

**AN INTELLIGENT FRAMEWORK FOR  
AN EFFECTIVE CLINICAL  
RECOMMENDATION SYSTEM TO  
PREDICT DISEASES FROM  
MULTIMODAL MEDICAL DATA**

**THESIS**

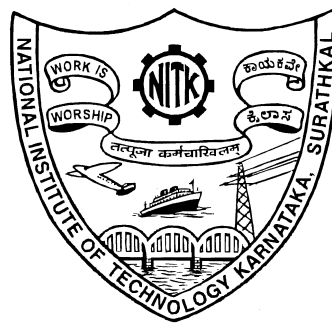
Submitted in partial fulfillment of the requirements  
for the award of the degree of

**DOCTOR OF PHILOSOPHY**

by

**SHASHANK**

(Reg. No.: 177087IT502)



DEPARTMENT OF INFORMATION TECHNOLOGY  
NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA  
SURATHKAL, MANGALORE - 575 025

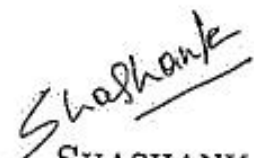
JULY 2023



# DECLARATION

I hereby declare that the Research Thesis entitled "AN INTELLIGENT FRAMEWORK FOR AN EFFECTIVE CLINICAL RECOMMENDATION SYSTEM TO PREDICT DISEASES FROM MULTIMODAL MEDICAL DATA" which is being submitted to National Institute of Technology Karnataka, Surathkal in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy in Information Technology is a bonafide report of the research work carried out by me. The material contained in this Research Thesis has not been submitted to any University or Institution for the award of any degree.

Place : NITK - Surathkal  
Date : 28<sup>th</sup> July 2023

  
SHASHANK  
Reg.No.: 177087IT502  
Department of IT,  
NITK Surathkal.

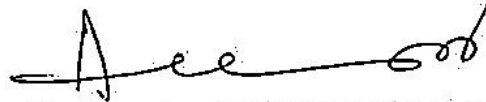




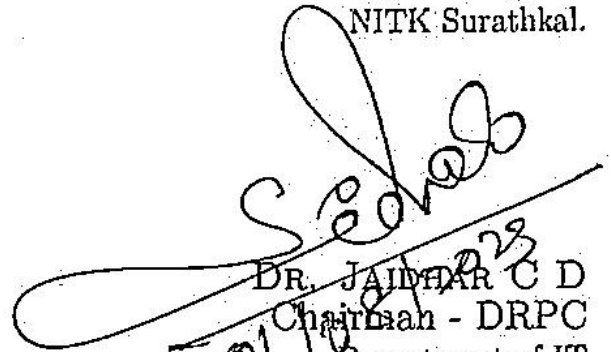
# CERTIFICATE

This is to certify that the Research Thesis entitled, "AN INTELLIGENT FRAME-  
WORK FOR AN EFFECTIVE CLINICAL RECOMMENDATION SYS-  
TEM TO PREDICT DISEASES FROM MULTIMODAL MEDICAL DATA"  
submitted by SHASHANK. (Reg. No. 177087IT502), as the record of research  
work carried out by him, is accepted as the Research Thesis submission in partial fulfil-  
ment of the requirements for the award of the degree of Doctor of Philosophy.

Place : NITK - Surathkal  
Date : 28<sup>th</sup> July 2023



PROF. ANANTHANARAYANA V S.  
Research Guide  
Professor  
Department of IT  
NITK Surathkal.



DR. JAIDEV C D  
Chairman - DRPC  
Department of IT  
NITK Surathkal  
Mangaluru 575 025, INDIA



*Dedicated to*

*My beloved mother and father, whose endless love,  
unwavering support, and sacrifices have made my  
academic journey possible, my dedicated guide whose  
mentorship and guidance have shaped me into a  
confident researcher, and all the special ones who have  
supported and inspired me throughout my PhD journey*



# Acknowledgements

At the very outset, I would like to convey my heartfelt thanks to my dear mentor, Prof. Ananthanarayana V. S., for his invaluable assistance, motivation, and backing during my doctoral research. His knowledge and perspectives have played a vital role in influencing my research work and enhancing my scholarly development. His unwavering support, patience, and constructive feedback have been critical in refining my thoughts, shaping my arguments, and advancing my research abilities. To work under the guidance of a distinguished mentor for my Ph.D. study is a privilege that I deeply appreciate.

I wish to express my utmost appreciation to the members of my RPAC committee, Prof. M. S. Bhat, Dept. of ECE, and Dr. Nagamma Patil, Dept. of IT, for their involvement in validating this research. Their valuable feedback and motivation throughout the research process have been incredibly valuable to me. I am also grateful to Dr. Ajit Mahale, Dept. of Radiology, KMC Hospital, Mangalore for all the enduring help during the medical data collection and expert advice during this research. His guidance and expertise were essential in ensuring the accuracy and reliability of my findings.

I want to express my heartfelt gratitude to the Head of the Department, faculties and staff members of the Dept. of IT, NITK for their valuable assistance whenever I needed it. I would also like to extend my thanks to NMAMIT, Nitte, for their unwavering backing and encouragement throughout my research work. Heartfelt thanks to KMC Hospital, Mangalore for continuous support during the data collection process.

I express my honor to my amazing parents, Mallika C Shetty and Chandrashekar Shetty, for the love, perseverance, and emotional support they have given me. I am deeply appreciative of their constant presence in my life and the invaluable role they played in helping me reach this point. I express my gratitude to my younger brother, Shaswath Shetty, and my wife, Nidhi Hegde, for their consistent accompaniment and presence during my Ph.D. journey.

I want to express my special thanks to my dear friends - Ankitha A. N., Krishna

P R., Puneeth R. P., Dr Sanjay S. B., Dr Karthik K., Dr Rathina R., Dr Manjunath K V., Dr Ashwin T. S., Archana B., Sanket S. S., Tulasi G. D., and others, for constant support and encouragement to move ahead during this journey. I also want to thank my fellow research scholars and lab-mates for fruitful research discussions and for keeping a fun-filled environment during the last five years.

Finally, I am indebted to my family and friends for their constant love, encouragement, and prayers, which have sustained me through the ups and downs of this challenging endeavor. Lastly, I extend my gratitude to all those who have assisted or provided aid in any form towards the successful completion of my research project.

