-rays are usually handled as a gray-scale image, therefore, a single channel is kept at the input resulting in a 2-channel image for the second U-Net where the attention map output from the initial U-Net constitutes the  $2^{nd}$  channel. In order to gauge the performance of the model during the training, the F1 score is used that is computed for the validation split and the best-performing model is kept.

### 5.1.3 Localisation Head

The localization head of the suggested framework uses the EfficientDet architecture, which combines EfficientNet’s scaling efficiency with effective bi-directional feature fusion.

#### 5.1.3.1 EfficientDet

While classification just needs a single class label for the entire image, localization needs the positional coordinates in addition to the class label. EfficientDet is one such one-stage model that has been built for efficiency without sacrificing performance.

The scale of the objects in an image that need to be localised can vary, making it necessary to simultaneously detect them at various scales. Feature fusion technique allows for the fusion of features that are obtained from different points from the network backbone and have been employed by [167, 240, 241]. The architecture in [12] proposed a new feature fusion module named Bidirectional Feature Pyramid Network (BiFPN) that performs the job of fusing features obtained from the backbone at different resolutions. Figure 5.8 shows the structure of the BiFPN module.

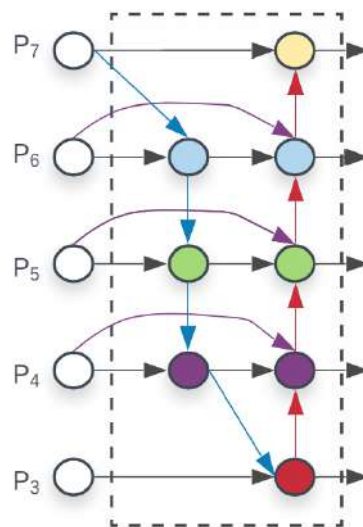


Figure 5.8: BiFPN block structure. The flow of the features is in both directions from top to bottom and from bottom to top. Image taken from [12].

The architecture of the BiFPN block offers a more efficient design as compared to [240] while also having an advantage over [167] which only uses a top-to-bottom approach. While it may be slightly less efficient than [241], the regularity of the structure of BiFPN block makes up for it. In contrast to FPN where  $P_6^{Output} = Conv(P_6^{input} + Resize(P_7^{Output}))$ , in BiFPN,  $P_6^{Output}$  is the normalised, weighted summation of three values  $P_6^{input}$ ,  $P_5^{output}$  and  $P_6^{topdown}$  as shown in equation 5.3, where  $P_6^{topdown}$  is given in 5.4.

$$P_6^{Output} = Conv\left(\frac{w'_1 \cdot P_6^{input} + w'_2 \cdot P_6^{topdown} + w'_3 \cdot Resize(P_5^{Output})}{w'_1 + w'_2 + w'_3 + \epsilon}\right) \quad (5.3)$$

$$P_6^{topdown} = Conv\left(\frac{w_1 \cdot P_6^{input} + w_2 \cdot Resize(P_7^{Output})}{w_1 + w_2 + \epsilon}\right) \quad (5.4)$$

The other building block of [12] is the compound scaling that was inspired by [11]. Using a single factor, the network is scaled in depth, width, and input resolution at the same time. In essence, [11] forms the feature extraction backbone of the EfficientDet. Figure 5.9 shows the complete architecture of EfficientDet. The combination of these approaches allows this architecture to provide excellent performance while utilising fewer parameters and less processing power.

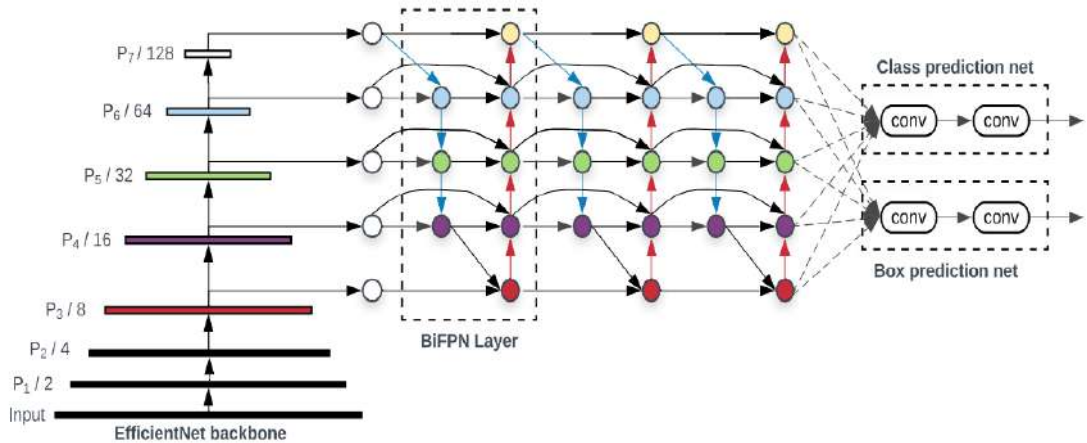


Figure 5.9: EfficientDet architecture with BiFPN blocks that can be repeated several times with EfficientNet [11] backbone for feature extraction. The networks at the end provide the positional coordinates along with the class label and confidence. Image taken from [12].

### **5.1.3.2 Training the Localisation Head**

After the lungs have been segmented, opacities in the lung regions can then be localised. In the proposed framework, just like the classification head, an ensemble of multiple localisers is used that have been trained at different input resolutions of 512, 640 and 768 pixels using the EfficientDet with D0, D1 and D2 [12]. The use of different backbone sizes and the variation in the input resolution allows the ensemble to be able to detect opacities at different scales.

In order to improve the performance of the ensemble, two techniques were used: Test Time Augmentation (TTA) and Weighted Box Fusion (WBF) [187]. TTA entails applying a number of data augmentation techniques to the test data that were applied to the training data. The final prediction is then created by combining the predictions from each augmented version. The bounding boxes predicted by each model are combined after being weighted according to their confidence scores in Weighted Box Fusion. WBF can efficiently decrease the number of false positives and false negatives by giving larger weights to forecasts that are more confident.

In order to gauge the performance of the model during the training, the mean average precision score is used that is computed for the validation split and the best-performing model was kept just as for other heads.

### **5.1.4 Severity Quantification through A Combination of Segmentation and Localisation**

Different techniques have been developed to evaluate the development and severity of various respiratory diseases. These techniques, which take into account the opacification of various lung regions, are quite useful in a variety of ways.

#### **5.1.4.1 Respiratory Infections' Severity Assessment**

Severity assessment of different respiratory pathologies can help medical care providers come up with a better treatment plan. In light of this, the severity assessment of COVID-19 CXR images using CNNs into different groups has been proposed by several researchers. [100] utilised nine publicly available CXR data sets with 3260 images in total. The disease severity score was based on an opacity score by two radiologists and based on that score the images were divided into the following groups: mild, moderate, severe, and critical.

Signoroni et al. [95] devised the Brixia score where each lung was divided into three equal parts for a total of six. Each of these six lung regions was graded on a scale of 0 to 3 where the level of lung compromise in COVID-19 cases was reflected by this score for each location, with 0 indicating no abnormalities and 3 indicating significant aberrations. Additionally, they proposed BS-Net, which utilised a pyramid technique to combine features gathered at various scales after automatically aligning the segmentation masks with a multi-feature region aligner using a multi-feature area aligner, and used a series of convolutional blocks to translate the input feature to the final Brixia Score.

In the same vein, [96] proposed a vision transformer-based framework for COVID-19 diagnosis and severity quantification in a multi-task learning approach by employing a large CXR data set for training the backbone model. The use of a large CXR data set allowed the model to learn low-level generalised features. The deep features from the backbone model were then used in conjunction with the vision transformer for the prediction of disease class and severity map for severity quantification. The researchers used a similar scoring technique as [95], however, they modified the scoring system such that each of the six lung regions could only have a maximum score of 1 with a total score of 6 for the entire CXR. While the transformer output a pixel-level severity map for each of the six lung regions, it was then converted to a 0 or 1 via max pooling. The use of a vision transformer achieved comparable performance to the state-of-the-art with a unified framework.



### 5.1.4.2 Mechanism for Quantifying Severity

The combination of opacity localisation along with lung segmentation provides a robust method for generating a severity score for a diseased CXR. Each lung segment is given a score of 0 or 1, indicating whether or not opacity is present in that region. It is vital to have a clear criterion when deciding whether a specific region should be labelled as containing opacity. Intersection over Union (IoU) is a metric that measures how much two regions overlap. By measuring the region's IoU with the opacity identified by the localisation head, the threshold for IoU can be determined. The threshold value of 30% produced the best results after a grid search was employed to find the best IoU between the lung region and the opacity.

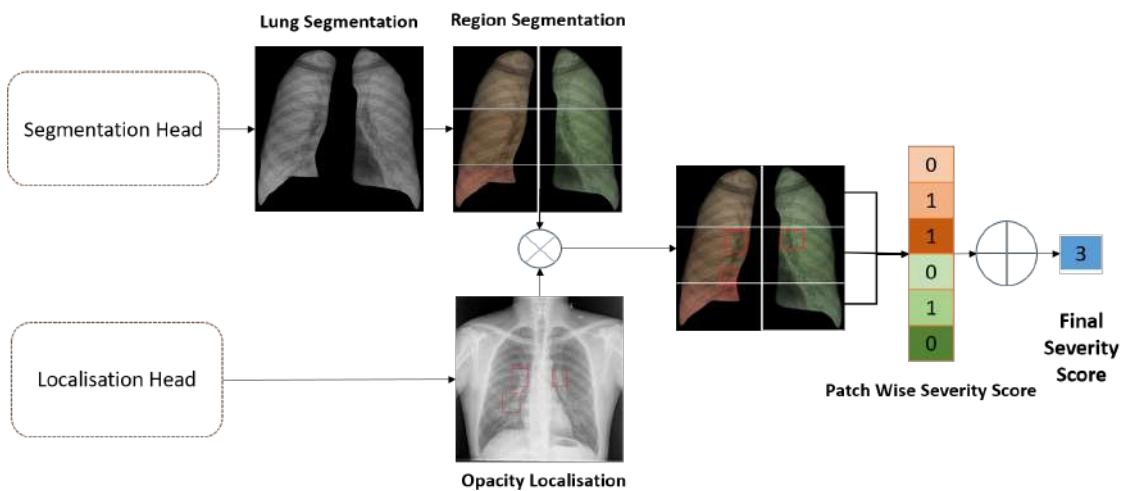


Figure 5.10: After being divided into six lung regions, the output from the segmentation head is multiplied with the output from the opacity localization head to provide the severity score for each region, which is then added together to produce the final severity score.

Because the severity score assigned to each lung segment can only be binary, even if there are multiple regions of opacity in the same lung segment and all regions meet the IoU criterion, the severity score assigned to that section remains 1. The total score for the CXR is determined by adding the results from each of the six regions. Consequently, the severity score might range from 0 to 6, with 0 denoting no opacity and 6 denoting opacity in all lung areas. Figure 5.10 shows the details of how the severity score is calculated using segmentation and localisation output.

Table 5.1: Hyperparameters of different output heads in the proposed framework

Output Head	Augmentations	Learning Rate	Learning Scheduler	Loss Function	Activation Function	Input Size	Input Dimensions	Optimizer	Batch Size	Test Time Augmentation	
Classification	Rotation, Flipping, Zoom and shearing, Random brightness, Random sharpness, Random blurring, CLAHE	$10^{-3}$ to $10^{-6}$	Decaying Cosine Annealing Scheduler	Categorical Crossentropy	Softmax	256 to 768	3	Adam	4	No	
	Horizontal Flip, Cut Out					512	1		2	No	
	Horizontal Flip, Rotation	$10^{-3}$	Constant	Focal Loss	Swish	512 to 768	3	AdamW	16	Yes	
Segmentation											
Localisation											

### 5.1.5 Training Parameters

The training parameters for various heads vary since the proposed framework comprises a variety of distinct heads. However, where possible, the hyperparameters have been kept the same across different output heads and training strategies. Table 5.1 summarizes the scores for each fold for all the models and the ensembles.

## 5.2 Results

We evaluated the performance of our proposed framework primarily on BRAX [2]. As [2] lacks a separate validation data set provided by the authors, a 5-fold cross-validation strategy was used for gauging the classification performance. Within this data set, the No Finding class significantly outweighs all other classes in terms of image count. To ensure class equilibrium while utilizing the BRAX [2] dataset, downsampling was mainly implemented for this class during both the training and the validation. This involved randomly selecting a subset of images thus limiting the representation of the aforementioned class. Two kinds of validation sets were used to assess the trained models: one contained all of the No Finding class' samples, while the other had a limited number of samples. The performance of segmentation was evaluated using two sets: the validation data set of the three combined data sets that were used to train the segmentation module and the manually annotated BRAX validation set.

As our severity scoring for different pathologies is based on the presence and prevalence of opacities in different lung regions, therefore, the results of opacity localisation and severity have been presented separately. For opacity localisation, SIIM [8] validation data set is used. For severity scoring, we present our results on the BRAX validation data set.

Every model in the network was trained at least three times, and the best-performing model was retained. The models were trained using TensorFlow 2.8 in Python on a system with 64 GB RAM and two Nvidia RTX 2070 GPUs. In order to train some models on

higher image resolution, we also made use of Google Cloud using Google TPUs (v2.8). The performance metrics that have been calculated for this framework include the AUROC score for classification, F1 score for segmentation, and mAP for opacity localisation, and a novel matching score for severity classification.

### **5.2.1 Classification**

Although EfficientNet B0 [11] serves as the foundation of the proposed framework, we also assessed other networks (B1 to B7) in the same network family. Since the EfficientNet B0 model performed the best among the group, the results presented here are based on this backbone. The performance of the models for the validation data sets is measured through AUROC. For the validation data set where the number of samples of No Findings class is limited, using an ensemble of the 4 trained networks, the average AUROC achieved ranges from 0.861 to 0.886, which is higher than any single model at any size as the single best performing models only achieved an AUROC of 0.875 at two different sizes of 384 and 768 pixels. The worst-performing single model achieved an AUROC of 0.826 with an input size of 256 pixels. Furthermore, it can be observed from Table 5.2 that, on average, the best-performing class is Pleural Effusion while the worst-performing class is Atelectasis. Table 5.2 summarizes the scores for each fold for all the models and the ensembles. The last column shows the average AUROC score.

Similar to Table 5.2, Table 5.3 shows the performance of the models on the validation data set where No Findings samples' are not limited to maintain equilibrium for the samples of all classes. Compared to the previous validation set, there is a slight improvement in the average AUROC, which varies from 0.863 to 0.892. However, there is a single model which outperforms the best ensemble in one of the validation folds. The worst-performing single model achieved an AUROC of 0.837 at an input size of 768 but it still managed to outperform the other validation set. The addition of samples to the validation set also results in another change where the worst-performing class becomes Consolidation while

the best-performing class becomes Edema as evident from Table 5.3.

Table 5.2: AUROC scores across 5 fold cross validation using amended progressive learning on BRAX [2] data set where the validation set contains a limited number of No Finding class

Fold, Model	AUROC Score						Avg. Score
	AT	CA	CO	ED	NF	PE	
1, 256	0.795	0.889	0.777	0.829	0.867	0.906	0.844
1, 384	0.784	0.862	0.78	0.822	0.859	0.904	0.835
1, 512	0.763	0.891	0.79	0.817	0.852	0.919	0.839
1, 768	0.807	0.877	0.781	0.793	0.871	0.915	0.841
1, Ens	<b>0.824</b>	0.898	0.793	0.834	0.883	<b>0.937</b>	0.861
2, 256	0.79	0.914	0.803	0.94	0.872	0.896	0.869
2, 384	0.786	0.907	0.782	0.913	0.841	0.901	0.855
2, 512	0.779	0.896	0.797	0.941	0.879	0.913	0.868
2, 768	0.802	0.882	0.821	0.91	0.881	0.906	0.867
2, Ens	0.819	<b>0.924</b>	0.823	0.923	0.889	0.932	0.885
3, 256	0.763	0.851	0.806	0.833	0.837	0.869	0.826
3, 384	0.787	0.906	0.814	0.909	0.834	0.896	0.858
3, 512	0.784	0.893	0.815	0.928	0.852	0.907	0.863
3, 768	0.81	0.878	0.827	0.934	0.841	0.896	0.864
3, Ens	0.818	0.903	0.842	0.947	0.874	0.918	0.884
4, 256	0.788	0.888	0.827	0.857	0.867	0.908	0.856
4, 384	0.806	0.908	0.814	0.923	0.874	0.926	0.875
4, 512	0.789	0.901	0.844	0.917	0.868	0.897	0.869
4, 768	0.802	0.909	0.835	0.926	0.874	0.903	0.875
4, Ens	0.821	0.923	<b>0.861</b>	0.891	<b>0.89</b>	0.93	<b>0.886</b>
5, 256	0.755	0.892	0.796	0.942	0.863	0.896	0.857
5, 384	0.798	0.916	0.811	<b>0.958</b>	0.873	0.898	0.875
5, 512	0.789	0.903	0.819	0.901	0.882	0.896	0.865
5, 768	0.722	0.894	0.794	0.821	0.866	0.899	0.855
5, Ens	0.802	0.92	0.833	0.941	<b>0.891</b>	0.922	0.885

In order to visualise the performance of the models for classification, ROC curves for all ensembles across folds are shown in Figure 5.11. Atelectasis performs poorly compared to other classes, according to the ROC curves, but it still outperforms a random classifier emphasizing that the model has some predictive power for this class. However, because there are only a few samples for each fold, the ROC curve for edema appears less uniform. Similarly, in Figure 5.12, the box plots for all six classes are shown for different input sizes and the ensemble as well. It can be seen from the Figure 5.12 that image size has

a significant impact on the AUROC. Generally, larger image sizes correspond to higher AUROC. Furthermore, the ensemble method outperforms any of the individual models.

Table 5.3: AUROC scores across 5 fold cross validation using amended progressive learning on BRAX [2] data set where there is no cap on the number of images of No Finding class

Fold, Model	AUROC Score						Avg. Score
	AT	CA	CO	ED	NF	PE	
1, 256	0.829	0.900	0.784	0.902	0.853	0.913	0.863
1, 384	0.792	0.868	0.743	0.902	0.842	0.900	0.841
1, 512	0.738	0.898	0.737	0.897	0.853	0.902	0.837
1, 768	0.830	0.882	0.760	0.882	0.849	0.910	0.852
1, Ens	0.825	0.900	0.747	0.905	0.870	0.929	0.863
2, 256	0.820	0.913	0.774	0.964	0.856	0.902	0.872
2, 384	0.825	0.895	0.725	0.940	0.824	0.912	0.853
2, 512	0.772	0.879	0.761	0.965	0.848	0.922	0.858
2, 768	0.796	0.867	0.802	0.961	0.836	0.900	0.860
2, Ens	0.821	0.905	0.766	0.958	0.868	0.932	0.875
3, 256	0.763	0.856	0.748	0.935	0.847	0.883	0.839
3, 384	0.810	0.913	0.836	0.969	0.830	0.916	0.879
3, 512	0.803	0.898	0.756	0.979	0.855	0.912	0.867
3, 768	0.831	0.857	0.771	0.979	0.816	0.913	0.861
3, Ens	0.817	0.897	0.780	0.984	0.863	0.926	0.878
4, 256	0.810	0.890	0.807	0.931	0.853	0.916	0.868
4, 384	<b>0.860</b>	<b>0.916</b>	<b>0.854</b>	0.971	0.859	<b>0.940</b>	<b>0.900</b>
4, 512	0.804	0.895	0.834	0.957	0.854	0.905	0.875
4, 768	0.840	0.907	0.845	0.967	0.861	0.926	0.891
4, Ens	0.840	0.914	0.842	0.946	<b>0.876</b>	0.936	0.892
5, 256	0.771	0.891	0.746	0.959	0.838	0.887	0.849
5, 384	0.830	0.906	0.779	<b>0.985</b>	0.838	0.884	0.870
5, 512	0.815	0.892	0.824	0.954	0.852	0.878	0.869
5, 768	0.731	0.887	0.773	0.894	0.837	0.900	0.837
5, Ens	0.814	0.906	0.787	0.968	0.866	0.897	0.873

The strength of the proposed framework lies in modified progressive learning. Using the weights from the preceding model trained at a smaller input size, the next model is trained from these initial weights instead of random initialisations. It can be observed from figure 5.13 that the average AUROC for each class shows an improvement from the smallest input size to the largest with Atelectasis experiencing the maximum jump of 2.21% and the average jump for all the six classes is close to 1%. While Cardiomegaly shows improvement from the initial size of 256 pixels, the AUROC score for the largest

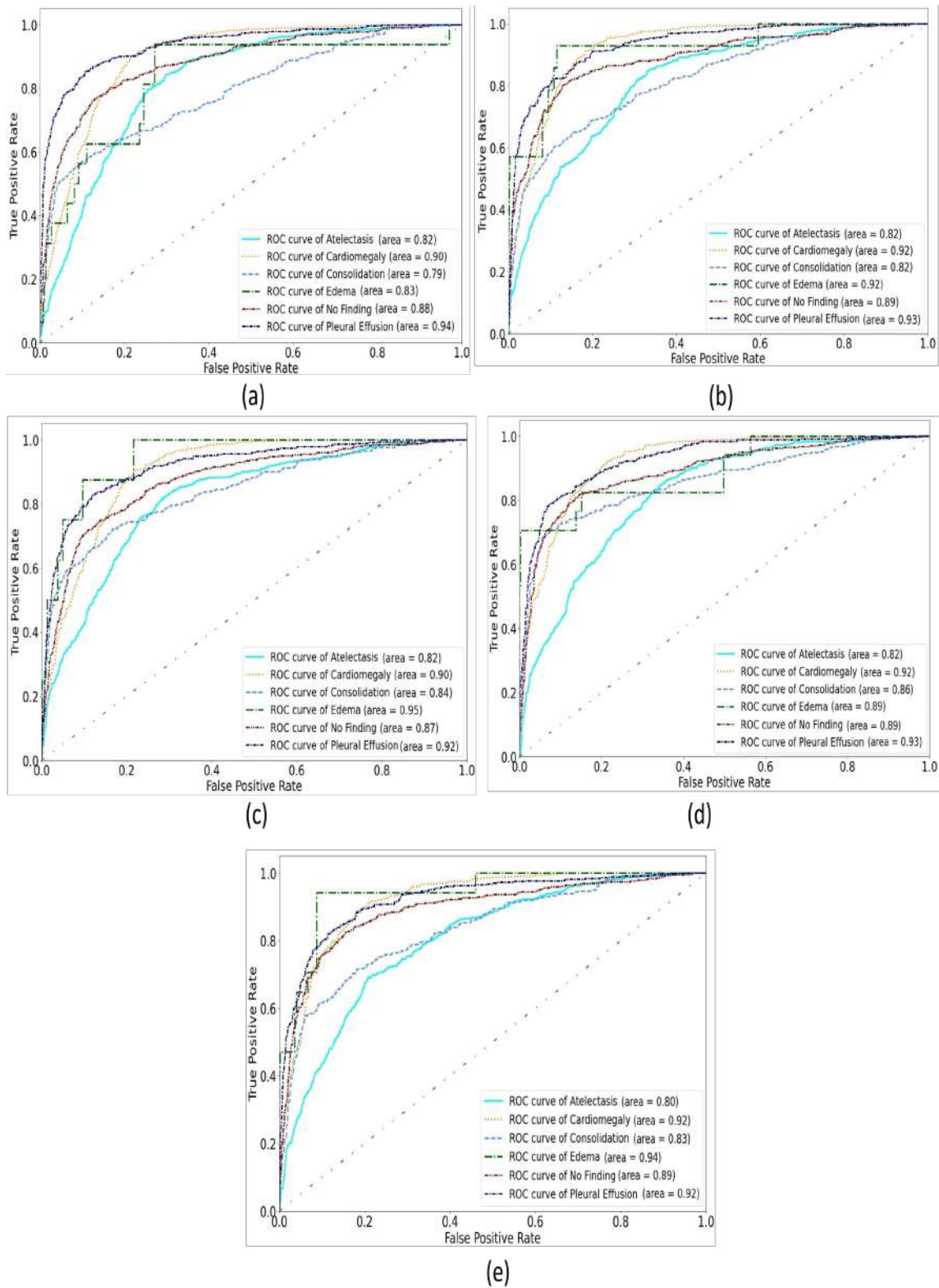


Figure 5.11: AUROC curves for the ensembles across five folds ((a) to (e)). Ensembling the models together increases the performance for each fold on BRAX [2] dataset.

size is the same as that of the other sizes, therefore resulting in no improvement at all.

