A MULTIMODAL FEATURE FUSION BASED THORACIC DISEASE CLASSIFICATION FRAMEWORK COMBINING MEDICAL DATA AND CHEST X-RAY IMAGES

by

Nusrat Binta Nizam
0421182003

MASTER OF SCIENCE
IN
BIOMEDICAL ENGINEERING

Department of Biomedical Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh

June, 2023

The thesis titled, **"A MULTIMODAL FEATURE FUSION BASED THORACIC DISEASE CLASSIFICATION FRAMEWORK COMBINING MEDICAL DATA AND CHEST X-RAY IMAGES"**, submitted by **Nusrat Binta Nizam**, Roll No.: 0421182003, Session: April 2021, has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Master of Science in Biomedical Engineering on 19th June, 2023.

BOARD OF EXAMINERS

Dr. Taufiq Hasan Al Banna
Associate Professor
Dept. of BME, BUET, Dhaka-1000

Chairman
(Supervisor)

Dr. Muhammad Tarik Arafat
Professor and Head
Dept. of BME, BUET, Dhaka-1000

Member
(Ex-Officio)

Dr. Jahid Ferdous
Associate Professor
Dept. of BME, BUET, Dhaka-1000

Member

Dr. Mohammad Ariful Haque
Professor
Dept. of EEE, BUET, Dhaka-1000

Member
(External)

Candidate's Declaration

This is to certify that the work presented in this thesis entitled, "A MULTIMODAL FEATURE FUSION BASED THORACIC DISEASE CLASSIFICATION FRAMEWORK COMBINING MEDICAL DATA AND CHEST X-RAY IMAGES", is the outcome of the research carried out by Nusrat Binta Nizam under the supervision of Dr. Taufiq Hasan Al Banna, Associate Professor, Department of Biomedical Engineering, Bangladesh University of Engineering and Technology (BUET), Dhaka-1000, Bangladesh.

It is also declared that neither this thesis nor any part thereof has been submitted anywhere else for the award of any degree, diploma, or other qualifications.

Signature of the Candidate

Nusrat Binta Nizam

Nusrat Binta Nizam
0421182003

Dedication

Dedicated to My Parents.

Acknowledgement

I am thankful to Almighty for His blessings for the successful completion of my thesis. I would like to express my sincere appreciation, immense gratitude, and utmost respect to my supervisor, Dr. Taufiq Hasan Al Banna, Associate Professor, Department of Biomedical Engineering (BME), Bangladesh University of Engineering and Technology. His consistent supervision, caring guidance, and unwavering support have been invaluable to me. I am truly grateful for his keen interest in the subject matter and valuable advice he provided, which played a significant role in the successful completion of my thesis.

I am especially grateful to Dr. Muhammad Tarik Arafat, Dr. Jahid Ferdous, and Dr. Mohammad Ariful Haque for their consent to be on my board of examiners. I appreciate the support from the Department of Biomedical Engineering (BME) of Bangladesh University of Engineering and Technology (BUET) during my M.Sc. courses and thesis work. Finally, I would like to thank my family members for supporting me through many sleepless nights during my academic career.

Abstract

Chest X-rays are commonly used in clinical settings to diagnose thoracic diseases, especially in low-resource settings. However, interpreting these images can be challenging, particularly in resource-constrained environment. Current AI-based methods focus solely on the X-ray images without considering relevant clinical information. To effectively assist with limited resources, it is important for a computerized system to generate decisions relevant to those of radiologists. This requires incorporating pertinent clinical details, such as medical history, symptoms, and demographic information, into image-based computerized systems to enhance their performance. The development of AI-based systems faces two main challenges: the limited availability of comprehensive medical image datasets suitable for machine learning and the difficulty in reproducing the advanced reasoning abilities of experienced radiologists, who have undergone extensive training and accumulated expertise. In this work, at first an unimodal anatomy aware network is proposed which provided about 11% relative improvement in mean square error (MSE) compared to existing methods when evaluated on a dataset for predicting the severity of COVID-19 pneumonia. This model also exhibits promising results on an unseen clinical evaluation dataset which provides evidence of the efficacy of anatomy-aware architecture for predicting the severity of COVID-19 disease. Additionally, this thesis proposes a multimodal feature fusion framework to improve disease classification by combining medical data and image information. Existing approaches rely on textual information, lacking anatomical details. An advanced multimodal feature fusion-based approach is needed to enhance disease classification accuracy. In this study, a comparison of incorporating clinical information demonstrates the substantial value of patient indication data (i.e., medical history, demographics, symptoms) in disease classification. Incorporating such information enables computer-aided systems to function more closely to radiologists. The proposed feature fusion-based framework, ResVCBERT and DenseVCBERT exhibit a significant improvement in accuracy compared to baseline architectures, even when there are errors in the textual information. The proposed DenseVCBERT provided significant improvement with an accuracy of about 88.44% using the OpenI dataset of radiological reports and chest X-rays. Including anatomical information in deep learning models through feature fusion enhances the accuracy of AI-based frameworks, as demonstrated in the analysis of COVID-19 pneumonia severity prediction. This approach aids disease diagnosis and severity prediction, benefiting radiologists in developed and underdeveloped nations.

Contents

Certification ii

Candidate's Declaration iii

Dedication iv

Acknowledgement v

Abstract vi

List of Figures x

List of Tables xiii

List of Abbreviations 1

1 Introduction 2
- 1.1 Introduction 2
- 1.2 Literature Review 5
  - 1.2.1 Vision Models for Disease Classification 5
  - 1.2.2 Language Models for Disease Classification 6
  - 1.2.3 Vision-Language Models for Medical Report Generation 7
  - 1.2.4 Vision-Language Models for Disease Classification 7
- 1.3 Motivation of the Work 8
- 1.4 Objectives of the Thesis 8
- 1.5 Thesis Outline 9

2 Background 11
- 2.1 Assessment of Different Thoracic Diseases 11
- 2.2 Overview of Machine Learning Classifiers 13
  - 2.2.1 Naive Bayes (NB) 13
  - 2.2.2 Stochastic Gradient Descent 14
  - 2.2.3 Logistic Regression (LR) 15

2.3 A Brief Introduction to Natural Language Processing (NLP) 16
2.3.1 Integration of Transformer in NLP 18
2.4 Medical Image Analysis 22
2.4.1 Medical Image Pre-processing 22
2.4.2 Deep Learning in Medical Image Processing 23

3 Medical Data Processing 25
3.1 Machine Learning Classifiers for Text Processing 25
3.2 NLP Based Techniques for Text Processing 27
3.2.1 Rule Based Techniques 27
3.2.2 Pre-trained NLP Models 28
3.3 Convolutional Neural Network (CNN) for Text Processing 32

4 Medical Image Processing for Disease Classification 35
4.1 Medical Image Pre-processing 35
4.2 Natural Language Processing (NLP) Based Techniques for Image Classification 35
4.3 Convolutional Neural Network (CNN) Based Techniques for Image Classification 37
4.3.1 ResNet-50 37
4.3.2 DenseNet-121 38
4.4 Anatomy Aware Neural Networks for Image Classification 38

5 Anatomy Aware Feature Fusion Based Framework 42
5.1 Motivation of Incorporating Anatomical Information 42
5.2 Data Resources and Organization 43
5.2.1 JSRT dataset 43
5.2.2 SCR dataset 43
5.2.3 Stanford Chexpert dataset 43
5.3 Anatomy Aware Network for Severity Prediction 44
5.4 Evaluate the Performance of Anatomy Aware Deep Learning Framework 47
5.5 Discussion 49

6 Effective Multimodal Approaches for Medical Data and Image Processing 51
6.1 Dataset Description 51
6.1.1 OpenI Dataset 51
6.1.2 MIMIC-CXR Dataset 52
6.1.3 CheXpert Labeler 54
6.2 Medical Data Pre-processing 54

6.2.1 Text Pre-processing 54
6.2.2 Word Token Generation 57
6.3 Chest X-ray Pre-processing and Feature Extraction 57
6.4 Baseline Architecture 58
6.4.1 VisualBERT 58
6.4.2 LXMERT 59
6.4.3 UNITER 59
6.5 Proposed System 60
6.5.1 VisualClinicalBert 60

7 Evaluate the Performance of Multimodal Approaches 63
7.1 Evaluation of Indication Information in Disease Classification 63
7.2 Evaluation of Disease Information in Chest X-ray Images 65
7.3 Evaluation of Disease Classification Using Medical Data and Chest X-ray Images 66
7.4 Evaluation of the Performance of Baseline Systems 66
7.5 Evaluation of Proposed Framework 67
7.6 Discussion 67
7.6.1 Significance of Indication Information in Thoracic Disease Classification 67
7.6.2 Analyzing the Robustness of the Medical Data 68
7.6.3 Analysis Combining Image Feature with Medical Data for Thoracic Disease Classification 70

8 Conclusions 74
8.1 Summary of the Present Work 74
8.2 Limitations of the Present Work 75
8.3 Future Prospects of the Present Work 76

References 77

List of Figures

1.1 Diagram of a multimodal approach for classification of disease using medical text and image information. 3

2.1 The heart and lungs are located within the thoracic cavity between the lungs in the mediastinum. 12
2.2 Key assumptions for implementing Logistic Regression. 16
2.3 A historical perspective on the growth of NLP. 17
2.4 Basic architecture of Transformer [1]. 19
2.5 Basic architecture of BERT Encoder [2]. 20
2.6 Basic architecture of GPT Decoder [1]. 21
2.7 Basic Chest X-ray processing using a Deep Learning architecture [3]. . 22

3.1 Machine Learning based approach for text processing. 25
3.2 Natural Language Processing for radiology text processing [4]. 27
3.3 BERT model specialized for pre-training and fine-tuning [5]. 29
3.4 Clinical BERT learns deep representations of clinical text using masked language modeling and next-sentence prediction. In masked language modeling, a fraction of input tokens are held out for prediction; in next sentence prediction, Clinical BERT predicts whether two input sentences are consecutive. 30
3.5 Med-BERT structure details. 32
3.6 Convolutional Neural Network (CNN) architecture for text classification purpose. 34

4.1 Basic diagram of a vision transformer architecture [6]. 36
4.2 Basic architecture of ResNet-50 [7]. 38
4.3 Basic architecture of DenseNet-121 [8]. 39

4.4 Overview of the proposed semi-supervised anatomy-aware attention-based thoracic disease classification framework. A semi-supervised technique is utilized to generate anatomy masks for unannotated CXR images. Then, with the help of our proposed novel anatomy-aware attention module, anatomical information is integrated into the classification network for pathology detection [9]. 40
4.5 Basic architecture of Anatomy-XNet [9]. 40

5.1 Dataset organization and overall flow diagram in the feature extractor training and regression training/testing phases. 44
5.2 The proposed Anatomically-Aware Network for COVID-19 severity prediction can be summarized in this figure. In path 1, Feature Pyramid Attention is employed to improve pixel-level attention, resulting in the generation of disease-level features for five specific diseases (Atelectasis, Consolidation, Edema, Pleural Effusion, and Consolidation). On the other hand, path 2 is utilized for generating class-wise heatmaps using PCAM pooling. These heatmaps, along with the lung masks, are used to calculate the lung involvement score. Finally, the disease-level features and lung involvement scores are utilized for regression analysis. 45
5.3 (a) Predicted value and ground truth of geographical extent score of selected chest X-ray images annotated by an In-house experienced radiologist, and (b) Predicted value and ground truth of opacity score of selected chest X-ray images annotated by an In-house experienced radiologist. 49

6.1 OpenI Chest X-ray image with corresponding medical report (XML). . . 52
6.2 MIMIC-CXR Chest X-ray image with corresponding medical report (XML). 53
6.3 The CheXpert predicts the probability of different observations from chest x-rays. 55
6.4 Chest X-ray pre-processing and feature extraction framework. 57
6.5 Configuration of VisualBERT Baseline Framework [10]. 58
6.6 Configuration of LXMERT Baseline Framework [11]. 59
6.7 Configuration of UNITER Baseline Framework [12]. 60
6.8 Proposed thoracic disease classification framework. 61
6.9 The diagram is a two-layer MLP. There are three inputs with three input nodes and the hidden layer has three nodes. The output layer gives two outputs as there are two output nodes. 61

7.1 OpenI radiological report indication information similarity. 69
7.2 Accuracy (%) analysis of baselines and proposed systems after adding some text errors to the indication information. 69
7.3 Result analysis of proposed DenseVCBERT after removing indication information from most of the reports. 70
7.4 Result comparison of proposed DenseVCBERT with indication information and without indication information (Only image). 70
7.5 Accuracy (%) analysis of baselines and proposed systems with both indication text information and medical image features. 71
7.6 Precision (%) analysis of baselines and proposed systems with both indication text information and medical image features. 71
7.7 Recall(%) analysis of baselines and proposed systems with both indication text information and medical image features. 72
7.8 F1 Score (%) analysis of baselines and proposed systems with both indication text information and medical image features. 72

List of Tables

5.1 Datasets required for the training and evaluation of the proposed framework. 44
5.2 Performance evaluation of the proposed COVID-19 severity prediction architecture compared to existing models. These results are also compared with the baseline scores from Cohen *et al.* [13].) 48
5.3 Evalution of the proposed COVID-19 severity prediction model on Selected Chest X-ray images annotated by an In-house experienced radiologist 49

6.1 An Overview of the Datasets Used in Different Experiments. 55

7.1 Experimental results on OpenI Medical Reports using Machine Learning Classifiers 64
7.2 Experimental results on OpenI Medical Reports using Different Language Models 64
7.3 Experimental results on MIMIC-CXR Medical Reports using Different Frameworks 65
7.4 Experimental results on OpenI Medical Images 65
7.5 Experimental results on OpenI (Text-Indication and Image Combined) . 66
7.6 Experimental results on OpenI Dataset of Baseline System (Text-Indication and Image Combined) 67
7.7 Experimental results on OpenI Dataset of Proposed System (Text-Indication and Image Combined) 67
7.8 Classwise Accuracy (%) Comaparison of Different Frameworks (Text-Indication and Image Combined) 68

List of Abbreviations

AA	Anatomy Aware
AI	Artificial Intelligence
ANN	Artificial Neural Networks
BERT	Bidirectional Encoder Representations from Transformer
CNN	Convolutional Neural Network
CT	Computed Tomography
CV	Computer Vision
CXR	Chest X-ray
EHR	Electronic Health Record
FPA	Feature Pyramid Attention
GELU	Gaussian Error Linear Unit
LDCT	Low Dose Computed Tomography
LEXIMER	Lexicon Mediated Entropy Reduction
LR	Logistic Regression
ML	Machine Learning
MRI	Magnetic Resonance Imaging
NB	Naive Bayes
NLP	Natural Language Processing
NLU	Natural Language Understanding
PET	Positron Emission Tomography
SGD	Stochastic Gradient Descent
UDA	Unsupervised Domain Adaptation
VL	Vision-Language
ViT	Vision Transformer

Chapter 1

Introduction

1.1 Introduction

The advancements achieved in Artificial Intelligence (AI) present a significant opportunity to revolutionize healthcare, particularly in the field of medical imaging used for diagnosing, prognosing, and treating diseases. Currently, state-of-the-art radiology techniques primarily focus on pixel-level details, neglecting the valuable clinical data and the patient's medical history. By incorporating this additional information, we can greatly improve the interpretation of imaging results, leading to more accurate diagnoses, better decision-making, and ultimately improved patient outcomes. The rise of advanced diagnostic tools has given rise to the increasing importance of multimodal fusion in the medical field. A recent approach in medical informatics involves fusing visual information from radiological images with associated textual descriptions. However, effectively handling the complexities of high dimensionality, heterogeneity, and biases inherent in such systems presents significant technical challenges. Leveraging multimodal approaches that combine vision and language can offer several benefits, including enhanced automated disease classification and support systems for generating medical reports. In Figure 1.1 a basic multimodal framework is shown. In a clinical setting, there is a growing demand for computer-based support systems capable of utilizing not only radiology images but also supplementary patient data, enabling the processing of multimodal information to make informed decisions. When it comes to the semantic understanding of medical texts, contextual word embeddings, particularly BERT, have demonstrated exceptional performance. Consequently, combining convolutional neural networks (CNN) and BERT has become a popular architecture for jointly processing both images and texts in an integrated manner.

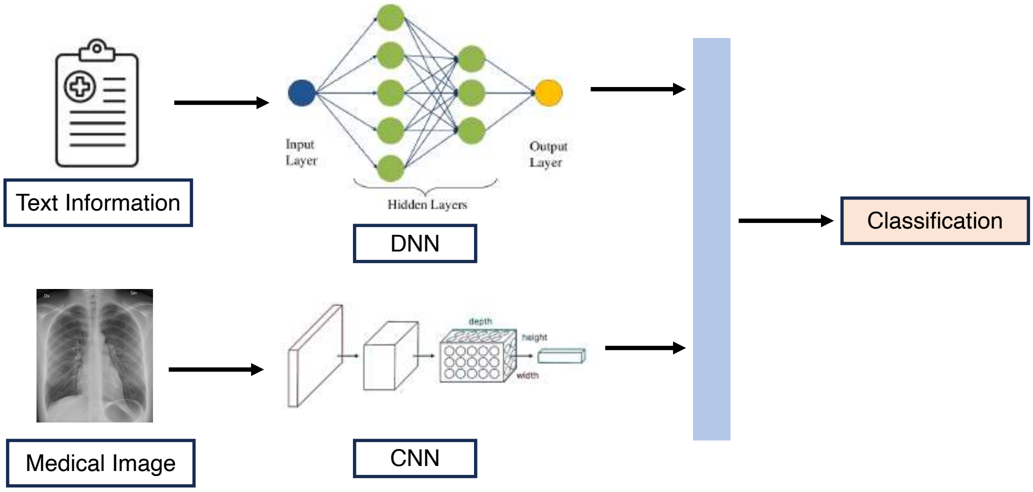


Figure 1.1: Diagram of a multimodal approach for classification of disease using medical text and image information.

The complete medical background of a patient, encompassing their overall health, past medical conditions, and prior treatments, provides valuable contextual information. This data allows physicians to understand the patient's health journey, identify patterns, and make well-informed diagnostic decisions. By considering the symptoms reported by the patient, doctors can utilize the medical history to assess the timeline and progression of symptoms, narrowing down potential diagnoses and focusing on relevant areas for further investigation. Moreover, the medical history plays a crucial role in evaluating the patient's susceptibility to specific diseases. It provides insights into genetic predispositions, family medical history, lifestyle habits, occupational exposures, and other relevant factors that contribute to the development of diseases. This valuable information helps doctors determine the likelihood of certain conditions and guides them in ordering appropriate diagnostic tests. In terms of differential diagnosis, where doctors consider various possible causes for the patient's symptoms, a comprehensive medical history is instrumental. By analyzing the medical history, doctors can prioritize potential diagnoses and order targeted tests, thereby saving time and resources. Understanding the patient's medical history allows doctors to tailor treatment plans based on the patient's specific circumstances. Previous treatment responses, medication allergies, and other relevant details inform doctors' decisions in selecting appropriate therapies and minimizing potential complications. Furthermore, the medical history establishes a foundation for long-term disease management. It enables doctors to track the progression of a disease, evaluate the effectiveness of treatments, and make necessary adjustments to optimize patient outcomes. In summary, the medical history is of utmost importance in disease diagnosis as it provides crucial insights into a patient's health background. It aids doctors in evaluating symptoms, assessing risk factors, conducting differential diagnosis, planning treatment, and effectively managing diseases.

Chest X-rays provide visual representations of the internal structures of the chest, including the heart, lungs, and surrounding tissues. They offer valuable insights into the presence of abnormalities such as tumors, infections, or lung diseases. As a diagnostic tool, they assist healthcare professionals in identifying and categorizing various diseases, including pneumonia, tuberculosis, lung cancer, congestive heart failure, and other pulmonary or cardiac disorders. Additionally, chest X-rays are commonly used for screening purposes, allowing for the early detection of potential abnormalities in individuals at risk or displaying disease-related symptoms. This early detection leads to improved treatment outcomes. Serial chest X-rays taken over time enable the monitoring of disease progression by comparing current and past images. This facilitates the evaluation of changes in lung structure size, shape, or density and helps assess disease advancement and treatment response. The information derived from chest X-rays is crucial in treatment planning, guiding physicians in determining appropriate actions such as prescribing medications, recommending surgical interventions, or referring patients for specialized evaluation. Furthermore, chest X-rays are frequently used to assess treatment effectiveness, monitor patient progress, identify possible complications, and make necessary adjustments to treatment strategies. Moreover, the findings from chest X-rays contribute to medical research and data analysis. Aggregating and analyzing chest X-ray data can lead to advancements in disease classification, prognostication, and the development of more accurate diagnostic algorithms. Chest X-ray information plays a significant role in disease classification. These imaging studies provide vital visual assessments, aid in diagnosis, serve as screening tools, assist in treatment planning, enable disease progression monitoring, and contribute to research and data analysis. They guide healthcare professionals in accurately classifying diseases and implementing appropriate management strategies to ensure optimal patient care.

Integrating various clinical data, including patient history, laboratory results, and symptoms, with the findings from chest X-rays offers a comprehensive understanding of the patient's condition. This comprehensive approach allows healthcare professionals to consider a broader range of factors and achieve more precise disease classifications. By combining multiple sources of information, the multimodal approach enhances the accuracy of disease classification. Clinical data provides supplementary insights into the patient's overall health, existing conditions, and risk factors, complementing the visual analysis of chest X-rays and leading to more accurate diagnoses. The amalgamation of clinical data and chest X-ray information enhances the sensitivity and specificity of disease classification. Clinical data aids in identifying subtle abnormalities or providing context for interpreting chest X-ray results, thereby improving overall diagnostic performance and reducing the likelihood of misclassification. The multimodal approach empowers healthcare professionals to make well-informed decisions regarding treatment plans and interventions. By considering both clinical data and chest X-ray information, they can customize treatment strategies to address the patient's specific needs, considering factors such as disease severity, comorbidities, and potential complications. The multimodal approach facilitates early detection and intervention in diseases. By analyzing both clinical data and chest X-ray findings, healthcare professionals can identify early indicators of diseases that may not be apparent using either modality alone. This allows for timely interventions, resulting in better patient outcomes and improved prognosis. Integrating clinical data and chest X-ray information supports the development of personalized medicine approaches. By taking into account individual patient characteristics and merging modalities, healthcare professionals can create tailored treatment plans that are more effective and minimize the risks of potential adverse events. The multimodal approach of combining clinical data and chest X-ray information contributes to advancements in medical research. By aggregating and analyzing data from diverse sources, new insights, patterns, and correlations can be uncovered, leading to the development of improved disease classification algorithms and decision support systems.