*SHASHANK*

# Abstract

Over the past few decades, the enormous expansion of medical data has led to a way for data analysis in the smart healthcare system. Data analytics in healthcare typically involves the use of statistical and machine learning algorithms to process and analyze clinical data in order to identify correlations and insights that can help enhance health outcomes - in terms of automated disease prediction with minimized human errors, a reduced readmission rate, improved clinical care at a lower cost, and optimized hospital operations. In this direction, over the years, there has been a significant study focusing on Health Information Systems (HIS), particularly Clinical Recommendation Systems (CRS). A CRS offers computer-generated suggestions and advice to healthcare professionals when making clinical decisions. These systems evaluate patient information and propose suitable treatment alternatives, considering clinical guidelines, evidence-based medicine, and other pertinent factors. Lately, a tremendous amount of clinical data has been acquired from various sources, including Electronic Health Records (EHRs), medical imaging, laboratory tests, wearable devices, health apps, telemedicine, and genomic data, which led to the concept of multimodality. Recent progress in deep learning and machine learning algorithms has facilitated the use of artificial intelligence techniques on multimodal medical data, helping to improve diagnostic predictions. Despite the considerable advantages offered by CRSs, their maximum potential can only be realized by effectively tackling several existing challenges. There is a considerable prospect of enhancing the predictive model's ability, particularly with respect to multimodal medical data.

The primary objective of the research work presented in this thesis is to develop an effective clinical recommendation system that can accurately predict abnormalities from diverse types of clinical data for personalized, data-driven recommendations to healthcare providers. This study explores multiple approaches for disease prediction using both unimodal and multimodal data sources, including diagnostic clinical notes and radiology images. The research also presents the cross-modal task of generating diagnostic reports from radiology images and analyzes the effec-

tiveness of different imaging sequences in predicting diseases. Radiology reports contain rich information about patients' health conditions; however, their unstructured format makes it challenging to retrieve this valuable information. Towards the unimodal task, we proposed an effective Unimodal Medical Text Embedding Subnetwork (UM-TES) that incorporates a knowledge base trained on a large corpus to extract the textual features and predict the pulmonary abnormalities from the unstructured radiology free-text reports. The benchmarking analysis revealed that UM-TES outperformed standard NLP and ML techniques in predicting pulmonary diseases from unstructured diagnostic reports. Diagnostic imaging plays a critical role in modern medicine, serving as an essential tool to aid in the prognosis and therapy of various health ailments, supporting essential applications of recommendation systems. The texture and shape of the tissues in the diagnostic images are essential aspects of diagnosis. The pulmonary diseases have irregular and different sizes; hence, several studies sought to add new components to existing deep learning techniques for acquiring multi-scale imaging features from diagnostic chest X-rays. Towards this unimodal task of leveraging diagnostic images for disease prediction, the explainable and lightweight Unimodal Medical Visual Encoding Subnetwork (UM-VES) is proposed to predict pulmonary abnormalities from the diagnostic chest X-ray images. The proposed model is tested with a publicly available Open-I Dataset and data collected from a private hospital. After the comprehensive assessment, it was observed that the performance of the designed approach showcased a 7% to 18% increase in accuracy compared to the existing method.

Many contemporary DL strategies for radiology focus on a single modality of data utilizing imaging features without considering the clinical context that provides more valuable complementary information for clinically consistent prognostic decisions. Towards this objective, the two novel multimodal medical fusion techniques: Compact Bilinear Pooling and Deep Hadamard Product is proposed to integrate textual and visual medical features from clinical text reports and Chest X-rays to predict abnormalities from multimodal data. A comprehensive analysis was conducted and compared the performance of unimodal and multimodal models. The proposed models were applied to standard augmented data and the synthetic data generated to check the model's ability to predict from the new and unseen data. The proposed multimodal models have given superior results compared to the unimodal models. There has been a significant contribution in the area of cross-modal medical description generation. In order to create accurate and reliable radiology reports, radiologists need to be experienced and dedicate



sufficient time to reviewing medical images. However, many radiology reports end with ambiguous conclusions, leading patients to undergo additional tests, such as pathology or advanced imaging. To address this, we propose an encoder-decoder-based deep learning framework to produce diagnostic radiology reports based on chest X-ray images. Additionally, we have developed a dynamic web portal that accepts chest X-rays as input and generates a radiology report as output. We conducted a thorough analysis and compared the performance of our model with other state-of-the-art deep learning approaches. Our results show that our proposed model outperforms existing models in terms of BLEU score on the Indiana University Dataset.

In the medical domain, the radiologist examines multiple imaging modalities to determine the disease outcome. Acute infarct is one such illness where radiologists utilize multiple MRI sequences like DWI, T2-Flair, ADC, and SWI to examine the prognosis. Currently, expert clinicians rely on manual interpretation of imaging methods for diagnosing diseases. However, with the rising number of chronic cases, this approach has become a burden on healthcare professionals, increasing their cognitive and diagnostic workload. Towards this multi-image fusion task, We introduce the DL framework, including contour-based brain segmentation techniques and two stacked multi-channel convolution neural networks, SMC-CNN-M and SMC-CNN-I, to predict the disease from both multiple and individual MRI sequences. We evaluate our proposed models on a medical dataset collected from a private hospital and compare their classification performance to that of state-of-the-art deep learning networks. Additionally, we conduct a quantitative, qualitative, and ablation study on different MRI sequences to assess their effectiveness and generate synthetic data using DCGAN to compare model performance.

**KEYWORDS:** *Unstructured Data Analysis, Multimodal Representation, Cross-modal Retrieval, Medical Image Fusion, Machine Learning, Deep Learning*



# Contents

List of Figures	xii
List of Tables	xv
List of Abbreviations	xix

## Part I - Introduction and Background

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Unstructured Medical Text Analysis . . . . .	9
1.2	Unstructured Medical Image Analysis . . . . .	11
1.3	Multimodal Medical Data Analysis . . . . .	13
1.3.1	Multimodal Medical Image-text Data Analysis . . . . .	15
1.3.2	Cross-Modal Medical Image-Text Analysis . . . . .	17
1.3.3	Multimodal Medical Image Analysis . . . . .	17
1.4	Prominent Obstacles and Concerns . . . . .	20
1.5	Summary . . . . .	23
1.6	Thesis Organization . . . . .	23
<b>2</b>	<b>Literature Review</b>	<b>25</b>
2.1	CRSs for Unstructured Medical Text Data Analysis . . . . .	27
2.1.1	Disease Prediction from Unstructured Radiology Reports . . . . .	27
2.2	CRSs for Unstructured Medical Image Data Analysis . . . . .	30
2.2.1	Disease Prediction from Unstructured Chest X-ray Images . . . . .	31
2.2.1.1	Disease Detection and Localization Task . . . . .	31
2.2.1.2	Disease Classification and Prediction Task . . . . .	33
2.2.1.3	Data Augmentation vs. Synthetic Data Generation . . . . .	34
2.3	CRSs for Multimodal Medical Data Analysis . . . . .	43
2.3.1	Multimodal Diagnostic Image and Text Analysis . . . . .	48
2.3.2	Cross-modal Medical Report Generation . . . . .	56