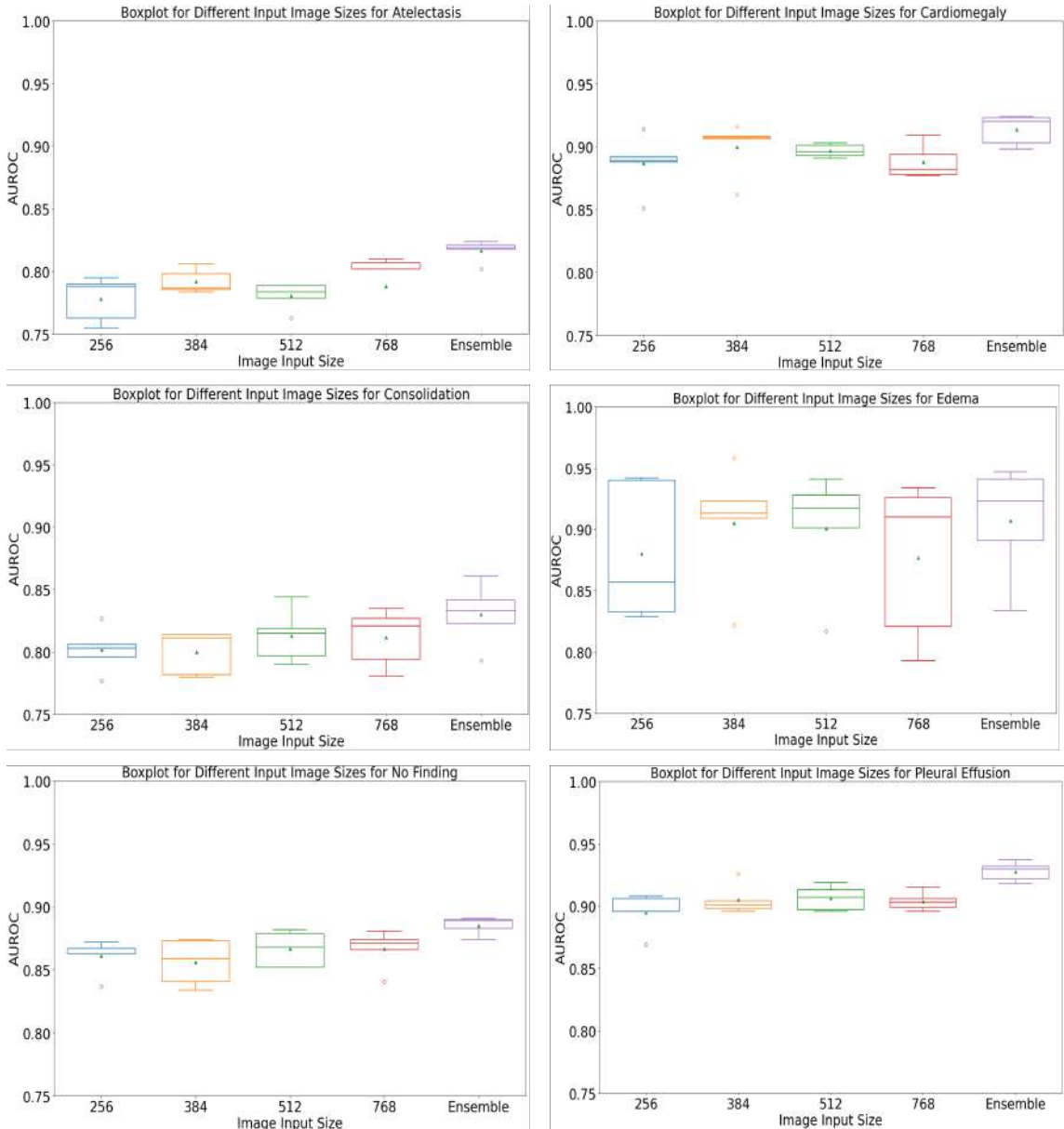


Figure 5.12: Box plots of AUROC for all six classes for the [2] dataset at different input image sizes. The ensemble method outperforms any individual model.

Taking inspiration from [96], we also included a pre-training step that made use of the large-scale, publicly available data set Chexpert [1] with modified progressive learning. Instead of training from scratch with random weights, pre-trained Imagenet weights were used. Results of this training strategy are given in Table 5.4. With pre-training on [1], it can be seen that the average AUROC across 5 folds is within 1.5% of the results that do



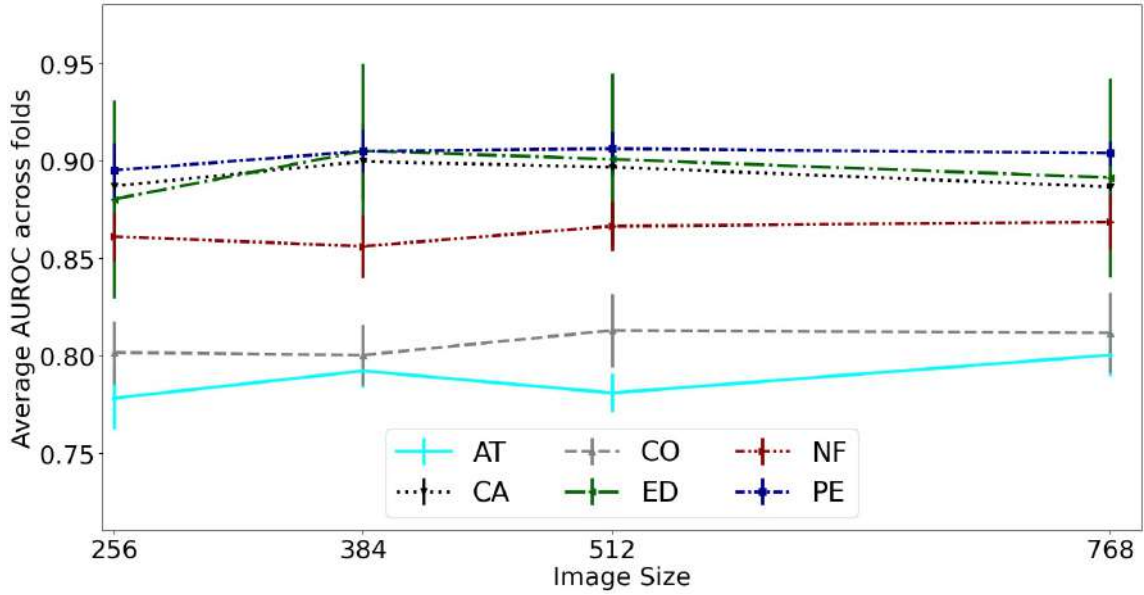


Figure 5.13: Average AUROC for each class across different input image sizes shows the utility of progressive learning.

not include this step indicating that the pre-training does not have overly adverse effects and may even improve results in some cases.

A comparison with other state-of-the-art techniques is undertaken to assess the efficacy of the proposed classification methodology with modified progressive learning for the six classes under consideration. These techniques range from improving generalisation through better normalisation [242] to alleviating the results of domain discrepancy by dividing the tasks into independent binary tasks [243] and from domain adaptation using Fourier [244] to adversarial style augmentation [245]. Table 5.5 shows the results of the proposed methodology for both the validation sets against different techniques.

Table 5.5: Comparison of AUROC scores for 6 classes using modified progressive learning with other techniques in the literature on BRAX [2] dataset. Equal and Unequal represents the validation sets where the samples are capped and uncapped respectively.

Technique	AUROC Score					
	AT	CA	CO	ED	NF	PE
Tang et al. [242]	0.7450	0.8669	0.6734	0.7510	0.7002	0.8505
Lou et al. [243]	0.7481	0.8712	0.6442	0.7489	0.7033	0.8580

**Table 5.5 continued from previous page**

Technique	AUROC Score					
	AT	CA	CO	ED	NF	PE
Yang et al. [244]	0.7644	0.8461	0.7001	0.7532	0.7042	0.8642
Nuriel et al. [246]	0.7511	0.8691	0.6772	0.7482	0.7067	0.8605
Zhong et al. [245]	0.7496	0.8699	0.6832	0.7563	0.7046	0.8608
Nam et al. [247]	0.7527	0.8753	0.6754	0.7417	0.7078	0.8615
Yamashita et al. [248]	0.7442	0.8653	0.6997	0.7647	0.7111	0.8677
Wang et al. [249]	0.7496	0.8590	0.6948	0.7550	0.7072	0.8649
Zhou et al. [250]	0.7537	0.8791	0.6770	0.7489	0.7124	0.8653
Wang et al. [251]	0.7584	0.8752	0.6919	0.7611	0.7117	0.8582
Zunaed et al. [252]	0.7777	0.8867	0.6805	0.7454	0.7217	0.8919
Proposed (Equal)	0.8168	<b><u>0.9133</u></b>	<b><u>0.8305</u></b>	0.9070	<b><u>0.8852</u></b>	<b><u>0.9278</u></b>
Proposed (Unequal)	<b><u>0.8233</u></b>	0.9044	0.7843	<b><u>0.9522</u></b>	0.8687	0.9238

## 5.2.2 Segmentation

Segmentation performance has been gauged for the validation data set of the training data set as well as the BRAX data set and Table 5.6 represents those results. The F1 score difference between our suggested methodology and U-Net on the validation data set is only 0.011 while employing a significantly less number of parameters, and it outperforms U-Net on the BRAX data set, where it achieves an F1 score of 0.9246 in contrast to an F1 score of 0.9162 achieved by U-Net.

In order to further improve the performance on BRAX, we fine-tune our trained model on a set of 100 images from the BRAX data set. To discover the bare minimum number of images required to outperform the prior model, we begin our fine-tuning with just 40 images and gradually increase the number of images by 10 in each re-training of the model. The performance of the model improves going from 0.9163 when 40 images are

Table 5.4: AUROC scores across 5 fold cross validation using modified progressive learning with pre-training on CheXpert [1] data set where the validation set contains a limited number of No Finding class

Fold, Model	AUROC Score						Avg. Score
	AT	CA	CO	ED	NF	PE	
1, 256	0.777	0.89	0.766	0.727	0.832	0.882	0.812
1, 384	0.727	0.889	0.766	0.837	0.83	0.86	0.818
1, 512	0.798	0.892	0.785	0.762	0.862	0.898	0.833
1, 768	0.784	0.898	0.79	0.772	0.875	0.883	0.834
1, Ens	<b>0.802</b>	0.925	0.804	0.813	0.882	0.912	0.856
2, 256	0.755	0.876	0.759	0.8	0.847	0.849	0.815
2, 384	0.765	0.914	0.774	0.835	0.87	0.905	0.844
2, 512	0.76	0.909	0.777	0.873	0.877	0.879	0.846
2, 768	0.757	0.879	0.774	0.851	0.842	0.893	0.833
2, Ens	0.801	0.926	0.804	<b>0.921</b>	0.88	0.924	0.876
3, 256	0.688	0.85	0.768	0.745	0.785	0.817	0.775
3, 384	0.759	0.887	0.808	0.91	0.839	0.893	0.849
3, 512	0.683	0.883	0.77	0.847	0.807	0.861	0.808
3, 768	0.748	0.892	0.819	0.84	0.828	0.874	0.834
3, Ens	0.767	0.915	<b>0.828</b>	0.878	0.85	0.891	0.855
4, 256	0.758	0.893	0.769	0.812	0.848	0.887	0.828
4, 384	0.775	0.912	0.822	0.828	0.884	0.931	0.859
4, 512	0.754	0.91	0.812	0.893	0.873	0.916	0.86
4, 768	0.752	0.9	0.81	0.907	0.854	0.907	0.855
4, Ens	0.785	0.931	0.826	0.907	0.883	<b>0.941</b>	<b>0.879</b>
5, 256	0.7	0.898	0.738	0.807	0.855	0.86	0.81
5, 384	0.771	0.911	0.763	0.815	0.868	0.887	0.836
5, 512	0.752	0.914	0.78	0.835	0.862	0.878	0.837
5, 768	0.764	0.903	0.77	0.916	0.868	0.879	0.85
5, Ens	0.773	<b>0.933</b>	0.785	0.878	<b>0.887</b>	0.91	0.861

Table 5.6: F1 score for validation data sets for different segmentation model architectures on JSRT, Montgomery and Shenzhen [3–7] data sets

Data Sets	Model	Thresh- olding	Parameters (Millions)	F1 Score	
				Validation	BRAX
JSRT,	U-Net		34.5	<b>0.9767</b>	0.9162
Montgomery,	LU-net	OTSU	0.032	0.9488	0.9114
Shenzhen [3–7]	LW-net		0.068	0.9656	<b>0.9246</b>

used to 0.9398 when all are used. Our fine-tuning strategy outperforms the previous model with just 50 images. Figure 5.14 demonstrates that increasing the number of images for fine-tuning the model can have a positive effect on the performance of the model. In fact, from 5.14, it is clear that there is close to a linear relationship between performance and the number of images.

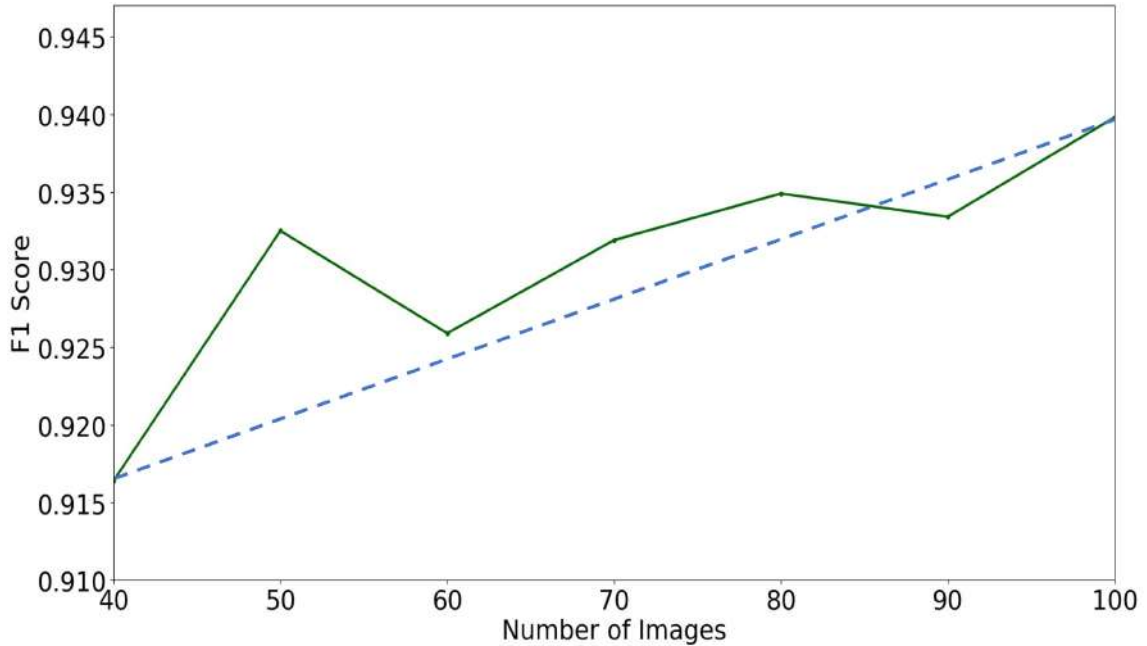


Figure 5.14: F1 score increases with an increase in the number of fine-tuning images from the new (BRAX [2]) data set showing that this technique can be used for continual learning of new data sets.

### 5.2.3 Opacity Localisation

The results for opacity localisation on SIIM [8] data set are shown in Table 5.7 which are reported using the PASCAL VOC at 0.5 threshold for mAP calculation. As mentioned earlier, Efficient Detector [12] D0 to D2 have been used at varying image sizes of 512, 640, and 768. The results demonstrate that performance improved not only by increasing the network size but by also combining the results using Weighted Box Fusion [187]. The weights for each classifier were determined empirically and are, for D0, D1, and D2, respectively, 0.2, 0.05, and 0.75. The scoring cut-off has been set for WBF [187] at 0.05 to maximise the score.

## 5.2.4 Severity Score

The presence or absence of opacity, denoted as 1 or 0, in each of the six lung regions is added to determine the severity score for each CXR and therefore the severity score for each CXR can vary between 0 to 6 where 0 represents normal lungs while 6 represents prevalent opacity in all regions. Its effectiveness is determined by calculating a second matching score that contrasts our framework’s performance with the radiologist’s in determining if there is opacity in a particular area across the full CXR. The greater the degree of agreement between the radiologist and the proposed framework on the presence or absence of opacity in each area, the higher this matching score will be which ranges from 0 to 1. The matching score will be 1 for a CXR where the suggested framework predicted opacity in 4 regions and the radiologist concurs that the opacity is present in only those 4 regions. The score will be 0.8 if the radiologist finds one more location in the CXR where opacity is present but the suggested framework was unable to detect it. This is because there are 4 correctly detected regions out of 5 total regions in the CXR where opacity is present. Table 5.8 shows the mean matching score on the validation data set for different models at different confidence values at 0.3 Intersection over Union. The value of Intersection over Union was chosen empirically as it maximised the mean matching score.

Table 5.7: mAP score for SIIM [8] validation dataset for different models at IoU threshold of 0.5

<b>Model</b>	<b>Image Size</b>	<b>mAP</b>
D0 [12]	512	0.4992
D1 [12]	640	0.4878
D2 [12]	768	0.5109
<b>WBF [187]</b>	-	<b><u>0.5239</u></b>

Out of the 100 BRAX images that were manually annotated by the radiologist, the distribution of matching scores for these images can shed light on the performance of the severity scoring part of the framework. Figure 5.15 provides insight into the performance of our suggested model. Figure 5.15 shows that the framework receives a perfect score of

Table 5.8: Mean matching score opacities localised by different architectures at IoU value of 0.3 on a subset of BRAX [2] data set that was manually annotated by a radiologist

<i>Confidence</i>	<b>Efficient Detector D0 [12]</b>	<b>Efficient Detector D1 [12]</b>	<b>Efficient Detector D2 [12]</b>	<b>WBF [187]</b>
0.1	<b>0.788</b>	<b>0.754</b>	<b>0.788</b>	<b>0.808</b>
0.15	0.75	0.746	0.746	0.738
0.2	0.671	0.733	0.717	0.688
0.25	0.671	0.692	0.675	0.65
0.3	0.621	0.658	0.629	0.592
0.35	0.6	0.621	0.588	0.563
0.4	0.575	0.583	0.571	0.542
0.45	0.55	0.529	0.525	0.508
0.5	0.5	0.504	0.508	0.496
0.55	0.483	0.475	0.492	0.483
0.6	0.471	0.471	0.479	0.467
0.65	0.467	0.471	0.467	0.463
0.7	0.463	0.463	0.467	0.463

1 for 28 out of 100 images. It is off by just one region over 48 images. In contrast, there are only 2 images that have a difference of 4 regions' labels and 12 images that differ for three lung regions. The validation set's distribution reveals that the framework performs reasonably well by correctly identifying opacity in 76 images, when the margin of error is set to just 1 region.

### 5.3 Discussion

The techniques presented here focus on several aspects, one of which is utilizing a small number of images from an unknown target data set for segmentation. It is evident from Figure 5.14 that with as few as 50 images, we were able to demonstrate how this could enhance performance on the target data set for which the model was not initially trained. By fine-tuning the model using a small number of images, this technique can produce a generalised model whose performance can be continuously improved. For data sets where objects do not significantly differ between different data sets, this type of continuous training can be utilised as an alternative to [235]. Retraining for the new data set does not

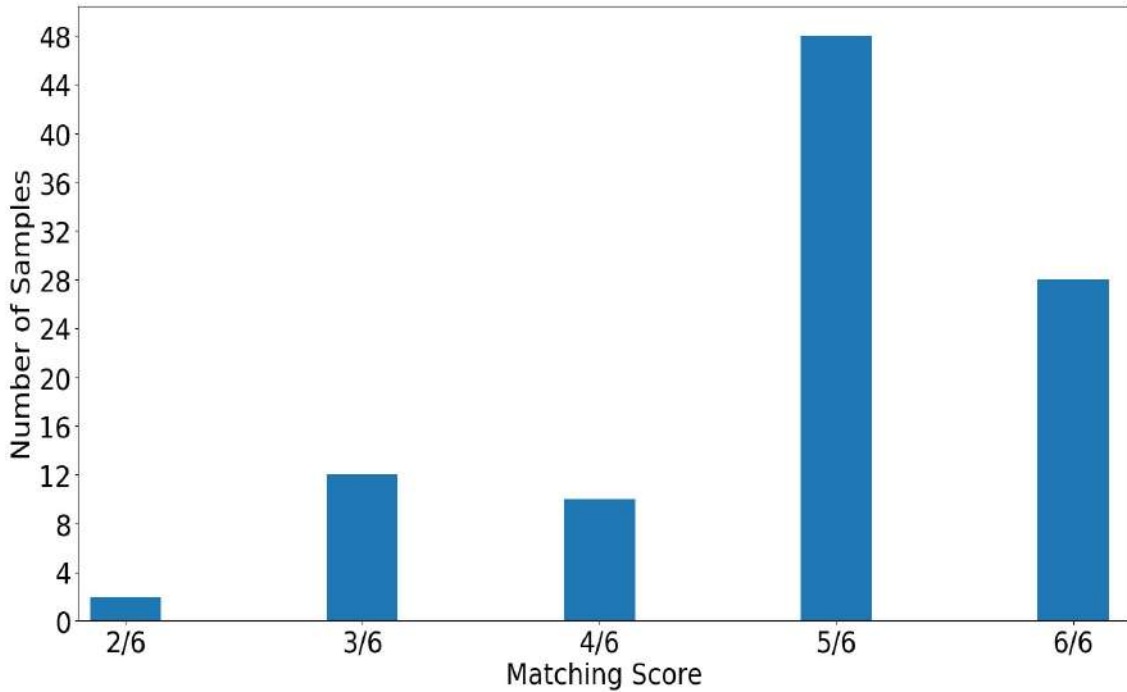


Figure 5.15: Distribution of BRAX [2] validation set according to the matching score. The majority of the images differ by 1 lung region at most from the markings of the radiologist.

result in a significant time cost due to the network’s small size.

In our experimentation, using backbones trained on large CXR data sets yielded lower performance than using Imagenet weights by 1.5% which can be seen from Table 5.4. However, some studies [95, 96] have shown the potential of pre-training using large data sets. One of the reasons for this behaviour might be the small number of epochs that were used for training. Even though, the models were allowed to converge during training, increasing the number of epochs during this step might increase the performance in comparison to training directly from Imagenet weights. Another reason that may have resulted in this lower performance might be that even during pre-training, the weights were not randomly initialised but were initialised from Imagenet. For all our classification models, we used a softmax activation at the output layer while the performance metric was AUROC. In other studies [1, 36, 95, 96], the researchers opted for sigmoid activation. As softmax scales the probabilities such that they add up to 1, the decision to use this activation at the final layer might have resulted in some loss of performance as AUROC

is concerned.