1.2 Literature Review

1.2.1 Vision Models for Disease Classification

Deep Convolutional Neural Networks (CNNs) have gained popularity for directly extracting feature representations from CXR images through supervised learning, demonstrating impressive effectiveness in classifying thoracic diseases [14–16]. Various techniques have been documented for feature learning, utilizing established CNN architectures such as ResNet [17] and DenseNet [18]. For example, Wang et al. [14] utilize AlexNet [19], VGG16 [20], ResNet50, and GoogLeNet [21] as backbone networks, pretraining them on the ImageNet dataset [22] and fine-tuning on specific CXR datasets. Similarly, Chen et al. [23] combine ResNet and DenseNet to effectively capture various abnormal features in CXR images. Notably, Chen et al. [23] introduce Graph Convolution Networks (GCNs) [24] for thoracic disease classification, exploring the interplay among different pathologies.

In recent years, the medical image analysis community has extensively utilized semi-supervised learning to tackle the challenge of limited image annotation. Several studies [25–27] have employed techniques that leverage unlabeled data to improve model predictions through consistency-enforcing methods. In the domain of thorax disease classification, various approaches have employed semi-supervised learning to optimize deep neural networks. For example, Aviles et al. [28] propose a graph-based optimization model to enhance collaboration between a small number of labeled samples and a large amount of unlabeled data. Despite achieving some degree of success, semi-supervised learning often becomes impractical when additional high-quality labeled CXR images are difficult to obtain, especially in real-world scenarios where expert radiologists are scarce.

To overcome the time-consuming process of manually labeling domain-specific data for training, extensive research has been conducted on Unsupervised Domain Adaptation (UDA) [29, 30]. The goal of UDA is to transfer discriminative feature representations from a labeled source domain to an unlabeled target domain. Current approaches primarily focus on directing feature learning to minimize the differences between the distributions of features in the source and target domains.

1.2.2 Language Models for Disease Classification

In the medical domain, BERT [31] and ELMo [32] have been widely employed for various tasks such as medical image analysis and natural language processing of electronic health records. The recent advancement in language modeling, ChatGPT [33], has demonstrated significant progress. While language models have found extensive applications in processing electronic health records and disease diagnosis [34–36], their utilization in medical imaging tasks, such as disease diagnosis, remains relatively limited [37–39].

Language models have gained significant attention and accessibility by surpassing previous approaches like RNN-based models [40, 41] across various tasks. These models can be broadly categorized into three types: autoregressive models (e.g., GPT), masked language models (such as BERT), and encoder-decoder models (e.g., BART [42] and T5 [43]). Recently, there has been a notable increase in the development of extremely large language models, including GPT-3 [44], Bloom [45], PaLM [46], and OPT [47].

1.2.3 Vision-Language Models for Medical Report Generation

The field of automated medical report generation has witnessed several advancements in order to achieve its intended objective. Initially, methods such as template filling, description retrieval, and manual construction of natural language generation techniques were employed. The central challenge can be defined as the transformation of images into sequences, where the input comprises pixel values arranged in a sequential format. Through visual encoding, these input patches are converted into feature vectors, ultimately generating a latent space vector that serves as the input for the subsequent language generation step. In this phase, the latent vector is decoded using a specific vocabulary, resulting in the production of a sequence of words or subwords as the final output.

In [48], a two-part model is proposed, consisting of an Image Encoder and a Captioning Decoder that does not employ recurrent connections. The KERP (Knowledge-driven Encode, Retrieve, Paraphrase) approach [49] integrates contemporary learning-based methodologies for report generation with knowledge and retrieval-based methods. In [50], a memory-driven transformer is proposed for report generation. This method incorporates a transformer architecture with a relational memory to store important information. A deep neural network is employed in [51] to predict tags and generate reports based on provided chest X-ray images. The tag embeddings are obtained using a convolutional neural network, followed by transformers that facilitate the learning of self and cross attention mechanisms. In [52], a CNN-based feature with an attention layer and LSTM are utilized to generate more reliable reports. A novel two-step model is introduced in [53] that extracts overarching concepts from images and transforms them into detailed and coherent textual representations using a transformer architecture. In [54], a deep learning architecture comprising a CNN model as the encoder and a Transformer model as the decoder is utilized. Chexnet is employed as the encoder to predict tags for images and generate a latent space vector.

1.2.4 Vision-Language Models for Disease Classification

Several CNN-RNN-based V-L (Vision-Language) models have been proposed for disease diagnosis using CXR images. A recent study introduces a novel approach called TNNT (Text-guided Neural Network Training) [55]. TNNT enhances the efficiency of training on V-L data by incorporating guidance from text report embeddings into the CNN model. The evaluation of TNNT on four V-L datasets, including the OpenI dataset, demonstrates the significance of text reports, as they contain crucial information that can improve the accuracy of diagnosis compared to models relying solely on visual information. TieNet [56] is another CNN-RNN-based V&L embedding model that integrates multi-level attention layers into an end-to-end CNN-RNN framework for disease diagnosis and radiology report generation tasks. In [39], the transferability of pre-trained V-L models is evaluated through fine-tuning. Additionally, a transformer-based model called BERTHop [57] is proposed. BERTHop combines PixelHop++ and VisualBERT to better capture the associations between the two modalities.

1.3 Motivation of the Work

The motivation for integrating a patient's medical history and symptoms with chest X-ray image features for thoracic disease classification arises from the understanding that a comprehensive approach can greatly improve the accuracy and effectiveness of disease diagnosis. By combining clinical data and symptoms, we can gain a more thorough comprehension of the patient's health status, pre-existing conditions, and risk factors, providing valuable context for interpreting chest X-ray findings. The multimodal approach, which merges clinical data and chest X-ray information, plays a vital role in disease classification as it enables a comprehensive assessment, enhances diagnostic precision, improves decision-making, facilitates early detection and intervention, supports personalized medicine, and drives research advancements.

By utilizing multiple sources of information, healthcare professionals can achieve more precise and customized disease classifications, leading to improved patient care and outcomes. Therefore, there is a need for a state-of-the-art multimodal feature fusion-based approach that integrates significant clinical information from reports and combines it with radiological findings to achieve more accurate classification. The current methods that solely rely on clinical findings for classification may exhibit biases, particularly when disease information is available. To address this issue, a framework is required that leverages the patient's medical history and symptomatic information along with image features to emulate the decision-making capabilities of a radiologist.

1.4 Objectives of the Thesis

The main objectives of the work are:

1. To propose a state-of-art multimodal feature fusion-based architecture using raw

medical text data and chest x-ray images for improving thoracic disease classification performance.

2. To address the limitations of current multimodal approaches combining patient's medical history and image-based significant features.
3. To propose a novel feature extraction framework for Tele-radiology platforms.
4. To compare and evaluate the performance of the proposed method with the baseline methods.
5. To develop an Artificial Intelligence (AI) based framework for assisting Radiologists.

1.5 Thesis Outline

The rest of this thesis is organized as follows: Chapter 2 begins with a background discussion of different thoracic diseases and their prevalence in the recent world. Then the fundamental theory of different machine learning classifiers, natural language processing (NLP) and medical image processing has been explained. The mechanism behind different existing algorithms is discussed thoroughly in this chapter.

In Chapter 3, different medical data processing is discussed. Two basic NLP-based techniques used in medical data processing are explained. The overview of convolutional neural networks (CNN) in text processing is discussed. In addition to this, a brief introduction to the machine learning techniques in text processing is provided.

Chapter 4 is basically on different medical image processing techniques that are used in this work. At first, different pre-processing techniques are explained. Then NLP and CNN-based classification procedures are discussed. Finally, the usage of anatomy-aware neural networks in disease classification and even in severity prediction is discussed.

Chapter 5 reports the motivation for incorporating anatomical information in the feature fusion approach. In addition to this, a study of unimodal feature extraction approach for COVID-19 severity prediction and an experimental study with results is provided in this section. Performance evaluation and a brief discussion are provided to explain the performance of this feature fusion-based approach.

Chapter 6 describes the multimodal framework designed for medical data and image processing. An explanation of baseline architectures and proposed framework is provided. In addition to this, dataset description, pre-processing techniques, and feature extraction techniques are explained in this section.

Chapter 7 reports the evaluation results and comparison study of different frameworks. First, an analysis of the significance of indication information is provided using different experimental analyses. The robustness of the proposed architecture is explained and analyzed with examples and experimental results. A comparative analysis with baselines is provided in this section.

Chapter 8 serves as the concluding section of the thesis, providing a comprehensive overview of the research conducted. It includes the key findings and emphasizes the potential influence of this thesis on forthcoming research. Furthermore, it offers a concise overview of the potential route for future experimental and theoretical research in this specific domain.

Chapter 2

Background

2.1 Assessment of Different Thoracic Diseases

According to global statistics from 2019, Chronic Respiratory Diseases (CRDs) are the third leading cause of mortality worldwide, resulting in 4.0 million deaths [58]. The prevalence of CRDs is estimated at 454.6 million cases globally. On the other hand, Cardiovascular diseases (CVDs) hold the highest position as the primary cause of mortality, responsible for approximately 17.9 million deaths annually [59].

Computer-aided diagnostic methods are essential in low-income settings to combat the high mortality rates associated with respiratory and cardiovascular diseases. Different imaging modalities, including chest X-ray, chest computed tomography (CT), MRI, positron emission tomography (PET), and others, are available for diagnostic purposes. Among these options, chest X-ray is the most widely used and cost-effective imaging technique for diagnosing various thoracic diseases in developing and underdeveloped countries. Thoracic diseases primarily consist of abnormal lung and heart conditions, such as Atelectasis, Cardiomegaly, Edema, Pleural effusion, Pneumonia, Pneumothorax, and more. Figure 2.1 provides a visual representation of the anatomical structure of the thoracic cage. Abnormalities can be observed in the frontal and lateral views of chest X-rays, providing valuable information for disease classification.

Radiologists employ a systematic approach to analyze chest X-rays and identify thoracic disease information. Here's a general outline of the process:

1. **Image Evaluation:** Radiologists begin by assessing the technical quality of the X-ray image, checking for appropriate positioning, exposure, and image clarity. If the image quality is inadequate, they may request a repeat X-ray.

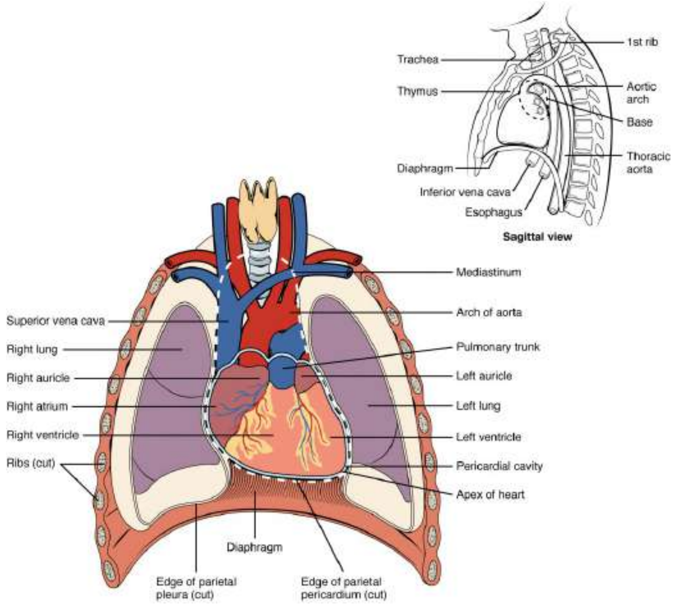


Figure 2.1: The heart and lungs are located within the thoracic cavity between the lungs in the mediastinum.

2. **Initial Observation:** Radiologists conduct a preliminary assessment of the overall appearance of the chest X-ray. They evaluate the lung fields, heart, ribs, and other structures to identify any gross abnormalities or artifacts. Lung Evaluation: Radiologists systematically examine the lung fields, assessing for any signs of abnormal lung parenchyma (tissue) or lung diseases. They observe the lung markings, looking for changes in density, nodules, masses, consolidation, or areas of collapse.

3. **Lung Evaluation:** Radiologists systematically examine the lung fields, assessing for any signs of abnormal lung parenchyma (tissue) or lung diseases. They observe the lung markings, looking for changes in density, nodules, masses, consolidation, or areas of collapse.

4. **Mediastinum and Heart Assessment:** Radiologists evaluate the mediastinum, which includes the heart, great vessels, and other structures in the central chest. They analyze the size, shape, and position of the heart, as well as the width of the mediastinum. They also search for signs of enlarged lymph nodes, masses, or abnormalities in the major blood vessels.

5. **Bones and Soft Tissues:** Radiologists examine the ribs, clavicles, and other bony structures for fractures, bone lesions, or evidence of trauma. They also assess the soft tissues, including the chest wall and subcutaneous tissues, looking for any abnormalities.

6. **Comparison:** If available, radiologists compare the current chest X-ray with any previous imaging studies to identify changes over time, which can be crucial for detecting disease progression or improvement.
7. **Reporting:** After evaluating the chest X-ray thoroughly, radiologists generate a structured radiology report. This report includes a detailed description of their findings, impression, and recommendations. The report is then shared with the referring physician to guide further diagnosis and treatment.

It is essential to acknowledge that the procedure may differ based on the particular clinical situation and the proficiency of the radiologist. Furthermore, in intricate cases or when additional investigation is required, supplementary imaging techniques such as computed tomography (CT) scans or magnetic resonance imaging (MRI) might be suggested.

The patient's clinical history and disease symptoms are significant to understand the state of health of the patient and to determine any acute complaints that can be directed towards diagnosis [60]. In the process of treating a patient, information obtained through various methods serves as a vital guide for providing appropriate care. Challenges often arise when patients are hesitant to disclose their complete medical history. However, for subsequent visits, a review of the medical history and any necessary updates can suffice. The medical history unveils pertinent chronic conditions and past diseases that may not be currently treated but have a lasting impact on the patient's health. Therefore, medical history plays a crucial role in directly formulating differential diagnoses [61–63]. When it comes to diagnosing heart and lung diseases, the medical history of patients holds great importance. For instance, individuals with a history of smoking are more prone to developing lung diseases, as smoking is a major contributing factor. Additionally, in cases of internal infections, the presence of fever may suggest an ongoing infection within the body. Therefore, the significance of medical history and symptomatic information cannot be overlooked when it comes to categorizing various thoracic diseases.

2.2 Overview of Machine Learning Classifiers

2.2.1 Naive Bayes (NB)

A simplistic learning technique called Naive Bayes (NB) [64, 65] makes use of Baye's rule and the fundamental presumption that the characteristics are conditionally independent given the class. Despite the fact that in real-world situations the assumption of independence is frequently broken, the Naive Bayes (NB) classifier routinely obtains equivalent classification accuracy. This, together with other characteristics, makes NB a well-liked option in real-world applications. NB enables the estimation of the posterior probability $P(y|x)$ for each class y given an object x using the information from sample data. This type of estimator proves beneficial in classification tasks and other decision-support applications.

Naïve Bayes is a form of Bayesian Network Classifier based on Bayes' rule together with an assumption that the attributes are conditionally independent given the class.

$$P(y|x) = P(y)\frac{P(x|y)}{P(x)} \tag{2.1}$$

For attribute-value data, this assumption can be represented as:

$$P(x|y) = \prod_{i=1}^{n} P(x_i|y) \tag{2.2}$$

where x_i is the value of the i^{th} attribute in x, and n is the number of attributes.

$$P(x) = \prod_{i=1}^{k} P(c_i)P(x|c_i) \tag{2.3}$$

where k is the number of classes and c_i is the i^{th} class.

Two naive Bayes variations are frequently used in text mining [66]. The multi-variate Bernoulli model, which represents each document as a vector of binary variables, uses the naive Bayes algorithm discussed above. These variables show if certain issues are present or not. However, only the words that are genuinely present in a document are taken into account for determining its probability.

2.2.2 Stochastic Gradient Descent

Let us consider an information processing system that receives a vector input signal: r and emits an output signal z. The system includes feed forward-type connections, but not feedback connections. It defines a mapping from the set $X = x$ of input signals to the set $Z = z$ of output signals,

$$z = F(x) \tag{2.4}$$

When the system includes a number of modifiable parameters $v = (v_1, ..., v_n)$, the input-output function 2.4 is specified by v,

$$z = F(x; v) \tag{2.5}$$

Here, we assume that F is differentiable with respect to v. When an input signal x is processed by a system specified by v, a loss is caused because the system might not be optimally tuned. The loss is denoted by $l(x; v)$. In some cases, a desired output y accompanies x. In this case, the loss is written as $l(x, y; v)$, denoting the loss when x with a desired or teacher signal y is processed by the network specified by v.

Let us assume that the input signal x is generated subject to a fixed but unknown probability distribution $p(x)$ each time independently. The accompanying desired output y is usually a function of x called the desired output. It is sometimes disturbed by noise. In this case, y is generated subject to the conditional probability $p(y|x)$ and the expectation of y,

$$y_d(x) = E[y|x] = \int yp(y|x)\,dy$$

where $E[y[x]]$ is the conditional expectation of y under the condition that the input x is the desired signal [67]. The stochastic Y is its noisy version. In the noiseless ease, y is written as:

$$p(y|x) = \delta(y - y_d(x)) \tag{2.7}$$

2.2.3 Logistic Regression (LR)

Logistic regression is the most appropriate regression analysis when dealing with a binary and dichotomous dependent variable [68–72]. It is a predictive analysis method used to describe data and explain the relationship between a dependent binary variable and one or more independent nominal, ordinal, interval, or ratio-level variables. The formula of LR is:

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \tag{2.8}$$

We all know the equation of the best-fit line in linear regression is:

$$y = \beta_0 + \beta_1 x \tag{2.9}$$

where x is the input and y is output. β_0 is bias or intercept term and β_1 is coefficient for input x. Given that the outcome is a probability with a range of 0 to 1, logistic regression's interpretation of weights is different from that of linear regression. The weights no longer have a linear effect on the likelihood in logistic regression. Instead, the weighted sum is converted into a probability by the logistic function. Using probabilities rather than a linear relationship, logistic regression may now represent the relationship between the predictors and the outcome. The key assumptions for Logistic regression can be observed in Figure 2.2.

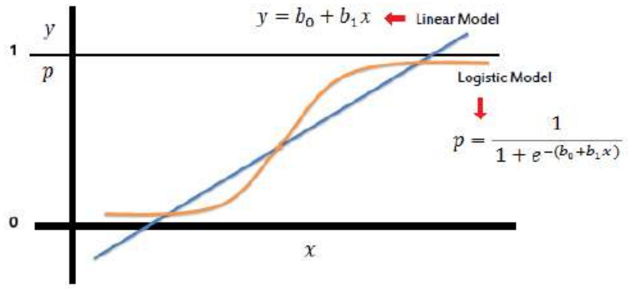


Figure 2.2: Key assumptions for implementing Logistic Regression.

2.3 A Brief Introduction to Natural Language Processing (NLP)

In the domain of Natural Language Processing (NLP), text analysis is performed using a range of theories and tools. It is a dynamic and progressive field of research that lacks a universally agreed-upon definition that would satisfy all experts. Nonetheless, there are certain components considered essential and would be included in the understanding of NLP by any knowledgeable individual.

Natural Language Processing (NLP) refers to a range of computer methods supported by a theoretical foundation. The historical perspective of NLP can be observed in Figure 2.3. These methods encompass various levels of linguistic analysis and are utilized for examining and representing naturally occurring texts. The ultimate objective of NLP is to develop language processing abilities that resemble those of a human. This broadens its applicability to a wide range of tasks and practical applications [73–75].

The aim of Natural Language Processing (NLP) is to attain language processing capabilities that closely resemble human abilities. It is important to note that the term

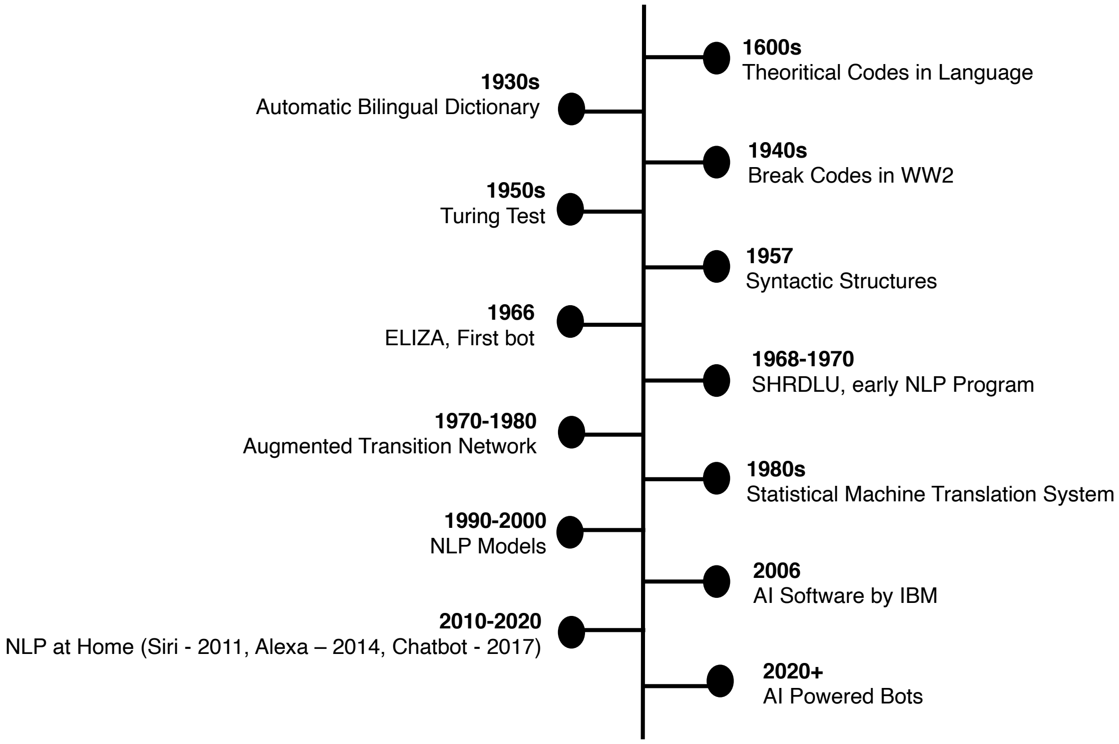

1600s
Theoritical Codes in Language
1930s
Automatic Bilingual Dictionary
1940s
Break Codes in WW2
1950s
Turing Test
1957
Syntactic Structures
1966
ELIZA, First bot
1968-1970
SHRDLU, early NLP Program
1970-1980
Augmented Transition Network
1980s
Statistical Machine Translation System
1990-2000
NLP Models
2006
AI Software by IBM
2010-2020
NLP at Home (Siri - 2011, Alexa – 2014, Chatbot - 2017)
2020+
AI Powered Bots

Figure 2.3: A historical perspective on the growth of NLP.