2.3.3	Multimodal Medical Image Analysis . . . . .	61
2.4	Outcome of Literature Review . . . . .	68
2.5	Summary . . . . .	72
<b>3</b>	<b>Problem Description</b>	<b>75</b>
3.1	Background . . . . .	75
3.2	Research Gaps . . . . .	75
3.3	Scope of the Work . . . . .	77
3.3.1	Problem Statement . . . . .	77
3.3.2	Research Objectives . . . . .	78
3.4	Brief Overview of Proposed CRS Framework . . . . .	80
3.4.1	Unimodal Medical Text Embedding Subnetwork . . . . .	80
3.4.2	Unimodal Medical Visual Encoding Subnetwork . . . . .	81
3.4.3	Deep Medical Multimodal Fusion Network . . . . .	82
3.4.4	Cross-modal Deep Learning Framework . . . . .	83
3.4.5	Multimodal Image Fusion Network . . . . .	84
3.5	Research Contributions . . . . .	85
3.6	Summary . . . . .	86
<b>Part II - Unimodal Unstructured Medical Data Analysis</b>		
<b>4</b>	<b>Unimodal Medical Text Embedding Subnetwork (UM-TES)</b>	<b>89</b>
4.1	Introduction . . . . .	89
4.1.1	Problem Statement . . . . .	91
4.2	Methodology . . . . .	92
4.2.1	Basic Pre-processing . . . . .	92
4.2.2	Clinical Knowledge-based Text Modelling . . . . .	95
4.2.3	Embedding Layer . . . . .	100
4.2.4	Discriminative Dimensionality Reduction using Convolution Neural Network (DDR-CNN) . . . . .	100
4.2.5	Network Structure of UM-TES: . . . . .	103
4.2.6	Fully connected Deep Neural Network (DNN) for Disease Prediction . . . . .	105
4.3	Comparison with State-of-the-art Text Modelling Strategies . . . . .	105
4.4	Experimental Results and Discussion . . . . .	109
4.4.1	Datasets and Cohort Selection . . . . .	109
4.4.2	Data Preparation and Augmentation Stage . . . . .	111

4.4.3	Evaluation Metrics . . . . .	111
4.4.4	Results and Discussions . . . . .	115
4.4.4.1	Performance Analysis with the State-of-the-art NLP Techniques . . . . .	115
4.4.4.2	Performance Analysis with the State-of-the-Art ML techniques . . . . .	116
4.4.4.3	Effect of Clinical Knowledge-based Text Modelling	118
4.4.4.4	Effect of CNN-based Discriminate Dimensionality Reduction . . . . .	120
4.5	Summary . . . . .	121
<b>5</b>	<b>Unimodal Medical Visual Encoding Subnetwork (UM-VES)</b>	<b>123</b>
5.1	Introduction . . . . .	123
5.1.1	Problem Statement . . . . .	124
5.2	Methodology . . . . .	126
5.2.1	Multi-Scale Dilation Layer . . . . .	127
5.2.2	Depthwise Separable Convolution Neural Network (DS-CNN)	130
5.2.3	Network Structure and Training of UM-VES . . . . .	135
5.2.4	Fully Connected Deep Neural Network for Abnormality Pre- diction . . . . .	137
5.2.5	Disease Visualization using Grad-CAM Technique . . . . .	138
5.3	Data Augmentation vs. Synthetic Data Generation: An Empirical Evaluation for Enhancing Radiology Image Classification . . . . .	142
5.3.1	Basic Data Augmentation . . . . .	143
5.3.2	RAD-DCGAN for Synthetic Data Generation . . . . .	144
5.3.3	Objective Function of RAD-DCGAN . . . . .	147
5.3.4	Loss Function of RAD-DCGAN . . . . .	147
5.4	Experimental setup . . . . .	148
5.4.1	Parameter Configurations . . . . .	148
5.4.2	Radiology Cohort Selection . . . . .	149
5.4.3	Data Augmentation Settings . . . . .	151
5.4.4	Evaluation Criteria . . . . .	151
5.5	Results and Discussions . . . . .	152
5.5.1	Quantitative Analysis of Proposed UM-VES with the Fine- tuned Pre-trained Deep Learning Models . . . . .	153
5.5.2	Performance Analysis of Proposed UM-VES with the Exist- ing State-of-the-art DL Strategies . . . . .	155

5.5.3	Qualitative Analysis of Proposed UM-VES . . . . .	157
5.6	Data Augmentation vs. Synthetic Data Generation . . . . .	160
5.6.1	An Empirical Evaluation for Enhancing Radiology Image Classification . . . . .	160
5.6.2	Cohort Selection . . . . .	161
5.6.3	Results and Discussions . . . . .	162
5.7	Summary . . . . .	165

### Part III - Multimodal Unstructured Medical Data Analysis

<b>6</b>	<b>Deep Medical Multimodal Fusion Networks (DMMFN)</b>	<b>169</b>
6.1	Introduction . . . . .	169
6.1.1	Problem Statement . . . . .	172
6.2	Methodology . . . . .	173
6.2.1	Compact Bilinear Pooling-based Medical Multimodal Fu- sion Network (CBP-MMFN) . . . . .	174
6.2.2	Deep Hadamard Product-based Medical Multimodal Fusion Network (DHP-MMFN) . . . . .	179
6.3	Experimental Setup and Evaluation . . . . .	182
6.3.1	Datasets and Cohort Selection . . . . .	182
6.3.2	Evaluation Criteria . . . . .	183
6.3.3	Data Preparation and Augmentation Stage . . . . .	183
6.3.3.1	Standard Data Augmentation . . . . .	183
6.3.3.2	Generation of Synthetic CXRs using DCGAN . . . . .	184
6.3.4	Network Configurations and Parameter Settings . . . . .	185
6.3.5	Ablation Study . . . . .	188
6.3.6	Performance Analysis of Unimodal and Multimodal Models . . . . .	190
6.3.7	Performance Analysis on Synthetic Data Generated . . . . .	191
6.3.8	Performance Comparison with the State-of-the-art Models . . . . .	193
6.4	Summary . . . . .	199
<b>7</b>	<b>Cross-Modal Deep Learning Framework for Report Generation</b>	<b>201</b>
7.1	Introduction . . . . .	201
7.1.1	Problem Statement . . . . .	202
7.2	Methodology . . . . .	203
7.2.1	Unimodal Medical Visual Encoding Subnetwork (UM-VES)	203
7.2.2	Unimodal Medical Text Embedding Subnetwork (UM-TES)	204