Basing the severity score of pathology on the opacities present in different lung zones enables the use of a single-digit score that can shed light on the progression of the disease. While the severity score is an indication of how many zones are affected by the disease, the matching score that has been proposed here can act as a confidence marker for the severity score. For the framework proposed here, it can be seen from Figure 5.15 that for the BRAX validation set, for 28% of the images, the matching score is 1 i.e. the framework and the radiologist are in agreement for all the six regions of the lungs. However, if this threshold is dropped by just 1 region, then 75% of the images have the same opacity markings for both the framework and the human grader. This large jump can be explained by the nature of the opacities that may spill over from one region to the next and the IoU or the confidence or a combination of both may remove some of the opacities from the final scoring. In addition, the matching score can not only be used for gauging the performance of such frameworks but can even be extended to different human graders.

The framework proposed here achieves excellent performance for segmentation with a lightweight model as evident by Table 5.6. This performance can be improved even further by utilising more images from the BRAX [2] data set during the training. The proposed modified progressive learning module has produced good classification results in the form of AUROC which can be seen from the comparison with other techniques in literature in Table 5.5. For severity scoring, from Table 5.8, it can be seen that a mean matching score of 80.8% is achieved indicating that there is still room for improvement. Improving the opacity localisation by using diverse data sets can lead to an even better severity scoring performance.

Due to the fact that the framework we present here is made up of three distinct sub-modules, the overall computational complexity of the framework is an accumulation of the complexity of these modules. For opacity localization, three different Efficient Detectors are used from D0 to D2 with 3.9 million, 6.6 million, and 8.1 million parameters and 2.54



billion, 6.1 billion, and 11 billion Floating Point Operations Per Second [12]. Similarly, the classification head is composed of quadruple EfficientNet B0 at varying sizes which has 5.3 million parameters and 0.4 billion FLOPS at 224x224 size and scales up for the other sizes that have been used [11]. The segmentation head is the smallest one with only 0.068 million parameters. In essence, for each image at inference, more than 20 billion FLOPS are required.

The suggested framework's combination of progressive learning with a large data set results in one of its primary limitations. The training time and hardware requirements might quickly rise when using this methodology in conjunction with a huge data set because progressive learning necessitates training across a range of input sizes. The model's current training set only includes a portion of the pulmonary disorders that are contained in the data set, which is another drawback. The performance may suffer if the number of diseases for classification is increased. The proposed matching score is also currently region-based rather than pixel-based, which may offer a greater resolution for the disease's progression over time.

## **5.4 Summary**

We present a multi-head deep learning framework for pulmonary disease detection and severity scoring using opacity detection and localisation from a single chest X-Ray and demonstrate reasonably good performance in all three components of the framework. The severity grading of pulmonary diseases helps to guide the subsequent processes as it can help determine what care should be provided to the patient based on the progression of the disease. We demonstrate the use of LW-Net for segmentation provides an efficient method for segmenting the lungs which can be continually improved for new target data sets by fine-tuning using a handful of images. Modified progressive learning which caps the augmentation rate of training images for subsequent models in the pyramid has proved to be useful for model performance for classification on the BRAX data set. The collaboration

with radiologists helped us in generating severity scores and to the best of our knowledge, the results of severity grading with the proposed techniques are the first of its kind along with the results for segmentation and classification. While our work aims to employ pre-training from a large scale data set, a future avenue that might be investigated is continual learning by training on other data sets to make the framework more adaptable and capable of handling X rays from varied populations. Lung segmentation performance, particularly for lungs affected by a disease to a great extent, can be improved. In addition, the number of pulmonary diseases covered in this work can also be increased in a continual fashion by making use of continual learning. The number of disorders for which a severity score can be awarded can also be expanded in the future along with increasing its resolution by providing a severity score for each affected pixel instead of a region.

## **Chapter 6**

# **Proposed Framework for Radiology Report Generation from a Singular Perspective using Transformers with Knowledge Distillation**

The accompanying report which describes all the findings that can be drawn from a chest X-Ray forms a crucial part of the examination of the chest cavity. The key elements of this process demand that the CXR be thoroughly studied so that various anomalies can be detected. By providing a report that is accurate, healthcare professionals can be enabled to make better decisions about the care being provided to the patient. To this end, our proposed end-to-end radiology report generation framework built on transformers is trained on text reports in conjunction with visual characteristics of the chest X-ray to generate a reliable report that astutely describes the condition of the thoracic organs.

## 6.1 Report Generation Framework

Our proposed framework generates a CXR report that provides the findings from a single CXR taken either from the Anterior-Posterior or Posterior-Anterior viewing position. An encoder and a decoder are employed in the report-generation module; the former splits the image into patches to create hidden states, while the latter uses the hidden encoded states to generate word probabilities, which are then used to build the final report. The training process also makes use of fine-tuning of a foundation model that is then used to perform Knowledge Distillation (KD) using the encoder. The sections below detail different parts of the framework. An overview of the entire framework is provided in Figure 6.1.

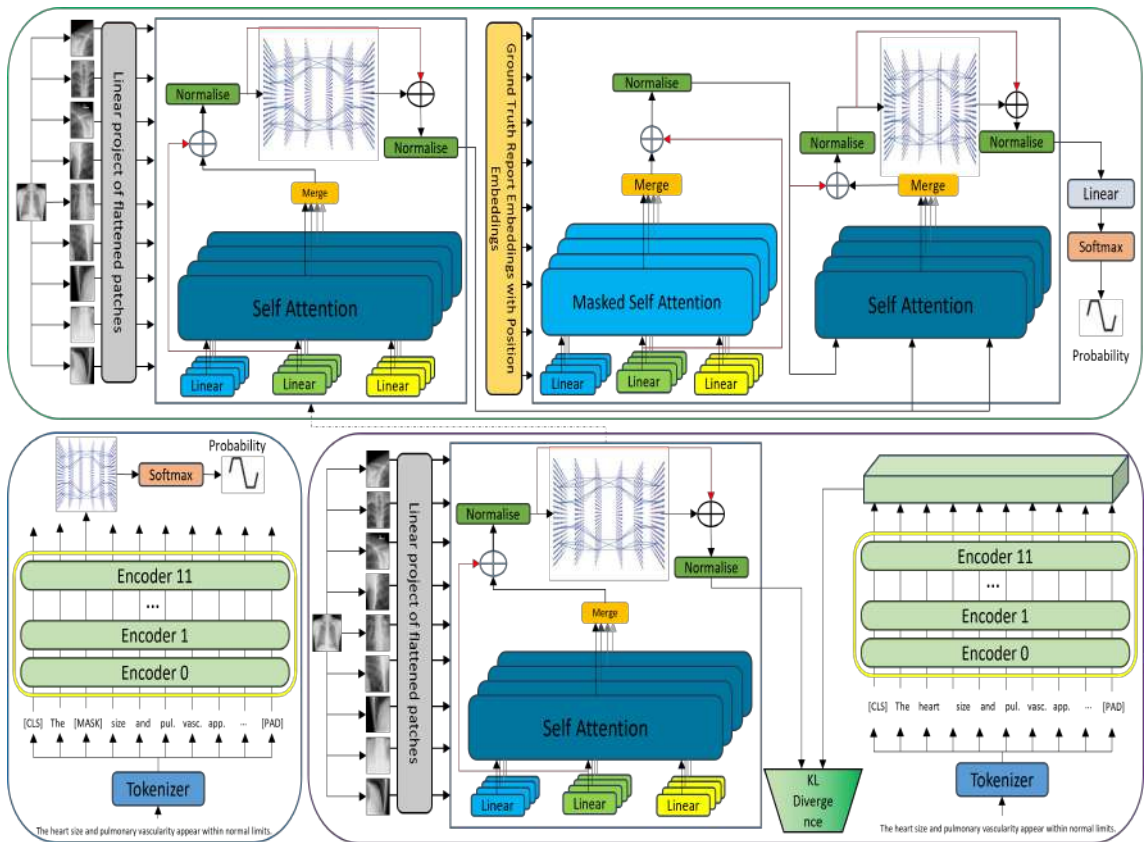


Figure 6.1: Proposed report generation architecture. In the first step, the teacher BioBERT model (Bottom Left) is fine-tuned on the target dataset. Knowledge distillation (Bottom Right) utilises the trained teacher model to reduce the loss in the ViT student model. Finally, this trained ViT is used as the Encoder in the report generation module (Top Centre) in combination with the decoder.

## 6.2 Encoder: Visual Feature Extractor

The first step in creating a report from a single image is to extract the features from the CXR. A Vision Transformer is used in place of a convolutional neural network due to the better performance of the Transformer for image classification tasks [253]. Keeping the architecture similar to the encoder architecture proposed by [137], the ViT preprocess the images by first flattening the patches and then creating embeddings by applying a linear projection of size  $d$  — the size of the latent vector for all the layers. To these embeddings, positional encoding embeddings are added to keep a record of the position of each patch. This resultant embedding is then fed to the encoder architecture from [137].

The Multi-head Self-Attention and a Multilayer Perceptron, also known as the Feed Forward Network, are the two separate sub-layers that make up one block in the encoder design. The output of each sub-layer is added to the input of the same layer by the use of skip connections. This is achieved by keeping the size of all the latent vectors constant. After each layer, layer normalisation [254] is applied. As opposed to the batch normalisation that is standard in CNNs, the approach used by transformers is that of normalising each feature dimension effectively, normalising the embeddings of each token in the input [255]. Different numbers of the aforementioned block result in different sizes of networks with different numbers of parameters. Figure 6.2 shows the ViT architecture along with the procedure for the Self-Attention mechanism.

### 6.2.1 Image Input Representation

In order to make use of the transformer encoder for two-dimensional images instead of text, the input image must be modified such that the image embeddings are similar to the text embeddings. To that end, the two-dimensional image is converted to a sequence of  $N$  patches where each patch has the size  $(P^2 \cdot C)$  where  $P$  represents the size of each patch while  $C$  is the number of channels. The total number of patches  $N$  equals  $(H * W)/P^2$

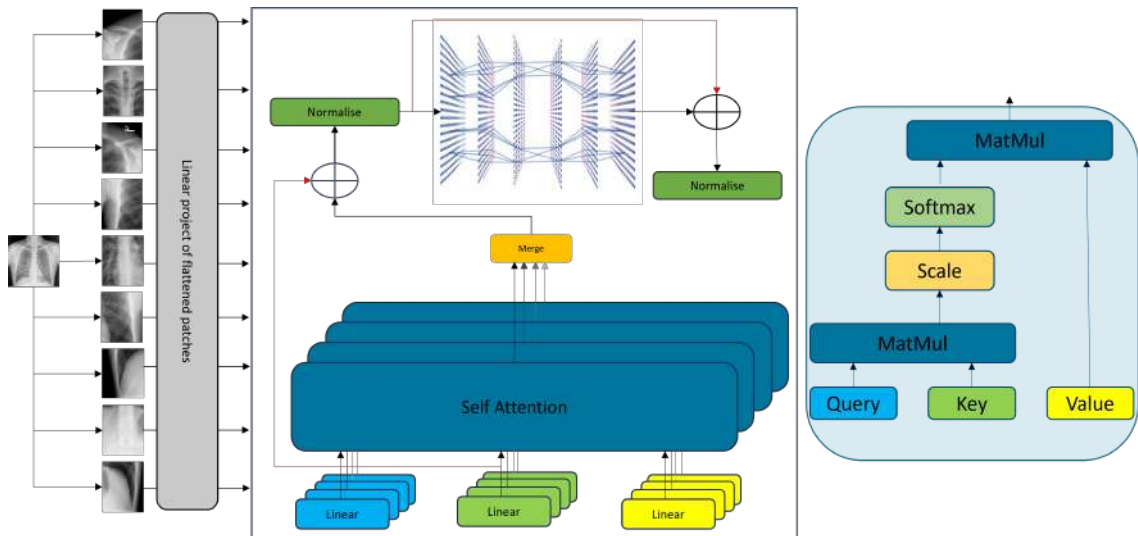


Figure 6.2: Vision Transformer is based on the original encoder model and has been modified for an image. The images are divided into 196 patches of 16x16 for the input size of 224x224. The encoder is used to extract the visual features through the use of Self-Attention which is the matrix product of linear projections of the input that is scaled and then passed through the Softmax activation.

where  $H$  and  $W$  are the height and width of the image and serves as the length of the input sequence to the transformer. To embed the two-dimensional patches in a higher dimension  $D$ , the patches are first flattened and then put through a linear projection transformation. These patch embeddings resemble the text embeddings that the transformer takes as input. To these patch embeddings, embeddings generated from positional encoding are added to preserve the positional information of each patch in the original image.

The ViT architecture was trained for classification tasks, therefore, with the input sequence, a learnable embedding representing the class of the image is prepended. This sequence of embeddings is then passed onto the transformer encoder containing Multi-Head Self-Attention.

## 6.2.2 Self Attention Mechanism

The Self Attention (SA) process remains unchanged because ViT is based on [137]. For SA, instead of using the input embeddings of each patch directly, these are used to generate the *Key* ( $K$ ), *Query* ( $Q$ ) and *Value* ( $V$ ) matrices by projecting the input with different

weight matrices. The similarity score between  $Q$  and  $K$  is calculated by 6.1 where the  $1/\sqrt{D_q}$  accounts for the scaling that is required to prevent the vanishing gradients.

$$\text{Similarity Score} = E = \frac{Q \cdot K^T}{\sqrt{D_q}} \quad (6.1)$$

The similarity score is used for the calculation of the attention weights by the application of Softmax non-linearity by 6.2. In order to obtain the output vector,  $Y$ , attention weights are multiplied by the  $V$  matrix as shown in equation 6.3.

$$\text{Attention Weights} = A = \text{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{D_q}}\right) \quad (6.2)$$

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{D_q}}\right) \cdot V \quad (6.3)$$

Self-attention is employed repeatedly in parallel via the use of multiple attention heads with various projections of  $K$ ,  $Q$ , and  $V$  derived from various weight matrices to enhance performance as shown in 6.4. The result from each of the attention heads is concatenated before being linearly projected using  $W^O$  before being passed onto the next encoder block in the architecture.

$$\begin{aligned} \text{Multi-Head Attention}(Q, K, V) &= \text{Concat}(\text{Head}_1, \text{Head}_2, \dots, \text{Head}_N) \cdot W^O \\ \text{where Head}_i &= \text{Softmax}\left(\frac{Q_i \cdot K_i^T}{\sqrt{D_q}}\right) \cdot V_i \end{aligned} \quad (6.4)$$

The use of the multiple attention heads allows for reducing the dimensions of the  $K$ ,  $Q$ , and  $V$  by the number of heads  $h$  by  $d_{model}/h$ . The concatenation of the output from each head results in the same number of dimensions as using a single self-attention head.

### 6.2.3 Positional Encoding

Sequential models like the RNN and LSTM have the advantage of knowing the position of the input token which, by the virtue of their design, transformers-based architectures lack due to the inherent lack of recurrence and convolutions. Therefore in order to make use of the order of each token in the sequence, this information has to be integrated with the input. Position encodings allow incorporating this information with the embeddings generated from the tokens by adding the encoding vector to the embeddings. This is accomplished by maintaining the same dimension for both the position encoding and the input encoding. [137] opted for a sine and a cosine function for generating position encodings given in 6.5 and 6.6.

$$\text{PE}(p, 2i) = \sin\left(\frac{p}{10000^{2i/d}}\right) \quad (6.5)$$

$$\text{PE}(p, 2i + 1) = \cos\left(\frac{p}{10000^{2i/d}}\right) \quad (6.6)$$

where  $L$  is the entire length of the sentence,  $p$  indicates the current position of the word in the sequence  $L$ , and  $i$  has a range of  $0 \leq i \leq L/2$ . The input embedding dimension is designated as  $d$ .

The use of the sine and the cosine functions allow for a continuous range while being a geometric progression. Furthermore, using these sinusoidal functions allow the models to learn better relative positions which can then be approximated by a linear function [137].

## 6.3 Decoder

The decoder architecture is relatively similar to the encoder architecture in that Masked Multi-Head Self Attention (MMHA), another sub-module, is included in addition to Multi-



Head Self Attention and Multilayer Perceptrons. In the Masked MHA layer, some of the values in the product of the Attention Weights and the Value  $V$  are masked by being set to negative infinity. This ensures that the prediction  $i$  is limited to being dependent on the outputs coming before  $i$  by offsetting the embeddings by one position. The MHA sub-module of the decoder computes the Self Attention utilising Keys  $K$  and Values  $V$  from the hidden state from the encoder block and Queries  $Q$  generated from the decoder input, another distinction between the encoder and the decoder. By adding the input to the output of a sub-layer and then normalising the output, skip connections and normalisation layers function similarly to the encoder. Positional encoding embeddings are used in a similar manner to the encoder.

In the proposed framework, the decoder from MiniLM [13] is used in conjunction with the ViT encoder. The transformer model in [13] has been trained using distillation by using a larger teacher model based on BERT [256]. This approach allows the student model to capitalise on the information learnt by the larger teacher at a fraction of the parameters of the teacher. The knowledge distillation is achieved by using Kullback-Leibler Divergence [257] between the scaled dot products of Query-Key and Value-Value from the last self attention module of both the teacher and the student model. Figure 6.3 shows the distillation process used to train the student model.

### **6.3.1 Text Input Representation**

The relevant input text must be provided to the model in a format that is appropriate for model processing. As a result, the model initially tokenizes the text reports from the CXR. This procedure entails breaking down the text into distinct components, ranging from syllables to whole words. Each of these tokens functions as a separate input entity, and tokens originating from a single input text are padded to preserve consistency across the dataset. The embeddings are produced using these tokenized representations. These embeddings are vector representations with length equal to the model dimensions that

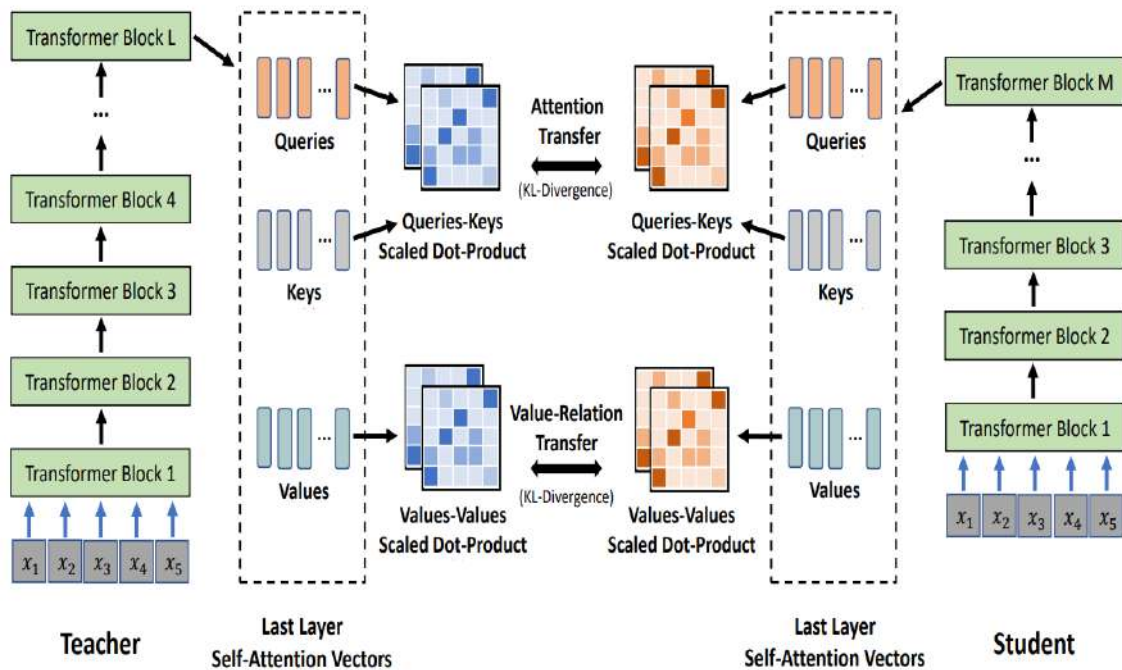


Figure 6.3: Query-Keys and Values-Values scaled dot product is calculated for both the teacher and the student model the similarity between them is increased using the KL Divergence. Image taken from [13]

were produced by the distinct vocabulary in the dataset. Positional information is added using the positional encoding once the embeddings have been generated. The full associated text is given to the decoder during training, but only the initial start token is given to produce the predicted report.

## 6.4 Foundation Model: BioBERT

In computer vision, a model can learn broad representations of various objects by being pre-trained on a big labeled dataset with a large number of samples for a large number of classes [258]. The same model can then be fine-tuned for a comparable task using these weights. In the same vein, it is possible to train large language representation models for upstream tasks using unlabeled data and then fine-tune them for a particular downstream task - an approach that has shown performance improvements [256, 259–261].

Keeping the above in view, we pre-train the Bidirectional Encoder Representations from

Transformers for Biomedical Text Mining (BioBERT) [262] on the CXR reports using the Masked Language Modelling approach which was inspired by [263]. BioBERT has applied the same approach to [256] by pre-training the aforementioned model on medical corpora from sources such as PubMed and PMC articles totaling 18 billion words [262]. It was then further fine-tuned for three downstream tasks namely biomedical named entity recognition, biomedical relation extraction, and biomedical question answering where it achieved better performance compared to BERT. Although BioBERT is already pre-trained on a medical corpus that is suited for the medical domain, this training corpus lacks CXR reports. Therefore, pre-training it on the CXR report corpus should improve the performance for the downstream task of knowledge distillation.

## 6.5 Knowledge Distillation Module via Teacher-Student

### Model

In order to further improve the performance of the proposed framework, features-based knowledge distillation was adopted. Using a similar approach to [264], the visual encoder in the report generation module acts as the student model while BioBERT which has been pre-trained on a particular dataset reports acts as the teacher model.

In contrast to the masked language model training of BioBERT [262], the complete associated text is tokenised and no token is masked. Concurrently, the corresponding image is passed through the ViT [253] to obtain the image embeddings. The last hidden state from BioBERT and the image embeddings are then used to calculate the Kullback-Leibler [257] divergence loss given by equation 6.7.

$$L_{KL}(y_{\text{pred}}, y_{\text{true}}) = y_{\text{true}} \cdot \log \left( \frac{y_{\text{pred}}}{y_{\text{true}}} \right) = y_{\text{true}} \cdot (\log y_{\text{true}} - \log y_{\text{pred}}) \quad (6.7)$$

$$L_{KL}(y_{pred}, y_{true}) = \text{Softmax}(y_{true}) \cdot (\log(\text{Softmax}(y_{true})) - \log(\text{Softmax}(y_{pred}))) \quad (6.8)$$

To ensure that the KL divergence is mathematically correct, a log operation is applied to the embeddings from each model preceded by a Softmax operation as shown in equation 6.8. The KL divergence loss makes sure that the hidden states, which are regarded as features, allow the reduction of the difference between the teacher's and the student's model activation, where the embeddings from the former are regarded as the target while those from the latter are regarded as input. Figure 6.4 shows the training time configuration of the Teacher-Student model. The teacher model is removed during inference and the trained student model is used as the encoder in training the encoder-decoder module in the report generation framework.