”processing” is deliberately chosen and should not be replaced with ”understanding.” Although NLP was initially known as Natural Language Understanding (NLU) in the early days of AI, it is now widely recognized that complete NLU has not yet been achieved. NLP methodologies can be classified into four primary types: symbolic, statistical, connectionist, and hybrid approaches.

Symbolic and statistical approaches have coexisted in the field of NLP since its early days. The emergence of connectionist NLP work dates back to the 1960s. Symbolic approaches have long dominated the field, involving detailed language analysis through explicit representations and established algorithms. The description of language analysis levels provided earlier in this text is presented from a symbolic standpoint. Symbolic systems heavily rely on rules and lexicons created by humans as the primary sources of evidence.

In contrast, statistical approaches employ a range of mathematical techniques and extensive text collections to construct approximate and generalized language models based on real-world examples from these collections. Unlike symbolic approaches, statistical methods primarily rely on observable data as evidence for their models.

Connectionist approaches also construct generalized models based on linguistic examples. What sets connectionism apart from other statistical methods is its integration of statistical learning with various representation theories. This integration enables connectionist models to handle transformations, inferences, and manipulations of logical formulas. Linguistic models within connectionist systems are more complex to observe due to the greater flexibility offered by connectionist architectures compared to statistical approaches.

2.3.1 Integration of Transformer in NLP

A transformer is a type of advanced machine learning model that sets itself apart by integrating self-attention, a mechanism that assigns different levels of importance to various elements within the input data, including recursive outputs. This model has extensive utility in the domains of natural language processing (NLP) and computer vision (CV).

As illustrated in Figure 2.4, the architecture of transformer models consists of encoders and decoders. BERT [76] exclusively employs encoders, while GPT [44,77] solely utilizes decoders. Both variants possess the ability to comprehend language, including its syntax and semantics. Particularly, the more recent generation of large-scale language models like GPT, which comprise billions of parameters, excels in this aspect.

The two models focus on different scenarios. However, since the field of foundation models is evolving, the differentiation is often fuzzier.

1. BERT (encoder): classification (e.g., sentiment), questions and answers, summarization, named entity recognition.
2. GPT (decoder): translation, generation (e.g., stories).

The outputs of the core models are different:

1. BERT (encoder): Embeddings representing words with attention to the information in a certain context.
2. GPT (decoder): Next words with probabilities.

Both models come pre-trained, eliminating the need for extensive training. Some models are openly accessible through platforms like Hugging Face, while others are commercially available. The ability to reuse pre-trained models is highly valuable, as training procedures often demand significant resources and expenses, limiting their feasibility to only a few companies.

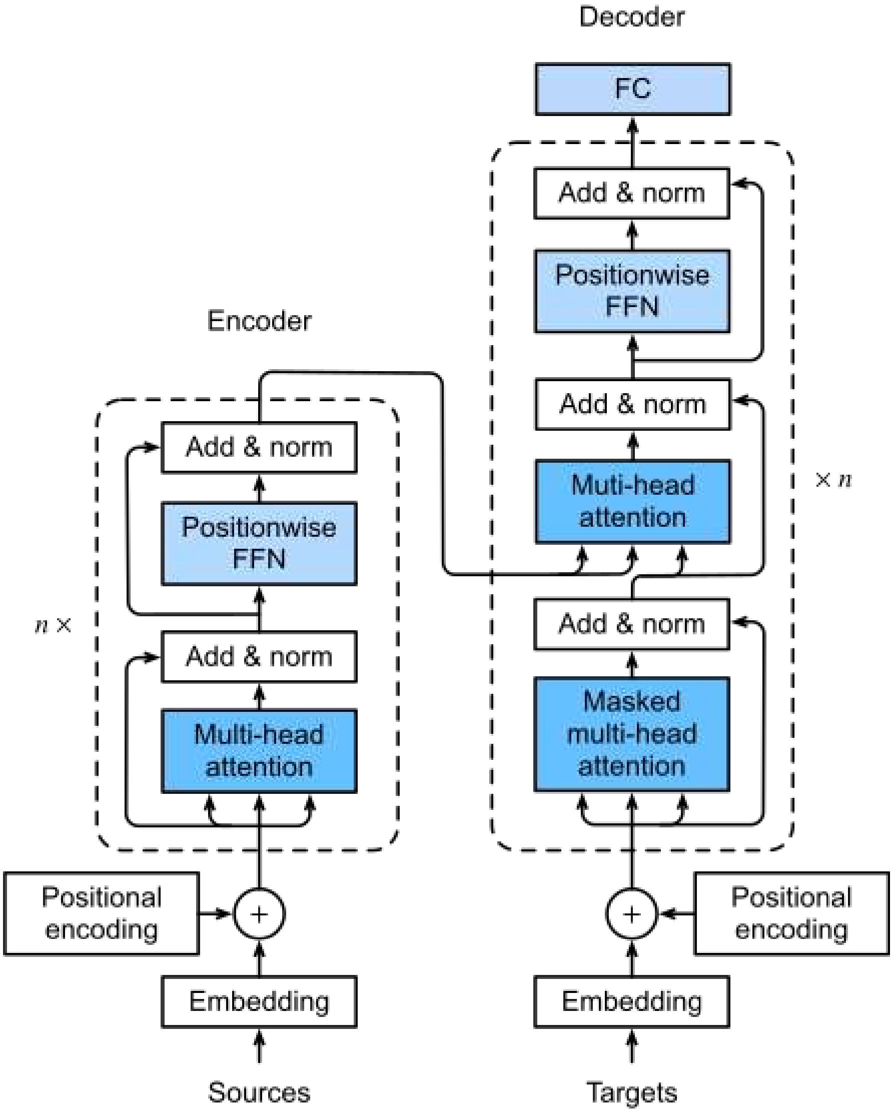


Figure 2.4: Basic architecture of Transformer [1].

The pre-trained models have the flexibility to be extended and tailored to suit various domains and specific tasks. In some cases, the existing layers can be directly reused, while additional layers can be added on top. However, if modifications to the layers are necessary, the process of retraining becomes more costly. This approach, known as transfer learning, allows the general model to be easily transferred and adapted to different domains according to specific requirements.

2.3.1.1 BERT Encoder

BERT utilizes the encoder component of the transformer architecture to comprehend both the semantic and syntactic elements of language. Unlike predicting subsequent words, BERT generates embeddings as its output. To make use of these embeddings, additional layer(s) need to be incorporated, such as for tasks like text classification or question answering.

BERT applies self-supervised learning, a technique that assigns labels to initially unlabeled data. This approach is particularly effective when dealing with large datasets. In Figure 2.5 the basic architecture of BERT encoder is provided.

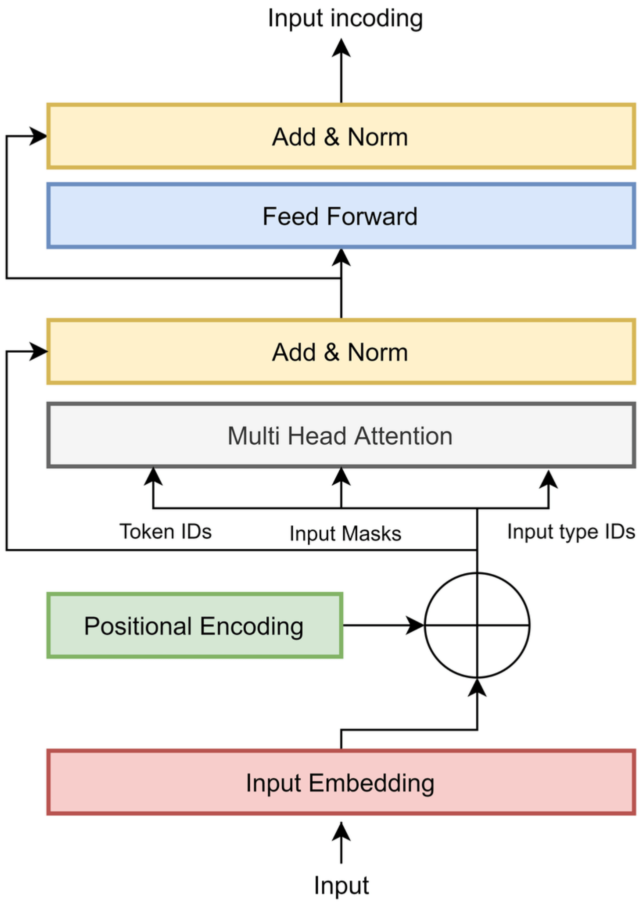


Figure 2.5: Basic architecture of BERT Encoder [2].

2.3.1.2 GPT Decoder

In language-related situations, decoders are employed to generate consecutive words, as seen in tasks like text translation or story generation. The decoder produces words along with their respective probabilities. The fundamental architecture of the GPT decoder is depicted in Figure 2.6.

Decoders also incorporate attention mechanisms, employing them twice in the process. During model training, Masked Multi-Head Attention is utilized, where only the initial words of the target sentence are provided. This approach ensures that the model learns without any form of ”cheating” or access to future information. This mechanism shares similarities with the MASK concept used in BERT.

Afterward, the decoder employs Multi-Head Attention, similar to the encoder. Transformer-based models that consist of both encoders and decoders utilize an intelligent approach to enhance efficiency. The output of the encoders, specifically the keys and values, is passed as input to the decoders. The decoders can generate queries to identify the most pertinent keys. This allows for tasks such as understanding the essence of the original sentence and translating it into different languages, even if the translation differs in terms of word count and order.

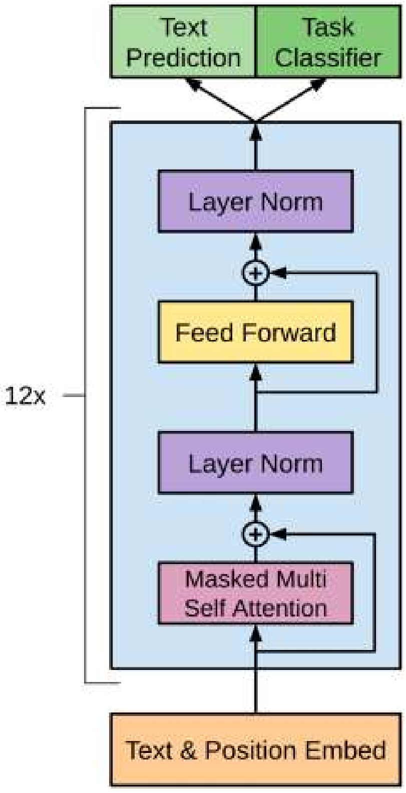


Figure 2.6: Basic architecture of GPT Decoder [1].

methods aim to enhance the image's clarity, resulting in an improved version that is more suitable for the specific application compared to the original image.

2.4.2 Deep Learning in Medical Image Processing

Deep learning has emerged as a robust methodology in the field of medical image processing. It encompasses the utilization of artificial neural networks, specifically deep neural networks consisting of multiple layers, to analyze and interpret medical images. Here are some key aspects of deep learning in medical image processing:

1. **Image Classification and Diagnosis:** Deep learning models have the capability to undergo training for the purpose of classifying medical images into various categories, including the identification of tumors, lesions, or specific diseases. These models possess the ability to learn intricate patterns and features directly from the images, facilitating precise diagnosis and automated detection of diseases [78–83].

2. **Segmentation:** Deep learning techniques can perform image segmentation, which involves dividing an image into meaningful regions or identifying specific structures within the image. This is particularly useful for tasks like organ segmentation, tumor delineation, or extracting anatomical features [84–88].

3. **Image Reconstruction:** Deep learning-based methods can reconstruct high-quality medical images from noisy or incomplete data. By leveraging large amounts of training data, deep learning models can fill in missing information or enhance low-resolution images, aiding in improved visualization and analysis [89–92].

4. **Image Registration:** Deep learning algorithms can be employed to align or register multiple medical images acquired from different modalities or time points. This enables precise comparison, fusion, or tracking of anatomical structures or pathological changes [90, 93–95].

5. **Disease Progression Modeling:** Deep learning models can learn temporal or sequential patterns from longitudinal medical images, facilitating disease progression modeling and prediction. This can aid in personalized treatment planning and monitoring the effectiveness of therapies [96–98].

6. **Transfer Learning and Pretrained Models:** Deep learning frameworks allow the transfer of knowledge from models pretrained on large-scale datasets to specific medical image analysis tasks. This helps overcome the challenge of limited

labeled medical data and improves the performance of models in medical image processing applications [99–102].

Deep learning in medical image processing has shown promising results and has the potential to revolutionize healthcare by enabling more accurate diagnoses.

Chapter 3

Medical Data Processing

3.1 Machine Learning Classifiers for Text Processing

Text classification in machine learning involves the automatic assignment of tags or categories to text data. To train a classifier using machine learning techniques, it is necessary to convert the text into a format that can be understood by the algorithm. In many cases, the text is vectorized, which means it is represented by a numerical vector that captures the frequency of words from a predefined list.

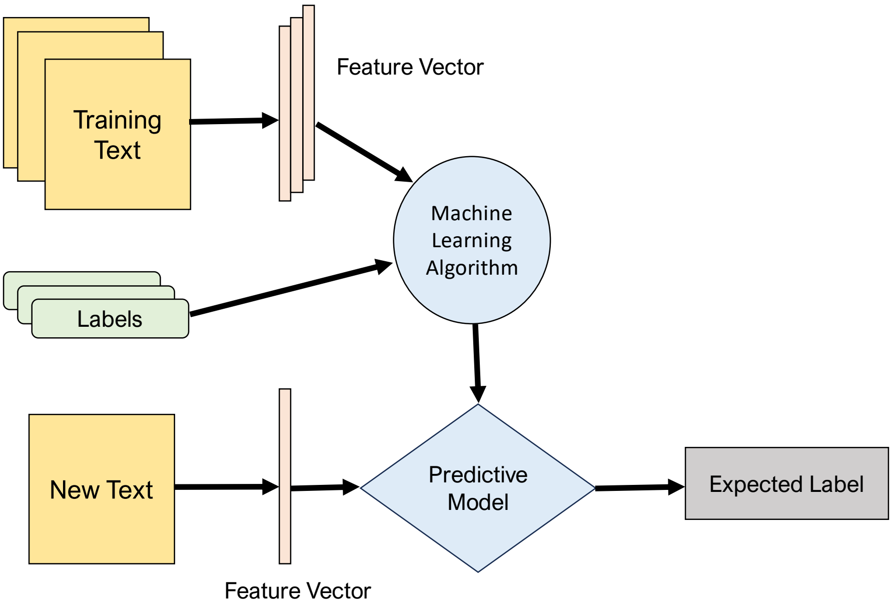


Figure 3.1: Machine Learning based approach for text processing.

The machine learning based text processing approach is shown in Figure 3.1. Once the text is vectorized, the text classifier is trained using a dataset that includes feature vectors for each text sample along with their corresponding tags. By providing enough training samples, the model can learn patterns and relationships between the text features and the assigned tags, enabling it to make accurate predictions on new, unseen data.

Machine learning algorithms have the capability to undergo training for the purpose of categorizing text into predefined classes or categories. For instance, sentiment analysis involves determining the sentiment of a text (positive, negative, or neutral), while spam detection aims to classify emails or messages as spam or non-spam. Named Entity Recognition (NER) is another application where machine learning models identify and classify named entities, such as names, organizations, locations, and dates, within a text. By training these models, various applications like information retrieval and question-answering systems can effectively extract and classify these entities. Machine learning models, particularly sequence models like recurrent neural networks (RNNs) or transformers, can be trained to generate text. This includes applications such as language modeling, dialogue generation, or text summarization, where the models learn to produce coherent and contextually relevant text based on training data. Additionally, machine learning algorithms can group similar documents together through text clustering, aiding in the organization of large collections of text and the identification of common themes or topics. Topic modeling algorithms like Latent Dirichlet Allocation (LDA) or Non-negative Matrix Factorization (NMF) can uncover latent topics within a collection of documents. By training these models, underlying topics or themes present in the documents can be revealed, providing insights into the content and facilitating further analysis. Furthermore, machine learning models can determine the sentiment expressed in a text, such as positive, negative, or neutral. This is valuable in applications such as social media analysis, customer feedback analysis, or brand monitoring, where understanding the sentiment of the text can provide valuable insights. Concise summaries of larger texts can be generated using machine learning techniques. These techniques involve extracting important sentences or phrases that capture the main essence of the original content, enabling the creation of meaningful and condensed summaries. Moreover, machine learning models, especially neural machine translation models, have revolutionized automated translation systems. These models learn to translate text from one language to another by training on large parallel corpora, improving the accuracy and fluency of automated translations.

Machine learning in text processing allows computers to process and understand textual data more efficiently, making it invaluable for a wide range of applications in natural language processing, information retrieval, content analysis, and many other fields.

3.2 NLP Based Techniques for Text Processing

Radiology reports are analyzed using different NLP based techniques to extract clinically significant findings [103–105]. There are different NLP based Methods available for text feature extraction in different classification, diagnosis, prognosis and even report generation for radiological purpose which is shown in Figure 3.2.

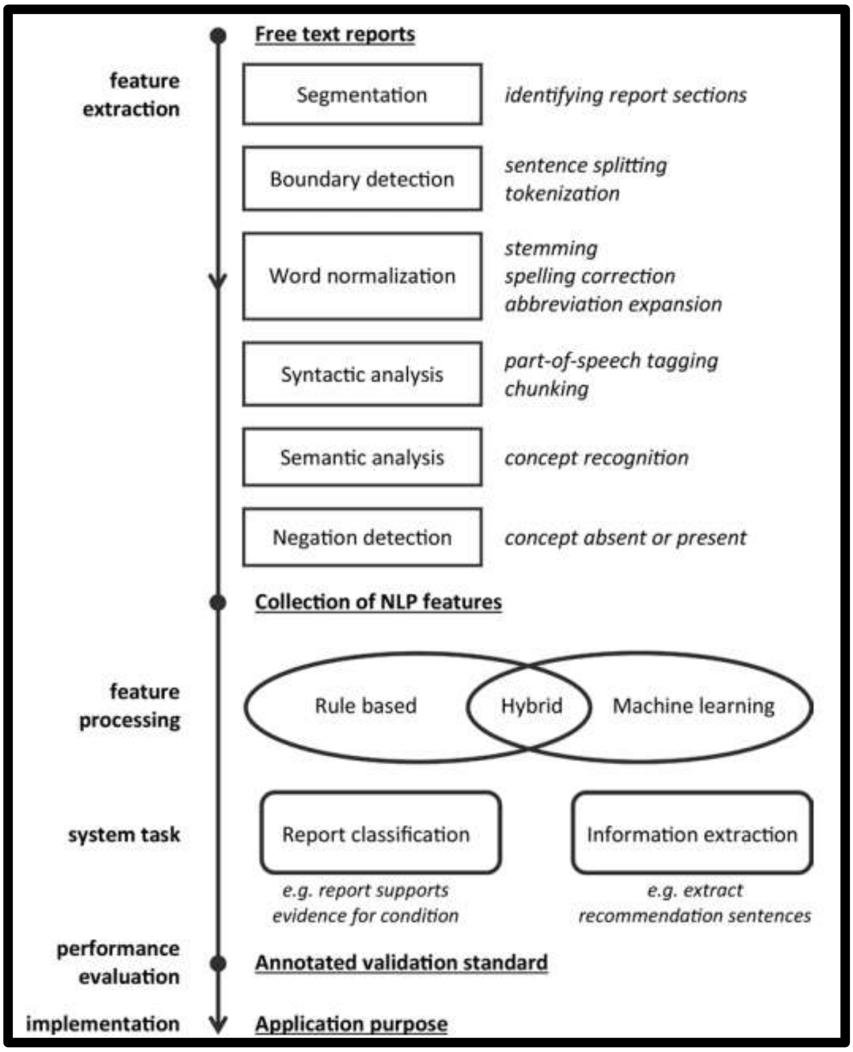


Figure 3.2: Natural Language Processing for radiology text processing [4].

3.2.1 Rule Based Techniques

Rule-based natural language processing methods require substantial manual effort and are not easily reusable. However, formal rule-based approaches prove valuable in extracting intricate and structured templates, yielding reliable results. An example of such a system is Lexicon Mediated Entropy Reduction (LEXIMER) [106], which employs lexicon-based hierarchical decision trees to extract and classify phrases. There are also general natural language processing (NLP) techniques that have been previously used for classification and information extraction from radiology reports. MedLEE [107], a rule-based system, utilizes semantic lexicons and the results of basic syntactic analysis. The automatic learning of complex structures poses challenges, especially when there is a limited availability of semantically annotated clinical data sources and restricted access to clinical data for training and evaluation purposes. Taira and Soderland developed a system [108, 109] that applies a maximum entropy classifier for sentence boundary identification, a lexical analyzer based on manually created lexicons with syntactic and semantic features, a statistical parser for generating dependency structures, two types of semantic interpreters, and a rule-based frame filler. The most promising approach involves combining symbolic and machine-learning techniques. MPLUS [110], for instance, integrates semantic analysis based on Bayesian networks and syntactic analysis based on context-free grammar. This technology has found applications in extracting medical data from Head CT reports.