7.2.3	Long Short-term Memory-based Report Generation . . . . .	205
7.3	Web-based Framework for Report Generation . . . . .	206
7.4	Experimental Setup and Evaluation . . . . .	207
7.5	Summary . . . . .	210
<b>8</b>	<b>Multimodal Image Fusion Network from MRI Images</b>	<b>211</b>
8.1	Introduction . . . . .	211
8.1.1	Problem Statement . . . . .	214
8.2	Materials . . . . .	215
8.2.1	Data Collection . . . . .	215
8.2.2	Standard Augmentation Techniques . . . . .	216
8.2.3	Synthetic Data Generated using DCGAN . . . . .	217
8.3	Methodology . . . . .	217
8.3.1	Contour-based Brain Segmentation for MRI Sequences . . . . .	218
8.3.2	Stacked Multi-Channel Convolution Neural Network (SMC-CNN) . . . . .	219
8.3.3	Acute Brain Infarct Visualization using Grad-CAM . . . . .	222
8.4	Experimental Setup . . . . .	227
8.4.1	Parameter Configuration of Proposed SMC-CNN and State-of-the-art Deep Learning Models . . . . .	227
8.4.2	Evaluation Metrics . . . . .	228
8.5	Results and Discussion . . . . .	231
8.5.1	Quantitative Analysis . . . . .	232
8.5.2	Ablation Study . . . . .	239
8.5.3	Qualitative Analysis . . . . .	240
8.6	Discussion . . . . .	244
8.7	Summary . . . . .	245
<b>9</b>	<b>Conclusion &amp; Future Work</b>	<b>247</b>
9.1	Conclusion . . . . .	247
9.2	Future Work . . . . .	255
<b>A</b>	<b>Publications based on Research Work</b>	<b>257</b>
	<b>References</b>	<b>259</b>





# List of Figures

1.1	Categorization of Clinical Recommendation Systems . . . . .	8
1.2	Sample Unstructured Medical Text . . . . .	10
1.3	Sample Unstructured Medical Image . . . . .	13
1.4	Sample Multimodal Medical Data . . . . .	16
1.5	Sample Multimodal Medical Image . . . . .	18
1.6	Essentials of Clinical Recommendation Systems . . . . .	19
2.1	Classification of Clinical Recommendation Systems . . . . .	26
2.2	General Architecture of Multimodal Medical Data Analysis . . . . .	44
3.1	Systematic Overview of the Intelligent CRS Framework . . . . .	79
3.2	CRS with Unstructured Free-text Reports for Disease Prediction . . . . .	81
3.3	CRS with Unstructured Diagnostic Images for Disease Prediction . . . . .	82
3.4	CRS with Multimodal Unstructured Clinical Data for Disease Prediction . . . . .	83
3.5	CRS with Multimodal Unstructured Clinical Data for Diagnostic Report Generation . . . . .	84
3.6	CRS with Multimodal Diagnostic Images for Disease Prediction . . . . .	85
4.1	Proposed Unimodal Medical Text Embedding Subnetwork (UM- TES) for Text Feature Extraction . . . . .	93
4.2	Overall workflow of various deep learning-based NLP framework for predicting pulmonary diseases from radiology free-text reports . . . . .	108
4.3	Data Augmentation of Radiology Data. . . . .	112
4.4	Performance analysis of UM-TES with state-of-the art NLP models on IU cohort . . . . .	117
4.5	Performance analysis of UM-TES with state-of-the art NLP models on KMC cohort . . . . .	118
4.6	Comparing AUROC performance of proposed Deep learning model w.r.t. State-of-art Machine Learning techniques. . . . .	119

4.7	Effectiveness of Customized Clinical Knowledge-based Text Modelling compared to the GloVe Embeddings . . . . .	120
4.8	Comparison of performance metrics with and without DDR-CNN . . . . .	121
5.1	Proposed Unimodal Medical Visual Encoding Subnetwork (UM-VES) to extract the imaging features . . . . .	128
5.2	Proposed Multi-Scale Dilation Layer (MSDL) . . . . .	131
5.3	Conventional convolution filters and Depthwise Separable filters . . . . .	132
5.4	Overall operation of Depthwise Separable Convolution Neural Network (DS-CNN) . . . . .	136
5.5	General process flow of the DS-CNN followed by Batch Normalization and ReLU . . . . .	137
5.6	Fully Connected Deep Neural Network for abnormality prediction . . . . .	139
5.7	Grad-CAM based visual explanation of the proposed UM-VES framework . . . . .	141
5.8	Schematic representation of the architecture used in this study for disease classification of radiology images using RAD-DCGAN and traditional data augmentation techniques. . . . .	142
5.9	Basic data augmentation techniques applied on X-ray and MR images: (a) original X-ray and MR images, (b) rotated images, (c) zoomed images, (d) after increase in brightness and (e) sheared images . . . . .	143
5.10	The proposed RAD-DCGAN for synthetic image generation from radiology images. . . . .	145
5.11	General architecture of generator and discriminator module of RAD-DCGAN . . . . .	146
5.12	Systematic data augmentation process flow of diagnostic CXRs . . . . .	152
5.13	Experimental observation of the loss and accuracy vs total number of epochs w.r.t 10-fold cross-validation for Open-I X-ray dataset . . . . .	155
5.14	Experimental observation of the loss and accuracy vs total number of epochs w.r.t 10-fold cross-validation for KMC Chest X-ray dataset . . . . .	155
5.15	Performance analysis of proposed UM-VES with the different baseline deep learning model for Open-I CXR dataset . . . . .	156
5.16	Performance analysis of proposed UM-VES with the different baseline deep learning model for KMC Hospital CXR dataset . . . . .	156
5.17	Disease Visualization with Grad-CAM technique . . . . .	159

5.18	The generation of synthetic data after every 20 and 50 epochs in X-ray and MR images, respectively . . . . .	161
5.19	Accuracy and loss during the training of discriminator and generator component in RAD-DCGAN on X-ray images . . . . .	163
5.20	Accuracy and loss during the training of discriminator and generator component in RAD-DCGAN on MR images . . . . .	163
6.1	The grid of normal (No diseases) and abnormal (pulmonary diseases) CXR from Indiana University dataset. . . . .	170
6.2	Proposed Multimodal Medical Tensor Fusion Network-based DL Framework for predicting Abnormality from the heterogeneous radiology CXR and text reports. . . . .	175
6.3	Proposed Compact Bilinear Pooling-based Medical Multimodal Fusion Network (CBP-MMFN) . . . . .	177
6.4	Proposed Deep Hadamard Product-based Medical Multimodal Fusion Network (DHP-MMFN) . . . . .	182
6.5	Architectural diagram of Deep Convolution Generative Adversarial Network . . . . .	186
6.6	Generator Module architecture of DCGAN . . . . .	187
6.7	Discriminator Module architecture of DCGAN . . . . .	187
6.8	Comparison of performance metrics of proposed unimodal vs multimodal models for Indiana dataset . . . . .	191
6.9	Comparison of performance metrics of proposed unimodal vs multimodal models for KMC dataset . . . . .	192
6.10	Discriminator Accuracy on real and fake samples during training of DCGAN . . . . .	193
6.11	Generator and Discriminator loss during the training of DCGAN . . . . .	193
6.12	The synthetic CXR images generated after every 20 epochs for KMC hospital dataset . . . . .	194
6.13	Comparison of performance metrics of proposed unimodal vs multimodal models on Actual data with synthetic data generated from Indiana University dataset . . . . .	195
6.14	Comparison of performance metrics of proposed unimodal vs multimodal models on Actual data with synthetic data generated from KMC Hospital dataset . . . . .	195
7.1	Overall architecture of the proposed cross-modal deep learning-based model for automatic report generation . . . . .	204