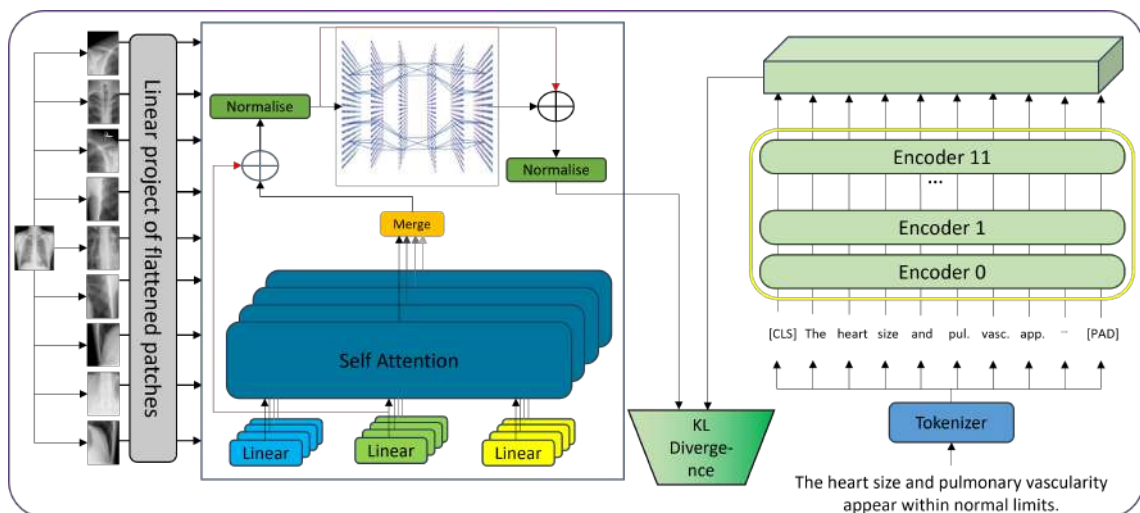


Figure 6.4: For a CXR-report pair, the latent space features are generated by both the ViT and the BioBERT which are used to as features to increase the similarity between the two by reducing the loss using KL Divergence.

## 6.6 Training the Report Generation Framework

The proposed framework was trained and evaluated on three different datasets: Indiana University Chest X-ray [9], MIMIC [10] and MIMIC-PRO [14]. In addition, the trained models underwent blind evaluation on two different data sets as well: a local dataset and a subset of BRAX [2]. For both these datasets, the reports were generated by a team of radiologists with X number combined experience. Three substeps, as shown in Figure 6.1, were used in training the framework.

### 6.6.1 Pre-Training of Foundation Model

The pre-training of the [262] model takes place on the appropriate dataset reports dataset using the original split using the original dataset split for MIMIC and MIMIC-PRO and the split used for Indiana University [9] is taken from [265]. For all the datasets, only the AP and PA view positions are kept while all other view positions are discarded.

The textual data must be cleaned as part of the standardisation process by getting rid of special characters, unused spaces, and redacted patient data. The removed personal information of the patients is represented by *xxxx* in [9] and *----* in [10]. Furthermore, after the removal of all unnecessary characters, the remaining characters are converted to lowercase. The standardisation process from [266] was adapted for the [10, 14] datasets while for the Indiana dataset, [265] was used along with the training, validation, and test split. In addition, the vocabulary for each of the datasets — generated from the entire text corpus including findings, impressions, and free text — was determined individually, and for [10] the words with an occurrence of  $\leq 3$  were dropped from the reports. This resulted in 9396 unique words for [10] and 5615 for [14]. The findings and impressions sections are combined for [10], and if either is missing, the other is used. The free text section is utilised if either or both are missing.

We make use of pre-trained BioBERT<sub>BASE</sub> having 137 million parameters where approx-

imately 108 million parameters are trainable. The model contains 12 layers with the size of the latent vector from the last hidden layer being 768 and it also incorporates a token classification layer. This model is trained with the maximum number of 512 tokens and 35% of the tokens are masked during this process. The learning rate is kept as  $5 \times 10^{-5}$  with a weight decay of 0.01. As the result of applying a softmax function to the full vocabulary, the model outputs the probability of the masked token. Figure 6.5 shows a part of a report is masked for training the foundation model.

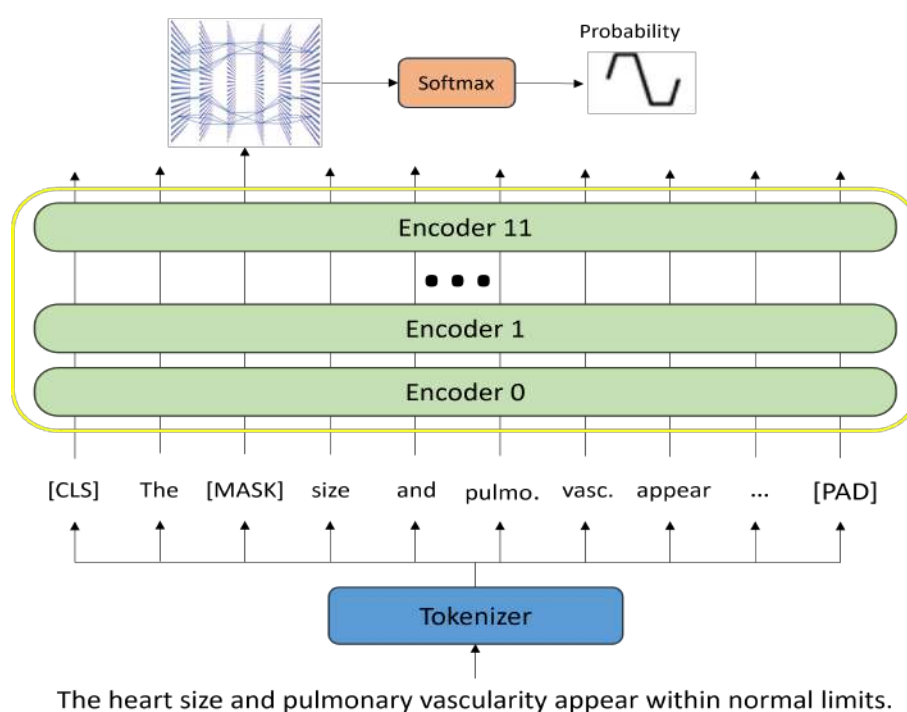


Figure 6.5: Part of the CXR report is masked to train the foundation model in an unsupervised manner. The trained foundation model is then used as the teacher in the Knowledge Distillation step.

## 6.6.2 Knowledge Distillation

The encoder model ViT [253], which behaves as the student, is trained using the foundation model, which serves as the teacher, once the foundation model has been trained using the processed dataset. The student and teacher models are provided, respectively, with a training image and its corresponding processed report. The KL divergence [257] uses the embeddings generated by the final hidden layer of each of these models, which have

the same dimensions. The weights of the models are then updated using backpropagation using the calculated loss. The encoder for training the report-generation sub-module is the student model from the training with the minimum loss.

For the student model, We make use of ViT-Base [253] architecture consisting of 12 layers and 12 heads and containing approximately 86 million parameters. The input image size for the ViT is set to 224x224 and 16 patches per image are used. The latent vector's dimensionality from ViT's final hidden layer mirrors that of BioBERT<sub>BASE</sub> at 768, eliminating the need for dimensional adjustments. The collective parameter count for the combined ViT and BioBERT<sub>BASE</sub> models reaches approximately 194 million, indicating a substantial capacity for learning and adaptation.  $5 \times 10^{-4}$  is the learning rate that is used for training with no weight decay and a batch size of 128.

### **6.6.3 Training of the Downstream Task: Report Generation**

Using the distilled student ViT-Base from the previous training step, the report generation sub-module is trained. ViT acts as the encoder that extracts the visual features while MiniLM [13] acts as the decoder that takes the latent space vectors from the encoder and uses it as the hidden state. The MiniLM that is employed in this configuration has 12 layers, a hidden state dimension of 384, and only 41 million parameters, bringing the total number of parameters for the sub-module close to 127 million. The tokenizer from MiniLM is the one used for tokenization with a varying maximum length for different datasets. Figure 6.6 shows the configuration of the report generation module.

AdamW [267] optimiser is used with a learning rate  $5 \times 10^{-5}$  and the training batch size is varied between 2 to 8. The performance metric on which the models are saved is the BLEU-1 score on the validation dataset. The same pre-processing that was used to train the foundation model was applied to the CXR reports for training the decoder.

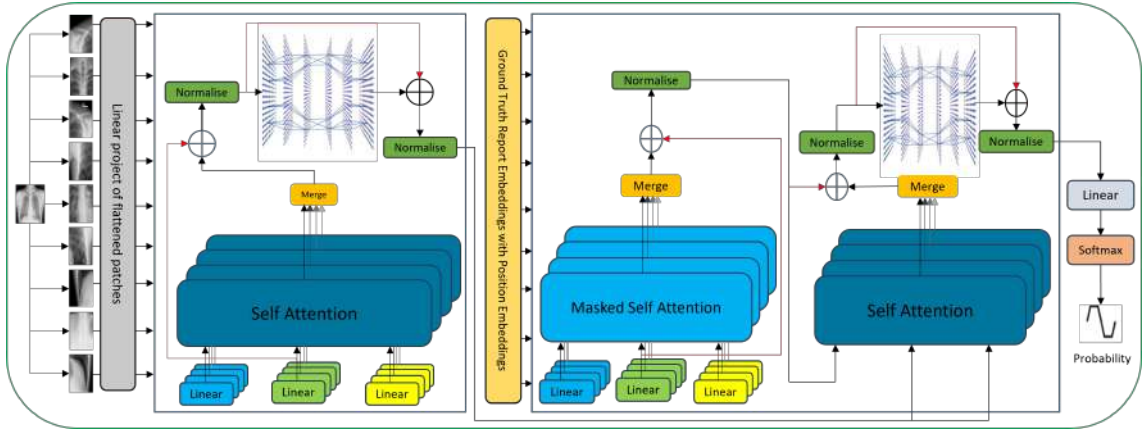


Figure 6.6: Multiple layers are used in both the encoder and the decoder with each layer containing multiple attention heads. Each successive layer in the encoder receives the input from the previous layer while the last layer passes the output to each decoder layer. The figure illustrates the last encoder and decoder layer.

## 6.7 Results

To gauge the performance of the proposed framework, every model in the framework was trained at least three times, and the best-performing model was retained. The models were trained using PyTorch in Python on two systems; one with 64 GB RAM and two Nvidia RTX 2070 GPUs and the other with 128 GB RAM and a single Nvidia RTX 3090 Ti GPU.

### 6.7.1 Evaluation Metrics

The performance metrics that have been calculated for this framework include both semantic similarity and lexical similarity. The BLEU score [219], a precision-based lexical similarity metric, divides the total number of words in the output by the number of n-grams that are present in both the predicted output and the ground truth. Equation 6.9 defines the BLEU score [268].

$$\text{BLEU-N} = BP \cdot \exp \left( \sum_{n=1}^N W_n \cdot \log(\text{precision}_n) \right) \quad (6.9)$$

Where:



$$BP = \begin{cases} 1 & \text{if } |p| > |r| \\ e^{1 - \frac{|r|}{|p|}} & \text{otherwise} \end{cases}$$

$$\text{precision}_n = \frac{\sum_{p \in \text{output}} \sum_{n\text{-gram} \in p} \text{Count}_{\text{clip}}(n\text{-gram})}{\sum_{p \in \text{output}} \sum_{n\text{-gram} \in p} \text{Count}(n\text{-gram})}$$

$$\text{Count}_{\text{clip}}(n\text{-gram}) = \min(\text{matched } n\text{-gram count}, \max_{r \in R}(n\text{-gram count in } r))$$

By computing the embeddings of each token in the generated report and the ground truth using the cosine similarity, BERTScore [269], in contrast to BLEU prioritises semantic similarity as shown in equation 6.10 [268].

$$\text{BERTscore} = F_{\text{BERT}} = \frac{2 \cdot P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \quad (6.10)$$

Where:

$$\text{Recall}_{\text{BERT}} = \frac{1}{|r|} \sum_{i \in r} \max_{j \in p} \vec{i}^T \vec{j}, \quad \text{Precision}_{\text{BERT}} = \frac{1}{|p|} \sum_{j \in p} \max_{i \in r} \vec{i}^T \vec{j}$$

Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L) [270] and Radiology Report Clinical Quality, (RADCliQ) [271] are also used. The former measures the sentence level structural similarity [268] while the latter measures the quality of the report as a combination of BLEU and RadGraph score. Equation 6.11 to 6.12 represent the aforementioned metrics.

$$\text{ROUGE-N} = \frac{\sum_{s_r \in \text{references}} \sum_{n\text{-gram} \in s_r} \text{Count}_{\text{match}}(n\text{-gram})}{\sum_{s_r \in \text{references}} \sum_{n\text{-gram} \in s_r} \text{Count}(n\text{-gram})} \quad (6.11)$$

$$\text{RadCliQ} = \alpha \cdot \text{BLEU-2} + (1 - \alpha) \cdot \text{RadGraph F1} \quad (6.12)$$

where  $\alpha$  controls the relative importance of BLEU-2 and RadGraph F1 [229].

Table 6.1: Hyperparameters of different stages of the report generation framework

Stage	Models	Layers	Parameters (Millions)	Latent Dimensions	Tokens	Learning Rate	Scheduler	Loss Function	Activation	Image Size	Input Dimensions	Optimizer	Batch Size
Foundation Model	BioBERT	12	137, 108	768	512	$5 \times 10^{-5}$	Constant	Cross- entropy	Softmax	-	-	AdamW	4
		12, 12	194 (108, 86)	768	512	$5 \times 10^{-4}$		KL Diverg- ence	-	224	3		128
Distillation Report	ViT Base	12,	127	768,	512,	$5 \times 10^{-5}$		Cross- entropy	GELU				2 to 8
		12	(86, 41)	384	50/80	$5 \times 10^{-5}$							
Generation	MiniLM												

## 6.7.2 MIMIC Dataset Evaluation

For the MIMIC dataset, the Findings and the Impressions sections were combined to form the reports for the chest X-rays. The use of this combination provides a two-fold advantage: it not only increases the length of the individual report but also increases the number of samples that would have been otherwise discarded due to the missing Findings or Impressions sections. Furthermore, only PA and AP views were kept resulting in 237,887 training samples and 1958 validation samples. The foundation model and the knowledge distillation sub-module were trained using the training and evaluated using the test samples. Table 6.2 shows the results of the proposed framework on only Findings, only Impressions, and Findings combined with Impressions.

Table 6.2: Performance of proposed on MIMIC [10] dataset. The framework is trained on a combination of Findings and Impressions sections while the results are computed on different sections of the report. F represents *Findings*, F+I represents *Findings & Impressions* while I represents *Impressions*. The absence of (AP & PA) denotes that all samples were used for testing.

Data	B1 ↑	B2 ↑	B3 ↑	B4 ↑	ROUGE ↑	RadCliQ ↓	BERT-Score ↑
F (AP & PA)	0.3104	0.1801	0.1167	0.0818	0.2213	3.6145	0.3484
F	0.3048	0.1762	0.1145	0.0803	0.2197	3.6361	0.3507
F & I (AP & PA)	0.3007	0.1720	0.1099	0.0763	0.2142	3.721	0.3192
F & I	0.2959	0.1689	0.1084	0.0725	0.2134	3.745	0.3207
I (AP & PA)	0.1625	0.0839	0.0488	0.0312	0.1404	4.1372	0.2405

**Table 6.2 continued from previous page**

<b>Data</b>	<b>B1</b> ↑	<b>B2</b> ↑	<b>B3</b> ↑	<b>B4</b> ↑	<b>ROUGE</b> ↑	<b>RadCliQ</b> ↓	<b>BERT-Score</b> ↑
I	0.1512	0.0767	0.044	0.0276	0.1344	4.2053	0.2344

Along with the quantitative results shown in Table 6.2, the qualitative results from randomly selected samples from the MIMIC dataset are shown in Figures 6.7, 6.8, 6.9 corresponding to results on Findings, Impressions, and combined Findings and Impressions sections. The differences are shown in red while the other colors are used for similarities. The order of the sentences generated varies from that of the ground truth. Furthermore, different verbiage is employed to express the same concept.

A comparison with state-of-the-art techniques is undertaken to assess the efficacy of the proposed report generation methodology. Table 6.3 shows the performance of the proposed methodology against different techniques for the under consideration evaluation metrics for just the Findings section of the reports.

Table 6.3: Performance of the proposed framework on the Findings section against different techniques in literature for MIMIC [10] data set. The absence of (AP & PA) denotes that all samples were used for testing.

<b>Technique</b>	<b>B1</b> ↑	<b>B2</b> ↑	<b>B3</b> ↑	<b>B4</b> ↑	<b>RL</b> ↑	<b>Rad-CliQ</b> ↓	<b>BERT-Score</b> ↑
Vinyals et al. [272]	0.256	0.157	0.102	0.070	0.249	–	–
Xu et al. [273]	0.304	0.177	0.112	0.077	0.249	–	–
Anderson et al. [274]	0.280	0.169	0.108	0.074	0.250	–	–
Liu et al. [275]	0.334	0.217	<b>0.140</b>	<b>0.097</b>	<b>0.281</b>	–	–
Lu et al. [276]	0.302	0.189	0.122	0.082	0.259	–	–
Rennie et al. [277]	0.314	0.199	0.126	0.087	0.265	–	–

Table 6.3 continued from previous page

Technique	B1 ↑	B2 ↑	B3 ↑	B4 ↑	RL ↑	Rad- CliQ ↓	BERT- Score ↑
Krause et al. [278]	0.321	0.203	0.128	0.089	0.266	–	–
Harzig et al. [279]	0.328	0.204	0.127	0.090	0.267	–	–
Jing et al. [216]	0.329	0.206	0.133	0.095	0.273	–	–
Huang et al. [280]	<b>0.337</b>	0.211	0.136	0.095	0.274	–	–
Jeong et al. [281]	–	<b>0.220</b>	–	–	–	<b>3.585</b>	<b>0.353</b>
Nicolson et al. [282]	–	0.196	–	–	–	<b>3.617</b>	<b>0.347</b>
Proposed (PA & AP)	0.310	0.180	0.116	0.081	0.221	3.614	0.348
Proposed	0.304	0.176	0.114	0.080	0.219	3.636	0.350

Similar to Table 6.3, Table 6.4 shows the performance of the proposed methodology against different techniques the Impressions section. As mentioned earlier, this section is a brief summary of the findings section.

Table 6.4: Performance of the proposed framework on the Impressions section against different techniques in literature for MIMIC [10] data set. The absence of (AP & PA) denotes that all samples were used for testing.

Technique	B1 ↑	B2 ↑	B3 ↑	B4 ↑	RL ↑	Rad- CliQ ↓	BERT- Score ↑
Ramesh et al. [14]	–	–	–	–	–	–	0.229
Jeong et al. [281]	–	0.084	–	–	–	<b>3.781</b>	<b>0.287</b>
Endo et al. [283]	–	0.055	–	–	–	4.121	0.193
Li et al. [284]	–	0.030	–	–	–	4.313	0.190
Miura et al. [285]	–	<b>0.087</b>	–	–	–	–	0.227
Chen et al. [221]	–	0.059	–	–	–	–	0.186

**Table 6.4 continued from previous page**

<b>Technique</b>	<b>B1</b> ↑	<b>B2</b> ↑	<b>B3</b> ↑	<b>B4</b> ↑	<b>RL</b> ↑	<b>Rad- CliQ</b> ↓	<b>BERT- Score</b> ↑
Yan et al. [286]	–	0.064	–	–	–	–	0.188
Nicolson et al. [282]	–	0.066	–	–	–	–	0.192
Proposed (PA & AP)	0.162	0.083	0.048	0.031	0.140	4.137	0.240
Proposed	0.151	0.076	0.044	0.027	0.134	4.205	0.234

Table 6.5 shows the comparison of the combined Findings and Impression sections with the literature. Only a few studies have chosen to use this combined strategy.

Table 6.5: Performance of the proposed framework on the combined Findings and Impressions section against different techniques in literature for MIMIC [10] data set. The absence of (AP & PA) denotes that all samples were used for testing.

<b>Technique</b>	<b>B1</b> ↑	<b>B2</b> ↑	<b>B3</b> ↑	<b>B4</b> ↑	<b>RL</b> ↑	<b>Rad- CliQ</b> ↓	<b>BERT- Score</b> ↑
Jeong et al. [281]	–	0.161	–	–	–	3.835	0.287
Yan et al. [286]	–	0.144	–	–	–	3.986	0.275
Chen et al. [129]	–	0.137	–	–	–	4.051	0.271
Proposed (PA & AP)	0.3007	<b><u>0.172</u></b>	0.1099	0.0763	0.2142	<b><u>3.7210</u></b>	0.3192
Proposed	0.1512	0.0767	0.0440	0.0276	0.1344	3.7450	<b><u>0.3207</u></b>

### 6.7.3 Indiana University CXR Dataset Evaluation

Keeping the same approach as that used for MIMIC, the proposed framework was also trained on the Indiana University chest X-rays and reports. However, instead of combining

<b>Actual</b>	pa and lateral views of the chest provided . lung volumes are somewhat low . allowing for this there is no focal consolidation effusion or pneumothorax . the cardiomeastinal silhouette is normal . imaged osseous structures are intact . no free air below the right hemidiaphragm is seen .
<b>Predicted</b>	pa and lateral views of the chest provided. there is no focal consolidation effusion or pneumothorax. the cardiomeastinal silhouette is normal. imaged osseous structures are intact. no free air below the right hemidiaphragm is seen. no acute intrathoracic process.
<b>Actual</b>	frontal and lateral views of the chest demonstrate low lung volumes which accentuate bronchovascular markings . there is no focal consolidation pleural effusion or pneumothorax . the hilar and mediastinal silhouettes are unremarkable . heart size is top normal . partially imaged upper abdomen demonstrates prominent air-filled large bowel loops .
<b>Predicted</b>	frontal and lateral views of the chest demonstrate low lung volumes which accentuate bronchovascular markings. there is no pleural effusion focal consolidation or pneumothorax. hilar and mediastinal silhouettes are unremarkable. heart size is normal. there is no pulmonary edema. partially imaged upper abdomen is unremarkable. a
<b>Actual</b>	pa and lateral views of the chest provided . there is no focal consolidation effusion or pneumothorax . the cardiomeastinal silhouette is normal . imaged osseous structures are intact . no free air below the right hemidiaphragm is seen . elevation of the right hemidiaphragm is unchanged from chest radiograph .
<b>Predicted</b>	pa and lateral views of the chest provided. lung volumes are low. there is no focal consolidation effusion or pneumothorax. the cardiomeastinal silhouette is normal. imaged osseous structures are intact. no free air below the right hemidiaphragm is seen. no acute intrathoracic process.
<b>Actual</b>	as compared to the previous radiograph the patient has received a new dobhoff tube . the tip of the tube projects over the middle parts of the stomach . the course of the tube is unremarkable . there is no evidence of complications notably no pneumothorax . otherwise the radiographic appearance of the thoracic organs is similar to the previous examination .
<b>Predicted</b>	as compared to the previous radiograph the patient has received a nasogastric tube. the course of the tube is unremarkable the tip of the tube projects over the middle parts of the stomach. there is no evidence of complications notably no pneumothorax. otherwise the radiograph is unchanged.