3.2.2 Pre-trained NLP Models

Transfer learning [111] was developed to address the challenge of utilizing representations that were initially trained on extensive unannotated datasets and subsequently fine-tuning them for specific tasks. A recent trend in transfer learning involves employing self-supervised learning on large general datasets to create a versatile pre-trained model that captures the underlying structure of the data [112–116]. Following the initial pre-training on a diverse dataset, this model can be further customized and fine-tuned to suit a specific task using a particular dataset. The effectiveness of this pre-training and fine-tuning approach has been prominently demonstrated in the fields of natural language processing (NLP) and, more recently, computer vision.

3.2.2.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) [117] is a Machine Learning (ML) model specifically designed for natural language processing tasks (Figure 3.3). Developed by researchers at Google AI Language in 2018, BERT stands out for its ability to pre-train deep bidirectional representations from unlabeled text, taking into account both left and right context information across all layers. This unique architecture enables the pre-trained BERT model to be fine-tuned with an additional output layer, allowing the creation of state-of-the-art models for various tasks such as question-answering and language inference. BERT builds upon previous advancements in pre-training contextual representations, drawing inspiration from works like Semi-supervised Sequence Learning, Generative Pre-Training, ELMo, and ULMFit. Notably, these earlier models were either unidirectional or had limited bi-directionality.

For using BERT for pre-training, in this experiment learning rate 2×10^{-5}, sparse categorical cross-entropy loss is used. It is fine-tuned for additional tasks with an additional classification layer on the above encoder.

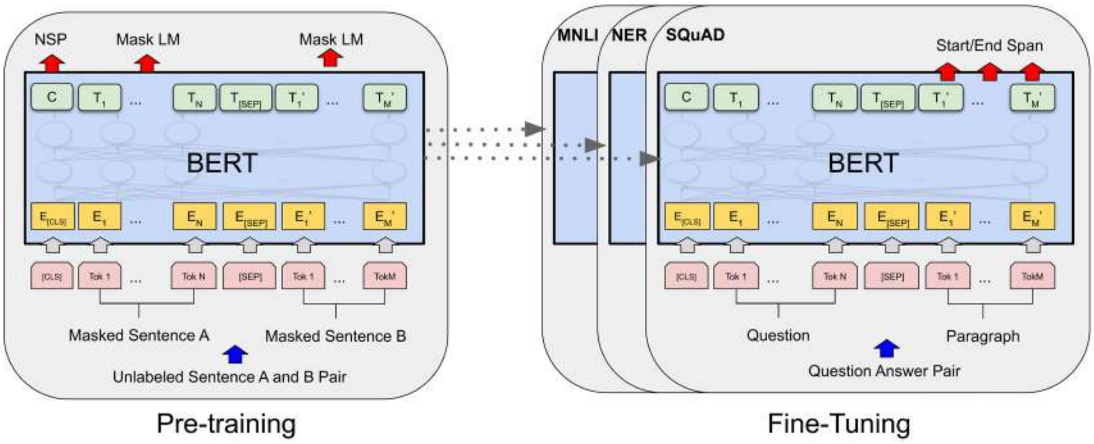


Figure 3.3: BERT model specialized for pre-training and fine-tuning [5].

3.2.2.2 ClinicalBERT

ClinicalBERT [32] is a customized version of the BERT model (Figure 3.4). It is specifically designed to learn representations from medical notes and utilize them for various clinical applications. During training, ClinicalBERT utilizes clinical notes and Electronic Health Records (EHR) data, allowing it to develop a comprehensive understanding of the qualitative relationships among different clinical concepts within a database of medical terms. This specialized training enables ClinicalBERT to capture the unique nuances and complexities of medical language, making it a valuable tool in the healthcare domain.

1. **Clinical Text Embeddings:** Clinical BERT processes clinical notes by taking an array of tokens as input. These tokens are obtained through a preprocessing step where the text is segmented into subword units [118]. Within Clinical BERT, each token in a clinical note is decomposed into three components: segment embedding, location embedding, and token embedding. The segment embedding indicates the specific sequence that the token belongs to when multiple token

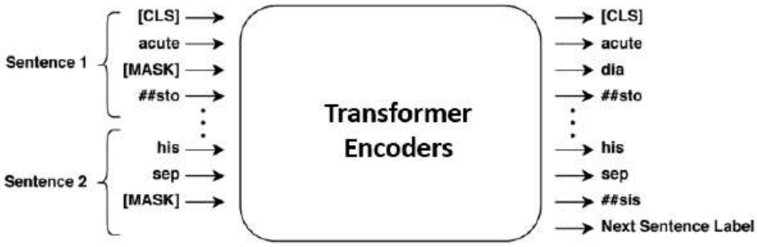


Figure 3.4: Clinical BERT learns deep representations of clinical text using masked language modeling and next-sentence prediction. In masked language modeling, a fraction of input tokens are held out for prediction; in next sentence prediction, Clinical BERT predicts whether two input sentences are consecutive.

sequences are inputted into Clinical BERT. The location embedding consists of learned parameters that encode the token's position in the input sequence (position embeddings are shared among all tokens). Furthermore, for classification tasks, a classification token [CLS] is inserted at the beginning of each input token sequence.

2. **Self-Attention Mechanism:** The embeddings corresponding to the input tokens are employed to compute the attention function on an input sequence. The attention function takes sets of queries, keys, and values as input. The queries, keys, and values are generated by multiplying the input embeddings with learned sets of weights. This type of attention, where the queries, keys, and values are derived from the same input, is referred to as "self-attention." The output of the attention function is a weighted combination of values for each query. The weight assigned to a value is determined by the corresponding query and key, enabling the model to capture relevant relationships and dependencies between different elements of the input sequence. The Attention function can be denoted such as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}}V) \tag{3.1}$$

   where Q = queries, K = keys, V = values and d = dimensionality of the queries, keys, and values. This function can be computed efficiently and can capture long-range interactions between any two elements of the input sequence [119].

3. **Pre-training ClinicalBERT:** The quality of learned text representations is heavily influenced by the nature of the text data used to train the model. BERT, as an example, is trained on BooksCorpus and Wikipedia, which differ significantly from clinical notes in terms of content and language style. In contrast, understanding clinical notes can be challenging without domain-specific knowledge

and expertise in the medical field. The specialized terminology, complex medical concepts, and unique linguistic patterns present in clinical notes require tailored approaches to effectively comprehend and process this type of text data.

Clinical BERT adopts the same pre-training tasks as described in Devlin et al. (2018) [5]. One of these tasks is masked language modeling, where certain tokens in the input are randomly masked, and the model is trained to predict the masked tokens based on the context provided by the surrounding tokens. Another task is next sentence prediction, where the model is presented with pairs of sentences and learns to predict whether they are consecutive in the original text. During pre-training, the objective function consists of the sum of the log-likelihood of the predicted masked tokens and the log-likelihood of a binary variable indicating the consecutiveness of the sentences. Specific parameters are used for pre-training, including a batch size of 32, a maximum sequence length of 256, and a learning rate of 5×10^{-5}.

4. **Fine-Tuning ClinicalBERT:** During the fine-tuning process, ClinicalBert is specifically adapted for the task of thoracic disease classification. The model parameters are adjusted to optimize the log-likelihood of the multiclass classification, aiming to accurately assign the input samples to their corresponding disease categories. In the classification output layer, there are 14 nodes, each representing a specific disease class, and the sigmoid activation function is applied to the outputs of these nodes. This allows for the prediction of the probability of each disease class independently, enabling the model to make predictions for multiple diseases simultaneously.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{3.2}$$

Here, σ is the activation function of the output layer.

3.2.2.3 Med-BERT

Med-BERT [120] is like BERT which has multi-level embeddings and bidirectional transformers (Figure 3.5). In [120] they have mentioned the adaptation of similar pre-training techniques using EHR. The original BERT model was designed to process a 1-D sequence of words as input. However, Electronic Health Records (EHR) used for pre-training contain complex multilayered and multi-relational information. Converting such structured EHR data into a flattened 1-D sequence and encoding the inherent ”structures” within the BERT transformer architecture poses challenges, as there are no established guidelines or rules for this process.

Med-BERT utilizes three distinct types of embeddings extracted from the Electronic Health Record (EHR) dataset as input features. These embeddings include code embeddings, serialization embeddings, and visit embeddings. Code embeddings are generated based on diagnosis codes found within the EHR, capturing the specific representation of each code. Serialization embeddings, on the other hand, encode the sequential order of codes within each visit, providing information about the temporal sequence of medical events. Lastly, visit embeddings represent the position of each visit within the dataset, allowing the model to understand the relative significance and context of each visit in relation to the entire dataset. By incorporating these three types of embeddings, Med-BERT captures both the diagnostic codes and the temporal structure of the EHR data, enabling comprehensive analysis and interpretation of patient records.

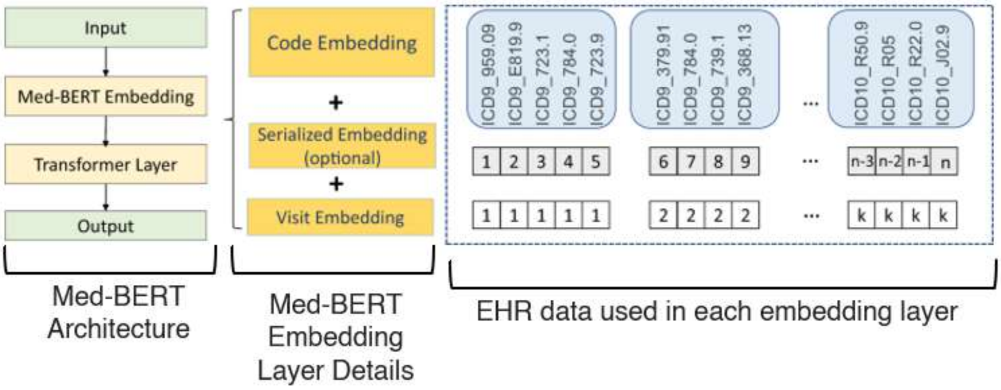


Figure 3.5: Med-BERT structure details.

3.3 Convolutional Neural Network (CNN) for Text Processing

Convolutional Neural Networks (CNNs) have primarily been utilized in the field of image processing, but they can also be applied to text processing tasks, such as text classification and sentiment analysis. When employed in text processing, CNNs operate on one-dimensional input instead of two-dimensional images. In text applications, each word or token is typically represented as a vector, such as word embeddings like Word2Vec or GloVe, which capture the semantic relationships between words. These vectors form the input sequence for the CNN.

In the text-based CNN architecture, convolutional operations are applied to the input sequence to extract local patterns or features. The convolutional layer employs multiple filters of fixed sizes that slide over the input sequence, performing element-wise multiplications and summations to generate feature maps. To introduce non-linearity, activation functions like ReLU (Rectified Linear Unit) are applied to the extracted features. Max pooling is commonly employed to reduce the dimensionality of the feature maps while preserving the most important information.

The resulting pooled feature maps are then flattened and passed through one or more fully connected layers, also known as dense layers. These layers learn higher-level representations by combining the extracted features. Finally, a softmax layer is typically used as the output layer, producing a probability distribution over pre-defined classes or labels in text classification tasks. The predicted label is determined by selecting the class with the highest probability.

By adapting CNNs to text processing, researchers have extended the capabilities of these models beyond image analysis, enabling them to effectively handle various natural language processing tasks.

CNNs for text processing offer several advantages. They can effectively capture local patterns and dependencies within the input sequence, allowing them to model text structures and identify important features. Additionally, by utilizing pre-trained word embeddings, they can leverage semantic information encoded in the word vectors.

However, it's worth noting that CNNs might not capture long-range dependencies or sequential information as effectively as recurrent neural networks (RNNs) or transformer-based models. Therefore, CNNs are often employed for tasks that primarily rely on local context and surface-level features.

To utilize text as input in a CNN architecture, it is necessary to treat the text data as sequential data, similar to time series data, forming a one-dimensional matrix. In such cases, a one-dimensional convolutional layer is employed. To effectively process text, a word embedding layer is utilized in conjunction with the one-dimensional convolutional neural network.

Word embeddings serve as representations of word density and can be generated using the Keras TensorFlow module. These embeddings are then used as input in the CNN architecture. Within the CNN architecture, the convolutional process involves extracting patches of input features with the size of the filter kernel. The dot product is computed between the weights of the filter and the multiplied values of the patch. The one-dimensional convnet layer exhibits invariance to translations, meaning that certain sequences can be recognized regardless of their position. This property can be advan-

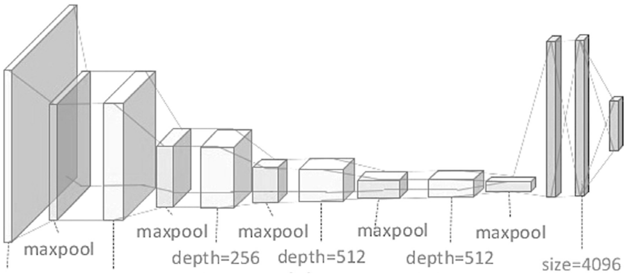


Figure 3.6: Convolutional Neural Network (CNN) architecture for text classification purpose.

tageous for identifying specific patterns within the text. The CNN structure that is used in experiments is showed in Figure 3.6.

Chapter 4

Medical Image Processing for Disease Classification

4.1 Medical Image Pre-processing

The initial phase of image analysis, known as the pre-processing step, holds significance in the overall scheme. Its purpose is to improve the original image by minimizing noise and eliminating undesired elements. To achieve this, histogram equalization is employed to expand the pixel's intensity range from its initial scale to a new scale ranging from 0 to 255. This expansion results in an enhanced image that exhibits a broader range of intensity and slightly increased contrast. Additionally, the images undergo cropping and resizing to dimensions of 224×224.

4.2 Natural Language Processing (NLP) Based Techniques for Image Classification

There are some transformer based architectures for medical image calssification. The Vision Transformer (ViT) [121] framework, introduced by Dosovitskiy et al. in 2020, represents a pioneering approach in utilizing Transformer architecture for achieving outstanding results in medical image classification tasks. Inspired by the BERT architecture initially designed for language understanding, ViT adapts this Transformer-based approach to the context of image classification. In ViT, images are segmented into rectangular patches, treating each patch as a token. Subsequently, embeddings are computed for these patches. To capture the spatial structure of the image, positional

architecture with transformers. By incorporating transformers, TransUNet improves modeling capabilities by effectively capturing long-range dependencies in the image. It combines the strengths of both convolutional neural networks (CNNs) and transformers to achieve enhanced performance.

TransFuse, presented by Zhang et al. [126], adopts a parallel approach by utilizing both transformers and CNNs for medical image segmentation. This architecture leverages the unique abilities of transformers to capture global context information, while also benefiting from the feature extraction capabilities of CNNs.

Segtran, described by Li et al. [127], is a medical image segmentation system that harnesses the power of transformers. By incorporating transformers into the architecture, Segtran is able to capture both the global context and fine-grained details within medical images. This comprehensive approach leads to superior segmentation performance.

In the field of image denoising, Zhang et al. [128] proposed Transct, a neural network architecture based on transformers. Specifically designed for low dose computed tomography (LDCT) images, Transct utilizes transformers to explore and address the long-range dependencies between pixels. By leveraging the capabilities of transformers, Transct aims to achieve effective denoising by effectively capturing the relationships and dependencies within the image.

4.3 Convolutional Neural Network (CNN) Based Techniques for Image Classification

4.3.1 ResNet-50

ResNet-50, introduced by Koonce et al. [129], is a convolutional neural network (CNN) architecture that consists of 34 weighted layers. It addresses the issue of vanishing gradients in deep CNNs by incorporating shortcut connections, which are based on two core principles. First, the number of filters within each layer remains constant, determined by the size of the output feature map. Second, when the dimensions of the feature map are reduced by half, the number of filters is doubled to maintain the time complexity of individual layers. These shortcut connections allow for the efficient flow of gradients during training, enabling the successful training of deep networks.

ResNet-50 is widely recognized as one of the most commonly used pre-trained models in computer vision tasks. It is frequently employed for feature extraction and fine-tuning purposes. Specifically, the pre-trained ResNet-50 model trained on the ImageNet dataset [22] is commonly utilized. The ImageNet dataset is a large-scale collection of 14,197,122 images with annotations and serves as a benchmark for image classification and object detection tasks, including the ImageNet Large Scale Visual Recognition Challenge (ILSVRC).

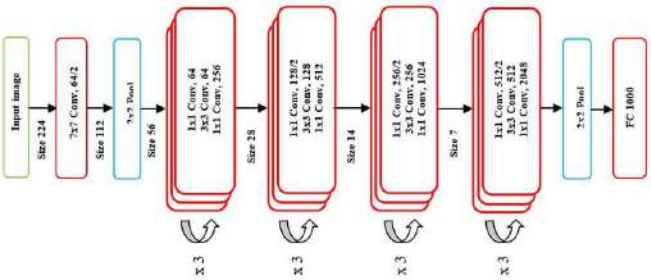


Figure 4.2: Basic architecture of ResNet-50 [7].

4.3.2 DenseNet-121

DenseNet-121, proposed by Rochmawanti et al. [8], is a convolutional neural network (CNN) architecture consisting of a total of 120 convolutional layers and 4 average pooling layers (Figure 4.3). The unique characteristic of DenseNet is the incorporation of DenseBlocks, which connect each layer to every other layer in a feed-forward manner. This dense connectivity pattern enables feature reuse and facilitates gradient flow, mitigating the vanishing gradient problem often encountered in deep neural networks.

Within the DenseNet architecture, the number of filters between feature maps is adjusted while maintaining consistent dimensions. Transition Layers are introduced to reduce the number of channels by half between these DenseBlocks. Specifically, DenseNet-121 comprises one 7×7 Convolution layer, fifty-eight 3×3 Convolution layers, sixty-one 1×1 Convolution layers, 4 Average pooling layers, and one fully connected layer.

By simplifying the connectivity pattern and promoting feature reuse, DenseNet-121 addresses the vanishing gradient problem and facilitates effective gradient flow during training.

4.4 Anatomy Aware Neural Networks for Image Classification

In the field of medical practice, the accurate interpretation of chest X-rays and other medical imaging modalities requires a comprehensive understanding of the underlying

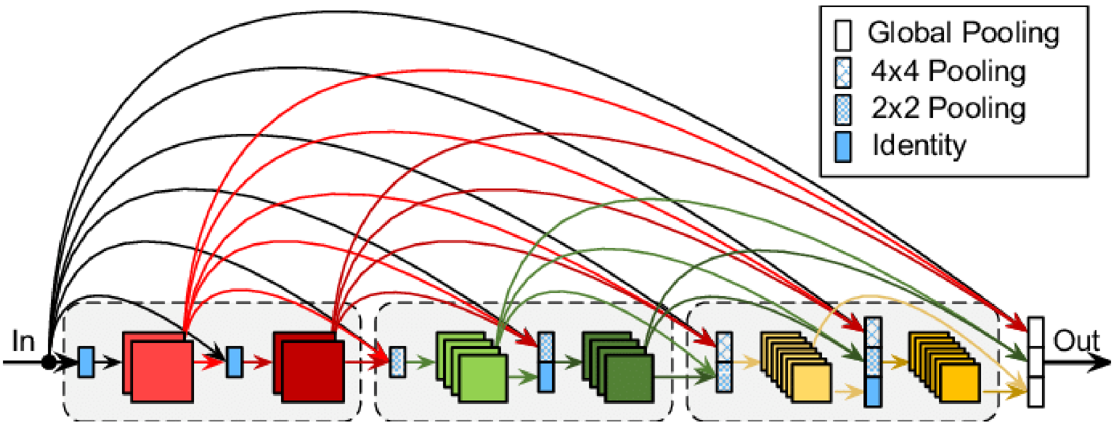


Figure 4.3: Basic architecture of DenseNet-121 [8].

human anatomy being captured. For instance, when analyzing chest X-rays, radiologists examine various anatomical structures such as the position of the trachea, the expansion of the lungs, the clarity of the lung fields, and the size of the heart. These observations form the foundation of chest X-ray interpretation based on human visual perception, emphasizing the significance of anatomical knowledge. However, previous research in automated chest X-ray analysis often overlooks this aspect and treats the problem as a standard computer vision task. Most existing studies adopt a global learning strategy or employ attention mechanisms to identify spatial regions that contribute more to the model's predictions.

To the best of our knowledge, there has been limited exploration of integrating anatomy information into automated chest X-ray analysis models. Unlike generic image classification, the problem of thoracic disease classification involves identifying lesion regions that are often associated with important anatomical structures. It is therefore desirable for the model to discover and focus on these salient lesion regions by leveraging prior knowledge of anatomy when making predictions.

For successful integration of deep learning techniques in the classification of thoracic diseases, it is imperative to achieve not only higher accuracy but also ensure interpretability. In real-world scenarios, radiologists typically identify and prioritize essential anatomical regions before assessing them for any abnormalities. While extensive research has been conducted in this area, most studies have focused on determining the spatial regions that contribute to the model's predictions. However, the scarcity of accurately annotated anatomical regions at a contour-level for large-scale datasets presents a significant challenge. The currently available annotated segmentation datasets are relatively small, limiting the generalizability of models trained using supervised learning approaches. To address these limitations, a novel architecture called Anatomy X-Net [9], which incorporates anatomical segmentation knowledge into different stages

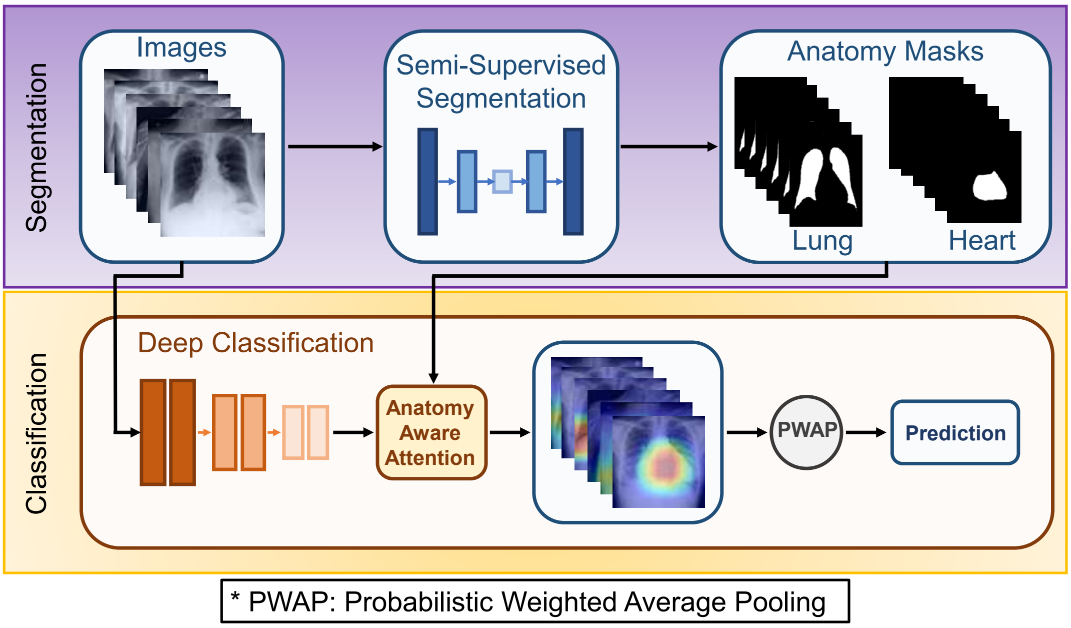


Figure 4.4: Overview of the proposed semi-supervised anatomy-aware attention-based thoracic disease classification framework. A semi-supervised technique is utilized to generate anatomy masks for unannotated CXR images. Then, with the help of our proposed novel anatomy-aware attention module, anatomical information is integrated into the classification network for pathology detection [9].