7.2	Long Short-term Memory Architecture . . . . .	205
7.3	Client-Server interaction used for predicting reports . . . . .	207
7.4	The dynamic web portal for automatic diagnostic report generation. . . . .	209
8.1	Contour-based brain segmentation of MRI sequence . . . . .	218
8.2	SMC-CNN-M: Stacked Multi-Channel Convolution Neural Network for Multiple MRI Sequences . . . . .	223
8.3	SMC-CNN-I: Stacked Multi-Channel Convolution Neural Network for Individual MRI Sequence . . . . .	224
8.4	Disease Visualization using Gradient-weighted Class Activation Map- ping (Grad-CAM) . . . . .	226
8.5	The learning curve of proposed SMC-CNN-I model . . . . .	238
8.6	Confusion Matrix of proposed SMC-CNN-M model on DWI, T2- flair, ADC and SWI MRI sequences . . . . .	240
8.7	Disease Visualization of Acute Infarct . . . . .	243

# List of Tables

2.1	List of some currently available diagnostic X-ray datasets for chest diseases. . . . .	32
2.2	Summary of Literature Survey - Disease Prediction from Unstructured Chest X-ray Images . . . . .	36
2.3	Summary of Literature Survey - Multimodal Diagnostic Image and Text Analysis . . . . .	52
2.4	Summary of Literature Survey. . . . .	58
2.5	Summary of Findings from Literature Review . . . . .	64
4.1	Cohort Statistics: Chest X-Ray with associated Radiology reports from two Institutions . . . . .	110
4.2	The various data augmentation techniques applied on the medical cohort with the ranges of each techniques . . . . .	113
4.3	Benchmarked performance analysis results of the proposed deep learning-based NLP technique with the state-of-the-art text modelling techniques on the diagnostic clinical free-text cohort collected from the publicly available Indiana University dataset . . . . .	116
4.4	Benchmarked performance analysis results of the proposed deep learning-based NLP technique with the state-of-the-art text modelling techniques on the diagnostic clinical free-text cohort collected from KMC hospital . . . . .	117
4.5	Benchmarking the proposed DNN model with and without Knowledge Base (KB) against the State-of-the-Art Machine Learning Model w.r.t. Indiana University and KMC Hospital Dataset . . . . .	119
5.1	Overall architecture of the proposed UM-VES: Multi-Scale Dilated Network with depthwise Separable convolution . . . . .	127
5.2	Parameter details of all the state-of-the-art Deep Learning Models and the proposed UM-VES . . . . .	149
5.3	Dataset Statistics: Detailed description of the CXR diagnostic images from two medical repositories . . . . .	150

5.4	Image augmentation settings . . . . .	151
5.5	Benchmarked Experimental results of proposed UM-VES Model with the state-of-the-art Deep Learning Model on Open-I Dataset. . . . .	154
5.6	Benchmarked Experimental results of proposed UM-VES Model with the state-of-the-art Deep Learning Model on KMC hospital Dataset. . . . .	154
5.7	Performance analysis of the proposed UM-VES with the existing state-of-the-art deep learning strategies on Open-I Dataset . . . . .	157
5.8	Classification performance metrics for Chest X-Ray Images . . . . .	163
5.9	Classification performance metrics for MRI T2-Flair sequences . . . . .	164
6.1	Cohort Statistics: CXR with associated clinical diagnostic notes from two clinical cohorts . . . . .	183
6.2	Experimental results of the proposed unimodal and multimodal model for abnormality prediction from CXR images and its associated radiology reports collected with standard augmented data from Indiana University and KMC hospital dataset. . . . .	189
6.3	Experimental results of the proposed unimodal and multimodal models for abnormality prediction on Actual data with Synthetic CXRs and radiology reports generated from Indiana University and KMC hospital datasets using DCGAN. . . . .	196
6.4	Comparing the performance of the proposed multimodal models against the existing state-of-the art medical multimodal fusion model for abnormality prediction from CXR and its associated radiology reports from the Indiana University dataset. The results of the state-of-the-art medical multimodal fusion model is taken from their published research work. . . . .	198
7.1	Performance analysis of the proposed model . . . . .	208
7.2	Performance analysis compared with existing work of report generation . . . . .	208
8.1	Cohort Statistics: Detailed description of the MRI sequences collected from KMC private hospital. . . . .	216
8.2	The detailed parameter configuration of proposed SMC-CNN and state-of-the-art Deep learning models . . . . .	229

8.3	Benchmarked performance evaluation of proposed SMC-CNN-I model and the baseline DL techniques for predicting acute brain infarct from DWI MRI sequences obtained from the KMC hospital. . . . .	233
8.4	Benchmarked performance evaluation of proposed SMC-CNN-I model and the baseline DL techniques for predicting acute brain infarct from T2-Flair MRI sequences obtained from the KMC hospital. . . . .	234
8.5	Benchmarked performance evaluation of proposed SMC-CNN-I model and the baseline DL techniques for predicting acute brain infarct from ADC MRI sequences obtained from the KMC hospital. . . . .	235
8.6	Benchmarked performance evaluation of proposed SMC-CNN-I model and the baseline DL techniques for predicting acute brain infarct from SWI MRI sequences obtained from the KMC hospital. . . . .	236
8.7	Ablation study of the proposed SMC-CNN-M by varying the MRI input sequence . . . . .	241





## List of Abbreviations

<i>AB</i>	AdaBoost
<i>AD</i>	Alzheimer’s disease
<i>ADC</i>	Apparent Diffusion Coefficient
<i>AI</i>	Artificial Intelligence
<i>AUPRC</i>	Area Under the Precision-Recall Curve
<i>AUROC</i>	Area Under the Receiver Operating Characteristics
<i>BCI</i>	Brain-Computer Interface
<i>BERT</i>	Bidirectional Encoder Representations from Transformers
<i>Bi – LSTM</i>	Bidirectional Long Short-Term Memory
<i>BLEU</i>	Bilingual Evaluation Understudy
<i>BN</i>	Batch Normalization
<i>BoW</i>	Bag-of-Words
<i>CBOW</i>	Continuous Bag of Words
<i>CBP – MMFN</i>	Compact Bilinear pooling-based Medical Multimodal Fusion Network
<i>cGAN</i>	Conditional Generative Adversarial Network
<i>CKB</i>	Clinical Knowledge Base
<i>CNN</i>	Convolutional Neural Network
<i>COPD</i>	Chronic Obstructive Pulmonary Disease
<i>COVID – 19</i>	COronaVirus Disease of 2019