Figure 6.7: The text highlighted in red represents portions of the text that are absent in the generated report for the Findings section of the MIMIC [10] data set. Conversely, the other highlighted sentences demonstrate the similarities between the ground truth and the generated report

<p><b>Actual</b></p> <p>as compared to the previous radiograph the patient has been intubated . the tip of the endotracheal tube projects approximately 48 cm above the carinal . the course of the nasogastric tube is unremarkable the tip projects over the middle parts of the stomach . the right internal jugular venous introduction sheet is unchanged . increasing atelectasis at the left lung bases . unchanged appearance of the right lung .</p>	<p><b>Predicted</b></p> <p>as compared to the previous radiograph the patient has been intubated . the tip of the endotracheal tube projects approximately 4 cm above the carina . the patient has also received a nasogastric tube . the course of the tube is unremarkable the tip of the tube is not included on the image . the right internal jugular vein catheter is</p>
<p><b>Actual</b></p> <p>in comparison with the study of there again are bilateral pleural effusions with compressive atelectasis much more prominent on the left . continued enlargement of the cardiac silhouette with mild elevation of pulmonary venous pressure . pacer leads are unchanged .</p>	<p><b>Predicted</b></p> <p>in comparison with the study of there is little overall change . again there is substantial enlargement of the cardiac silhouette with pulmonary edema and bilateral pleural effusions with compressive basilar atelectasis more prominent on the left . pacer device remains in place .</p>
<p><b>Actual</b></p> <p>as compared to the previous radiograph the patient has received a nasogastric tube . the tip projects over the middle parts of the stomach . no complications notably no pneumothorax . the apical lung regions are not included on the image .</p>	<p><b>Predicted</b></p> <p>as compared to the previous radiograph the patient has received a nasogastric tube . the course of the tube is unremarkable the tip of the tube projects over the middle parts of the stomach . there is no evidence of complications notably no pneumothorax . unchanged appearance of the lung parenchyma and the cardiac silhouette .</p>
<p><b>Actual</b></p> <p>portable ap radiograph of the chest was reviewed in comparison to . the et tube tip is 83 cm above the carina . the right picc line tip is at the level of low svc . cardiomeastinal silhouette is unchanged . extensive consolidations throughout the lungs are demonstrated with slight improvement at the level of mid left lung . bilateral pleural effusions are unchanged .</p>	<p><b>Predicted</b></p> <p>portable ap radiograph of the chest was reviewed in comparison to . the et tube tip is 5 cm above the carina . the right picc line tip is at the level of mid svc . heart size and mediastinum are stable . bilateral pleural effusions and bibasal consolidations are unchanged . there is no evidence of pneumothorax .</p>

Figure 6.8: The impression section is briefer than the findings section. The differences have been marked in red for the MIMIC [10] data set. The model is unable to correctly identify the distance of different tubes in the image.



<b>Actual</b>	no focal consolidation is seen there is no pleural effusion or pneumothorax . the cardiac and mediastinal silhouettes are unremarkable . no acute cardiopulmonary process .
<b>Predicted</b>	the lungs are clear without focal consolidation. no pleural effusion or pneumothorax is seen. the cardiac and mediastinal silhouettes are unremarkable. no acute cardiopulmonary process.
<b>Actual</b>	cardiac mediastinal and hilar contours are normal . pulmonary vasculature is normal . no focal consolidation pleural effusion or pneumothorax is present . there are no acute osseous abnormalities . no acute cardiopulmonary process .
<b>Predicted</b>	heart size is normal. the mediastinal and hilar contours are normal. the pulmonary vasculature is normal. lungs are clear. no pleural effusion or pneumothorax is seen. there are no acute osseous abnormalities. no acute cardiopulmonary abnormality.
<b>Actual</b>	the lungs are clear without focal consolidation . no pleural effusion or pneumothorax is seen . cardiac and mediastinal silhouettes are unremarkable . suggestion of mitral anulus calcification is seen . no acute cardiopulmonary process .
<b>Predicted</b>	the lungs are clear without focal consolidation. no pleural effusion or pneumothorax is seen. the cardiac and mediastinal silhouettes are stable. no pulmonary edema is seen. no acute cardiopulmonary process.
<b>Actual</b>	as compared to the previous radiograph the monitoring and support devices are constant . there is unchanged evidence of extensive bilateral parenchymal opacities . the extent and the severity of these opacities have not changed . unchanged appearance of the cardiac silhouette .
<b>Predicted</b>	as compared to the previous radiograph there is no relevant change. the monitoring and support devices are constant. the extensive bilateral parenchymal opacities are constant in extent and severity. no new opacities. unchanged size of the cardiac silhouette.

Figure 6.9: The combination of Findings and Impression sections not only improves the number of unique words in vocabulary but also improves the quality of the generated reports as well. The results shown above are for the MIMIC [10] data set.

the Findings and the Impressions, only Findings were used in training. In addition, only PA and AP views of the CXRs are kept while others are discarded. Images that do not contain the findings section are also discarded. 3200 training samples, 500 validation samples, and 300 test samples were produced as a result. Table 6.6 shows the results of the proposed framework with the inclusion of Knowledge Distillation.

Table 6.6: Performance of proposed framework on Indiana University [9] dataset. The framework is trained on just the Findings sections.

<b>B1</b> ↑	<b>B2</b> ↑	<b>B3</b> ↑	<b>B4</b> ↑	<b>RL</b> ↑	<b>RadCliQ</b> ↓	<b>BERT-Score</b> ↑
0.4848	0.3183	0.228	0.167	0.3562	2.767	0.4725

Along with the quantitative results and comparison with different techniques in literature, the qualitative results from randomly selected samples from the IU dataset are shown in Figures 6.10, The differences are shown in red.

### 6.7.3.1 MIMIC Pre-training for Report Generation

In order to utilise a large amount of data and account for the lower number of samples used in the Indiana University dataset, two experiments were conducted. In the first experiment (*Exp 1*), the foundation model was trained on the larger MIMIC dataset while the knowledge distillation and the report generation sub-module were trained using the IU dataset. In the other experiment (*Exp 2*), only the report generation sub-module was trained on the IU dataset, and the training of the foundation model and knowledge distillation were both performed using the MIMIC dataset. Table 6.7 shows the results of this approach. As before, the split given in [265] has been used.

<b>Actual</b>	the cardiomeastinal silhouette is within normal limits for size and contour the lungs are normally inflated without evidence of focal airspace disease pleural effusion or pneumothorax no acute <b>bone</b> abnormality
<b>Predicted</b>	the cardiomeastinal silhouette is within normal limits for size and contour the lungs are normally inflated without evidence of focal airspace disease pleural effusion or pneumothorax no acute <b>bony</b> abnormality
<b>Actual</b>	the cardiomeastinal silhouette is within normal limits for size and contour the lungs are normally inflated without evidence of focal airspace disease pleural effusion or pneumothorax <b>no acute osseus abnormality</b>
<b>Predicted</b>	the cardiomeastinal silhouette is within normal limits for size and contour the lungs are normally inflated without evidence of focal airspace disease pleural effusion or pneumothorax <b>osseous structures are intact</b>
<b>Actual</b>	the cardiomeastinal silhouette is within normal limits for size and contour the lungs are normally inflated without evidence of focal airspace disease pleural effusion or pneumothorax <b>no acute osseus abnormality</b>
<b>Predicted</b>	the cardiomeastinal silhouette is within normal limits for size and contour the lungs are normally inflated without evidence of focal airspace disease pleural effusion or pneumothorax <b>osseous structures are within normal</b>
<b>Actual</b>	the cardiomeastinal silhouette is within normal limits for size and contour the lungs are normally inflated without evidence of <b>focal airspace disease</b> pleural effusion or pneumothorax <b>no acute osseus abnormality</b>
<b>Predicted</b>	the cardiomeastinal silhouette is within normal limits for size and contour the lungs are clear without evidence of <b>focal consolidation</b> pneumothorax or pleural effusion

Figure 6.10: In the majority of cases for IU [9] dataset, the generated report closely resembles the ground truth. The discrepancies, marked in red, involve the use of different words, but they still convey the same meaning.

Table 6.7: Performance of proposed framework on Indiana University [9] data set with pre-training on MIMIC [10] data set for different sub-modules.

<b>Exp</b>	<b>B1</b> ↑	<b>B2</b> ↑	<b>B3</b> ↑	<b>B4</b> ↑	<b>RL</b> ↑	<b>RadCliQ</b> ↓	<b>BERTScore</b> ↑
Exp 1	0.5572	0.4032	0.3028	0.2172	0.4211	2.1943	0.5800
Exp 2	0.5476	0.3397	0.3450	0.2218	0.4291	2.1600	0.5730

From Table 6.7, it can be seen that both types of pre-training result in improvement of scores across all evaluation metrics. Just using the foundation model trained on MIMIC dataset and re-training the Knowledge Distillation module on IU dataset results in higher BLEU-1, BLEU-2, and BERTScore as compared to the other methodology. Conversely, BLEU-3, BLEU-4, ROUGE, and RadCliQ show improvement when the ViT distilled using MIMIC foundation model is used.

A comparison with state-of-the-art techniques is undertaken to assess the efficacy of the proposed report generation methodology. Table 6.8 shows the performance of the proposed methodology against different techniques for the under-consideration evaluation metrics.

Table 6.8: Performance of the proposed framework on the Findings section against different techniques in literature for Indiana University [9] data set.

<b>Technique</b>	<b>B1</b> ↑	<b>B2</b> ↑	<b>B3</b> ↑	<b>B4</b> ↑	<b>RL</b> ↑	<b>Rad- CliQ</b> ↓	<b>BERT- Score</b> ↑
Liu et al. [227]	0.483	0.315	0.224	0.168	0.376	–	–
Liu et al. [287]	0.492	0.314	0.222	0.169	0.381	–	–
Liu et al. [275]	0.473	0.305	0.217	0.162	0.378	–	–
You et al. [288]	0.484	0.313	0.255	0.173	0.379	–	–
Noorala- hzadeh et al. [289]	0.486	0.317	0.232	0.173	0.390	–	–

Table 6.8 continued from previous page

Technique	B1 ↑	B2 ↑	B3 ↑	B4 ↑	RL ↑	Rad- CliQ ↓	BERT- Score ↑
Yang et al. [266]	0.496	0.317	0.232	0.173	0.390	–	–
Kale et al. [290]	0.423	0.256	0.194	0.165	<b>0.444</b>	–	–
Qin et al. [291]	0.492	0.321	0.235	0.181	0.384	–	–
Wang et al. [292]	0.497	0.357	0.279	<b>0.225</b>	0.414	–	–
Proposed	<b>0.557</b>	<b>0.403</b>	<b>0.302</b>	0.217	0.421	2.194	0.58

#### 6.7.4 MIMIC-PRO Dataset Evaluation

For MIMIC-PRO, only the Impression section has been rewritten to remove references to previous reports which result in an overall reduced length of the report [14]. In a similar fashion to the other datasets, only the PA and AP views were used in this dataset for training. There are 2065 samples in the test dataset which are either PA or AP and 2188 samples if all views are considered, whereas 199,396 samples make up the training dataset. Table 6.9 shows the results of the proposed framework on the test set.

Table 6.9: Performance of proposed on MIMIC-PRO [14] dataset. The framework is trained on Impressions from which prior references have been removed. I represents *Impressions* and the absence of (AP & PA) denotes that all samples were used for testing.

Data	B1 ↑	B2 ↑	B3 ↑	B4 ↑	RL ↑	RadCliQ ↓	BERT- Score ↑
I (AP & PA)	0.0988	0.0496	0.0274	0.0162	0.1064	4.0684	0.2074
I	0.0972	0.0481	0.0262	0.0154	0.1034	4.0766	0.2085

In addition with the quantitative results in Table 6.9, the qualitative results from randomly

selected samples from the MIMIC-PRO are shown in Figure 6.11. Following the earlier convention, the differences are shown in red while the other colors are used for similarities.

#### 6.7.4.1 MIMIC Pre-training for Report Generation

While the difference between the number of samples in MIMIC and MIMIC-PRO is only around 16%, the difference in unique words in the text corpus between the two datasets is nearly 40%. Therefore, keeping this in view, in a similar manner to the pre-training performed on the Indiana University dataset, the same two experiments are conducted as before. *Exp 1*, only the foundation model trained on the MIMIC dataset is used while in *Exp 2*, only Report Generation sub-module is retrained. Table 6.10 shows the results of this approach.

Table 6.10: Performance of proposed framework on MIMIC-PRO [14] data set with pre-training on MIMIC [10] data set for different sub-modules.

Exp	Data	B1 ↑	B2 ↑	B3 ↑	B4 ↑	RL ↑	RadCliQ ↓	BERT Score ↑
Exp 1	I (AP & PA)	0.1181	0.0602	0.0329	0.0192	0.118	3.9468	0.2243
	I	0.1157	0.0581	0.0314	0.018	0.1139	3.9590	0.2245
Exp 2	I (AP & PA)	0.1098	0.0555	0.0303	0.0172	0.1130	3.9393	0.2235
	I	0.1081	0.0542	0.0294	0.0165	0.1099	3.948	0.2245

Similar to Indiana [9], both types of pre-training result in an improvement of scores for all metrics over the baseline as evident from Table 6.10.

A comparison with state-of-the-art techniques is undertaken to assess the efficacy of the proposed report generation methodology. Table 6.11 shows the performance of the pro-

<b>Actual</b>	There are small bilateral pleural effusions, which combine with signs of mild pulmonary edema. Findings suggest small bilateral pleural effusions, pulmonary vascular congestion and bibasilar atelectasis.
<b>Predicted</b>	mild pulmonary edema with small bilateral pleural effusions and bibasilar atelectasis.
<b>Actual</b>	Moderate congestive heart failure with moderate pulmonary edema, small bilateral pleural effusions, and bibasilar airspace opacities likely reflective of atelectasis. Findings compatible with acute pulmonary edema with possible layering pleural effusions.
<b>Predicted</b>	bilateral lower lobe opacities and enlargement of the cardiac silhouette consistent with moderate pulmonary edema. probable small bilateral pleural effusions and bibasilar atelectasis.
<b>Actual</b>	Moderate bilateral effusions and moderate interstitial edema. Constant moderate bilateral pleural effusions with signs of mild to moderate pulmonary edema as well as mild cardiomegaly.
<b>Predicted</b>	lower lung opacities concerning for atelectasis and moderate pleural effusions. cardiomegaly and interstitial pulmonary edema.
<b>Actual</b>	At least moderate right pleural effusion is present and right basal dense opacity. Moderate right pleural effusion with severe right lower lobe atelectasis.
<b>Predicted</b>	small right pleural effusion and right basilar atelectasis.

Figure 6.11: The text highlighted in red represents portions of the text that are absent in the generated report for MIMIC-PRO [14] data set. Conversely, the other highlighted sentences demonstrate the similarities between the ground truth and the generated report.

posed methodology against different techniques for the under-consideration evaluation metrics.

Table 6.11: Performance of the proposed framework on the against different techniques in literature for MIMIC-PRO [14] data set. For the different techniques used in the proposed methodology, the absence of (AP & PA) denotes that all samples were used for testing. \* denotes the results without pre-training on the MIMIC dataset while <sup>+</sup> and <sup>#</sup> refer to *Exp 1* and *Exp 2* from 6.10 respectively.

Technique	B1 ↑	B2 ↑	B3 ↑	B4 ↑	S <sub>emb</sub> ↑	Rad Graph F1 ↑	BERT- Score ↑
Report							
Retrieval	–	–	–	–	0.3601	0.0925	0.2160
[14]							
Sentence							
Retrieval	–	–	–	–	0.3967	0.0864	0.2159
(k=1)							
[14]							
Sentence							
Retrieval	–	–	–	–	0.3859	0.1056	<b>0.2351</b>
(k=2)							
[14]							
Sentence							
Retrieval	–	–	–	–	0.3779	0.1112	0.2254
(k=3)							
[14]							
Proposed							
(AP & PA)*	0.0988	0.0496	0.0274	0.0162	0.3739	0.0983	0.2074
Proposed*	0.0.0972	0.0481	0.0262	0.0154	0.3735	0.0904	0.2085



Table 6.11 continued from previous page

Technique	B1 ↑	B2 ↑	B3 ↑	B4 ↑	$S_{emb}$ ↑	Rad Graph F1 ↑	BERT- Score ↑
Proposed (AP & PA) <sup>+</sup>	0.1098	0.0555	0.0303	0.0172	<b>0.4069</b>	0.1125	0.2235
Proposed <sup>+</sup>	0.1081	0.0542	0.0294	0.0165	0.4059	0.1048	0.2245
Proposed (AP & PA) <sup>#</sup>	<b>0.1181</b>	<b>0.0602</b>	<b>0.0329</b>	<b>0.0192</b>	0.4000	<b>0.1165</b>	0.2243
Proposed <sup>#</sup>	0.1157	0.0581	0.0314	0.0180	0.3988	0.1083	0.2245

### 6.7.5 Local Dataset Evaluation

Without any fine-tuning or retraining, the trained models were used on the locally gathered dataset to assess the generalisability of the proposed framework. The models trained on both the MIMIC [10] and Indiana [9] dataset were used for the local dataset where the model trained on IU dataset performed better than the one trained on MIMIC. Using this trained model, a BLEU-1 score of 0.3827 and a BERTScore of 0.4392 was achieved on the local dataset as shown in Table 6.12.

Table 6.12: Performance of proposed framework on Local data set. The framework trained on the Indiana University [9] data set was used.

B1 ↑	B2 ↑	B3 ↑	B4 ↑	RL ↑	RadCliQ ↓	BERT- Score ↑
0.3827	0.2367	0.1624	0.1153	0.2745	2.5256	0.4392

## 6.7.6 BRAX Dataset Evaluation

Just like with the local dataset, the trained models were evaluated on the BRAX subset of 100 images, for which reports were provided by a radiologist. The models were used without any retraining or fine-tuning on the target dataset to assess its performance. Table 6.13 shows the metrics under consideration for three models that were trained on [9], [10] and [14] respectively.

Table 6.13: Performance of proposed framework on BRAX [2] subset. Models trained on three different datasets (Indiana University [9], MIMIC [10] and MIMIC-PRO [14]) were used.

<b>Model Trained On</b>	<b>B1</b> ↑	<b>B2</b> ↑	<b>B3</b> ↑	<b>B4</b> ↑	<b>RL</b> ↑	<b>Rad CliQ</b> ↓	<b>BERT-Score</b> ↑
Indiana [9]	0.1671	0.0816	0.0422	0.0169	0.1085	4.3325	0.2186
MIMIC [10]	0.1295	0.057	0.0261	0.0139	0.1057	4.0717	0.207
MIMIC-PRO [14]	0.1013	0.0391	0.0129	0.0009	0.0716	4.2896	0.1875

From Table 6.13, it can be seen that the model trained on Indiana dataset outperforms the models trained on the other two datasets in all metrics except RadCliQ. This difference in performance can be explained by the overwhelming ratio of normal samples in the Indiana dataset. Furthermore, the model trained on MIMIC outperforms both models for RadCliQ by a significant margin. For BERTScore, the difference between models trained on Indiana and MIMIC dataset is just 1.16% which is significantly less than the difference observed in BLEU-1, BLEU-2 and BLEU-3 scores. Similarly, for BLEU-4 score, the difference between the model trained on Indiana dataset and the the model trained of MIMIC dataset is far less.

Another thing to note is that the variation in the generated reports is least from the model that has been trained on Indiana dataset which is due to less variability of the Indiana corpus itself. The reports that have been generated by the model trained on MIMIC have a tendency to have hallucinatory references to prior examinations as such reports are quite common in the dataset corpus. This problem is largely solved by the use of model trained on MIMIC-PRO which does not contain such references. However, as the MIMIC-PRO text corpus focuses on just the *Impressions* portion of the report, the generated reports tend to be relatively brief and follow the structure of the *Impressions* section instead of the *Findings* section which more closely resembles the structure of BRAX reports. Figure 6.12 shows the radiologist report and the reports generated by MIMIC-PRO [14] model.

### 6.7.7 Ablation Study

The strength of the proposed framework stems from Knowledge Distillation where what the teacher model has learnt can be used to improve the performance of a smaller student model despite their architectural differences and distinct purposes. In order to evaluate the effect of Knowledge Distillation, using the IU dataset [9], the proposed framework was trained with and without utilising this training step. Table 6.14 shows the effects of using this strategy. The BERTScore and BLEU-1 scores both increased with the use of knowledge distillation, going from 0.4725 to 0.5482 and 0.4449 to 0.4848, respectively, as seen in Table 6.14.

Table 6.14: Effects of using Knowledge Distillation in training by using the foundation model as teacher and ViT as the student model on the Indiana University [9] data set.

Technique	B1 ↑	B2 ↑	B3 ↑	B4 ↑	RL ↑	RadCliQ ↓	BERT-Score ↑
Without Knowledge Distillation	0.4449	0.2921	0.207	0.1518	0.3562	2.767	0.4725

**Table 6.14 continued from previous page**

<b>Proposed Framework</b>	<b>B1</b>	<b>B2</b>	<b>B3</b>	<b>B4</b>	<b>ROUGE</b>	<b>RadCliQ</b>	<b>BERT-Score ↑</b>
With Knowledge Distillation	0.4848	0.3183	0.228	0.167	0.3573	2.3954	0.5482

## 6.8 Discussion

One of the limitations of the framework presented here lies in the hardware and the time requirements for training the complete framework as it contains several sub-modules. However, at the time of inference, the architecture is simplified which reduces the number of parameters required. This, in addition to the proposed framework not relying on report retrieval, makes the framework a relatively simple approach at the time of inference.

Due to most of the lungs having a normal presentation in the MIMIC [10], BRAX [2], MIMIC-PRO [14] and the Indiana [9] datasets and the findings section following a systematic structure, most of the associated reports are repetitive and similar thus lacking information regarding different abnormalities. While this allows the machine learning approaches to learn better representations of normal images and is inline with the natural distribution of the samples in most of the CXR datasets [2, 9, 10, 14], critical findings can be missed in the report due to the lack of abnormal samples. This issue also plagues the proposed framework here as evidenced by the sample generated reports present Figures 6.7, 6.8 and 6.9. One attempt to address this is the use of only abnormal radiology reports in the MIMIC dataset to train a model as proposed by [293]. To improve the model’s ability to identify abnormalities in CXRs, it may be helpful to fine-tune a model that was trained on the complete MIMIC corpus using this aberrant-only dataset.

Knowledge Distillation is shown to improve the performance as evident by the improve-

<b>Actual</b>	sternotomy sutures in place mild cardiomegaly with increased cardiothoracic ratio mild increased perihilar bronchovascularity may suggest early interstitial edema both cp angles sharp .
<b>Predicted</b>	mild cardiomegaly with interstitial edema.
<b>Actual</b>	heart size is enlarged with well defined margins may suggest moderate pericardial effusion however need correlation with echocardiogram lungs are clear cp angles are sharp.
<b>Predicted</b>	massive cardiomegaly with history of pericardial effusion. echocardiogram is recommended for further evaluation.
<b>Actual</b>	raised right hemidiaphragm no significant mediastinal shift lungs appear normal needs ultrasound chest evaluation to rule out eventration or subpulmonic effusion .
<b>Predicted</b>	the right hemidiaphragm is elevated. cardiomeastinal silhouette is within normal limits. there is no focal consolidation pleural effusion or pulmonary edema. there are no pneumothoraces.
<b>Actual</b>	lungs are clear heart size appear normal normal study .
<b>Predicted</b>	no acute cardiopulmonary pathology.

Figure 6.12: The reports for BARX [2] subset have been generated without any retraining or fine-tuning. Red highlights the differences between the generated report and the report by the radiologist. Similarities are shown in different colors. It can be observed that different verbiage has been used for the same statement in the actual and the predicted reports.

ment in performance on Indiana dataset in Table 6.14. Furthermore, it has been demonstrated that pre-training for report generation enhances performance on smaller datasets relative to the bigger pre-training dataset, in addition to Knowledge Distillation. The performance for both [9] and [14] datasets increases – as evident by Tables 6.7 and 6.10 — as a result of pre-training on [10]. The larger corpus of the MIMIC dataset, both in terms of samples and vocabulary allows the framework to learn better representations. This is especially evident in the results of the Indiana dataset where the pre-training step significantly increases all metrics as shown in Table 6.7.

The use of Knowledge Distillation allows for a model with a relatively small size to perform reasonably well. In contrast to our proposed strategy, the architecture proposed in [283] makes use of a report retrieval strategy which necessitates that the embeddings from a large CXR reports corpus are extracted and then compared with the embeddings extracted from the test image. This results in large memory and time requirements which can be reduced by a compression technique [283]. Even though the proposed framework takes less computational resources [294] than the retrieval approach [283], it is still able to outperform the in  $s_{emb}$  and RadGraph F1 score. Even in BERTScore, the difference is a just little over 1% as can be seen in Table 6.11.