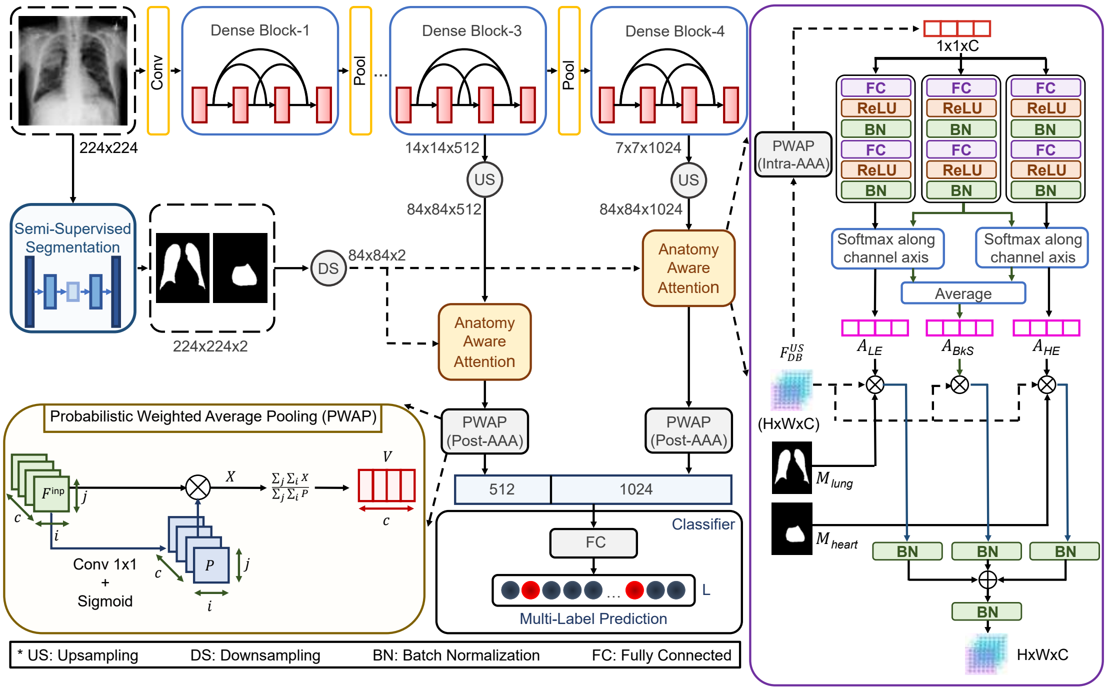


Figure 4.5: Basic architecture of Anatomy-XNet [9].

of the network, has been proposed. By utilizing information from anatomical segmentation, this architecture effectively prioritizes the spatial regions that are commonly associated with various pathologies. Additionally, the hierarchical feature-fusion-based network, guided by the anatomical segmentation information, learns to leverage both coarse-grained and fine-grained features, resulting in improved classification performance.

Anatomy-XNet [9] is a thoracic disease classification network that incorporates anatomical awareness and attention-based mechanisms (Figure 4.5). It gives importance to spatial characteristics by utilizing pre-identified regions of anatomy. To address the lack of organ-level annotations in extensive datasets, a semi-supervised learning method is utilized. This approach leverages limited organ-level annotations available to localize the anatomy regions. The network architecture incorporates a pre-trained DenseNet-121 as its foundation, alongside two structured modules: Anatomy Aware Attention (AAA) and Probabilistic Weighted Average Pooling (PWAP). These modules synergistically work together within a unified framework to facilitate the learning of anatomical attention. Through its implementation, Anatomy-XNet achieves impressive accuracy in classifying thoracic diseases using chest X-ray images.

Chapter 5

Anatomy Aware Feature Fusion Based Framework

5.1 Motivation of Incorporating Anatomical Information

A deep learning model with an anatomy-aware (AA) approach is employed to extract generic features from x-ray images, taking into account the anatomical information present. By utilizing a pre-trained model and lung segmentation masks, the model generates a feature vector that encompasses features related to diseases and scores indicating the extent of lung involvement.

COVID-19 pneumonia primarily affects the density of the lungs, resulting in areas of increased whiteness in radiography images, which varies depending on the severity of the pneumonia. When hazy gray areas partially obscure the dark lung markings in chest X-rays (CXR), this is known as ”ground-glass opacity”. Ground-glass opacity refers to a hazy increase in attenuation in the interstitial and alveolar processes of the lungs. Additionally, linear opacities, such as peripheral, coarse, horizontal white lines, bands, or reticular changes, may be present alongside ground-glass opacity. In severe cases, the lung markings may be completely obscured, leading to a condition known as consolidation, characterized by a complete whiteness. These changes are typically observed in the peripheral and lower zones of the lungs, although the entire lung can be affected. Bilateral lung involvement is commonly observed in COVID-19 cases. The presence of nodules, pneumothorax, or pleural effusion may be incidental findings in COVID-19 cases.

Given that COVID-19 infection primarily affects the lungs, a trained radiologist naturally approaches the interpretation of a COVID-19 CXR image by first identifying the anatomical structure of the lung. Similarly, an algorithm informed by lung anatomy can analyze CXR images for more accurate analysis and subsequent prediction of disease severity. For instance, if the lung regions are pre-identified, deep learning models can be trained using higher-resolution images specific to the relevant portion of the images. This type of approach is referred to as "anatomy-aware," where the algorithm takes into account the anatomical context of the lungs to enhance the analysis of CXR images [130].

5.2 Data Resources and Organization

Four different datasets are used for the training and evaluation of the proposed anatomy-aware framework. These include chest x-ray image datasets including anatomy segmentation masks and disease labels. The datasets as follows:

5.2.1 JSRT dataset

The JSRT dataset consists of 247 images (154 nodule and 93 non-nodule images), with a resolution of 2048×2048. This dataset also includes patients information such as: age, gender, diagnosis (malignant or benign), X and Y coordinates of nodule, simple diagram of nodule location.

5.2.2 SCR dataset

SCR dataset is a database of posterior-anterior chest radiographs where the manual segmentation of lungs, heart, and clavicles are provided. This dataset includes chest radiographs of 247 subjects where annotations of the anatomical structures of the images of the JSRT database, e.g., left lung, right lung, heart, left clavicle, right clavicle are included.

5.2.3 Stanford Chexpert dataset

The Chexpert dataset [131] is a large public dataset for chest x-ray interpretation. This dataset contains $224,316$ radio-graphic images of $65,240$ patients labeled with 14 observations such as: Atelectasis, Cardiomegaly, Consolidation, Edema, Pleural Effusion, Pneumonia, Pneumothorax, Enlarged Cardiom., Lung Lesion, Lung Opacity, Pleural Other, Fracture, Support Devices, No Finding.

Table 5.1: Datasets required for the training and evaluation of the proposed framework.

Dataset	Number of CXRs
JSRT Dataset [132]	247 CXRs
SCR Dataset [133]	247 annotated CXRs
Stanford Chexpert Dataset [131]	224,316 CXRs
COVID-19 Pneumonia Severity Dataset [13]	94 images with scores
In-house Dataset	12 independent annotated CXRs

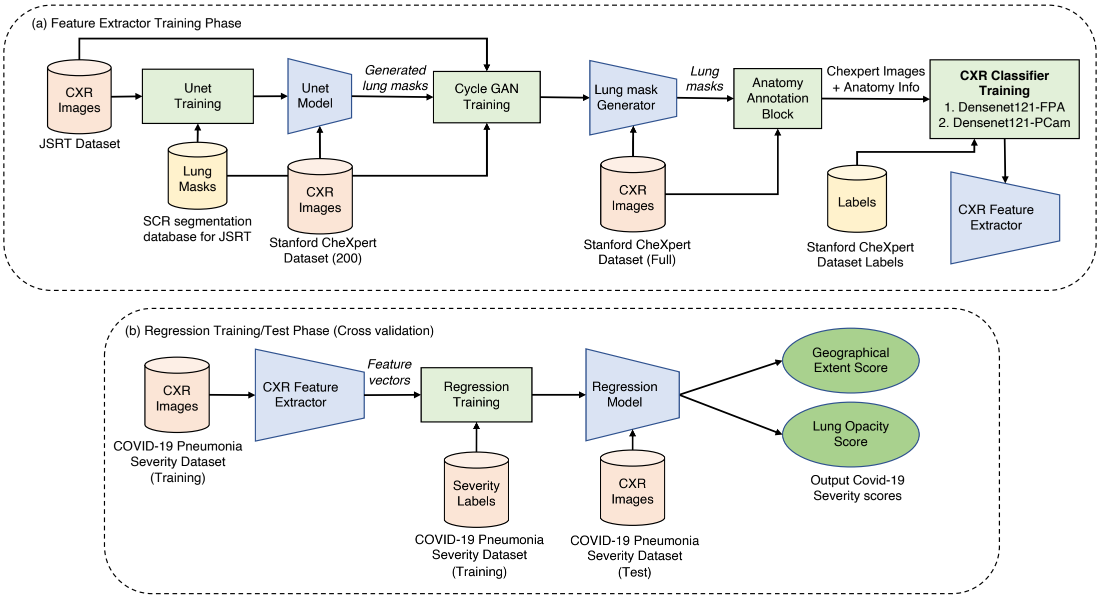

(a) Feature Extractor Training Phase
CXR Images
JSRT Dataset
Unet Training
Lung Masks
SCR segmentation database for JSRT
Unet Model
Generated lung masks
CXR Images
Stanford CheXpert Dataset (200)
Cycle GAN Training
Lung mask Generator
CXR Images
Stanford CheXpert Dataset (Full)
Lung masks
Anatomy Annotation Block
Chexpert Images + Anatomy Info
CXR Classifier Training
1. Densenet121-FPA
2. Densenet121-PCam
Labels
Stanford CheXpert Dataset Labels
CXR Feature Extractor
(b) Regression Training/Test Phase (Cross validation)
CXR Images
COVID-19 Pneumonia Severity Dataset (Training)
CXR Feature Extractor
Feature vectors
Regression Training
Severity Labels
COVID-19 Pneumonia Severity Dataset (Training)
Regression Model
CXR Images
COVID-19 Pneumonia Severity Dataset (Test)
Geographical Extent Score
Lung Opacity Score
Output Covid-19 Severity scores

Figure 5.1: Dataset organization and overall flow diagram in the feature extractor training and regression training/testing phases.

5.3 Anatomy Aware Network for Severity Prediction

The proposed anatomy-aware (AA) model comprises two main components: the pre-processing model and the backbone model. The pre-processing model, called the Anatomy Annotation model, incorporates the anatomical structure information from chest X-rays (CXRs), while the backbone model is responsible for dense pooling. Subsequently, the feature map undergoes a gating system that consists of two paths.

One path utilizes the Feature Pyramid Attention (FPA) [134] module to enhance pixel-level attention and extract features. This processed feature map is then pooled to obtain disease-level features (feat 1). The other path employs Probabilistic Class Activation Map (P-CAM) pooling [135] which has proven to possess excellent localization capability. With P-CAM pooling, disease-specific heatmaps are generated, and the degree of lung involvement for each disease is calculated as a decimal value. These disease-level features and the lung coverage percentages obtained from the heatmaps are used as feature vectors and concatenated to form the final feature vector. Finally, a linear regression model is applied to this feature vector to derive COVID-19 severity scores.

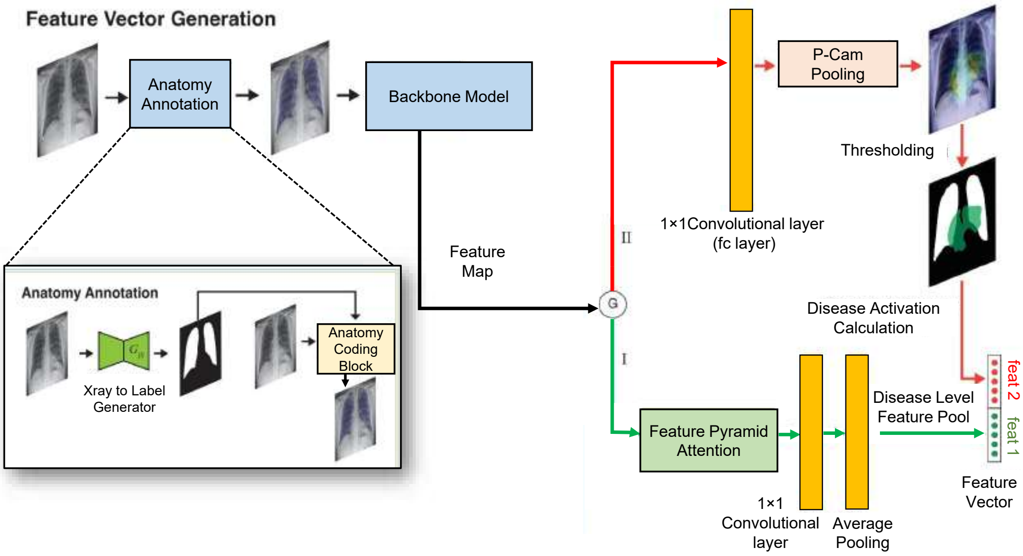


Figure 5.2: The proposed Anatomically-Aware Network for COVID-19 severity prediction can be summarized in this figure. In path 1, Feature Pyramid Attention is employed to improve pixel-level attention, resulting in the generation of disease-level features for five specific diseases (Atelectasis, Consolidation, Edema, Pleural Effusion, and Consolidation). On the other hand, path 2 is utilized for generating class-wise heatmaps using PCAM pooling. These heatmaps, along with the lung masks, are used to calculate the lung involvement score. Finally, the disease-level features and lung involvement scores are utilized for regression analysis.

Anatomy Annotation block

To incorporate the anatomical features in the AA model it is important to emphasize on different anatomical structures, e.g., lungs, heart, ribs, clavicles, diaphragm, etc. In this work, we present a novel pre-processing method termed as *Anatomy Informed Annotation*.

To perform lung segmentation, we employ a Cycle-GAN-based semi-supervised method that has shown superior performance compared to current methods for this task [136].

Let $\mathbf{x}$ be a chest X-ray image and $\mathbf{y}$ be the generated segmentation mask. The output of the neural network, which performs anatomy-informed segmentation, can be expressed as a function:

$$\mathbf{y}_{i,j} = \prod_{i,j} \mathbf{f}_{ks}\big((\mathbf{x}_{(s)i+(\delta)i,(s)j+(\delta)j})_{0\leq(\delta)i,(\delta)j\leq k}\big) \tag{5.1}$$

where k is the kernel size, s is the stride or subsampling factor, and $\mathbf{f}_{ks}$ is the layers of the neural network which is determined by the layer type. The output $\mathbf{y}$ is used for merging anatomy information in the original CXR image, $\mathbf{x}$. We can denote the RGB vector image as three separate column vectors, like, $\mathbf{x}$=[$\mathbf{x}^0$ $\mathbf{x}^1$ $\mathbf{x}^2$]. These column vectors can be denoted as three RGB channel column matrices. We can write as, $[\mathbf{x}^0\ \mathbf{x}^1\ \mathbf{x}^2] = [\mathbf{X}^R\ \mathbf{X}^G\ \mathbf{X}^B]$. Then this data vector is infiltrated by a gray-scale segmentation mask, $\mathbf{Y}$=[$\mathbf{y}_{ij}$], which is composed of column vectors generated by the neural network shown in Eqn. 5.1.

$$\tilde{\mathbf{X}} = \mathbf{X} + \mathbf{I} \cdot \mathbf{P}^T \cdot \mathbf{Y} \tag{5.2}$$

We can find the anatomy-informed image using Eqn. 5.2 where, $\mathbf{I}$ is the identity matrix, $\mathbf{P}$ is a hyper-parameter matrix for controlling the measures of infiltration in a specific RGB channel. In order to keep the actual information of the chest radiographs as much as possible, only the Blue channel has been infiltrated with the anatomical information in a small amount, so the hyper-parameter matrix, $\mathbf{P}$ can be written as $\mathbf{P} = [1\ 1\ p]$, $0 < p < 1$. This anatomy informed image, $\tilde{\mathbf{x}} = [\tilde{\mathbf{X}}^R\ \tilde{\mathbf{X}}^G\ \tilde{\mathbf{X}}^B]$, is then fed into the backbone model.

Backbone Model

The backbone model is a traditional deep-learning classification network used in transfer learning. In this case, we use Densenet-121 as its dense block has been well-known for its *feature-reuse* capability during feature extraction. Training this model using the Anatomy-annotated images, $\tilde{\mathbf{x}}$, ensures that the model is aware of the chest radiograph anatomy.

Feature Pyramid Attention

To produce improved pixel-level attention, we incorporate a Feature Pyramid Attention (FPA) module [137] into the system. The Pyramid Attention module first fuses feature from two different branches: three $n \times n$ pyramid scales convolution (n=3,5,7) and origin features from CNNs going through a 1×1 convolution. Then, a global average pooling branch feature is added with the output features to select the discriminative multi-resolution feature representation. When these features overlap significantly, final pooling is applied to extract the disease-level features (pre-softmax or pre-sigmoid output).

P-CAM Pooling

We use P-CAM pooling for improved heatmap generation. P-CAM pooling explicitly leverages the excellent localization ability of CAM during training in a probabilistic fashion. The backbone network first processes the input CXR image and generates a feature map. Then, for a particular disease label, such as 'Consolidation,' each feature embedding within the feature map goes through a fully connected (FC) layer implemented as a 1×1 convolutional layer and generates the class activation score that monotonically measures the disease likelihood of each embedding. It is further bounded with the sigmoid function and interpreted as the disease probability of each embedding. Finally, the output probability map is normalized to the attention weights of each embedding, following the multiple-instance learning (MIL) framework, which is used to pool the original feature map by weighted average pooling. The pooled embedding goes through the same FC layer introduced above and generates the image-level disease probability for training. During inference time, the probability map is directly used for localization. Then, we apply simple hyperparameter thresholding to obtain disease regions. Finally, for the disease d, we define its activation score as follows:

$$\mathrm{A}_d = \frac{\sum_{0,0}^{h,w} [L_{i,j} \bigcap R_{i,j}^d]}{\sum_{0,0}^{h,w} L_{i,j}} \tag{5.3}$$

where $L_{i,j}$ is the lung segment (right and left both stacked), $R_{i,j}^d$ is the d region of the disease created from heatmap. The summation is across the segment and region's height and width. According to Eq. (5.3), disease activation will be between 0 to 1.

5.4 Evaluate the Performance of Anatomy Aware Deep Learning Framework

Our baseline model using Densenet121-FPA without the proposed anatomy-aware (AA) block provides very similar results as reported in [13]. In terms of geographical extent score, Densenet121-FPA and [13] provide MSE scores of 1.93 ± 0.63 and 2.06 ± 0.34, respectively (p-value = $0.001 < 0.01$). In the case of lung opacity score, Densenet121-FPA and [13] provide MSE scores of 1.08 ± 0.22 and 0.86 ± 0.11 (p-value = $0.0 < 0.01$), respectively. Thus, we can conclude that the baseline model consisting of the Densenet121-FPA pipeline is equivalent to [13] in terms of performance and statistical significance. Table 5.2 also shows that the overall performance improves after including the anatomy-aware block. The best-performing model is Densenet121-FPA, with the AA block included in geographic extent and opacity scores. In terms of geographical extent score, the MSE improves from 1.93 ± 0.63 to 1.85 ± 0.29 after including the AA block over the baseline architecture. On the other hand, for the lung opacity score, the MSE improves from 1.08 ± 0.22 to 0.97 ± 0.23 after including the AA block along with the baseline model. However, Table 5.2 also shows that including the disease activation score does not provide the best result for the prediction of lung opacity score.

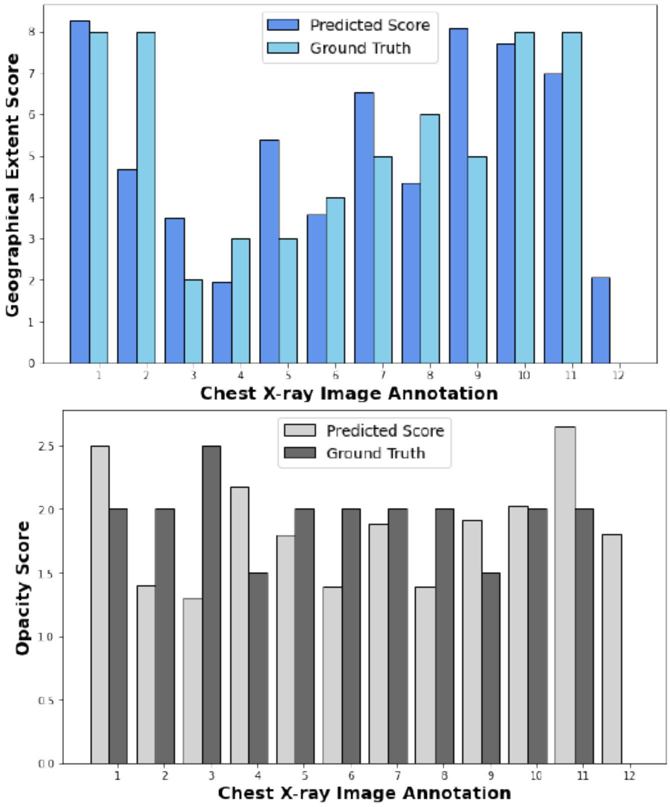
Table 5.2: Performance evaluation of the proposed COVID-19 severity prediction architecture compared to existing models. These results are also compared with the baseline scores from Cohen *et al.* [13].)