<i>CPOE</i>	Computerized Physician Order Entry
<i>CPT</i>	Current Procedural Terminology
<i>CRS</i>	Clinical Recommendation System
<i>CSF</i>	Signal of Cerebrospinal Fluid
<i>CT</i>	Computed Tomography
<i>CXR</i>	Chest X-Ray
<i>DCGAN</i>	Deep Convolutional Generative Adversarial Network
<i>DCN</i>	Densely Connected Network
<i>DDI</i>	Drug-Drug Interactions
<i>DDI</i>	Intensive Care Units
<i>DDR – CNN</i>	Discriminative Dimensionality Reduction using CNN
<i>DenseNet – 121</i>	Densely-connected-convolutional Networks-121
<i>DF</i>	Deep Forest
<i>DHP – MMFN</i>	Deep Hadamard Product-based Medical Multimodal Fusion Network
<i>DL</i>	Deep Learning
<i>DMMFN</i>	Deep Medical Multimodal Fusion Network
<i>DNN</i>	Deep Neural Network
<i>DS – CNN</i>	Depthwise Separable Convolution Neural Network
<i>DWI</i>	Diffusion weighted imaging
<i>ECG</i>	Electrocardiogram
<i>EEG</i>	Electroencephalogram
<i>EHR</i>	Electronic Health Record
<i>EMR</i>	Electronic Medical Record

<i>FN</i>	False Negative
<i>FP</i>	False Positive
<i>GAN</i>	Generative Adversarial Network
<i>GAP</i>	Global Average Pooling
<i>GBDT</i>	Gradient Boosting Decision Tree
<i>GloVe</i>	Global Vector
<i>GPU</i>	Graphics Processing Unit
<i>Grad – CAM</i>	Gradient-weighted Class Activation Mapping
<i>GUI</i>	Graphical User Interface
<i>GUI</i>	Open Computing Language
<i>HITECH</i>	Health Information Technology for Economic and Clinical Health
<i>IBM</i>	International Business Machines
<i>ICD</i>	International Statistical Classification of Diseases
<i>ICU</i>	Drug-Drug Interactions
<i>IEC</i>	Institutional Ethics Committee
<i>IST</i>	International Stroke Trial
<i>IU</i>	Indiana University
<i>KB</i>	Knowledge Base
<i>KGAE</i>	Knowledge Graph Auto-Encoder
<i>KMC</i>	Kasturba Medical College
<i>KNN</i>	K-Nearest Neighbour
<i>LiDAR</i>	Light Detection and Ranging
<i>LR</i>	Logistic Regression

<i>LSTM</i>	Long Short-Term Memory
<i>MCC</i>	Matthews Correlation Coefficient
<i>ML</i>	Machine Learning
<i>MLP</i>	Multilayer Perceptron
<i>MRI</i>	Magnetic Resonance Imaging
<i>MSDL</i>	Multi-Scale Dilation Layer
<i>NB</i>	Naive Bayes
<i>NIH</i>	National Institutes of Health
<i>NLP</i>	Natural Language Processing
<i>PD</i>	Parkinson's Disease
<i>PET</i>	Positron Emission Tomography
<i>RAD – DCGAN</i>	RADiology Deep Convolutional Generative Adversarial Network
<i>RAM</i>	Random-Access Memory
<i>RBC</i>	Red Blood Cell
<i>ReLU</i>	Rectified Linear Activation Unit
<i>ResNet</i>	Residual neural Network-50
<i>RF</i>	Random Forest
<i>RNN</i>	Recurrent Neural Network
<i>RoI</i>	Regions of Interest
<i>RPMS</i>	Remote Patient Monitoring Systems
<i>SGD</i>	Stochastic Gradient Descent
<i>SMC – CNN – I</i>	Stacked Multi-Channel Convolutional Neural Networks for Individual image

<i>SMC – CNN – M</i>	Stacked Multi-Channel Convolutional Neural Networks for Multiple images
<i>SMV</i>	Simple Majority Voting
<i>SNOMED</i>	Systematized Nomenclature of Medicine Clinical Terms
<i>SPECT</i>	Single-Photon Emission Computerized Tomography
<i>SVM</i>	Support Vector Machine
<i>SWI</i>	Susceptibility Weighted Imaging
<i>T2 – Flair</i>	T2 Fluid-Attenuated Inversion Recovery
<i>tf – idf</i>	Term Frequency-Inverse Document Frequency
<i>TN</i>	True Negative
<i>TP</i>	True Positive
<i>TREC</i>	Text REtrieval Conference
<i>UM – TES</i>	Unimodal Medical Text Embedding Subnetwork
<i>UM – VES</i>	Unimodal Medical Visual Encoding Subnetwork
<i>WHO</i>	World Health Organization
<i>WSI</i>	Whole Slide Imaging



# PART I

## Introduction and Background





# Chapter 1

## Introduction

The primary objective of the healthcare system is to offer “*healthcare services that are available, affordable, and of superior quality to individuals and communities, with a focus on promoting health, well-being, preventing and curing illnesses*”. High-quality healthcare is crucial as it significantly enables individuals to prevent, diagnose, and treat illnesses effectively, resulting in better health outcomes and improved quality of life with reduced hospital expenses. The World Health Organization (WHO) believes that a healthcare system that operates effectively should be available and accessible to every individual, regardless of their socio-economic background<sup>1</sup>. WHO advocates that the healthcare system should emphasise preventive measures, timely detection, and prompt treatment of illnesses to reduce the disease burden and prevent avoidable deaths<sup>2</sup>. Many healthcare systems around the world are currently facing multiple challenges that negatively affect the quality, accessibility, and affordability of healthcare services. Some of the major issues are listed below:

- *Limited Resources*: Many healthcare systems struggle with inadequate funding, shortages of medical personnel, and insufficient medical supplies, which may impact the quality of the healthcare provided. The impact of the COVID-19 pandemic on healthcare delivery and health outcomes is explored by [Anesi and Kerlin \(2021\)](#), with a particular focus on how the shortage of medical personnel, supplies, and funding has led to significant challenges.
- *Unequal Access*: The availability of necessary healthcare services is not evenly distributed, and regions that are far from urban areas and susceptible

---

<sup>1</sup>WHO the global health Observatory. Online: <https://www.who.int/data/gho/data/themes/topics/health-systems-strengthening>

<sup>2</sup>WHO Strengthening health information systems. Online: <https://apps.who.int/iris/rest/bitstreams/1092654/retrieve>

groups often have restricted access to them. Access to healthcare in some regions of the country is a privilege that only the wealthy can afford, while the poor resort to visiting inadequately resourced private healthcare providers, often paying beyond their means rather than utilizing the available public healthcare facilities (Barik and Thorat, 2015).