Another thing to consider is the report template discrepancy that can be found in the gathered local dataset and the international datasets that were used to train the framework. The difference in the style of how the findings are provided in a radiology report can impact the performance of automated frameworks. This difference can be seen in Table 6.13 for the BRAX reports created by local experts. The length of the reports in the different datasets can have greater impact on metrics like BLEU scores while metrics like BERTScore are affected less by it.

One more thing to consider is that the BLEU-1 score captures semantic similarity. For reports that are closer in content, the BLEU-1 score is also high. However, this is affected by the ratio of the normal reports in the testing data and the structure of the findings

section that is followed for a dataset. During experimentation, some models, while underperforming on the BLEU-1 metric (approximate BLEU-1 score of 0.41 on Indiana [9] dataset) had a greater number of unique generated reports with more variability in the verbiage of the reports. This variability resulted in a lower BLEU-1 score, despite having more unique words and less repetition.

While pre-training has been performed for the Foundation model to accommodate for the lack of CXR reports in the original training corpus, pre-trained weights have been used in case of BioBERT and ViT while MiniLM has been trained from scratch. Training these larger models from scratch on the CXR corpus might improve the overall performance of the proposed framework. However, such a change to the training strategy will increase the training time significantly.

## **6.9 Summary**

We present a radiology report generation framework from a single perspective using a transformer encoder-decoder with features-based knowledge distillation and demonstrate reasonably good performance on all the datasets. We also demonstrate that using pre-training and knowledge distillation improves the report generation results which is evident from an increase in metrics such as the BLEU score. Our method also shows that it is possible to achieve variation in the generated reports while still doing well on the aforementioned measures. Masked language pre-training of the teacher language model with a particular dataset allows for better domain and dataset representation. The parameters of the student models are then updated based on loss calculated using the latent space features of the student and the teacher model. Generating an automated report that represents an accurate picture of the condition of the chest cavity is an important step in providing better care to patients.

# Chapter 7

## Conclusion and Future Work

### 7.1 Conclusion

In particular, in Low- and Middle-Income Countries, Chronic Respiratory Diseases, including Chronic Obstructive Pulmonary Disease, offer substantial challenges. One of the most susceptible human body systems is the respiratory system, and respiratory disorders are a major worldwide burden, particularly in LMICs. The COVID-19 pandemic and the inadequate healthcare infrastructure make these problems even more severe. Due to these compelling factors, an automated system that is capable of quickly and accurately diagnosing a variety of pulmonary disorders using chest X-rays is needed. In addition to disease classification, disease severity, and automatic radiological report generation are also required as the former alone is insufficient.

Building models that are not only able to classify but also to offer radiological reports and severity grading has become possible owing to the availability of diverse datasets like BRAX, MIMIC, JSRT, and SIIM, among others. In order to train several sub-modules and validate the output of the trained models, over 400,000 CXR images were employed in the research presented in this thesis.

In this research endeavor, one of the sub-framework that has been presented is the multi-



head deep learning framework aimed at diagnosing pulmonary diseases through chest X-ray analysis. This sub-framework is structured into multiple specialized "heads", each serving a specific function: image-level classification, opacity localization, and severity quantification. The classification head leverages the power of EfficientNet architecture and also takes advantage of the modified version of progressive learning where the augmentation is capped for each increasing size. Furthermore, an ensemble-based approach further improves the results. The segmentation head employs a W-Net architecture that is an amalgamation of two U-Nets concatenated after one another. This architecture is particularly effective for isolating regions of interest, namely the lungs in CXRs. The use of a composite dataset ensures that the model is machine-agnostic, and can therefore be applied across different datasets with high efficacy. The localization head adopts the EfficientDet architecture, which provides the means to identify opacities in the lung region. Ensemble methods similar to the one used for classification and techniques like Test Time Augmentation and Weighted Box Fusion further enhance its performance, particularly in identifying opacities of various sizes. A novel approach to quantify severity by leveraging both segmentation and localization heads is also introduced. This approach not only detects the presence of pathology but also gives a measure of its severity, which is indispensable for treatment planning. Performance metrics such as Area under the ROC curve, F1 score, and mean average precision were used to rigorously evaluate each head of the model, with the best-performing versions retained for each task.

The second sub-framework is for automated radiology report generation leveraging transformer-based architectures. Specifically, our method utilizes Vision Transformers as encoders for image feature extraction and transformer decoders for decoding these features into comprehensive radiology reports. The sub-framework is further bolstered by implementing Knowledge Distillation, effectively transferring the learned 'knowledge' from a larger, more complex model (BioBERT) to a less computationally demanding student model. The use of BioBERT, pre-trained on a vast corpus of medical texts, enabled our framework to capture the nuanced language used in radiology reports effectively. The exper-

iments conducted across multiple datasets (Indiana and MIMIC) demonstrated that the proposed framework either achieved comparable performance or improved the results for performance metrics such as BLEU, ROUGE-L, and RADcliQ. While the training involves training of three different sub-modules resulting in increased time complexity and resource requirement, at the time of the inference, the report generation sub-module has a simple architecture with a relatively small number of parameters. In addition, ablation studies verified the value of incorporating knowledge distillation, marking a clear increase in metrics like BERTScore and BLEU-1.

Several factors have contributed to the performance achieved by the proposed methodology. It has been shown that modifying the progressive learning approach yields better results than relying solely on progressive learning. Additionally, many augmentation methods that are more effective than others have been identified, particularly for CXR images. Moreover, it is also observed that using relatively smaller models performs better than using larger models, indicating that the latent space feature extraction from the CXR images is negatively impacted by the larger networks' increased complexity. This may indicate that there might be a limit on the number of network parameters that result in good performance on CXR images. Similarly, for segmentation of lungs, using a smaller architecture that employs an attention based. As evident by the literature as well, pre-training on larger datasets than the target dataset can improve performance. This can be attributed to the model learning general features from the larger datasets. This pre-training step not only allows for performance improvements in classification of diseases but also works well for generating lung masks for new datasets and complements the report generation for datasets that lack a large vocabulary of unique words and a variety of report structures.

Although the proposed framework performs well for the various aspects required for CXR analysis, there are still several limitations of the proposed methodology. For disease classification, the scope of this research work is limited as only six diseases have been addressed for pulmonary classification. In the same manner, while this research expands the

gamut of severity detection beyond just COVID-19 which has been the focus of recent studies, the single-digit severity scoring mechanism has been used for only four chronic diseases. Other diseases have been left out that can also benefit from such a scoring system. Another limitation is the lack of segmentation of other anatomical structures in the CXR. Separating these structures can help with the identification of different diseases in addition to providing more details about pulmonary pathologies. The datasets utilised to generate reports have a tendency to be significantly skewed towards normal reports, which can lead to a bias in the framework towards normal reports as well. Although this represents the datasets' natural distribution, the model's performance may suffer in situations when there is a higher likelihood of abnormalities in the test images. The inherent imbalance of the datasets has not been specifically addressed in this thesis. Furthermore, the computational resources necessary for training the proposed framework, particularly with pre-training on large datasets, can be prohibitive in increasing the number of pathologies that this framework targets. Efficient training strategies and resource optimisation can provide a workaround for this limitation. Such strategies can also allow to circumvent limitations of using pre-trained weights by the different models in the framework and can allow to train from scratch potentially leading to better performance. Finally, more research and discussion are required on the ethical issues and legal compliance related to AI-based CXR analysis.

The major contribution of this research work is the development and validation of a comprehensive framework for automated CXR analysis encompassing disease classification, severity grading, and report generation. These achievements can be summarised as follows:

1. **Comprehensive Framework:** We propose a single framework consisting of two modules i.e. CXR manifestation analysis and radiology report generation. Compared to the literature, this integrated approach provides a holistic solution. for CXR analysis.

2. **Modified Progressive Learning:** We propose a single sub-framework consisting of disease classification using modified progressive learning and severity grading for different pulmonary disorders using opacity localisation. The modified progressive learning approach achieves an improvement of approximately 9% over the methods in literature depicting that this approach provides a performance improvement.
3. **Single-Digit Severity Scoring:** A single-digit severity score is introduced for 4 pulmonary diseases enabling concise and quantifiable assessment of disease progression.
4. **Segmentation Masks with Severity Scores:** We provide segmentation masks with severity grades for a validation data set from the BRAX [2] data set that has been validated by a radiologist.
5. **Efficient Segmentation Network:** A segmentation network sub-module is utilized that despite its relatively small size is able to perform relatively close to large architectures such as U-Net. This demonstrates that high performance at a comparatively small size is achievable.
6. **Fine Tuning for Segmentation on Unseen Datasets:** We experimentally show that while good performance can be achieved in segmentation using publicly available data sets, fine-tuning on just a small number of samples from the target data set can actually improve the segmentation performance even further.
7. **Foundation Model Fine-Tuning with Knowledge Distillation:** We propose a report generation sub-framework employing foundation model fine-tuning on CXR reports for use as a Teacher model to train a smaller Student model. In addition, we also propose employing Knowledge Distillation so that a smaller Student model can learn better CXR representation. It has been demonstrated that applying Knowledge Distillation can enhance performance as shown by the results of Indiana University [9] data set.

8. **Pre-Training on CXR Datasets with Reports:** Pre-training on larger CXR datasets with reports is shown to improve performance when used in conjunction with Knowledge Distillation providing reasonable performance at a relatively small size and simple architecture Using the MIMIC [10] data set for pre-training, the performance improvements are gauged on the Indiana data set by the application of this strategy validating the approach
9. **Local CXR Images Dataset with Findings:** Gathered local CXR images dataset with findings (reports) from a local hospital.
10. **Radiology Reports for Validation:** We also provide radiology reports generated by a radiologist for a validation data set from the BRAX [2] data set.

## 7.2 Future Work

For initial screening, automated CXR analysis for disease classification and report generation can be a useful tool for reducing the workload on radiologists. However, the performance of such systems can vary depending on how and what data such systems are trained on. There remain several unanswered questions and obstacles that need to be overcome.

Segmentation of the lungs plays a very important role in the automated analysis of lung diseases as it can lead to better classification, particularly for complex cases where the lungs are affected by multiple diseases. Therefore, incorporating a method to automatically segment the lungs can improve the performance of any intelligent framework. However, such a method should be robust to variations in the CXRs that can result from differences in patient anatomy or acquisition methods. Moreover, segmenting additional thoracic cavity anatomical features can enhance the process of deriving insights from the CXR that extend beyond lung disorders. Reliable segmentation that works equally well for image and anatomical variation can be achieved by incorporating weakly supervised

learning mechanisms or attention-based mechanisms similar to transformers.

While few chronic pulmonary diseases tend to have presentations that can lend themselves to assigning a single-digit score that not only relays the information about the severity but can also be used to track the progression, such a score for chronic diseases can provide insightful information. Thus, any intelligent framework's performance can be enhanced by adding a way to provide a severity score as it provides a more precise risk stratification and treatment guidance for patients. Furthermore, such a severity score can also improve the factual quality of a generated radiology report. However, the consistency and accuracy of such a scoring system is crucial. Such a system would also need to be able to adapt to different CXR views such as lateral as opposed to just frontal view. This approach can be enhanced by developing disease-specific severity scoring and leveraging existing knowledge for transfer learning for related diseases. Explainable AI can also be used to provide reasoning for the assigned severity scores.

As discussed earlier, CNNs have been used effectively for various tasks such as pulmonary disease classification, segmentation, opacity localisation, and even report generation. Transformers-based architectures have also been used for similar tasks and both of these approaches offer different advantages and disadvantages. The use of both in a single framework results in a solution where the sub-modules are well-suited for a particular sub-problem. This amalgamation of these methodologies can be improved to improve the performance of intelligent frameworks such as the one proposed in this research work by exploration of multi-stage approaches where CNNs are used as feature-extractors and transformers are used for higher-level reasoning for classification and report generation.

With an emphasis on the lungs and their many diseases, radiology reports are utilized to present a holistic view of the organs in the chest cavity. This information can also be effectively provided via automatically generated reports alongside image-level classification and other metrics from a single framework. However, these reports can contain inaccuracies regarding information that can not be directly retrieved from just an image

such as the length of various chest tubings. This propensity for errors can be reduced by the incorporation of associated information at the time of training in the form of specialised embedding tokens for transformer-based architecture or by incorporating another CNN-based sub-module that is adept at the translation of such values.

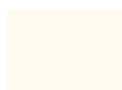
Large-scale CXR datasets have made it possible for deep learning architectures to create reliable automated solutions for obtaining various insights from the CXR. Nevertheless, these datasets are typically unbalanced, which can impair any artificial intelligence framework's performance. In addition to this, variations in radiology report formats across hospitals can impact such a system's capacity for effective generalisation. In order to have an unbiased framework, it is necessary to employ different class-balancing techniques and text-preprocessing techniques as has been done in this research work. Improvement in such techniques will also improve the performance of the complete framework.

In light of this, the following points, summarised below, could be addressed to further enhance the functionality and usability of this research work:

1. The segmentation of other anatomical structures e.g. heart among others present in CXR can be incorporated for better classification and report generation.
2. Severity scoring can be expanded to other pulmonary diseases as well that can be graded over a severity scale.
3. Recent CNNs architectures can be used as feature extractors for transformers-based frameworks for classification and report generation instead of relying just on inherent linear projections in the transformer architecture.
4. Associated knowledge from reports such as centimeters-to-pixels can be incorporated in the Encoder/Decoder architecture in Transformers for better classification and report generation.
5. Data imbalance with respect to clinical findings such as one class having a lot more samples is prevalent in CXR datasets and reports which needs to be addressed.

# Bibliography

- [1] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.
- [2] Eduardo P Reis, Joselisa PQ de Paiva, Maria CB da Silva, Guilherme AS Ribeiro, Victor F Paiva, Lucas Bulgarelli, Henrique MH Lee, Paulo V Santos, Vanessa M Brito, Lucas TW Amaral, et al. Brax, brazilian labeled chest x-ray dataset. *Scientific Data*, 9(1):1–8, 2022.
- [3] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodaera, and Kunio Doi. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology*, 174(1):71–74, 2000.
- [4] Stefan Jaeger, Sema Candemir, Sameer Antani, Yi-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475, 2014.





- [5] Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Les Folio, Jenifer Siegelman, Fiona Callaghan, Zhiyun Xue, Kannappan Palaniappan, Rahul K Singh, Sameer Antani, et al. Automatic tuberculosis screening using chest radiographs. *IEEE transactions on medical imaging*, 33(2):233–245, 2013.
- [6] Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P Musco, Rahul K Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE transactions on medical imaging*, 33(2):577–590, 2013.
- [7] Sergii Stirenko, Yuriy Kochura, Oleg Alienin, Oleksandr Rokovyi, Yuri Gordienko, Peng Gang, and Wei Zeng. Chest x-ray analysis of tuberculosis by deep learning with segmentation and augmentation. In *2018 IEEE 38th International Conference on Electronics and Nanotechnology (ELNANO)*, pages 422–428. IEEE, 2018.
- [8] SIIM-FISABIO-RSNA. SIIM-FISABIO-RSNA COVID-19 Detection. <https://www.kaggle.com/c/siim-covid19-detection>, 2021. Accessed: 26-Aug-2022.
- [9] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.
- [10] A Johnson, M Lungren, Y Peng, Z Lu, R Mark, S Berkowitz, and S Horng. Mimic-cxr-jpg-chest radiographs with structured labels (version 2.0. 0). *PhysioNet*, 2019.
- [11] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

- [12] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.
- [13] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020.
- [14] Vignav Ramesh, Nathan A Chi, and Pranav Rajpurkar. Improving radiology report generation systems by removing hallucinated references to non-existent priors. In *Machine Learning for Health*, pages 456–473. PMLR, 2022.
- [15] Xuting Jin, Jiajia Ren, Ruohan Li, Ya Gao, Haoying Zhang, Jiamei Li, Jingjing Zhang, Xiaochuang Wang, and Gang Wang. Global burden of upper respiratory infections in 204 countries and territories, from 1990 to 2019. *EClinicalMedicine*, 37:100986, 2021.
- [16] Wassim W Labaki and MeiLan K Han. Chronic respiratory diseases: a global view. *The Lancet Respiratory Medicine*, 8(6):531–533, 2020.
- [17] Alarcos Cieza, Kate Causey, Kaloyan Kamenov, Sarah Wulf Hanson, Somnath Chatterji, and Theo Vos. Global estimates of the need for rehabilitation based on the global burden of disease study 2019: a systematic analysis for the global burden of disease study 2019. *The Lancet*, 396(10267):2006–2017, 2020.
- [18] Patrik Danielsson, Inga Sif Ólafsdóttir, Bryndis Benediktsdóttir, Thórarinn Gíslason, and Christer Janson. The prevalence of chronic obstructive pulmonary disease in uppsala, sweden—the burden of obstructive lung disease (bold) study: cross-sectional population-based study. *The clinical respiratory journal*, 6(2):120–127, 2012.

- [19] Filip Mejza, Louisa Gnatiuc, A Sonia Buist, William M Vollmer, Bernd Lamprecht, Daniel O Obaseki, Pawel Nastalek, Ewa Nizankowska-Mogilnicka, and Peter GJ Burney. Prevalence and burden of chronic bronchitis symptoms: results from the bold study. *European Respiratory Journal*, 50(5), 2017.
- [20] RJ Halbert, JL Natoli, Anacleto Gano, E Badamgarav, A Sonia Buist, and DM Mannino. Global burden of copd: systematic review and meta-analysis. *European Respiratory Journal*, 28(3):523–532, 2006.
- [21] Chronic obstructive pulmonary disease (copd). [https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-\(copd\)](https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd)). Accessed: 03-Jan-2023.
- [22] Abid Haleem, Mohd Javaid, and Raju Vaishya. Effects of covid-19 pandemic in daily life. *Current medicine research and practice*, 10(2):78, 2020.
- [23] Godfrey N Hounsfield. Computerized transverse axial scanning (tomography): Part 1. description of system. *The British journal of radiology*, 46(552):1016–1022, 1973.
- [24] Donald W McRobbie, Elizabeth A Moore, Martin J Graves, and Martin R Prince. *MRI from Picture to Proton*. Cambridge university press, 2017.
- [25] Wafaa Abd-Elsalam, Shaimaa Waheed Zahra, Marwa Ahmed Abogabal, and Wafaa Madhy Atia. Accuracy of point of care lung ultrasound in diagnosis of ventilator associated pneumonia in intensive care unit. *Sohag Medical Journal*, 26(2):101–107, 2022.
- [26] Dale L Bailey, Michael N Maisey, David W Townsend, and Peter E Valk. *Positron emission tomography*, volume 2. Springer, 2005.
- [27] Jamilah Meghji, Kevin Mortimer, Alvar Agusti, Brian W Allwood, Innes Asher, Eric D Bateman, Karen Bissell, Charlotte E Bolton, Andrew Bush, Bartolome

- Celli, et al. Improving lung health in low-income and middle-income countries: from challenges to solutions. *The Lancet*, 397(10277):928–940, 2021.
- [28] Emerson Augusto Baptista, Sudeshna Dey, and Soumya Pal. Chronic respiratory disease mortality and its associated factors in selected asian countries: evidence from panel error correction model. *BMC Public Health*, 21:1–11, 2021.
- [29] Joan B Soriano, Parkes J Kendrick, Katherine R Paulson, Vinay Gupta, Elissa M Abrams, Rufus Adesoji Adedoyin, Tara Ballav Adhikari, Shailesh M Advani, Anurag Agrawal, Elham Ahmadian, et al. Prevalence and attributable health burden of chronic respiratory diseases, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet Respiratory Medicine*, 8(6):585–596, 2020.
- [30] Vizhub-gbd results. <https://vizhub.healthdata.org/gbd-results/>. Accessed: 03-Jan-2023.
- [31] World life expectancy. <https://www.worldlifeexpectancy.com/country-health-profile/pakistan,2023>. Accessed: 20-Feb-2023.
- [32] Murtaza Gowa, Irfan Habib, Amber Tahir, Uzair Yaqoob, and Sadaf Junejo. Disease spectrum and frequency of illness in pediatric emergency: a retrospective analysis from karachi, pakistan. *Ochsner Journal*, 19(4):340–346, 2019.
- [33] Ioannis D Apostolopoulos and Tzani A Mpesiana. Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. *Physical and engineering sciences in medicine*, 43:635–640, 2020.
- [34] Worldometers. <https://www.worldometers.info/coronavirus/>, 2023. Accessed: 14-Feb-2023.
- [35] World health statistics 2010. <http://www.who.int/whosis/whostat/2010/en/index.html,2010>. Accessed: 20-Feb-2023.

- [36] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.
- [37] Asad Khan, Muhammad Usman Akram, and Sajid Nazir. Automated grading of chest x-ray images for viral pneumonia with convolutional neural networks ensemble and region of interest localization. *PLoS One*, 18(1):e0280352, 2023.
- [38] Asad Mansoor Khan, Muhammad Usman Akram, Sajid Nazir, Taimur Hassan, Sajid Gul Khawaja, and Tatheer Fatima. Multi-head deep learning framework for pulmonary disease detection and severity scoring with modified progressive learning. *Biomedical Signal Processing and Control*, 85:104855, 2023.
- [39] Thoracic cavity. <https://www.britannica.com/science/thoracic-cavity#/media/1/593184/99769>, 2020. Accessed: 07-Sep-2022.
- [40] Lung. <https://www.britannica.com/science/lung>, 2020. Accessed: 07-Sep-2022.
- [41] Gas exchange in the lung. <https://www.britannica.com/science/lung#/media/1/351473/107200>, 2020. Accessed: 07-Sep-2022.
- [42] Ruicheng Hu, Qing Ouyang, Aiguo Dai, Shuangxiang Tan, Zhiqiang Xiao, and Cene Tang. Heat shock protein 27 and cyclophilin a associate with the pathogenesis of copd. *Respirology*, 16(6):983–993, 2011.
- [43] Luciano Gattinoni, Davide Chiumello, Pietro Caironi, Mattia Busana, Federica Romitti, Luca Brazzi, and Luigi Camporota. Covid-19 pneumonia: different respiratory treatments for different phenotypes? *Intensive care medicine*, 46:1099–1102, 2020.

- [44] William D Schweickert, Mark C Pohlman, Anne S Pohlman, Celerina Nigos, Amy J Pawlik, Cheryl L Esbrook, Linda Spears, Megan Miller, Mietka Franczyk, Deanna Deprizio, et al. Early physical and occupational therapy in mechanically ventilated, critically ill patients: a randomised controlled trial. *The Lancet*, 373(9678):1874–1882, 2009.
- [45] Marcelo BP Amato, Maureen O Meade, Arthur S Slutsky, Laurent Brochard, Eduardo LV Costa, David A Schoenfeld, Thomas E Stewart, Matthias Briel, Daniel Talmor, Alain Mercat, et al. Driving pressure and survival in the acute respiratory distress syndrome. *New England Journal of Medicine*, 372(8):747–755, 2015.
- [46] Timothy Baird, Caroline L Cooper, Richard Wong, Naomi Runnegar, and Gregory Keir. Pulmonary schistosomiasis mimicking igg4-related lung disease. *Respirology Case Reports*, 6(1):e00276, 2018.
- [47] Veysel Garani SOYLU, Öztürk TAŞKIN, Ufuk DEMİR, and Yunus YAŞAR. The importance of thoracic tomography in prognosis in critical covid 19 patients. *Journal of Contemporary Medicine*, 11(3):317–322.
- [48] Feng Pan, Tianhe Ye, Peng Sun, Shan Gui, Bo Liang, Lingli Li, Dandan Zheng, Jiazheng Wang, Richard L Hesketh, Lian Yang, et al. Time course of lung changes at chest ct during recovery from coronavirus disease 2019 (covid-19). *Radiology*, 295(3):715–721, 2020.
- [49] Lionel Piroth, Jonathan Cottenet, Anne-Sophie Mariet, Philippe Bonniaud, Mathieu Blot, Pascale Tubert-Bitter, and Catherine Quantin. Comparison of the characteristics, morbidity, and mortality of covid-19 and seasonal influenza: a nationwide, population-based retrospective cohort study. *The Lancet Respiratory Medicine*, 9(3):251–259, 2021.
- [50] Gustavo Matute-Bello, Charles W Frevert, and Thomas R Martin. Animal models of acute lung injury. *American Journal of Physiology-Lung Cellular and Molecular*

*Physiology*, 295(3):L379–L399, 2008.