Task	Method	Features	MSE	MAE	R^2
Geographical Extent Score	Densenet-FPA without AA	3 diseases	2.25 ± 0.62	1.22 ± 0.12	0.59 ± 0.10
		4 diseases	1.93 ± 0.63	1.16 ± 0.15	0.63 ± 0.11
		single disease	3.16 ± 0.72	1.45 ± 0.18	0.41 ± 0.04
	Densenet-FPA with AA	3 diseases	1.90 ± 0.45	1.14 ± 0.12	0.64 ± 0.05
		4 diseases	1.85 ± 0.29	1.21 ± 0.16	0.63 ± 0.10
		single disease	3.45 ± 1.56	1.34 ± 0.20	0.38 ± 0.15
	Densenet-FPA with AA and Disease Activation features	3 diseases	1.87 ± 0.51	1.15 ± 0.09	0.63 ± 0.10
		4 diseases	1.90 ± 0.39	1.12 ± 0.11	0.63 ± 0.05
		single disease	3.52 ± 1.24	1.34 ± 0.20	0.35 ± 0.10
	Cohen et al. [13]	lung opacity	2.06 ± 0.34	1.14 ± 0.10	0.60 ± 0.09
Opacity Score	Densenet-FPA without AA	3 diseases	1.20 ± 0.14	0.76 ± 0.04	0.57 ± 0.09
		4 diseases	1.08 ± 0.22	0.80 ± 0.10	0.61 ± 0.07
		single disease	1.37 ± 0.47	0.94 ± 0.16	0.39 ± 0.10
	Densenet-FPA with AA	3 diseases	0.94 ± 0.20	0.82 ± 0.04	0.55 ± 0.09
		4 diseases	0.97 ± 0.23	0.85 ± 0.11	0.56 ± 0.11
		single disease	1.54 ± 0.47	1.10 ± 0.12	0.32 ± 0.08
	Densenet-FPA with AA and Disease Activation features	3 diseases	0.94 ± 0.21	0.81 ± 0.05	0.46 ± 0.20
		4 diseases	0.96 ± 0.24	1.15 ± 0.31	0.45 ± 0.20
		single disease	1.45 ± 0.50	1.52 ± 0.21	0.20 ± 0.36
	Cohen et al. [13]	lung opacity	0.86 ± 0.11	0.78 ± 0.05	0.58 ± 0.09

Overall, analyzing the results of Table 5.2 implies that including additional anatomical information to the competitive baseline model further increases the ability of the system for disease severity prediction.

Table 5.3: Evalution of the proposed COVID-19 severity prediction model on Selected Chest X-ray images annotated by an In-house experienced radiologist

Scoring Method	**MAE**	**MSE**
Geographical Extent Score	1.55±0.98	3.35±3.51
Opacity Score	0.62±0.48	0.59±0.89



[a]

[b]

Figure 5.3: (a) Predicted value and ground truth of geographical extent score of selected chest X-ray images annotated by an In-house experienced radiologist, and (b) Predicted value and ground truth of opacity score of selected chest X-ray images annotated by an In-house experienced radiologist.

5.5 Discussion

Previous research on chest X-ray image analysis has primarily focused on traditional deep-learning architectures commonly used for image classification. However, there have been limited studies specifically addressing COVID-19 severity prediction. These studies typically employ pre-trained models without considering anatomical information [13]. Furthermore, there is a requirement for case-by-case comparisons to identify potential sources of error and enhance the image analysis architecture. Additionally, there is a scarcity of chest X-rays available for segmentation tasks, which can sometimes make it challenging, and anomalies may exist in the findings.

However, the existing models lack specific awareness of the anatomical structure present in typical chest X-ray images. In contrast, experienced radiologists first identify the thoracic organs before examining disease markers. Consequently, incorporating anatomical information into existing models could enhance their performance in predicting disease severity. Furthermore, anatomical information can reduce computational complexity by introducing disease-specific features, enabling the model to learn more precisely.

In this study, a semi-supervised model is employed to automatically generate lung segmentation masks, which are then fused with chest X-ray images. Our best-performing model has shown an 11% relative improvement in mean square error (MSE) compared to existing methods when evaluated on a dataset for predicting the severity of COVID-19 pneumonia. Experimental comparisons between systems with and without integrated anatomy information clearly demonstrate the effectiveness of the proposed method. Furthermore, this model exhibits promising results on an unseen clinical evaluation dataset that was annotated by an experienced radiologist. These experimental evaluations provide evidence of the efficacy of our anatomy-aware architecture for predicting the severity of COVID-19 disease.

Chapter 6

Effective Multimodal Approaches for Medical Data and Image Processing

6.1 Dataset Description

6.1.1 OpenI Dataset

OpenI [138] is a publicly available dataset designed to facilitate research and development in the field of medical imaging analysis. It offers a comprehensive collection of medical imaging data, specifically chest X-rays and radiology reports. The dataset encompasses a diverse range of images accompanied by corresponding textual information. This open-access resource enables researchers and developers to explore and innovate in the domain of medical imaging analysis, fostering advancements in diagnostic techniques and healthcare technologies.

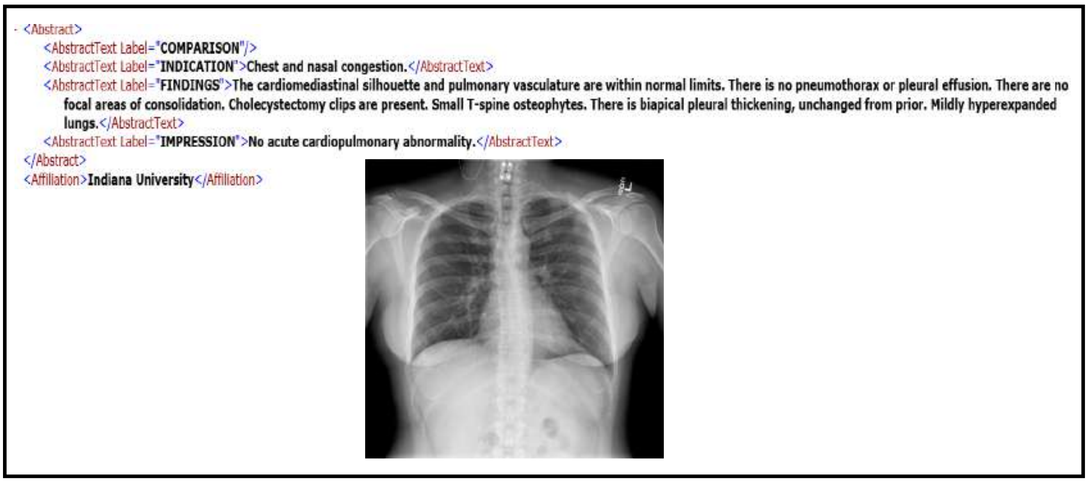
The chest X-ray dataset within OpenI comprises a vast collection of radiographs obtained from multiple medical institutions. These images provide visual representations of the internal structures within the chest, such as the heart, lungs, and adjacent tissues. The dataset encompasses a diverse range of medical conditions and diseases, offering a comprehensive resource for studying and analyzing various chest-related abnormalities.

Apart from the chest X-ray images, OpenI also offers access to radiology reports. These reports consist of written descriptions created by radiologists, offering comprehensive details regarding the observations made in the corresponding images. The reports encompass valuable information regarding any detected abnormalities, diagnoses, recommended treatment approaches, and other pertinent clinical information that aids in patient management and care.

The dataset has undergone meticulous curation to guarantee the quality and accuracy of the data. It includes a substantial collection of images and reports that have been carefully labeled and annotated, making it well-suited for a wide range of research endeavors. Researchers can leverage this dataset for tasks such as disease classification, lesion detection, and medical natural language processing, among others.

The OpenI dataset offers a valuable resource for researchers and developers engaged in various endeavors, including training and evaluating machine learning models, creating computer-aided diagnostic systems, and conducting research in medical imaging and radiology. The dataset's unique feature of containing both images and textual data enables multimodal analysis and the exploration of synergistic approaches that leverage both visual and textual information. This integration enhances disease diagnosis and treatment planning by leveraging the complementary nature of visual and textual data.

The OpenI dataset comprises chest X-ray images obtained from the Indiana University hospital network. The dataset is organized into two folders, one containing X-ray images and the other containing XML reports associated with the radiography. Each report may correspond to multiple images. In total, the dataset consists of 7,470 chest X-rays and 3,955 radiology reports. It is important to note that the dataset is unlabeled, and the segmentation ground truth for the images is not provided.



```
- <Abstract>
    <AbstractText Label="COMPARISON"/>
    <AbstractText Label="INDICATION">Chest and nasal congestion.</AbstractText>
    <AbstractText Label="FINDINGS">The cardiomediastinal silhouette and pulmonary vasculature are within normal limits. There is no pneumothorax or pleural effusion. There are no
       focal areas of consolidation. Cholecystectomy clips are present. Small T-spine osteophytes. There is biapical pleural thickening, unchanged from prior. Mildly hyperexpanded
       lungs.</AbstractText>
    <AbstractText Label="IMPRESSION">No acute cardiopulmonary abnormality.</AbstractText>
  </Abstract>
  <Affiliation>Indiana University</Affiliation>
```

Figure 6.1: OpenI Chest X-ray image with corresponding medical report (XML).

6.1.2 MIMIC-CXR Dataset

The MIMIC-CXR (Medical Information Mart for Intensive Care Chest X-ray) dataset [139–141] is a component of the larger MIMIC (Medical Information Mart for Intensive Care) database, which is a publicly accessible clinical database extensively used

for researchers, clinicians, and data scientists interested in various facets of medical imaging analysis, including computer-aided diagnosis, image classification, disease detection, and natural language processing of radiology reports.

It is important to note that the MIMIC-CXR dataset adheres to appropriate data usage agreements and ethical considerations to safeguard patient privacy and comply with relevant regulations. Access to the dataset necessitates proper approvals and adherence to specified terms of use to uphold patient confidentiality and ensure data integrity.

The MIMIC-CXR dataset is a vast and publicly accessible collection of chest radiographs accompanied by free-text radiology reports. It comprises 377,110 images representing 227,835 radiographic studies conducted at the Beth Israel Deaconess Medical Center in Boston, MA. The radiology reports associated with the images were identified and extracted from the hospital's electronic health record (EHR) system. This dataset provides a valuable resource for studying and developing algorithms and systems in the field of chest radiography analysis.

6.1.3 CheXpert Labeler

The CheXpert dataset is a extensive collection comprising 224,316 chest radiographs from 65,240 patients. These radiographs were obtained from the Stanford Hospital between October 2002 and July 2017, encompassing both inpatient and outpatient settings. In the study by Irvin et al. [131], a labeler was developed to automatically identify 14 specific observations in radiology reports. The authors explored different approaches to utilize the uncertainty labels generated by the labeler in training convolutional neural networks, which in turn produced probabilities for these observations. Additionally, an automated rule-based labeler was developed to extract observations from free-text radiology reports, enabling the creation of structured labels for the corresponding radiographic images.

6.2 Medical Data Pre-processing

6.2.1 Text Pre-processing

Text preprocessing for radiology reports involves a series of steps to clean and transform the raw text data before further analysis or natural language processing tasks. The following are common techniques used in text preprocessing for radiology reports:

tasks.

5. **Numerical value normalization:** Standardizing numerical values by replacing them with placeholders or converting them to a consistent format. For example, replacing specific lab test values with generic labels or normalizing dates and times.
6. **Spell checking and correction:** Identifying and correcting spelling errors to ensure accurate analysis and improve the quality of the data.
7. **Lemmatization or stemming:** Reducing words to their base or root form to consolidate similar words and reduce dimensionality. Lemmatization considers the context and part of speech, while stemming applies simple rules to truncate words.
8. **Removal of irrelevant information:** Eliminating non-textual information or sections that are not relevant to the analysis, such as headers, footers, or boilerplate text.
9. **Specialized domain-specific pre-processing:** Performing additional steps based on the specific requirements of radiology reports. This may include handling abbreviations, acronyms, or medical terminology unique to the field.
10. **Normalization:** Bringing the text data into a consistent format, such as converting different date formats to a standardized representation or ensuring consistent use of terminology and abbreviations.

Text preprocessing for radiology reports aims to clean and standardize the text data, making it ready for further analysis, information extraction, or machine learning algorithms. The specific preprocessing steps may vary depending on the objectives of the analysis and the characteristics of the radiology reports being processed.

The OpenI and MIMIC-CXR reports are divided into three segments: Findings, Impression, and Indication, some examples of processing steps of radiology reports are provided.

1. All values like ”1” and ”2” etc are removed.
2. All special characters except for full stop are removed.
3. Words like ”XXXX” are removed.
4. The words that occur multiple times but not necessary are removed.
5. Unwanted spaces are also removed.

6.2.2 Word Token Generation

First, the sentences undergo tokenization, which involves breaking them down into smaller units. BERT utilizes a WordPiece tokenizer that splits words into subwords and introduces special tokens. To perform tokenization, the Transformers library by Hugging Face is employed in Python. BERT necessitates the inclusion of specific tokens to indicate the sequence's beginning and end. At the start, the [CLS] token is appended, while the [SEP] token is added at the end. Both the [CLS] token and the [SEP] token are placed at the beginning and end of the sentence, respectively. Additionally, padding is applied to the sentences using the [PAD] token, ensuring their length matches the maximum length requirement. Finally, the tokens are converted into token IDs, which align with the input format expected by the BERT model.

6.3 Chest X-ray Pre-processing and Feature Extraction

For the pre-processing of chest x-rays, at first the images are resized to 224×224 pixels. Then, histogram equalization is applied on the chest x-rays for image enhancement.

In this method, ResNet-50 [129] and DenseNet-121 [142] architectures are used for feature extraction. First, the networks, pretrained on ImageNet, are fine tuned to classify chest x-rays. The outputs of the pre-trained architectures are reshaped to be used as the input of the VisualClinicalBERT. Image embeddings are generated to give input to the framework for classification purpose.

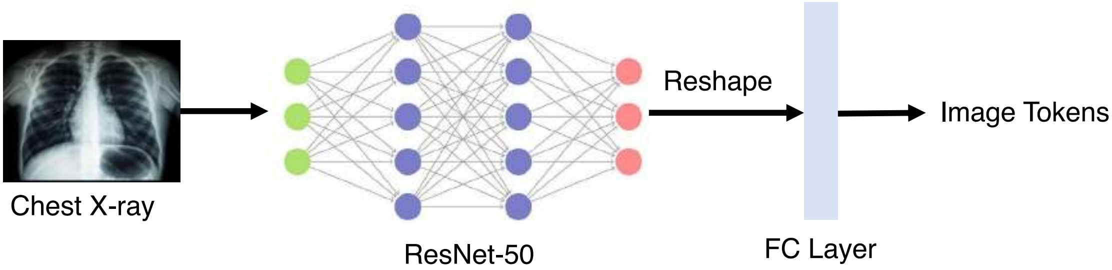


Figure 6.4: Chest X-ray pre-processing and feature extraction framework.

6.4 Baseline Architecture

6.4.1 VisualBERT

VisualBERT is a versatile framework designed to address various vision-and-language tasks in a straightforward and adaptable manner [10]. The framework utilizes a stack of Transformer layers, which employ self-attention to establish implicit alignments between elements in a given input text and regions within an associated input image. The COCO dataset [143] serves as the pre-training data for VisualBERT, primarily used for image captioning [144]. In a study by Li et al. [39], the combination of test findings and image embeddings is utilized for thoracic disease classification. The approach achieves an average accuracy of 98.7% in identifying seven specific thoracic diseases using the OpenI dataset [138].

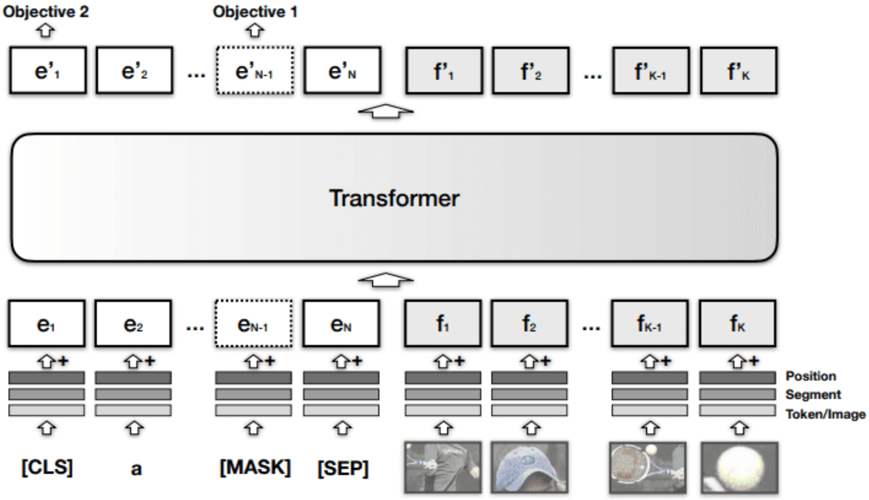


Figure 6.5: Configuration of VisualBERT Baseline Framework [10].

The proposed framework incorporates all the components of BERT and introduces a new set of visual embeddings, denoted as F, to effectively model an image. Each embedding f in F corresponds to a specific bounding region in the image, which is obtained from an object detector. Each embedding in F is computed by summing three embeddings:

- f_o: This represents the visual feature representation of the corresponding bounding region. It is computed using a convolutional neural network.
- f_s: This indicates a segment embedding, distinguishing it as an image embedding rather than a text embedding.

of the sentence as inputs. The Image Embedder utilizes a faster R-CNN to extract the visual features of each region, while the Text Embedder tokenizes the input sentence into WordPieces.

In a related study [39], the combination of joined test findings and image embeddings is utilized for thoracic disease classification. The approach achieves an average accuracy of 98.2% in identifying seven thoracic diseases using the OpenI dataset [138]. The model is fine-tuned with a learning rate of 5×10^{-5}, a batch size of 32, and the parameters are fine-tuned for a duration of 4 epochs for the experiments.

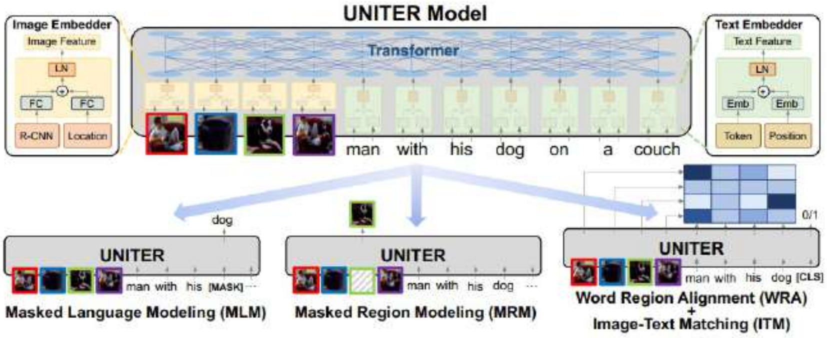


Figure 6.7: Configuration of UNITER Baseline Framework [12].

6.5 Proposed System

6.5.1 VisualClinicalBert

6.5.1.1 Backbone Architecture

ClinicalBERT [32] which is a pre-trained transformer encoder stack pre-trained with MIMIC-III [145], EHR and discharge information, is used as the backbone architecture here. In the classification head a two layered multilayer perceptron [146] with a dimention of 768 and Gaussian Error Linear Unit (GELU) activations [147] is added with layer normalisation.

Finally, there are 14 output nodes where sigmoid function is applied for classification. A binary cross-entropy loss function is used here for loss calculation. For experiment, the number of epoch is 14 with a batch size of 128. Adam optimiser is used here for optimisation with a learning rate of 5×10^{-5}. There are a total of 512 input tokens and 12 layers in the VisualClinicalBERT (VCBERT). Each vector is made up of 768 float

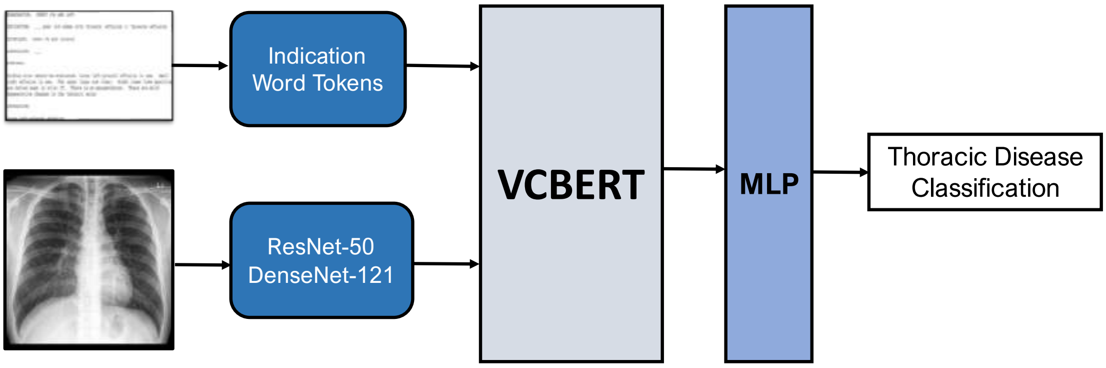


Figure 6.8: Proposed thoracic disease classification framework.

numbers (hidden units). The encoder block here uses the self-attention mechanism to enrich each token (embedding vector) with information.