- *Increased Healthcare cost:* Rising healthcare costs are creating challenges for many individuals to afford essential medical treatment and procedures. Although increasing healthcare costs are a significant issue in many high-income countries, attempts by political measures to reduce costs have been unsuccessful and have negatively impacted patients and citizens' best interests (Sturmberg and Bircher, 2019).
- *Aging Population and surge in chronic diseases:* As the population ages, the utilization of age-related procedures and treatments is increasing, leading to higher healthcare costs. The healthcare systems of many nations are under pressure to provide care for chronic and age-related illnesses due to the aging population (Cristea *et al.*, 2020).

A robust healthcare system can effectively address the aforementioned challenges by providing quality patient care and making a valuable contribution to the development of healthcare in a country (Croon *et al.*, 2021). A Clinical Recommendation System is a critical component of modern healthcare delivery systems that is necessary for providing high-quality healthcare. CRS is “*a health information system that assists clinicians in making well-informed decisions about patient care by utilizing patient data, including medical history, current medications, and symptoms, to provide enhanced evidence-based recommendations to clinicians in real-time*” (Berner, 2010). In the 1970s, Computerized CRS were prevalent but had certain limitations, such as inadequate system integration, a time-consuming process, and were mostly restricted to academic research (Shortliffe and Buchanan, 1975). The application of computer technology in the field of medicine has given rise to both ethical and legal issues, particularly with regard to the extent of physician autonomy and accountability for the imperfect nature of the system's recommendations (Sittig *et al.*, 2016). In recent years, CRS has adopted web-based applications integrated with electronic medical records (EMR) as a means of streamlining the data collection process and improving patient care. The use of a CRS allows medical professionals to enhance the precision of their diagnoses, reduce mistakes, and optimize treatment strategies, resulting in improved patient

outcomes (Sreejith *et al.*, 2022). The five key usages of CRS are as follows (Sutton *et al.*, 2020):

1. *Patient Safety:* CRS is often utilized in approaches aimed at decreasing medication errors. Mistakes related to drug-drug interactions (DDI) are frequently reported and avoidable, with as many as 65% of hospitalized patients being subjected to one or more combinations that have the potential to cause harm (Vonbach *et al.*, 2008). CRS can provide guidance on medication prescriptions, specifically regarding DDI and potential overdosing errors. (Zhou *et al.*, 2021). The notifications or alerts produced by these systems are among the most frequently employed categories of decision support tools (Koutkias and and, 2018). The research has identified considerable inconsistencies in how notifications for DDIs are presented (such as passive or active/disruptive), which interactions are given priority (Phansalkar *et al.*, 2012), and the methods employed to detect DDIs (McEvoy *et al.*, 2016). These systems frequently generate alerts that are not relevant, and there is no established guideline for the most effective way to present alerts to healthcare providers. Towards health information systems, the United States government has created a catalog of high-priority DDIs for recommendation systems, which are adopted by other countries like Belgium (Cornu *et al.*, 2018) and Korea (Cho *et al.*, 2016). Several clinical recommendation systems like Computerized Physician Order Entry (CPOE) (Helmons *et al.*, 2015), Remote Patient Monitoring Systems (RPMS) (Boikanyo *et al.*, 2023), and Telemedicine Systems (Mackintosh *et al.*, 2016) can be connected to patient monitoring devices such as blood glucose meters, blood pressure monitors, pulse oximeters, and many more, allowing it to alert clinicians about any emergencies or changes in a patient's condition (Chien *et al.*, 2022). There are several cases where a CRS implemented in the ICU for measuring blood glucose levels has resulted in a reduction in the frequency of hypoglycemic events (Eslami *et al.*, 2012). In general, CRS that aim to improve patient safety by implementing CPOE and other related systems have been quite effective in minimizing prescribing and dosage mistakes (Moghadam *et al.*, 2021).
2. *Healthcare management:* Research has indicated that the use of CRS can lead to an improvement in adherence to medical procedures and guidelines (Kwok *et al.*, 2009). This holds importance because conventional clinical protocols and treatment approaches have demonstrated poor implementation in

real-world scenarios due to limited compliance from healthcare professionals (Cabana *et al.*, 1999). Nonetheless, the regulations that are implicitly embedded in guidelines can be precisely encoded into recommendation systems. CRS can manifest in various ways, such as predefined order sets for a specific medical scenario, notifications for a particular protocol relevant to the patients, prompts for testing, and so on. In addition, CRS can aid in the management of patients who are following therapeutic guidelines (Lip-ton *et al.*, 2011). It can also keep track of orders and referrals, follow up on them, and ensure that preventative care measures are taken (Salem *et al.*, 2018). CRS can notify healthcare providers to contact patients who have not adhered to their treatment plans or require further monitoring and facilitate the identification of patients who meet particular criteria for research studies (Jimmy and Jose, 2011). The Cleveland clinic (USA) has developed and applied CRS that generates notification to clinicians during patient care if the medical history of any case meets the clinical trial standards (Embi *et al.*, 2005).

3. *Expense Management:* The use of CRS can result in cost savings for health-care systems by enabling clinical interventions (Calloway *et al.*, 2013) that can decrease the length of hospital stays for patients (Pichardo-Lowden *et al.*, 2022), propose cheaper medication options via CPOE-integrated systems (Schaut *et al.*, 2022), and minimize unnecessary duplication of medical tests (Hak *et al.*, 2022). A regulation was implemented in an intensive care unit (ICU) for treating pediatric cardiac conditions that restricted the scheduling of the blood count and other tests to once every 24 hours using a CPOE system (Algaze *et al.*, 2016). Implementing this policy led to a decrease in the use of laboratory resources, resulting in an estimated annual cost reduction of \$717,538, without any increase in length of stay or mortality rates. The use of CRS can provide users with information regarding lower-cost medication options and medical conditions that are eligible for coverage by insurance providers. It is common in German hospitals for inpatients to receive medications that are listed on the hospital's approved list of prescription drugs, known as the drug formulary. Nevertheless, a study discovered that 20% of the medication substitutions made from the hospital's drug formulary were inaccurate. In response, Heidelberg Hospital developed a drug-switch algorithm and incorporated it into their CPOE system to improve the accuracy of medication management (Pruszydlo *et al.*, 2012). The

utilization of the CRS facilitated the automated switching of 91.6% of 202 medication prescriptions without encountering any errors, which resulted in enhanced safety, diminished workload, and decreased expenses for healthcare providers.