- [51] Wenbo Zheng, Lan Yan, Chao Gou, Zhi-Cheng Zhang, Jun J Zhang, Ming Hu, and Fei-Yue Wang. Learning to learn by yourself: Unsupervised meta-learning with self-knowledge distillation for covid-19 diagnosis from pneumonia cases. *International Journal of Intelligent Systems*, 36(8):4033–4064, 2021.
- [52] CP Sharma, D Behera, AN Aggarwal, D Gupta, and SK Jindal. Radiographic patterns in lung cancer. *Indian Journal of Chest Diseases and Allied Sciences*, 44(1):25–30, 2002.
- [53] Arnab Saha, Kaushik Saha, Santanu Ghosh, Mrinmoy Mitra, Prabodh Panchadhayee, Aditya P Sarkar, et al. Chest x-ray of lung cancer: Association with pathological subtypes. *The Journal of Association of Chest Physicians*, 5(2):76, 2017.
- [54] Giovanni Volpicelli, Mahmoud Elbarbary, Michael Blaivas, Daniel A Lichtenstein, Gebhard Mathis, Andrew W Kirkpatrick, Lawrence Melniker, Luna Gargani, Vicki E Noble, Gabriele Via, et al. International evidence-based recommendations for point-of-care lung ultrasound. *Intensive care medicine*, 38:577–591, 2012.
- [55] I Simkova and J Urbanova. Pulmonary functions alterations after correction of mitral stenosis. *Bratislavske lekarske listy*, 102(6):278–281, 2001.
- [56] Adam Bernheim, Xueyan Mei, Mingqian Huang, Yang Yang, Zahi A Fayad, Ning Zhang, Kaiyue Diao, Bin Lin, Xiqi Zhu, Kunwei Li, et al. Chest ct findings in coronavirus disease-19 (covid-19): relationship to duration of infection. *Radiology*, 295(3):685–691, 2020.
- [57] Xi Xu, Chengcheng Yu, Jing Qu, Lieguang Zhang, Songfeng Jiang, Deyang Huang, Bihua Chen, Zhiping Zhang, Wanhua Guan, Zhoukun Ling, et al. Imaging and clinical features of patients with 2019 novel coronavirus sars-cov-2. *European journal of nuclear medicine and molecular imaging*, 47:1275–1280, 2020.

- [58] Jose M Porcel and Richard W Light. Diagnostic approach to pleural effusion in adults. *American family physician*, 73(7):1211–1220, 2006.
- [59] Susan Mertin, Jo-Ann V Sawatzky, and William L Diehl-Jones. Getting to the heart of pleural effusions: A case study. *Journal of the American Academy of Nurse Practitioners*, 21(9):506–512, 2009.
- [60] Essa M AlGhunaim, Hafsa AlNajem, and Derar S AlShehab. Spontaneous bilateral hemothorax as a case of epithelioid hemangioendothelioma (ehe). *Case Reports in Oncological Medicine*, 2019, 2019.
- [61] Quanlei Bao, Yaping Xu, Ming Ding, and Ping Chen. Identification of differentially expressed mirnas in differentiating benign from malignant pleural effusion. *Hereditas*, 157(1):1–8, 2020.
- [62] Zheng Ye, Yun Zhang, Yi Wang, Zixiang Huang, and Bin Song. Chest ct manifestations of new coronavirus disease 2019 (covid-19): a pictorial review. *European radiology*, 30:4381–4389, 2020.
- [63] Naoyuki Miyashita, Hiroto Akaike, Hideto Teranishi, Kazunobu Ouchi, and Niro Okimoto. Macrolide-resistant mycoplasma pneumoniae pneumonia in adolescents and adults: clinical findings, drug susceptibility, and therapeutic efficacy. *Antimicrobial agents and chemotherapy*, 57(10):5181–5185, 2013.
- [64] Catherine A Gordon, Johanna Kurscheid, Malcolm K Jones, Darren J Gray, and Donald P McManus. Soil-transmitted helminths in tropical australia and asia. *Tropical medicine and infectious disease*, 2(4):56, 2017.
- [65] Ahmad Khalid Omeri, Fumito Okada, Shoko Takata, Asami Ono, Tomoko Nakayama, Yumiko Ando, Haruka Sato, Kazufumi Hiramatsu, and Hiromu Mori. Comparison of high-resolution computed tomography findings between pseudomonas aeruginosa pneumonia and cytomegalovirus pneumonia. *European radiology*, 24:3251–3259, 2014.



- [66] A Ono, F Okada, S Takata, K Hiramatsu, Y Ando, T Nakayama, T Maeda, and H Mori. A comparative study of thin-section ct findings between seasonal influenza virus pneumonia and streptococcus pneumoniae pneumonia. *The British Journal of Radiology*, 87(1039):20140051, 2014.
- [67] Pneumonia causes and risk factors. <https://www.nhlbi.nih.gov/health/pneumonia/causes>. Accessed: 03-Mar-2023.
- [68] D Lichtenstein and Gilbert Mezière. A lung ultrasound sign allowing bedside distinction between pulmonary edema and copd: the comet-tail artifact. *Intensive care medicine*, 24:1331–1334, 1998.
- [69] Na Zhu, Dingyu Zhang, Wenling Wang, Xingwang Li, Bo Yang, Jingdong Song, Xiang Zhao, Baoying Huang, Weifeng Shi, Roujian Lu, et al. A novel coronavirus from patients with pneumonia in china, 2019. *New England journal of medicine*, 2020.
- [70] Farzaneh Rostamzadeh, Hamid Najafipour, Rostam Yazdani, Samira Nakhaei, and Ahmad Alinaghi Langari. Changes in serum levels of apelin and nitric oxide in hospitalized patients with covid-19: association with hypertension, diabetes, obesity, and severity of disease. *European Journal of Medical Research*, 27(1):1–8, 2022.
- [71] Kang Zhao, Jucun Huang, Dan Dai, Yuwei Feng, Liming Liu, and Shuke Nie. Serum iron level as a potential predictor of coronavirus disease 2019 severity and mortality: a retrospective study. In *Open forum infectious diseases*, volume 7, page ofaa250. Oxford University Press US, 2020.
- [72] Ani Nalbandian, Kartik Sehgal, Aakriti Gupta, Mahesh V Madhavan, Claire McGroder, Jacob S Stevens, Joshua R Cook, Anna S Nordvig, Daniel Shalev, Tejasav S Sehrawat, et al. Post-acute covid-19 syndrome. *Nature medicine*, 27(4):601–615, 2021.

- [73] Severe Covid-19 GWAS Group. Genomewide association study of severe covid-19 with respiratory failure. *New England Journal of Medicine*, 383(16):1522–1534, 2020.
- [74] Jerrold T Bushberg and John M Boone. *The essential physics of medical imaging*. Lippincott Williams & Wilkins, 2011.
- [75] Mohammad Rahimzadeh, Abolfazl Attar, and Seyed Mohammad Sakhaei. A fully automated deep learning-based network for detecting covid-19 from a new and large lung ct scan dataset. *medRxiv*, 2020.
- [76] Peter G Hagan, Christoph A Nienaber, Eric M Isselbacher, David Bruckman, Dean J Karavite, Pamela L Russman, Arturo Evangelista, Rossella Fattori, Toru Suzuki, Jae K Oh, et al. The international registry of acute aortic dissection (irad): new insights into an old disease. *Jama*, 283(7):897–903, 2000.
- [77] Alexander Ziegler, Martin Kunth, Susanne Mueller, Christian Bock, Rolf Pohmann, Leif Schröder, Cornelius Faber, and Gonzalo Giribet. Application of magnetic resonance imaging in zoology. *Zoomorphology*, 130:227–254, 2011.
- [78] Toemme Noesselt, Steve A Hillyard, Marty G Woldorff, Ariel Schoenfeld, Tilman Hagner, Lutz Jäncke, Claus Tempelmann, Hermann Hinrichs, and Hans-Jochen Heinze. Delayed striate cortical activation during spatial attention. *Neuron*, 35(3):575–587, 2002.
- [79] Adrienne E Campbell-Washburn, Anthony F Suffredini, and Marcus Y Chen. High-performance 0.55-t lung mri in patient with covid-19 infection. *Radiology*, 299(2):E246–E247, 2021.
- [80] Ashish Saraogi. Lung ultrasound: Present and future. *Lung India: Official Organ of Indian Chest Society*, 32(3):250, 2015.
- [81] Nuclear medicine. <http://hyperphysics.phy-astr.gsu.edu/hbase/NucEne/nucmed.html>. Accessed: 16-Mar-2023.

- [82] BJ Casey, Jay N Giedd, and Kathleen M Thomas. Structural and functional brain development and its relation to cognitive development. *Biological psychology*, 54(1-3):241–257, 2000.
- [83] Kazuhiro Kitajima, Hiroshi Doi, Tomonori Kanda, Tomohiko Yamane, Tetsuya Tsujikawa, Hayato Kaida, Yukihiisa Tamaki, and Kozo Kuribayashi. Present and future roles of fdg-pet/ct imaging in the management of lung cancer. *Japanese Journal of Radiology*, 34:387–399, 2016.
- [84] Anas M Tahir, Muhammad EH Chowdhury, Amith Khandakar, Tawsifur Rahman, Yazan Qiblawey, Uzair Khurshid, Serkan Kiranyaz, Nabil Ibte haz, M Sohel Rahman, Somaya Al-Maadeed, et al. Covid-19 infection localization and severity grading from chest x-ray images. *Computers in biology and medicine*, 139:105002, 2021.
- [85] Sayan Manna, Jill Wruble, Samuel Z Maron, Danielle Toussie, Nicholas Voutsinas, Mark Finkelstein, Mario A Cedillo, Jamie Diamond, Corey Eber, Adam Jacobi, et al. Covid-19: a multimodality review of radiologic techniques, clinical utility, and imaging features. *Radiology: Cardiothoracic Imaging*, 2(3), 2020.
- [86] Radiology chest x-ray normal. <https://www.ebmconsult.com/articles/radiology-chest-xray-normal>. Accessed: 19-Mar-2023.
- [87] Mehmet Ali GEDİK and Ayşe KEVEN. *Radyoloji Başucu Serisi Pediatrik Radyoloji*. Akademisyen Kitabevi, 2020.
- [88] The radiology assistant chest x-ray lung disease. <https://radiologyassistant.nl/chest/chest-x-ray/lung-disease>. Accessed: 19-Mar-2023.
- [89] Daniel Lichtenstein, Ivan Goldstein, Eric Mourgeon, Philippe Cluzel, Philippe Grenier, and Jean-Jacques Rouby. Comparative diagnostic performances of auscultation, chest radiography, and lung ultrasonography in acute respiratory distress syn-

- drome. *The Journal of the American Society of Anesthesiologists*, 100(1):9–15, 2004.
- [90] C Craig Blackmore, William C Black, Robert V Dallas, and Harte C Crow. Pleural fluid volume estimation: a chest radiograph prediction rule. *Academic radiology*, 3(2):103–109, 1996.
- [91] Jonathan Corne and Iain Au-Yong. *Chest X-ray made easy E-book*. Elsevier Health Sciences, 2022.
- [92] Robert A Novelline and Lucy Frank Squire. *Squire’s fundamentals of radiology*. La Editorial, UPR, 2004.
- [93] Hector Rodriguez, Tina V Hartert, Tebeb Gebretsadik, Kecia N Carroll, and Emma K Larkin. A simple respiratory severity score that may be used in evaluation of acute respiratory infection. *BMC research notes*, 9(1):1–4, 2016.
- [94] Jingyao Liu, Wanchun Sun, Xuehua Zhao, Jiashi Zhao, and Zhengang Jiang. Deep feature fusion classification network (dffcnnet): Towards accurate diagnosis of covid-19 using chest x-rays images. *Biomedical Signal Processing and Control*, 76:103677, 2022.
- [95] Alberto Signoroni, Mattia Savardi, Sergio Benini, Nicola Adami, Riccardo Leonardi, Paolo Gibellini, Filippo Vaccher, Marco Ravanelli, Andrea Borghesi, Roberto Maroldi, et al. Bs-net: Learning covid-19 pneumonia severity on a large chest x-ray dataset. *Medical Image Analysis*, 71:102046, 2021.
- [96] Sangjoon Park, Gwanhyun Kim, Yujin Oh, Joon Beom Seo, Sang Min Lee, Jin Hwan Kim, Sungjun Moon, Jae-Kwang Lim, and Jong Chul Ye. Multi-task vision transformer using low-level chest x-ray feature corpus for covid-19 diagnosis and severity quantification. *Medical Image Analysis*, 75:102299, 2022.
- [97] Jocelyn Zhu, Beiyi Shen, Almas Abbasi, Mahsa Hoshmand-Kochi, Haifang Li, and Tim Q Duong. Deep transfer learning artificial intelligence accurately stages covid-

- 19 lung disease severity on portable chest radiographs. *PloS one*, 15(7):e0236621, 2020.
- [98] Joseph Paul Cohen, Lan Dao, Karsten Roth, Paul Morrison, Yoshua Bengio, Almas F Abbasi, Beiyi Shen, Hoshmand Kochi Mahsa, Marzyeh Ghassemi, Haifang Li, et al. Predicting covid-19 pneumonia severity on chest x-ray with deep learning. *Cureus*, 12(7), 2020.
- [99] Ho Yuen Frank Wong, Hiu Yin Sonia Lam, Ambrose Ho-Tung Fong, Siu Ting Leung, Thomas Wing-Yan Chin, Christine Shing Yen Lo, Macy Mei-Sze Lui, Jonan Chun Yin Lee, Keith Wan-Hang Chiu, Tom Wai-Hin Chung, et al. Frequency and distribution of chest radiographic findings in patients positive for covid-19. *Radiology*, 296(2):E72–E78, 2020.
- [100] Emrah Irmak. Covid-19 disease severity assessment using cnn model. *IET image processing*, 15(8):1814–1824, 2021.
- [101] Andrea Borghesi and Roberto Maroldi. Covid-19 outbreak in italy: experimental chest x-ray scoring system for quantifying and monitoring disease progression. *La radiologia medica*, 125(5):509–513, 2020.
- [102] Medicalmnemonics. [http://www.medicalmnemonics.com/pdf/2002\\_06\\_pack\\_unabr\\_a4.pdf](http://www.medicalmnemonics.com/pdf/2002_06_pack_unabr_a4.pdf). Accessed: 19-Mar-2023.
- [103] Chest x-ray abnormalities - hilar abnormalities. [https://www.radiologymasterclass.co.uk/tutorials/chest/chest\\_pathology/chest\\_pathology\\_page2](https://www.radiologymasterclass.co.uk/tutorials/chest/chest_pathology/chest_pathology_page2). Accessed: 19-Mar-2023.
- [104] Yoshiko Kaneko, Norihiro Kikuchi, Yukio Ishii, Yoshinori Kawabata, Hiroshi Moriyama, Masaki Terada, Eiichi Suzuki, Masayoshi Kobayashi, Kouichi Watanabe, and Nobuyuki Hizawa. Upper lobe-dominant pulmonary fibrosis showing deposits of hard metal component in the fibrotic lesions. *Internal Medicine*, 49(19):2143–2145, 2010.

- [105] Ruwani Abeyratne, Peter Wills, and Simon William Dubrey. Asbestos-related pleural plaques: significance and associations. *Case Reports*, 2013:bcr2013008928, 2013.
- [106] Yongwon Cho, Young-Gon Kim, Sang Min Lee, Joon Beom Seo, and Namkug Kim. Reproducibility of abnormality detection on chest radiographs using convolutional neural network in paired radiographs obtained within a short-term interval. *Scientific Reports*, 10(1):17417, 2020.
- [107] Ali Narin. Accurate detection of covid-19 using deep features based on x-ray images and feature selection methods. *Computers in Biology and Medicine*, 137:104771, 2021.
- [108] Tim Dall, Ryan Reynolds, Kari Jones, Ritashree Chakrabarti, and W Iacobucci. The complexities of physician supply and demand: projections from 2017 to 2032. *Association of American Medical Colleges*, page 86, 2019.
- [109] Suhail Raoof, David Feigin, Arthur Sung, Sabiha Raoof, Lavanya Irugulpati, and Edward C Rosenow III. Interpretation of plain chest roentgenogram. *Chest*, 141(2):545–558, 2012.
- [110] Ali Mohammad Alqudah, Shoroq Qazan, and Amin Alqudah. Automated systems for detection of covid-19 using chest x-ray images and lightweight convolutional neural networks. 2020.
- [111] Elena Velichko, Faridoddin Shariaty, Mahdi Orooji, Vitalii Pavlov, Tatiana Pervunina, Sergey Zavjalov, Raziieh Khazaei, and Amir Reza Radmard. Development of computer-aided model to differentiate covid-19 from pulmonary edema in lung ct scan: Edecovid-net. *Computers in Biology and Medicine*, 141:105172, 2022.
- [112] Lingma Sun, Zhuoran Wang, Hong Pu, Guohui Yuan, Lu Guo, Tian Pu, and Zhenming Peng. Attention-embedded complementary-stream cnn for false positive

- reduction in pulmonary nodule detection. *Computers in Biology and Medicine*, 133:104357, 2021.
- [113] Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. Deep learning applications for covid-19. *Journal of big Data*, 8(1):1–54, 2021.
- [114] A Khan, A Sohail, U Zahoor, and AS Qureshi. A survey of the recent architectures of deep convolutional neural networks. arxiv 2019. *arXiv preprint arXiv:1901.06032*, 2019.
- [115] Wenwu Ye, Jin Yao, Hui Xue, and Yi Li. Weakly supervised lesion localization with probabilistic-cam pooling. *arXiv preprint arXiv:2005.14480*, 2020.
- [116] Alaa S Al-Waisy, Shumoos Al-Fahdawi, Mazin Abed Mohammed, Karrar Hameed Abdulkareem, Salama A Mostafa, Mashaal S Maashi, Muhammad Arif, and Begonya Garcia-Zapirain. Covid-chexnet: hybrid deep learning framework for identifying covid-19 virus in chest x-rays images. *Soft computing*, pages 1–16, 2020.
- [117] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9:611–629, 2018.
- [118] Rachna Jain, Preeti Nagrath, Gaurav Kataria, V Sirish Kaushik, and D Jude Hemant. Pneumonia detection in chest x-ray images using convolutional neural networks and transfer learning. *Measurement*, 165:108046, 2020.
- [119] Yongwon Cho, Beomhee Park, Sang Min Lee, Kyung Hee Lee, Joon Beom Seo, and Namkug Kim. Optimal number of strong labels for curriculum learning with convolutional neural network to classify pulmonary abnormalities in chest radiographs. *Computers in Biology and Medicine*, 136:104750, 2021.
- [120] Qingji Guan, Yaping Huang, Zhun Zhong, Zhedong Zheng, Liang Zheng, and Yi Yang. Thorax disease classification with attention guided convolutional neural network. *Pattern Recognition Letters*, 131:38–45, 2020.

- [121] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [122] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [123] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald 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*, pages 2097–2106, 2017.
- [124] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9049–9058, 2018.
- [125] Morteza Heidari, Seyedehnafiseh Mirniaharikandehei, Abolfazl Zargari Khuzani, Gopichandh Danala, Yuchen Qiu, and Bin Zheng. Improving the performance of cnn to predict the likelihood of covid-19 using chest x-ray images with preprocessing algorithms. *International journal of medical informatics*, 144:104284, 2020.
- [126] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018.
- [127] Daniel Kermany, Kang Zhang, Michael Goldbaum, et al. Labeled optical coherence tomography (oct) and chest x-ray images for classification. *Mendeley data*, 2(2):651, 2018.



- [128] Muhammad EH Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, et al. Can ai help in screening viral and covid-19 pneumonia? *IEEE Access*, 8:132665–132676, 2020.
- [129] Nanshan Chen, Min Zhou, Xuan Dong, Jieming Qu, Fengyun Gong, Yang Han, Yang Qiu, Jingli Wang, Ying Liu, Yuan Wei, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study. *The lancet*, 395(10223):507–513, 2020.
- [130] Cohen Joseph Paul, Morrison Paul, Dao Lan, et al. Covid-19 image data collection. *arXiv preprint arXiv:2003.11597*, 2020.
- [131] Maram Mahmoud A Monshi, Josiah Poon, Vera Chung, and Fahad Mahmoud Monshi. Covidxraynet: Optimizing data augmentation and cnn hyperparameters for improved covid-19 detection from cxr. *Computers in biology and medicine*, 133:104375, 2021.
- [132] Lin Wang et al. Wang l., lin zq, wong a. *Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images*, *Scientific Reports*, 10(1):1–12, 2020.
- [133] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning*, pages 10096–10106. PMLR, 2021.
- [134] Unais Sait, KG Lal, S Prajapati, Rahul Bhaumik, Tarun Kumar, S Sanjana, and Kriti Bhalla. Curated dataset for covid-19 posterior-anterior chest radiography images (x-rays). *Mendeley Data*, 1, 2020.
- [135] Prashant Bhardwaj and Amanpreet Kaur. A novel and efficient deep learning approach for covid-19 detection using x-ray imaging modality. *International Journal of Imaging Systems and Technology*, 31(4):1775–1791, 2021.

- [136] Marwa Ben Jabra, Anis Koubaa, Bilel Benjdira, Adel Ammar, and Habib Hamam. Covid-19 diagnosis in chest x-rays using deep learning and majority voting. *Applied Sciences*, 11(6):2884, 2021.
- [137] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [138] Maria De La Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García-García, et al. Bimev covid-19+: a large annotated dataset of rx and ct images from covid-19 patients. *arXiv preprint arXiv:2006.01174*, 2020.
- [139] Gabriel Iluebe Okolo, Stamos Katsigiannis, and Naeem Ramzan. Ievit: An enhanced vision transformer architecture for chest x-ray image classification. *Computer Methods and Programs in Biomedicine*, 226:107141, 2022.
- [140] Tawsifur Rahman, Amith Khandakar, Muhammad Abdul Kadir, Khandaker Rejaul Islam, Khandakar F Islam, Rashid Mazhar, Tahir Hamid, Mohammad Tariqul Islam, Saad Kashem, Zaid Bin Mahbub, et al. Reliable tuberculosis detection using chest x-ray with deep learning, segmentation and visualization. *IEEE Access*, 8:191586–191601, 2020.
- [141] Bekir Aksoy and Osamah Khaled Musleh Salman. Detection of covid-19 disease in chest x-ray images with capsul networks: application with cloud computing. *Journal of Experimental & Theoretical Artificial Intelligence*, 33(3):527–541, 2021.
- [142] P Mooney. Data of chest x-ray kaggle. *Tersedia melalui: Kaggle*, 2019.
- [143] Chung A Actualmed COVID. chest x-ray data initiative; 2020. URL <https://github.com/agchung/Actualmed-COVID-chestxray-dataset> (19), 2019.