6.5.1.2 Classification Head

A multi-layer perceptron has one input layer and for each input, there is one neuron(or node), it has one output layer with normalization. GELU relates to stochastic regularizers as it is a modification to Adaptive Dropout. GELU activation function is defined:

$$GELU(x) = xP(X \leq x) = x\phi(x) \tag{6.1}$$

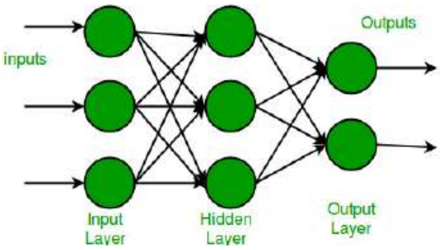


Figure 6.9: The diagram is a two-layer MLP. There are three inputs with three input nodes and the hidden layer has three nodes. The output layer gives two outputs as there are two output nodes.

It can be approximated through μ and σ which are learnable hyperparameters.

$$0.5x(1 + \tanh[\sqrt{2/\pi}(x + 0.044715x^3]) \tag{6.2}$$

Sigmoid activation function is applied for the final classification. In this experiment, binary cross-entropy loss is used as a loss function. The Binary cross-entropy loss function actually calculates the average cross-entropy. The formula of this loss function:

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot (log(1 - p(y_i))) \tag{6.3}$$

Chapter 7

Evaluate the Performance of Multimodal Approaches

7.1 Evaluation of Indication Information in Disease Classification

The medical reports provided with chest x-ray consists of three types of information. The indication is basically the symptoms and patient's medical history, the findings is the information of radiological findings provided by the radiologist, and the impression is the information of the intuition of the radiologists according to the findings. To evaluate the significance of the indication as a classification information, at first different machine learning classifiers are used.

For textual information, Naive Bayes (NB) [148], Stochastic Gradient Descent (SGD) [149], and Logistic Regression (LR) [150] are mostly used [151–153]. In Table 7.1, the results of different machine learning classifiers using different textual information is provided. The performance of NB is higher than the other classifiers. Three classes of information are pre-processed to be fed into the classifiers. It is obvious that the findings include the disease information which matches with the label and thus the accuracy with findings is the highest. The impression also includes direct disease selective information which caused higher accuracy. On the other hand, the indication information does not contain any direct disease information, only includes patients medical history and symptoms.

Using indication as a textual feature to the classifiers decreased the accuracy by 14.41%, precision by 22.49%, and F1-score by 11.35% for NB classifier than findings informa-

Table 7.1: Experimental results on OpenI Medical Reports using Machine Learning Classifiers

Using Only Findings Information			
Machine Learning Classifiers	**Accuracy (%)**	**Precision (%)**	**F1-Score (%)**
Naive Bayes (NB)	71.44	70.1	66.78
Stochastic Gradient Descent (SGD)	70.21	68.35	64.8
Logistic regression (LR)	58.43	60.76	53.9
Using Only Impression Information			
Naive Bayes (NB)	71.27	68.54	67.8
Stochastic Gradient Descent (SGD)	69.31	65.42	63.75
Logistic regression (LR)	60.21	58.6	53.9
Using Only Indication Information			
Naive Bayes (NB)	61.14	54.33	59.2
Stochastic Gradient Descent (SGD)	56.2	61.55	62.31
Logistic regression (LR)	52.1	55.34	50.4

tion. The accuracy is about 61.14% using NB classifier and indication information proves that there maybe some significant information in the indication that can be used for the classification of thoracic diseases.

From Table 7.2, an analysis of different language models using the findings, impression, and indication information is provided. The CNN model, mentioned in Section 3.3 is used for classification. The accuracy for CNN architecture using findings is 11.81% higher than the accuracy using indication information. Whereas, using pre-trained language models improved the classification performance. Using ClinicalBERT as a pre-trained model increased the accuracy by 3.09% as it is pre-trained on clinical text information. Only indication information reported about 72.89% accuracy on OpenI dataset proving the effectiveness of patient's history and symptoms for disease classification purpose.

Table 7.2: Experimental results on OpenI Medical Reports using Different Language Models

Frameworks	**Accuracy (%)** (Using Findings)	**Accuracy (%)** (Using Impression)	**Accuracy (%)** (Using Indication)
CNN	65.22	61.47	58.33
BERT	78.21	72.32	70.70
ClinicalBert	80.55	75.10	72.89

To analyze the information in indication, MIMIC-CXR text reports are evaluated using different machine learning and deep learning frameworks. From Table 7.3, we can see that Naive Bayes classifier's performance degraded. On the other hand, with the increase of number of data, the performance of ClinicalBERT improved. Using only indication information the accuracy is about 73.96% for the MIMIC-CXR dataset.

Table 7.3: Experimental results on MIMIC-CXR Medical Reports using Different Frameworks

Framework	**Accuracy (%)** (Using Findings)	**Accuracy (%)** (Using Impression)	**Accuracy (%)** (Using Indication)
Naive Bayes	65.21	62.35	57.66
BERT	80.23	75.66	72.45
ClinicalBert	83.78	77.24	73.96

7.2 Evaluation of Disease Information in Chest X-ray Images

To evaluate the significance of image features of chest x-ray images for thoracic disease classification, ResNet-50, and DenseNet-121 pre-trained architectures are used. These models are fine-tuned for classification. From Table 7.4, it can be observed that the accuracy with only x-rays is about 82.56% for ResNet-50 and 83.25% for DenseNet-121 architectures. Using the complex DenseNet-121 architecture the accuracy improved. This proves that the image features can increase the classification performance of thoracic diseases.

Table 7.4: Experimental results on OpenI Medical Images

Frameworks	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
ResnNet-50	82.56	65.34	71.89	68.57
DenseNet-121	83.25	68.47	73.81	71.17

7.3 Evaluation of Disease Classification Using Medical Data and Chest X-ray Images

To evaluate the performance of multimodal approach in thoracic disease classification, three approaches are tried. Firstly, CNN mentioned in Section 3.3 is used for textual feature extraction. Pre-trained ResNet-50 is used as an image feature extraction. After that, the textual and image features are concatenated and classified using Logistic regression. Using this feature extraction scheme and machine learning classifiers about 78.45% accuracy is found which is about 25.64% improvement using only textual indication features to CNN.

Table 7.5: Experimental results on OpenI (Text-Indication and Image Combined)

Frameworks	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
CNN+ResNet-50+LR	78.45	58.23	54.21	55.16
BERT+ResNet-50	84.45	61.56	62.67	62.11
BERT+DenseNet-121	84.99	60.21	62.58	61.37

In the second approach, the pre-trained BERT model is used as the backbone and ResNet-50 is used as a feature extractor. The accuracy is improved by 7.10% than the first approach. Similarly, in the third approach DenseNet-121 is used as the image feature extractor and pre-trained BERT is used as the backbone architecture. This approach increased the accuracy by about 0.64%. After evaluating these three approaches, it can be found that multimodal approaches can increase classification performance by incorporating both textual and image information.

7.4 Evaluation of the Performance of Baseline Systems

The baseline frameworks mentioned in Section 6.4 are evaluated. All of these architectures are pre-trained and used for thoracic disease classification purpose. From Table 7.6, we can notice that the performance of VisualBERT is higher than the other two baseline architechtures. The performance of LXMERT is the lowest among three baselines. The performance of LXMERT is decreased than the BERT+ResNET-50 and BERT+DesneNet-121 frameworks. On the other hand, the performance of UNITER is about 6.78% higher than LXMERT using both textual and image information.

Table 7.6: Experimental results on OpenI Dataset of Baseline System (Text-Indication and Image Combined)

Frameworks	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
VisualBERT	85.02	71.67	69.24	70.43
LXMERT	79.34	73.44	67.32	70.24
UNITER	84.72	67.98	70.22	69.08

7.5 Evaluation of Proposed Framework

Table 7.7: Experimental results on OpenI Dataset of Proposed System (Text-Indication and Image Combined)

Frameworks	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
ResVCBERT	88.29	72.12	69.62	70.84
DenseVCBERT	88.44	72.45	71.89	72.18

The proposed framework is experimented with using two image feature extraction schemes. At first, the image features are extracted using ResNet-50 and modified pre-trained ClinicalBERT is used as the backbone for thoracic disease classification. In this approach, the accuracy is about 88.29% (From Table 7.7). In addition to it, about 5% of text information error is introduced to the indication information, and the performance is decreased by 6.67%. Proposed DenseVisualClinicalBERT provided significant improvement with an accuracy of about 88.44%.

From Table 7.8, we can visualize the classwise performance of different frameworks using indication as textual information with image features. The performance of the Support devices class is the highest among all the 14 classes. Also, this framework can detect Pneumothorax, No Finding with an accuracy above 85%.

7.6 Discussion

7.6.1 Significance of Indication Information in Thoracic Disease Classification

In Figure 7.1, a comparison of two medical reports of OpenI dataset is provided. Dyspnea is a significant symptom of various kinds of lung disease. This indication information has high importance in disease diagnosis. When we use only radiology images for diagnostic purposes, we miss this indication information for diagnosis. While disease

Table 7.8: Classwise Accuracy (%) Comaparison of Different Frameworks (Text-Indication and Image Combined)

Thoracic Diseases	**Visual BERT**	**LXMERT**	**UNITER**	**ResVC BERT**	**DenseVC BERT**
Atelectasis	73.26	71.34	73.44	75.8	76.00
Cardiomegaly	75.33	72.19	73.22	78.6	77.25
Consolidation	75.48	72.29	75.89	77.1	78.40
Edema	79.00	76.44	78.24	81.3	80.34
Enlarged Card.	70.24	68.79	70.56	72.9	74.56
Fracture	73.67	71.24	73.89	75.6	75.78
Lung Lesion	68.78	65.40	67.29	71.5	72.44
Lung Opacity	81.88	80.39	82.34	82.3	84.76
No Finding	85.67	82.34	84.56	87.1	87.29
Ple. Effusion	80.88	78.90	79.65	80.2	81.34
Pleural Other	73.67	71.89	72.88	75.2	76.33
Pneumonia	73.49	72.19	75.38	76.3	77.45
Pneumothorax	88.78	83.29	86.56	88.9	89.80
Support Devices	91.65	88.76	89.34	92.2	92.36

diagnosis, doctors are not only dependent on radiological image information. They are also concerned about the patient's medical history and symptoms for better diagnosis.

In this situation, integrating indication information into computer-aided diagnosis systems can greatly enhance the accuracy of disease diagnosis, improve the ability to predict disease severity, and even provide insights into prognosis. In Section 7.1, the significance of various medical text information for thoracic disease classification is proved. The indication information provided about 61.14% accuracy using the NB classifiers. Whereas, using ClinicalBERT, the accuracy is increased by about 19.21%. As ClinicalBERT is pre-trained in medical EHR, it already understands the clinical information. Thus, in this case, the accuracy is improved. After the analysis and also looking into the dataset, it is evident that indication information can add significant value in thoracic disease diagnosis purposes.

7.6.2 Analyzing the Robustness of the Medical Data

To analyze the robustness of medical data, errors are added to text while coding involves introducing intentional errors to the data and then assessing the impact of those errors on downstream tasks or analyses. In this case, text swap, changing the text adding numbers, adding additional texts, etc are used to add some noise to the indication information. After that, this changed indication information is processed and fed into the

can be utilized even when there is no text information and only image information is available for classification.

From the whole analysis and comparison it is evident that among all the frameworks the proposed ResVCBERT and DenseVCBERT performed well. DenseVCBERT is the best performing model using this multimodal feature fusion based approach in all evaluation metrices.

Chapter 8

Conclusions

8.1 Summary of the Present Work

This work proposed novel feature fusion-based approaches to classify different thoracic diseases. In the first approach, indication information infusion with image features provided significant improvement in disease detection. The comparative analysis with different machine learning classifiers, NLP-based methods, baselines, and proposed framework proved the importance of clinical information in computer-aided diagnostic methods for better classification. In general, for the diagnosis of any disease, besides radiological images, the patient's clinical information has a higher impact on the final decision-making by the doctors/radiologists. In this work, the indication information provided about 61.14% accuracy using the NB classifiers. Whereas, using ClinicalBERT, the accuracy is increased by about 19.21% as it is pre-trained on medical EHR. The robustness of the framework is proved by the accuracy above 80% which is about 83.12% after the addition of errors to the clinical indication information. The accuracy comparison of different frameworks provides about 4.02% improvement comparing VisualBERT and DenseVisualClinicalBERT. Among the baselines, the performance of LXMERT is the lowest. The performance improvement after using DenseNet-121 instead of ResNet-50 is about 0.16% which is quite negligible. Finally, the proposed DenseVisualClinicalBERT provided significant improvement with an accuracy of about 88.44% using the OpenI dataset of radiological reports and chest X-rays.

In this thesis, the novel anatomy-aware deep-learning framework is also proposed for COVID-19 disease severity prediction from chest X-ray images. While traditional methods generally do not specifically consider anatomical information for medical image analysis, expert radiologists tend to always consider their human anatomy knowledge before making a diagnostic decision. In this work, a semi-supervised model is utilized for automatically generating lung segmentation masks that are subsequently fused within the chest X-ray images. Here the best-performing model has provided a relative improvement of 11% in MSE compared to existing methods when evaluated on a COVID-19 pneumonia severity prediction dataset. Experimental comparisons between systems with and without anatomy information integrated clearly show the effectiveness of the proposed method. This model also shows promising results on an unseen in-house clinical evaluation dataset that an experienced radiologist has annotated. The experimental evaluations demonstrate the effectiveness of the proposed anatomy-aware architecture for COVID-19 disease severity prediction.

To summarize, chest radiography is among the major radiological diagnostic methods in different low-income regions. Hence, there is a need for computer-aided diagnostic methods to provide healthcare facilities in under-served communities where the proposed feature fusion-based approach can provide assistance to healthcare professionals. Thus, the extension of the proposed method with anatomical information with clinical indications can be used for thoracic disease classification and even severity prediction in low-resource settings.

8.2 Limitations of the Present Work

Though this work shows promising performance using the feature fusion approaches, this work also has some limitations. Some of them are:

1. The proposed feature fusion-based approach needs to be validated on real-world test datasets. In this case, there is a need for data collection and annotation from expert radiologists.
2. There is a need for a combined framework and analysis using the clinical information and anatomical information for specific kinds of diseases.
3. There is a need for annotated dataset to validate the improvement and performance of the proposed framework. Currently, the available datasets are labeled using Chexpert labeler which may not perform well for all the available datasets. There is a need for cross-matching of the dataset annotation.
4. There is a need for a high-performance GPU and server for doing the computation works faster. The training of medical images takes a lot of time for computation.

Thus, high computational devices, servers, and storage systems can increase efficiency.

5. This proposed feature fusion approach separately proves the improvement of performance using clinical indication information and anatomical information. There is a need for further analysis of this combined approach using both the indication information with image features.
6. There is a need for a comparative analysis using different segmentation approaches, different text and image datasets, and different pre-trained NLP-based approaches to increase and validate the performance of feature fusion-based approaches.

8.3 Future Prospects of the Present Work

There are lots of opportunities for further research. The considered scopes are:

1. This work claims that clinical indication information and anatomical information with chest X-rays can be used in computer-aided diagnostic systems. This claim can be verified in the real world by implementing a software version in an AI-based disease diagnosis platform.
2. Data collection and data annotation from experts can open pathways to explore more about thoracic disease diagnosis, severity prediction, and even prognosis.
3. More sophisticated frameworks can be developed using patient's history, symptoms, and anatomical information with radiological image findings. There are scopes of analyzing other multimodal approaches using different types of imaging modalities and also using physiological signals.
4. There are scopes of analyzing different available frameworks and addressing their limitations in different disease classification tasks.
5. There are scopes to work with larger datasets with high computational power to validate the performance of developed frameworks.
6. More hyperparameter tuning on pre-trained models can be analyzed. There are scopes of generating loss function, and model optimization to find better performance of the architectures.
7. Analyze anatomical and clinical information from various imaging modalities (e.g., CT, MRI, PET, OCT) and explore feature fusion-based approaches to extract valuable information for AI-based system development.

References

[1] Perez, L., Ottens, L. and Viswanathan, S. “Automatic code generation using pre-trained language models.” *arXiv preprint arXiv:2102.10535*, 2021

[2] Rahali, A. and Akhloufi, M.A. “Malbert: Using transformers for cybersecurity and malicious software detection.” *arXiv preprint arXiv:2103.03806*, 2021

[3] Suganyadevi, S., Seethalakshmi, V. and Balasamy, K. “A review on deep learning in medical image analysis.” *International Journal of Multimedia Information Retrieval*, volume 11, no. 1:pp. 19–38, 2022

[4] Pons, E., Braun, L.M., Hunink, M.M. and Kors, J.A. “Natural language processing in radiology: a systematic review.” *Radiology*, volume 279, no. 2:pp. 329–343, 2016

[5] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. “Bert: Pre-training of deep bidirectional transformers for language understanding.” *arXiv preprint arXiv:1810.04805*, 2018

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X. and Unterthiner, T. “Transformers for image recognition at scale.” *arXiv preprint arXiv:2010.11929*, 2020

[7] Theckedath, D. and Sedamkar, R. “Detecting affect states using vgg16, resnet50 and se-resnet50 networks.” *SN Computer Science*, volume 1:pp. 1–7, 2020

[8] Rochmawanti, O. and Utaminingrum, F. “Chest x-ray image to classify lung diseases in different resolution size using densenet-121 architectures.” In “6th International Conference on Sustainable Information Engineering and Technology 2021,” pp. 327–331, 2021

[9] Kamal, U., Zunaed, M., Nizam, N.B. and Hasan, T. “Anatomy-xnet: An anatomy aware convolutional neural network for thoracic disease classification in

chest x-rays." *IEEE Journal of Biomedical and Health Informatics*, volume 26, no. 11:pp. 5518–5528, 2022

[10] Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J. and Chang, K.W. "Visualbert: A simple and performant baseline for vision and language." *arXiv preprint arXiv:1908.03557*, 2019

[11] Tan, H. and Bansal, M. "Lxmert: Learning cross-modality encoder representations from transformers." *arXiv preprint arXiv:1908.07490*, 2019

[12] Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y. and Liu, J. "Uniter: Universal image-text representation learning." In "Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX," pp. 104–120. Springer, 2020

[13] Cohen, J.P., Dao, L., Roth, K., Morrison, P., Bengio, Y., Abbasi, A.F., Shen, B., Mahsa, H.K., Ghassemi, M., Li, H. et al. "Predicting covid-19 pneumonia severity on chest x-ray with deep learning." *Cureus*, volume 12, no. 7, 2020

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

[15] Wang, G., Liu, X., Shen, J., Wang, C., Li, Z., Ye, L., Wu, X., Chen, T., Wang, K., Zhang, X. et al. "A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and covid-19 pneumonia from chest x-ray images." *Nature biomedical engineering*, volume 5, no. 6:pp. 509–521, 2021

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

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

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

[19] Krizhevsky, A., Sutskever, I. and Hinton, G.E. "Imagenet classification with deep convolutional neural networks." *Communications of the ACM*, volume 60, no. 6:pp. 84–90, 2017

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

[21] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. "Going deeper with convolutions." In "Proceedings of the IEEE conference on computer vision and pattern recognition," pp. 1–9, 2015

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

[23] Chen, B., Li, J., Guo, X. and Lu, G. "Dualchexnet: dual asymmetric feature learning for thoracic disease classification in chest x-rays." *Biomedical Signal Processing and Control*, volume 53:p. 101554, 2019

[24] Zhang, S., Tong, H., Xu, J. and Maciejewski, R. "Graph convolutional networks: a comprehensive review." *Computational Social Networks*, volume 6, no. 1:pp. 1–23, 2019

[25] Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P.M. and Rueckert, D. "Semi-supervised learning for network-based cardiac mr image segmentation." In "Medical Image Computing and Computer-Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II 20," pp. 253–260. Springer, 2017

[26] Nie, D., Gao, Y., Wang, L. and Shen, D. "Asdnet: attention based semi-supervised deep networks for medical image segmentation." In "Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11," pp. 370–378. Springer, 2018

[27] Cheplygina, V., de Bruijne, M. and Pluim, J.P. "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis." *Medical image analysis*, volume 54:pp. 280–296, 2019

[28] Aviles-Rivero, A.I., Papadakis, N., Li, R., Sellars, P., Fan, Q., Tan, R.T. and Schönlieb, C.B. "Graphx^\small net-net-chest x-ray classification under extreme minimal supervision." In "Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22," pp. 504–512. Springer, 2019

[29] Perone, C.S., Ballester, P., Barros, R.C. and Cohen-Adad, J. "Unsupervised domain adaptation for medical imaging segmentation with self-ensembling." *NeuroImage*, volume 194:pp. 1–11, 2019

[30] Zhang, Y., Wei, Y., Wu, Q., Zhao, P., Niu, S., Huang, J. and Tan, M. "Collaborative unsupervised domain adaptation for medical image diagnosis." *IEEE Transactions on Image Processing*, volume 29:pp. 7834–7844, 2020

[31] Balagopalan, A., Eyre, B., Rudzicz, F. and Novikova, J. "To bert or not to bert: comparing speech and language-based approaches for alzheimer's disease detection." *arXiv preprint arXiv:2008.01551*, 2020

[32] Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T. and McDermott, M. "Publicly available clinical bert embeddings." *arXiv preprint arXiv:1904.03323*, 2019

[33] Biswas, S.S. "Role of chat gpt in public health." *Annals of Biomedical Engineering*, volume 51, no. 5:pp. 868–869, 2023

[34] Pan, S., Tian, Z., Lei, Y., Wang, T., Zhou, J., McDonald, M., Bradley, J.D., Liu, T. and Yang, X. "Cvt-vnet: convolutional-transformer model for head and neck multi-organ segmentation." In "Medical Imaging 2022: Computer-Aided Diagnosis," volume 12033, pp. 914–921. SPIE, 2022

[35] Li, W., Song, H., Li, Z., Lin, Y., Shi, J., Yang, J. and Wu, W. "Orbitnet—a fully automated orbit multi-organ segmentation model based on transformer in ct images." *Computers in Biology and Medicine*, volume 155:p. 106628, 2023

[36] Petit, O., Thome, N., Rambour, C., Themyr, L., Collins, T. and Soler, L. "U-net transformer: Self and cross attention for medical image segmentation." In "Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12," pp. 267–276. Springer, 2021

[37] Hu, M., Pan, S., Li, Y. and Yang, X. "Advancing medical imaging with language models: A journey from n-grams to chatgpt." *arXiv preprint arXiv:2304.04920*, 2023

[38] Wang, S., Zhao, Z., Ouyang, X., Wang, Q. and Shen, D. “Chatcad: Interactive computer-aided diagnosis on medical image using large language models.” *arXiv preprint arXiv:2302.07257*, 2023

[39] Li, Y., Wang, H. and Luo, Y. “A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports.” In “2020 IEEE international conference on bioinformatics and biomedicine (BIBM),” pp. 1999–2004. IEEE, 2020

[40] Cai, X., Liu, S., Han, J., Yang, L., Liu, Z. and Liu, T. “Chestxraybert: A pre-trained language model for chest radiology report summarization.” *IEEE Transactions on Multimedia*, 2021

[41] Bazi, Y., Rahhal, M.M.A., Bashmal, L. and Zuair, M. “Vision–language model for visual question answering in medical imagery.” *Bioengineering*, volume 10, no. 3:p. 380, 2023

[42] Tan, Y.V. and Roy, J. “Bayesian additive regression trees and the general bart model.” *Statistics in medicine*, volume 38, no. 25:pp. 5048–5069, 2019

[43] Bird, J.J., Ekárt, A. and Faria, D.R. “Chatbot interaction with artificial intelligence: human data augmentation with t5 and language transformer ensemble for text classification.” *Journal of Ambient Intelligence and Humanized Computing*, volume 14, no. 4:pp. 3129–3144, 2023

[44] Floridi, L. and Chiriatti, M. “Gpt-3: Its nature, scope, limits, and consequences.” *Minds and Machines*, volume 30:pp. 681–694, 2020

[45] Scao, T.L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., Gallé, M. et al. “Bloom: A 176b-parameter open-access multilingual language model.” *arXiv preprint arXiv:2211.05100*, 2022

[46] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S. et al. “Palm: Scaling language modeling with pathways.” *arXiv preprint arXiv:2204.02311*, 2022

[47] Kautz, D.D., Kuiper, R., Pesut, D.J. and Williams, R.L. “Using nanda, nic, and noc (nnn) language for clinical reasoning with the outcome-present state-test (opt) model.” *International Journal of Nursing Terminologies and Classifications*, volume 17, no. 3:pp. 129–138, 2006

[48] Xiong, Y., Du, B. and Yan, P. "Reinforced transformer for medical image captioning." In "Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10," pp. 673–680. Springer, 2019

[49] Li, C.Y., Liang, X., Hu, Z. and Xing, E.P. "Knowledge-driven encode, retrieve, paraphrase for medical image report generation." In "Proceedings of the AAAI Conference on Artificial Intelligence," volume 33, pp. 6666–6673, 2019

[50] Chen, Z., Song, Y., Chang, T.H. and Wan, X. "Generating radiology reports via memory-driven transformer." *arXiv preprint arXiv:2010.16056*, 2020

[51] Srinivasan, P., Thapar, D., Bhavsar, A. and Nigam, A. "Hierarchical x-ray report generation via pathology tags and multi head attention." In "Proceedings of the Asian Conference on Computer Vision," , 2020

[52] Sirshar, M., Paracha, M.F.K., Akram, M.U., Alghamdi, N.S., Zaidi, S.Z.Y. and Fatima, T. "Attention based automated radiology report generation using cnn and lstm." *Plos one*, volume 17, no. 1:p. e0262209, 2022

[53] Nooralahzadeh, F., Gonzalez, N.P., Frauenfelder, T., Fujimoto, K. and Krauthammer, M. "Progressive transformer-based generation of radiology reports." *arXiv preprint arXiv:2102.09777*, 2021

[54] Alfarghaly, O., Khaled, R., Elkorany, A., Helal, M. and Fahmy, A. "Automated radiology report generation using conditioned transformers." *Informatics in Medicine Unlocked*, volume 24:p. 100557, 2021

[55] Zhang, Z., Chen, P., Shi, X. and Yang, L. "Text-guided neural network training for image recognition in natural scenes and medicine." *IEEE transactions on pattern analysis and machine intelligence*, volume 43, no. 5:pp. 1733–1745, 2019

[56] Wang, X., Peng, Y., Lu, L., Lu, Z. and Summers, R.M. "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," pp. 9049–9058, 2018

[57] Monajatipoor, M., Rouhsedaghat, M., Li, L.H., Jay Kuo, C.C., Chien, A. and Chang, K.W. "Berthop: An effective vision-and-language model for chest x-ray disease diagnosis." In "Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V," pp. 725–734. Springer, 2022

[58] Viegi, G., Maio, S., Fasola, S. and Baldacci, S. "Global burden of chronic respiratory diseases." *Journal of Aerosol Medicine and Pulmonary Drug Delivery*, volume 33, no. 4:pp. 171–177, 2020

[59] Santulli, G. "Epidemiology of cardiovascular disease in the 21st century: Updated updated numbers and updated facts." *Journal of Cardiovascular Disease Research*, volume 1, no. 1, 2013

[60] Porter, R. "The patient's view: doing medical history from below." *Theory and society*, volume 14:pp. 175–198, 1985

[61] Hampton, J.R., Harrison, M., Mitchell, J.R., Prichard, J.S. and Seymour, C. "Relative contributions of history-taking, physical examination, and laboratory investigation to diagnosis and management of medical outpatients." *Br Med J*, volume 2, no. 5969:pp. 486–489, 1975

[62] Waller, K.C. and Fox, J. "Importance of health history in diagnosis of an acute illness." *The Journal for Nurse Practitioners*, volume 16, no. 6:pp. e83–e86, 2020

[63] Muhrer, J.C. "The importance of the history and physical in diagnosis." *The Nurse Practitioner*, volume 39, no. 4:pp. 30–35, 2014

[64] Webb, G.I., Keogh, E. and Miikkulainen, R. "Naïve bayes." *Encyclopedia of machine learning*, volume 15:pp. 713–714, 2010

[65] Jiang, L., Zhang, H. and Cai, Z. "A novel bayes model: Hidden naive bayes." *IEEE Transactions on knowledge and data engineering*, volume 21, no. 10:pp. 1361–1371, 2008

[66] McCallum, A., Nigam, K. et al. "A comparison of event models for naive bayes text classification." In "AAAI-98 workshop on learning for text categorization," volume 752, pp. 41–48. Madison, WI, 1998

[67] White, H. "Learning in artificial neural networks: A statistical perspective." *Neural computation*, volume 1, no. 4:pp. 425–464, 1989

[68] Wright, R.E. "Logistic regression.", 1995

[69] Kleinbaum, D.G., Dietz, K., Gail, M., Klein, M. and Klein, M. *Logistic regression*. Springer, 2002

[70] DeMaris, A. "A tutorial in logistic regression." *Journal of Marriage and the Family*, pp. 956–968, 1995

[71] Hosmer Jr, D.W., Lemeshow, S. and Sturdivant, R.X. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013

[72] Hilbe, J.M. *Logistic regression models*. CRC press, 2009

[73] Yang, H., Luo, L., Chueng, L.P., Ling, D. and Chin, F. "Deep learning and its applications to natural language processing." *Deep learning: Fundamentals, theory and applications*, pp. 89–109, 2019

[74] Shaalan, K., Hassanien, A.E. and Tolba, F. *Intelligent natural language processing: trends and applications*, volume 740. Springer, 2017

[75] Jensen, K., Heidorn, G.E. and Richardson, S.D. *Natural language processing: the PLNLP approach*, volume 196. Springer Science & Business Media, 2012

[76] Tsai, H., Riesa, J., Johnson, M., Arivazhagan, N., Li, X. and Archer, A. "Small and practical bert models for sequence labeling." *arXiv preprint arXiv:1909.00100*, 2019

[77] Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z. and Tang, J. "Gpt understands, too." *arXiv preprint arXiv:2103.10385*, 2021

[78] Goyal, M., Knackstedt, T., Yan, S. and Hassanpour, S. "Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities." *Computers in Biology and Medicine*, volume 127:p. 104065, 2020

[79] Venmathi, A., Ganesh, E. and Kumaratharan, N. "A review of medical image classification and evaluation methodology for breast cancer diagnosis with computer aided mammography." *Int'l Journal of Applied Engineering Research*, volume 10, no. 11:pp. 30045–30054, 2015

[80] Yadav, S.S. and Jadhav, S.M. "Deep convolutional neural network based medical image classification for disease diagnosis." *Journal of Big data*, volume 6, no. 1:pp. 1–18, 2019

[81] Okuboyejo, D.A., Olugbara, O.O. and Odunaike, S.A. "Automating skin disease diagnosis using image classification." In "proceedings of the world congress on engineering and computer science," volume 2, pp. 850–854, 2013

[82] Jeyaraj, P.R. and Samuel Nadar, E.R. "Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm." *Journal of cancer research and clinical oncology*, volume 145:pp. 829–837, 2019

[83] Chen, E.L., Chung, P.C., Chen, C.L., Tsai, H.M. and Chang, C.I. "An automatic diagnostic system for ct liver image classification." *IEEE transactions on biomedical engineering*, volume 45, no. 6:pp. 783–794, 1998

[84] Guo, Z., Li, X., Huang, H., Guo, N. and Li, Q. "Deep learning-based image segmentation on multimodal medical imaging." *IEEE Transactions on Radiation and Plasma Medical Sciences*, volume 3, no. 2:pp. 162–169, 2019

[85] Liu, X., Song, L., Liu, S. and Zhang, Y. "A review of deep-learning-based medical image segmentation methods." *Sustainability*, volume 13, no. 3:p. 1224, 2021

[86] Masubuchi, S., Watanabe, E., Seo, Y., Okazaki, S., Sasagawa, T., Watanabe, K., Taniguchi, T. and Machida, T. "Deep-learning-based image segmentation integrated with optical microscopy for automatically searching for two-dimensional materials." *npj 2D Materials and Applications*, volume 4, no. 1:p. 3, 2020

[87] Hesamian, M.H., Jia, W., He, X. and Kennedy, P. "Deep learning techniques for medical image segmentation: achievements and challenges." *Journal of digital imaging*, volume 32:pp. 582–596, 2019

[88] Tripathi, M. "Analysis of convolutional neural network based image classification techniques." *Journal of Innovative Image Processing (JIIP)*, volume 3, no. 02:pp. 100–117, 2021

[89] Raj, A., Bresler, Y. and Li, B. "Improving robustness of deep-learning-based image reconstruction." In "International Conference on Machine Learning," pp. 7932–7942. PMLR, 2020

[90] de Haan, K., Rivenson, Y., Wu, Y. and Ozcan, A. "Deep-learning-based image reconstruction and enhancement in optical microscopy." *Proceedings of the IEEE*, volume 108, no. 1:pp. 30–50, 2019

[91] Antun, V., Renna, F., Poon, C., Adcock, B. and Hansen, A.C. "On instabilities of deep learning in image reconstruction and the potential costs of ai." *Proceedings of the National Academy of Sciences*, volume 117, no. 48:pp. 30088–30095, 2020

[92] Tatsugami, F., Higaki, T., Nakamura, Y., Yu, Z., Zhou, J., Lu, Y., Fujioka, C., Kitagawa, T., Kihara, Y., Iida, M. et al. "Deep learning–based image restoration algorithm for coronary ct angiography." *European radiology*, volume 29:pp. 5322–5329, 2019

[93] Haskins, G., Kruger, U. and Yan, P. “Deep learning in medical image registration: a survey.” *Machine Vision and Applications*, volume 31:pp. 1–18, 2020

[94] Fu, Y., Lei, Y., Wang, T., Curran, W.J., Liu, T. and Yang, X. “Deep learning in medical image registration: a review.” *Physics in Medicine & Biology*, volume 65, no. 20:p. 20TR01, 2020

[95] Chen, X., Diaz-Pinto, A., Ravikumar, N. and Frangi, A.F. “Deep learning in medical image registration.” *Progress in Biomedical Engineering*, volume 3, no. 1:p. 012003, 2021

[96] De Bruijne, M. “Machine learning approaches in medical image analysis: From detection to diagnosis.”, 2016

[97] Liu, T., Siegel, E. and Shen, D. “Deep learning and medical image analysis for covid-19 diagnosis and prediction.” *Annual Review of Biomedical Engineering*, volume 24:pp. 179–201, 2022

[98] Rehouma, R., Buchert, M. and Chen, Y.P.P. “Machine learning for medical imaging-based covid-19 detection and diagnosis.” *International Journal of Intelligent Systems*, volume 36, no. 9:pp. 5085–5115, 2021

[99] Kim, H.E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M.E. and Ganslandt, T. “Transfer learning for medical image classification: A literature review.” *BMC medical imaging*, volume 22, no. 1:p. 69, 2022

[100] Kora, P., Ooi, C.P., Faust, O., Raghavendra, U., Gudigar, A., Chan, W.Y., Meenakshi, K., Swaraja, K., Plawiak, P. and Acharya, U.R. “Transfer learning techniques for medical image analysis: A review.” *Biocybernetics and Biomedical Engineering*, volume 42, no. 1:pp. 79–107, 2022

[101] Chen, S., Ma, K. and Zheng, Y. “Med3d: Transfer learning for 3d medical image analysis.” *arXiv preprint arXiv:1904.00625*, 2019

[102] Morid, M.A., Borjali, A. and Del Fiol, G. “A scoping review of transfer learning research on medical image analysis using imagenet.” *Computers in biology and medicine*, volume 128:p. 104115, 2021

[103] Dreyer, K.J., Kalra, M.K., Maher, M.M., Hurier, A.M., Asfaw, B.A., Schultz, T., Halpern, E.F. and Thrall, J.H. “Application of recently developed computer algorithm for automatic classification of unstructured radiology reports: validation study.” *Radiology*, volume 234, no. 2:pp. 323–329, 2005

[104] Dreyer, K.J. “Information theory entropy reduction program.”, June 17 2014. US Patent 8,756,234

[105] Yetisgen-Yildiz, M., Gunn, M.L., Xia, F. and Payne, T.H. “A text processing pipeline to extract recommendations from radiology reports.” *Journal of biomedical informatics*, volume 46, no. 2:pp. 354–362, 2013

[106] Cho, J., Lee, K., Shin, E., Choy, G. and Do, S. “Medical image deep learning with hospital pacs dataset.” *arXiv preprint arXiv:1511.06348*, 2015

[107] Friedman, C., Shagina, L., Socratous, S.A. and Zeng, X. “A web-based version of medlee: A medical language extraction and encoding system.” In “Proceedings of the AMIA Annual Fall Symposium,” p. 938. American Medical Informatics Association, 1996

[108] Soderland, S. “Building a machine learning based text understanding system.” In “Proceedings of IJCAI Workshop on Adaptive Text Extraction and Mining,” pp. 64–70. Citeseer, 2001

[109] Taira, R.K., Soderland, S.G. and Jakobovits, R.M. “Automatic structuring of radiology free-text reports.” *Radiographics*, volume 21, no. 1:pp. 237–245, 2001

[110] Christensen, L., Haug, P. and Fiszman, M. “Mplus: a probabilistic medical language understanding system.” In “Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain,” pp. 29–36, 2002

[111] Pan, S.J. and Yang, Q. “A survey on transfer learning.” *IEEE Transactions on knowledge and data engineering*, volume 22, no. 10:pp. 1345–1359, 2010

[112] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J. “Distributed representations of words and phrases and their compositionality.” *Advances in neural information processing systems*, volume 26, 2013

[113] Pennington, J., Socher, R. and Manning, C.D. “Glove: Global vectors for word representation.” In “Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP),” pp. 1532–1543, 2014

[114] Ilić, S., Marrese-Taylor, E., Balazs, J.A. and Matsuo, Y. “Deep contextualized word representations for detecting sarcasm and irony.” *arXiv preprint arXiv:1809.09795*, 2018

[115] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. et al. “Improving language understanding by generative pre-training.”, 2018

[116] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. et al. “Language models are unsupervised multitask learners.” *OpenAI blog*, volume 1, no. 8:p. 9, 2019

[117] Sun, C., Qiu, X., Xu, Y. and Huang, X. “How to fine-tune bert for text classification?” In “Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18,” pp. 194–206. Springer, 2019

[118] Sennrich, R., Haddow, B. and Birch, A. “Neural machine translation of rare words with subword units.” *arXiv preprint arXiv:1508.07909*, 2015

[119] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. “Attention is all you need.” *Advances in neural information processing systems*, volume 30, 2017

[120] Rasmy, L., Xiang, Y., Xie, Z., Tao, C. and Zhi, D. “Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction.” *NPJ digital medicine*, volume 4, no. 1:p. 86, 2021

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

[122] Okolo, G.I., Katsigiannis, S. and Ramzan, N. “Ievit: An enhanced vision transformer architecture for chest x-ray image classification.” *Computer Methods and Programs in Biomedicine*, volume 226:p. 107141, 2022

[123] Taslimi, S., Taslimi, S., Fathi, N., Salehi, M. and Rohban, M.H. “Swinchex: Multi-label classification on chest x-ray images with transformers.” *arXiv preprint arXiv:2206.04246*, 2022

[124] Jun, E., Jeong, S., Heo, D.W. and Suk, H.I. “Medical transformer: Universal brain encoder for 3d mri analysis.” *arXiv preprint arXiv:2104.13633*, 2021

[125] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L. and Zhou, Y. “Transunet: Transformers make strong encoders for medical image segmentation.” *arXiv preprint arXiv:2102.04306*, 2021

[126] Zhang, Y., Liu, H. and Hu, Q. “Transfuse: Fusing transformers and cnns for medical image segmentation.” In “Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France,

September 27–October 1, 2021, Proceedings, Part I 24," pp. 14–24. Springer, 2021

[127] Li, S., Sui, X., Luo, X., Xu, X., Liu, Y. and Goh, R. "Medical image segmentation using squeeze-and-expansion transformers." *arXiv preprint arXiv:2105.09511*, 2021

[128] Zhang, Z., Yu, L., Liang, X., Zhao, W. and Xing, L. "Transct: dual-path transformer for low dose computed tomography." In "Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24," pp. 55–64. Springer, 2021

[129] Koonce, B. and Koonce, B. "Resnet 50." *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*, pp. 63–72, 2021

[130] Koo, H.J., Lim, S., Choe, J., Choi, S.H., Sung, H. and Do, K.H. "Radiographic and ct features of viral pneumonia." *Radiographics*, volume 38, no. 3:pp. 719–739, 2018

[131] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K. 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, pp. 590–597, 2019

[132] Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K.i., Matsui, M., Fujita, H., Kodera, Y. and Doi, K. "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*, volume 174, no. 1:pp. 71–74, 2000

[133] Van Ginneken, B., Stegmann, M.B. and Loog, M. "Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database." *Medical image analysis*, volume 10, no. 1:pp. 19–40, 2006

[134] Li, H., Xiong, P., An, J. and Wang, L. "Pyramid attention network for semantic segmentation." *CoRR*, volume abs/1805.10180, 2018

[135] Ye, W., Yao, J., Xue, H. and Li, Y. "Weakly supervised lesion localization with probabilistic-cam pooling." *arXiv preprint arXiv:2005.14480*, 2020

[136] Mütze, A., Rottmann, M. and Gottschalk, H. “Semi-supervised domain adaptation with cyclegan guided by downstream task awareness.” *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2023

[137] Li, H., Xiong, P., An, J. and Wang, L. “Pyramid attention network for semantic segmentation.” *arXiv preprint arXiv:1805.10180*, 2018

[138] Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R. and McDonald, C.J. “Preparing a collection of radiology examinations for distribution and retrieval.” *Journal of the American Medical Informatics Association*, volume 23, no. 2:pp. 304–310, 2016

[139] Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G. and Horng, S. “Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports.” *Scientific data*, volume 6, no. 1:p. 317, 2019

[140] Johnson, A., Lungren, M., Peng, Y., Lu, Z., Mark, R., Berkowitz, S. and Horng, S. “Mimic-cxr-jpg-chest radiographs with structured labels.” *PhysioNet*, 2019

[141] Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L. and Mark, R.G. “Mimic-iii, a freely accessible critical care database.” *Scientific data*, volume 3, no. 1:pp. 1–9, 2016

[142] Hastuti, E.T., Bustamam, A., Anki, P., Amalia, R. and Salma, A. “Performance of true transfer learning using cnn densenet121 for covid-19 detection from chest x-ray images.” In “2021 IEEE International Conference on Health, Instrumentation & Measurement, and Natural Sciences (InHeNce),” pp. 1–5. IEEE, 2021

[143] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L. “Microsoft coco: Common objects in context.” In “Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13,” pp. 740–755. Springer, 2014

[144] Li, L.H., You, H., Wang, Z., Zareian, A., Chang, S.F. and Chang, K.W. “Unsupervised vision-and-language pre-training without parallel images and captions.” *arXiv preprint arXiv:2010.12831*, 2020

[145] Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L. and Mark, R.G. “Mimic-iii, a freely accessible critical care database.” *Scientific Data*, volume 3, no. 1, 2016

[146] Ramchoun, H., Ghanou, Y., Ettaouil, M. and Janati Idrissi, M.A. "Multilayer perceptron: Architecture optimization and training.", 2016

[147] Hendrycks, D. and Gimpel, K. "Gaussian error linear units (gelus).", 2020

[148] Webb, G.I., Keogh, E. and Miikkulainen, R. "Naïve bayes." *Encyclopedia of machine learning*, volume 15:pp. 713–714, 2010

[149] Bottou, L. "Stochastic gradient descent tricks." *Neural Networks: Tricks of the Trade: Second Edition*, pp. 421–436, 2012

[150] LaValley, M.P. "Logistic regression." *Circulation*, volume 117, no. 18:pp. 2395–2399, 2008

[151] Ikonomakis, M., Kotsiantis, S. and Tampakas, V. "Text classification using machine learning techniques." *WSEAS transactions on computers*, volume 4, no. 8:pp. 966–974, 2005

[152] Sebastiani, F. "Machine learning in automated text categorization." *ACM computing surveys (CSUR)*, volume 34, no. 1:pp. 1–47, 2002

[153] Agarwal, B. and Mittal, N. "Text classification using machine learning methods-a survey." In "Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012," pp. 701–709. Springer, 2014