- [144] Khalid El Asnaoui and Youness Chawki. Using x-ray images and deep learning for automated detection of coronavirus disease. *Journal of Biomolecular Structure and Dynamics*, 39(10):3615–3626, 2021.
- [145] Isoon Kanjanasurat, Kasi Tenghongsakul, Boonchana Purahong, and Attasit Lasakul. Cnn–rnn network integration for the diagnosis of covid-19 using chest x-ray and ct images. *Sensors*, 23(3):1356, 2023.
- [146] Mei-Ling Huang and Yu-Chieh Liao. A lightweight cnn-based network on covid-19 detection using x-ray and ct images. *Computers in Biology and Medicine*, page 105604, 2022.
- [147] Maede Maftouni. Large covid-19 ct scan slice dataset. <https://www.kaggle.com/dsv/2321803>, 2021.
- [148] Paul Mooney. Chest x-ray images (pneumonia). <https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia>, 2018.
- [149] Amith Khandakar Tawsifur Rahman, Muhammad Chowdhury. Covid-19 radiography database, 2018.
- [150] Dina M Ibrahim, Nada M Elshennawy, and Amany M Sarhan. Deep-chest: Multi-classification deep learning model for diagnosing covid-19, pneumonia, and lung cancer chest diseases. *Computers in biology and medicine*, 132:104348, 2021.
- [151] Abhir Bhandary, G Ananth Prabhu, Venkatesan Rajinikanth, K Palani Thanaraj, Suresh Chandra Satapathy, David E Robbins, Charles Shasky, Yu-Dong Zhang, João Manuel RS Tavares, and N Sri Madhava Raja. Deep-learning framework to detect lung abnormality—a study with chest x-ray and lung ct scan images. *Pattern Recognition Letters*, 129:271–278, 2020.
- [152] Cohen Joseph Paul, P Morrison, Lan Dao, K Roth, Tim Q Duong, and M Ghassemi. Covid-19 image data collection: prospective predictions are the future. *arXiv preprint arXiv:2006.11988*, 2020.

- [153] Adnane Ait Nasser and Moulay A Akhloufi. Deep learning methods for chest disease detection using radiography images. *SN Computer Science*, 4(4):388, 2023.
- [154] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
- [155] Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist’s annotations. *Scientific Data*, 9(1):429, 2022.
- [156] Yogesh H Bhosale and K Sridhar Patnaik. Puldi-covid: Chronic obstructive pulmonary (lung) diseases with covid-19 classification using ensemble deep convolutional neural network from chest x-ray images to minimize severity and mortality rates. *Biomedical Signal Processing and Control*, 81:104445, 2023.
- [157] Ronald Summers. Cxr8-national institutes of health-clinical center. 12, 2021.
- [158] Mohamed Abdel-Basset, Victor Chang, Hossam Hawash, Ripon K Chakraborty, and Michael Ryan. Fss-2019-ncov: A deep learning architecture for semi-supervised few-shot segmentation of covid-19 infection. *Knowledge-Based Systems*, 212:106647, 2021.
- [159] Adrian Galdran, André Anjos, José Dolz, Hadi Chakor, Hervé Lombaert, and Ismail Ben Ayed. The little w-net that could: state-of-the-art retinal vessel segmentation with minimalistic models. *arXiv preprint arXiv:2009.01907*, 2020.
- [160] Min Seob Kwak, Jae Myung Cha, Jung Won Jeon, Jin Young Yoon, and Jong Wook Park. Artificial intelligence-based measurement outperforms current methods for colorectal polyp size measurement. *Digestive Endoscopy*, 2022.

- [161] Yukun Zhou, Siegfried K Wagner, Mark A Chia, An Zhao, Moucheng Xu, Robbert Struyven, Daniel C Alexander, Pearse A Keane, et al. Automorph: Automated retinal vascular morphology quantification via a deep learning pipeline. *Translational vision science & technology*, 11(7):12–12, 2022.
- [162] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.
- [163] Johnatan Carvalho Souza, João Otávio Bandeira Diniz, Jonnison Lima Ferreira, Giovanni Lucca França da Silva, Aristofanes Correa Silva, and Anselmo Cardoso de Paiva. An automatic method for lung segmentation and reconstruction in chest x-ray using deep neural networks. *Computer methods and programs in biomedicine*, 177:285–296, 2019.
- [164] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.
- [165] Anas M. Tahir, Muhammad E. H. Chowdhury, Yazan Qiblawey, Amith Khandakar, Tawsifur Rahman, Serkan Kiranyaz, Uzair Khurshid, Nabil Ibtehaz, Sakib Mahmud, and Maymouna Ezeddin. Covid-qu, 2021.
- [166] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018.

- [167] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.
- [168] James Devasia, Hridayanand Goswami, Subitha Lakshminarayanan, Manju Rajaram, and Subathra Adithan. Deep learning classification of active tuberculosis lung zones wise manifestations using chest x-rays: a multi label approach. *Scientific Reports*, 13(1):887, 2023.
- [169] Yujin Oh, Sangjoon Park, and Jong Chul Ye. Deep learning covid-19 features on cxr using limited training data sets. *IEEE transactions on medical imaging*, 39(8):2688–2700, 2020.
- [170] Pedro HT Gama, Hugo Oliveira, and Jefersson A dos Santos. Learning to segment medical images from few-shot sparse labels. In *2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 89–96. IEEE, 2021.
- [171] Alex Taranov. Openist: A set of open source tools (c classes and cmdutils) for image segmentation and classification.
- [172] You-Bao Tang, Yu-Xing Tang, Jing Xiao, and Ronald M Summers. Xlsor: A robust and accurate lung segmentor on chest x-rays using criss-cross attention and customized radiorealistic abnormalities generation. In *International Conference on Medical Imaging with Deep Learning*, pages 457–467. PMLR, 2019.
- [173] Hugo Oliveira, Virginia Mota, Alexei MC Machado, and Jefersson A dos Santos. From 3d to 2d: Transferring knowledge for rib segmentation in chest x-rays. *Pattern Recognition Letters*, 140:10–17, 2020.
- [174] John Suckling. The mammographic images analysis society digital mammogram database. In *Exerpta Medica. International Congress Series, 1994*, volume 1069, pages 375–378, 1994.

- [175] Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria Joao Cardoso, and Jaime S Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248, 2012.
- [176] Aysen Degerli, Serkan Kiranyaz, Muhammad EH Chowdhury, and Moncef Gabbouj. Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2306–2310. IEEE, 2022.
- [177] Serkan Kiranyaz, Junaid Malik, Habib Ben Abdallah, Turker Ince, Alexandros Iosifidis, and Moncef Gabbouj. Self-organized operational neural networks with generative neurons. *Neural Networks*, 140:294–308, 2021.
- [178] Aysen Degerli, Mete Ahishali, Mehmet Yamac, Serkan Kiranyaz, Muhammad EH Chowdhury, Khalid Hameed, Tahir Hamid, Rashid Mazhar, and Moncef Gabbouj. Covid-19 infection map generation and detection from chest x-ray images. *Health information science and systems*, 9(1):15, 2021.
- [179] Melissa A Warren, Zhiguou Zhao, Tatsuki Koyama, Julie A Bastarache, Ciara M Shaver, Matthew W Semler, Todd W Rice, Michael A Matthay, Carolyn S Calfee, and Lorraine B Ware. Severity scoring of lung oedema on the chest radiograph is associated with clinical outcomes in ards. *Thorax*, 73(9):840–846, 2018.
- [180] Vinaya S Karkhanis and Jyotsna M Joshi. Pleural effusion: diagnosis, treatment, and management. *Open access emergency medicine: OAEM*, 4:31, 2012.
- [181] EDWARD A Gaensler and CHARLES B Carrington. Peripheral opacities in chronic eosinophilic pneumonia: the photographic negative of pulmonary edema. *American Journal of Roentgenology*, 128(1):1–13, 1977.
- [182] Aritoshi Hattori, Takeshi Matsunaga, Kazuya Takamochi, Shiaki Oh, and Kenji Suzuki. Prognostic impact of a ground glass opacity component in the clinical t

- classification of non-small cell lung cancer. *The Journal of Thoracic and Cardiovascular Surgery*, 154(6):2102–2110, 2017.
- [183] Aritoshi Hattori, Kenji Suzuki, Kazuya Takamochi, Masashi Wakabayashi, Keiju Aokage, Hisashi Saji, Shun-ichi Watanabe, Yasuhiro Tsutani, Hiroshige Yoshioka, Shiono Satoshi, et al. Prognostic impact of a ground-glass opacity component in clinical stage ia non-small cell lung cancer. *The Journal of Thoracic and Cardiovascular Surgery*, 161(4):1469–1480, 2021.
- [184] Van-Tien Pham, Cong-Minh Tran, Stanley Zheng, Tri-Minh Vu, and Shantanu Nath. Chest x-ray abnormalities localization via ensemble of deep convolutional neural networks. In *2021 International Conference on Advanced Technologies for Communications (ATC)*, pages 125–130. IEEE, 2021.
- [185] Ultralytics. Yolov5. <https://github.com/ultralytics/yolov5>.
- [186] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.
- [187] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107:104117, 2021.
- [188] Ilyas Sirazitdinov, Maksym Kholiavchenko, Tamerlan Mustafaev, Yuan Yixuan, Ramil Kuleev, and Bulat Ibragimov. Deep neural network ensemble for pneumonia localization from a large-scale chest x-ray database. *Computers & electrical engineering*, 78:388–399, 2019.
- [189] Radiological Society of North America. Rsnna pneumonia detection challenge. <https://www.kaggle.com/c/rsna-pneumonia-detection-challenge>.



- [190] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
- [191] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [192] I-Yun Chang and Teng-Yi Huang. Deep learning-based classification for lung opacities in chest x-ray radiographs through batch control and sensitivity regulation. *Scientific Reports*, 12(1):17597, 2022.
- [193] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.
- [194] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.
- [195] Joy Wu, Yaniv Gur, Alexandros Karargyris, Ali Bin Syed, Orest Boyko, Mehdi Moradi, and Tanveer Syeda-Mahmood. Automatic bounding box annotation of chest x-ray data for localization of abnormalities. In *2020 IEEE 17th international symposium on biomedical imaging (ISBI)*, pages 799–803. IEEE, 2020.
- [196] Alfredo J Selim, Graeme Fincke, Xinhua S Ren, William Rogers, Austin Lee, and Lewis Kazis. A symptom-based measure of the severity of chronic lung disease: results from the veterans health study. *Chest*, 111(6):1607–1614, 1997.
- [197] Gabrielle B McCallum, Peter S Morris, Clare C Wilson, Lesley A Versteegh, Linda M Ward, Mark D Chatfield, and Anne B Chang. Severity scoring systems: are they internally valid, reliable and predictive of oxygen use in children with acute bronchiolitis? *Pediatric pulmonology*, 48(8):797–803, 2013.

- [198] Emma Taylor, Kathryn Haven, Peter Reed, Ange Bissielo, Dave Harvey, Colin McArthur, Cameron Bringans, Simone Freundlich, R Joan H Ingram, David Perry, et al. A chest radiograph scoring system in patients with severe acute respiratory infection: a validation study. *BMC medical imaging*, 15(1):1–10, 2015.
- [199] Matthew D Li, Nishanth T Arun, Mishka Gidwani, Ken Chang, Francis Deng, Brent P Little, Dexter P Mendoza, Min Lang, Susanna I Lee, Aileen O’Shea, et al. Automated assessment and tracking of covid-19 pulmonary disease severity on chest radiographs using convolutional siamese neural networks. *Radiology: Artificial Intelligence*, 2(4):e200079, 2020.
- [200] Danielle Toussie, Nicholas Voutsinas, Mark Finkelstein, Mario A Cedillo, Sayan Manna, Samuel Z Maron, Adam Jacobi, Michael Chung, Adam Bernheim, Corey Eber, et al. Clinical and chest radiography features determine patient outcomes in young and middle-aged adults with covid-19. *Radiology*, 297(1):E197, 2020.
- [201] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019.
- [202] Anna Majkowska, Sid Mittal, David F Steiner, Joshua J Reicher, Scott Mayer McKinney, Gavin E Duggan, Krish Eswaran, Po-Hsuan Cameron Chen, Yun Liu, Sreenivasa Raju Kalidindi, et al. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*, 294(2):421–431, 2020.
- [203] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020.

- [204] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.
- [205] A Haghanifar, MM Majdabadi, Y Choi, S Deivalakshmi, and S Ko. Covid-cxnet: Detecting covid-19 in frontal chest x-ray images using deep learning, arxiv preprint arxiv: 2006.13807. 2020.
- [206] IS Radiology. M and i. italian society of medical and interventional radiology, 2020.
- [207] Hinrich B Winther, Hans Laser, Svetlana Gerbel, Sabine K Maschke, Jan B Hinrichs, Jens Vogel-Claussen, Frank K Wacker, Marius M Höper, and Bernhard C Meyer. Covid-19 image repository. *Figshare (Dataset)*, 2020.
- [208] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.
- [209] Allen Institute For AI. Covid-19 open research dataset challenge (cord-19). <https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge>.
- [210] Zhang Li, Zheng Zhong, Yang Li, Tianyu Zhang, Liangxin Gao, Dakai Jin, Yue Sun, Xianghua Ye, Li Yu, Zheyu Hu, et al. From community-acquired pneumonia to covid-19: a deep learning–based method for quantitative analysis of covid-19 on thick-section ct scans. *European radiology*, 30:6828–6837, 2020.
- [211] Zhenyu Tang, Wei Zhao, Xingzhi Xie, Zheng Zhong, Feng Shi, Jun Liu, and Dinggang Shen. Severity assessment of coronavirus disease 2019 (covid-19) using quantitative features from chest ct images. *arXiv preprint arXiv:2003.11988*, 2020.

- [212] Lu-shan Xiao, Pu Li, Fenglong Sun, Yanpei Zhang, Chenghai Xu, Hongbo Zhu, Feng-Qin Cai, Yu-Lin He, Wen-Feng Zhang, Si-Cong Ma, et al. Development and validation of a deep learning-based model using computed tomography imaging for predicting disease severity of coronavirus disease 2019. *Frontiers in bioengineering and biotechnology*, 8:898, 2020.
- [213] Zekuan Yu, Xiaohu Li, Haitao Sun, Jian Wang, Tongtong Zhao, Hongyi Chen, Yichuan Ma, Shujin Zhu, and Zongyu Xie. Rapid identification of covid-19 severity in ct scans through classification of deep features. *Biomedical engineering online*, 19(1):1–13, 2020.
- [214] Alysson Roncally S Carvalho, Alan Guimarães, Gabriel Madeira Werberich, Stephane Nery de Castro, Joana Sofia F Pinto, Willian Rebouças Schmitt, Manuela França, Fernando Augusto Bozza, Bruno Leonardo da Silva Guimarães, Walter Araujo Zin, et al. Covid-19 chest computed tomography to stratify severity and disease extension by artificial neural network computer-aided diagnosis. *Frontiers in Medicine*, 7:577609, 2020.
- [215] Ying Zhang, Huawei Wu, Haitao Song, Xiaoqian Li, Shiteng Suo, Yan Yin, and Jianrong Xu. Covid-19 pneumonia severity grading: test of a trained deep learning model. 2020.
- [216] Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2577–2586, Melbourne, Australia, July 2018. Association for Computational Linguistics.
- [217] Yi Hong and Charles E Kahn. Content analysis of reporting templates and free-text radiology reports. *Journal of digital imaging*, 26:843–849, 2013.
- [218] Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings*

- of the AAAI Conference on Artificial Intelligence, volume 33, pages 6666–6673, 2019.
- [219] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
- [220] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. *Advances in neural information processing systems*, 31, 2018.
- [221] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020.
- [222] Mehreen Sirshar, Muhammad Faheem Khalil Paracha, Muhammad Usman Akram, Norah Saleh Alghamdi, Syeda Zainab Yousuf Zaidi, and Tatheer Fatima. Attention based automated radiology report generation using cnn and lstm. *Plos one*, 17(1):e0262209, 2022.
- [223] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [224] Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, Xu Sun, et al. Auto-encoding knowledge graph for unsupervised medical report generation. *Advances in Neural Information Processing Systems*, 34:16266–16279, 2021.
- [225] Preethi Srinivasan, Daksh Thapar, Arnav Bhavsar, and Aditya Nigam. Hierarchical x-ray report generation via pathology tags and multi head attention. In *Proceedings of the Asian Conference on Computer Vision*, 2020.
- [226] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The*

*Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

- [227] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13753–13762, 2021.
- [228] Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical image analysis*, 80:102510, 2022.
- [229] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021.
- [230] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.
- [231] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.
- [232] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11558–11567, 2023.
- [233] Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the*

- IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19809–19818, 2023.
- [234] Abhaya Agarwal and Alon Lavie. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. *Proceedings of WMT-08*, 2007.
- [235] Mehreen Sirshar, Taimur Hassan, Muhammad Usman Akram, and Shoab Ahmed Khan. An incremental learning approach to automatically recognize pulmonary diseases from the multi-vendor chest radiographs. *Computers in Biology and Medicine*, 134:104435, 2021.
- [236] AG Farman and RP Lapp. Image file interoperability for data protection, communication and trans-system connectivity. *Orthodontics & Craniofacial Research*, 6:151–155, 2003.
- [237] Afshin Shoeibi, Navid Ghassemi, Jonathan Heras, Mitra Rezaei, and Juan M Gorriz. Automatic diagnosis of myocarditis in cardiac magnetic images using cyclegan and deep pretrained models. In *International Work-Conference on the Interplay Between Natural and Artificial Computation*, pages 145–155. Springer, 2022.
- [238] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [239] Xide Xia and Brian Kulis. W-net: A deep model for fully unsupervised image segmentation. *arXiv preprint arXiv:1711.08506*, 2017.
- [240] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018.

- [241] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7036–7045, 2019.
- [242] Zhiqiang Tang, Yunhe Gao, Yi Zhu, Zhi Zhang, Mu Li, and Dimitris N Metaxas. Crossnorm and selfnorm for generalization under distribution shifts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 52–61, 2021.
- [243] Luyang Luo, Lequan Yu, Hao Chen, Quande Liu, Xi Wang, Jiaqi Xu, and Pheng-Ann Heng. Deep mining external imperfect data for chest x-ray disease screening. *IEEE transactions on medical imaging*, 39(11):3583–3594, 2020.
- [244] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4085–4095, 2020.
- [245] Zhun Zhong, Yuyang Zhao, Gim Hee Lee, and Nicu Sebe. Adversarial style augmentation for domain generalized urban-scene segmentation. *Advances in Neural Information Processing Systems*, 35:338–350, 2022.
- [246] Oren Nuriel, Sagie Benaim, and Lior Wolf. Permuted adain: Reducing the bias towards global statistics in image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9482–9491, 2021.
- [247] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8690–8699, 2021.
- [248] Rikiya Yamashita, Jin Long, Snikitha Banda, Jeanne Shen, and Daniel L Rubin. Learning domain-agnostic visual representation for computational pathology using medically-irrelevant style transfer augmentation. *IEEE Transactions on Medical Imaging*, 40(12):3945–3954, 2021.



- [249] Hongyu Wang and Yong Xia. Domain-ensemble learning with cross-domain mixup for thoracic disease classification in unseen domains. *Biomedical Signal Processing and Control*, 81:104488, 2023.
- [250] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021.
- [251] Yue Wang, Lei Qi, Yinghuan Shi, and Yang Gao. Feature-based style randomization for domain generalization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5495–5509, 2022.
- [252] Mohammad Zunaed, Md Aynal Haque, and Taufiq Hasan. Learning to generalize towards unseen domains via a content-aware style invariant framework for disease detection from chest x-rays. *arXiv preprint arXiv:2302.13991*, 2023.
- [253] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [254] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- [255] Aritra Roy Gosthipaty and Ritwik Raha. A deep dive into transformers with TensorFlow and Keras: Part 2. In Puneet Chugh, Susan Huot, Kseniia Kidriavsteva, and Abhishek Thanki, editors, *PyImageSearch*. 2022. <https://pyimg.co/pzu1j>.
- [256] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [257] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

- [258] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [259] Andrew M Dai and Quoc V Le. Semi-supervised sequence learning. *Advances in neural information processing systems*, 28, 2015.
- [260] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning. 2018.
- [261] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.
- [262] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.
- [263] Wilson L Taylor. “cloze procedure”: A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433, 1953.
- [264] Ivona Najdenkoska, Xiantong Zhen, Marcel Worrying, and Ling Shao. Uncertainty-aware report generation for chest x-rays by variational topic inference. *Medical Image Analysis*, 82:102603, 2022.
- [265] Mashood Mohammad Mohsan, Muhammad Usman Akram, Ghulam Rasool, Norah Saleh Alghamdi, Muhammad Abdullah Aamer Baqai, and Muhammad Abbas. Vision transformer and language model based radiology report generation. *IEEE Access*, 11:1814–1824, 2022.
- [266] Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical image analysis*, 80:102510, 2022.

- [267] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [268] Ziyou Yan. Patterns for building llm-based systems and products. *eugeneyan.com*, Jul 2023.
- [269] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.
- [270] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- [271] Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, et al. Evaluating progress in automatic chest x-ray radiology report generation. *medRxiv*, pages 2022–08, 2022.
- [272] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.
- [273] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.
- [274] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.
- [275] Fenglin Liu, Shen Ge, and Xian Wu. Competence-based multimodal curriculum learning for medical report generation. In *Proceedings of the 59th Annual Meet-*

*ing of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3001–3012, Online, August 2021. Association for Computational Linguistics.

- [276] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383, 2017.
- [277] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017.
- [278] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–325, 2017.
- [279] Philipp Harzig, Yan-Ying Chen, Francine Chen, and Rainer Lienhart. Addressing data bias problems for chest x-ray image report generation. *arXiv preprint arXiv:1908.02123*, 2019.
- [280] Xin Huang, Fengqi Yan, Wei Xu, and Maozhen Li. Multi-attention and incorporating background information model for chest x-ray image report generation. *IEEE Access*, 7:154808–154817, 2019.
- [281] Jaehwan Jeong, Katherine Tian, Andrew Li, Sina Hartung, Fardad Behzadi, Juan Calle, David Osayande, Michael Pohlen, Subathra Adithan, and Pranav Rajpurkar. Multimodal image-text matching improves retrieval-based chest x-ray report generation. *arXiv preprint arXiv:2303.17579*, 2023.
- [282] Aaron Nicolson, Jason Dowling, and Bevan Koopman. Improving chest x-ray report generation by leveraging warm starting. *Artificial Intelligence in Medicine*,

page 102633, 2023.

- [283] Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y Ng, and Pranav Rajpurkar. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Machine Learning for Health*, pages 209–219. PMLR, 2021.
- [284] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.
- [285] Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. *arXiv preprint arXiv:2010.10042*, 2020.
- [286] An Yan, Zexue He, Xing Lu, Jiang Du, Eric Chang, Amilcare Gentili, Julian McAuley, and Chun-Nan Hsu. Weakly supervised contrastive learning for chest x-ray report generation. *arXiv preprint arXiv:2109.12242*, 2021.
- [287] Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. Contrastive attention for automatic chest X-ray report generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 269–280, Online, August 2021. Association for Computational Linguistics.
- [288] Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 72–82. Springer, 2021.
- [289] Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. Progressive transformer-based generation of

radiology reports. *arXiv preprint arXiv:2102.09777*, 2021.

- [290] Kaveri Kale, Pushpak Bhattacharyya, Milind Gune, Aditya Shetty, and Rustom Lawyer. KGVL-BART: Knowledge graph augmented visual language BART for radiology report generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3401–3411, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.
- [291] Han Qin and Yan Song. Reinforced cross-modal alignment for radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 448–458, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [292] Yixin Wang, Zihao Lin, Zhe Xu, Haoyu Dong, Jiang Tian, Jie Luo, Zhongchao Shi, Yang Zhang, Jianping Fan, and Zhiqiang He. Trust it or not: Confidence-guided automatic radiology report generation. *arXiv preprint arXiv:2106.10887*, 2021.
- [293] Jianmo Ni, Chun-Nan Hsu, Amilcare Gentili, and Julian McAuley. Learning visual-semantic embeddings for reporting abnormal findings on chest x-rays. *arXiv preprint arXiv:2010.02467*, 2020.
- [294] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.