4. *Organizational Management:* In addition to their clinical applications, recommendation systems can also be utilized in various administrative functions like supply chain management (Singh and Parida, 2022), staffing and scheduling (Güler and Geçici, 2020), facility management (Abdellatif *et al.*, 2021), and financial management (Jia *et al.*, 2022) within a hospital. CRS assists in various clinical tasks such as coding prognosis, ordering tests and procedures, and prioritizing patients. Physicians can be assisted in selecting the most appropriate diagnostic codes by the customized computational procedure that suggests a more accurate and refined list of codes. The creation of a CRS was intended to tackle the issue of inaccurate International Statistical Classification of Diseases (ICD) coding in emergency admissions of patients patient (Higgins *et al.*, 2020). The quality of clinical documentation can be enhanced directly through the use of CRS. A CRS for obstetrics included an improved system for prompting, which led to a notable increase in the accuracy of documenting reasons for inducing labor and estimating the weight of the fetus (Haberman *et al.*, 2009). Having precise documentation is crucial, as it can directly assist in the implementation of clinical procedures. An instance of the implementation of a CRS was seen in the context of ensuring proper vaccination of patients who have undergone splenectomy, which is essential in mitigating the heightened risk of infections such as pneumococcal and meningococcal associated with spleen removal. The authors discovered that 71% of cases with “splenectomy” mentioned in their EHR did not have it recorded in their problem list, which is the crucial criterion for activating the CRS alert (McEvoy *et al.*, 2018). To sum up, recommendation systems and clinical decision support systems such as CRS can be utilized in a range of clinical and administrative tasks within a hospital, resulting in improved precision and effectiveness in tasks such as clinical documentation, coding, and the implementation of clinical procedures.
5. *Prognosis aid:* CRS helps healthcare professionals make precise diagnoses by offering tailored recommendations that rely on individual patient data. The recommendation systems that utilize specialized knowledge can assist healthcare professionals in diagnosing intricate cases. Clinicians can receive

guidance and recommendations based on a wealth of expertise and knowledge, allowing them to diagnose better and treat patients (Dramburg *et al.*, 2020). CRS can be beneficial in regions lacking experienced healthcare professionals and improve the quality of care by optimizing the available healthcare resources. Various recommendation systems have demonstrated diagnostic capabilities that are comparable to those of human experts (Stivaros *et al.*, 2010). It can find and extract relevant and precise data that would be used for the prognosis of specific diseases (Saxena *et al.*, 2021). Primary care requires specialized CRS and IT solutions due to the frequent occurrence of prognosis errors (Singh *et al.*, 2016). Kunhimangalam *et al.* (2014) developed an effective CRS for diagnosing peripheral neuropathy with fuzzy logic and achieved a 93% accuracy rate using 24 input fields. Although this system is valuable in areas with limited access to clinical experts, there is still a need for diagnostic tools that can support specialist diagnostics.

These CRS systems can be broadly categorized into two types based on the algorithms used to generate recommendations.

- *Rule-based CRS*: These systems rely on a set of established rules that are derived from clinical guidelines or expert opinions. Clinical experts define and develop the set of protocols, which are programmed into the system. The system examines patient data and utilizes the applicable rules to produce a diagnostic result. While rule-based CRS can offer precise and consistent outcomes, the extent of their diagnostic abilities is restricted by the number of rules incorporated into the system. Figure 1.1a illustrates that rule-based systems consist of predetermined rules programmed as knowledge bases. An interface engine applies these algorithms to patient data to generate a predictive output. Rule-based imaging CRS are commonly employed for image ordering in radiology. Rule-based CRS can be utilized with imaging data for prognostic aid, assisting radiologists in making diagnoses or providing recommendations based on imaging data. These CRS can issue reminders of the best practice guidelines or alerts to potential risks or limitations. Research carried out at Virginia Mason Medical Center showed that implementing a CRS for image ordering resulted in a substantial decrease in the utilization rate of sinus Computed Tomography (CT) for sinusitis, lumbar Magnetic Resonance Imaging (MRI) for lumbar discomfort, and head MRI for headache (Blackmore *et al.*, 2011). An example of a commercially available system is RadWise, which helps clinicians choose the most appropriate

imaging test by examining clinical signs of the patient and comparing them to a vast database of possible diagnoses<sup>3</sup>.

- *Artificial Intelligence (AI)-based CRS*: These systems employ advanced algorithms and machine learning techniques to scrutinize patient data and produce diagnostic results. AI-based CRS have the ability to examine enormous amounts of data and detect patterns that may not be readily discernible to humans. In addition, they can learn from new data and enhance their diagnostic accuracy as time progresses. Nonetheless, AI-based CRS can be intricate and necessitate substantial computational power. AI-based systems, depicted in Figure 1.1b, comprise statistical or machine learning algorithms that are trained on a substantial amount of expert-annotated data. An AI-powered interface engine utilizes these advanced algorithms to scrutinize patient data and produce a prognostic output. The field of radiology is experiencing a surge of interest in AI-based CRS that aims to improve imaging and precision radiology, commonly referred to as *radiomics* (McCague *et al.*, 2023). Manual interpretation can become burdensome as medical images continue to represent a more significant portion of healthcare data. Therefore, healthcare providers require technologies that can assist them in processing, displaying, and analyzing these images (Kelly *et al.*, 2022). AI-based CRS have demonstrated their ability to provide insights into data that surpass the capabilities of humans (Greenspan *et al.*, 2016). These technologies employ sophisticated Deep Learning (DL) algorithms to identify abnormalities in the images (Wang *et al.*, 2023). Various companies like IBM Watson Health, Microsoft, NVIDIA, and Google are leading the way in developing innovative products for detecting tumours (Williams *et al.*, 2021), diagnosing diabetic retinopathy (Gulshan *et al.*, 2016), Alzheimer’s diagnosis (Suzuki and Chen, 2018) and many more.

During the training phase, an AI-based CRS makes use of sophisticated statistical methods, such as Machine Learning or Deep Learning, to identify patterns within the clinical data housed in the Electronic Health Record (EHR). It then utilizes this knowledge to generate diagnostic output through the interface in the testing phase without depending on pre-established rules, as is the case with rule-based CRS. Healthcare data is expanding at a fast pace, and this includes information such as the clinical traits of patients, clinical notes that are associated

---

<sup>3</sup>DSS Inc. Radiology Decision Support (RadWise). Online: <https://www.dssinc.com/products/integrated-clinical-products/radwise-radiology-decision-support/>.



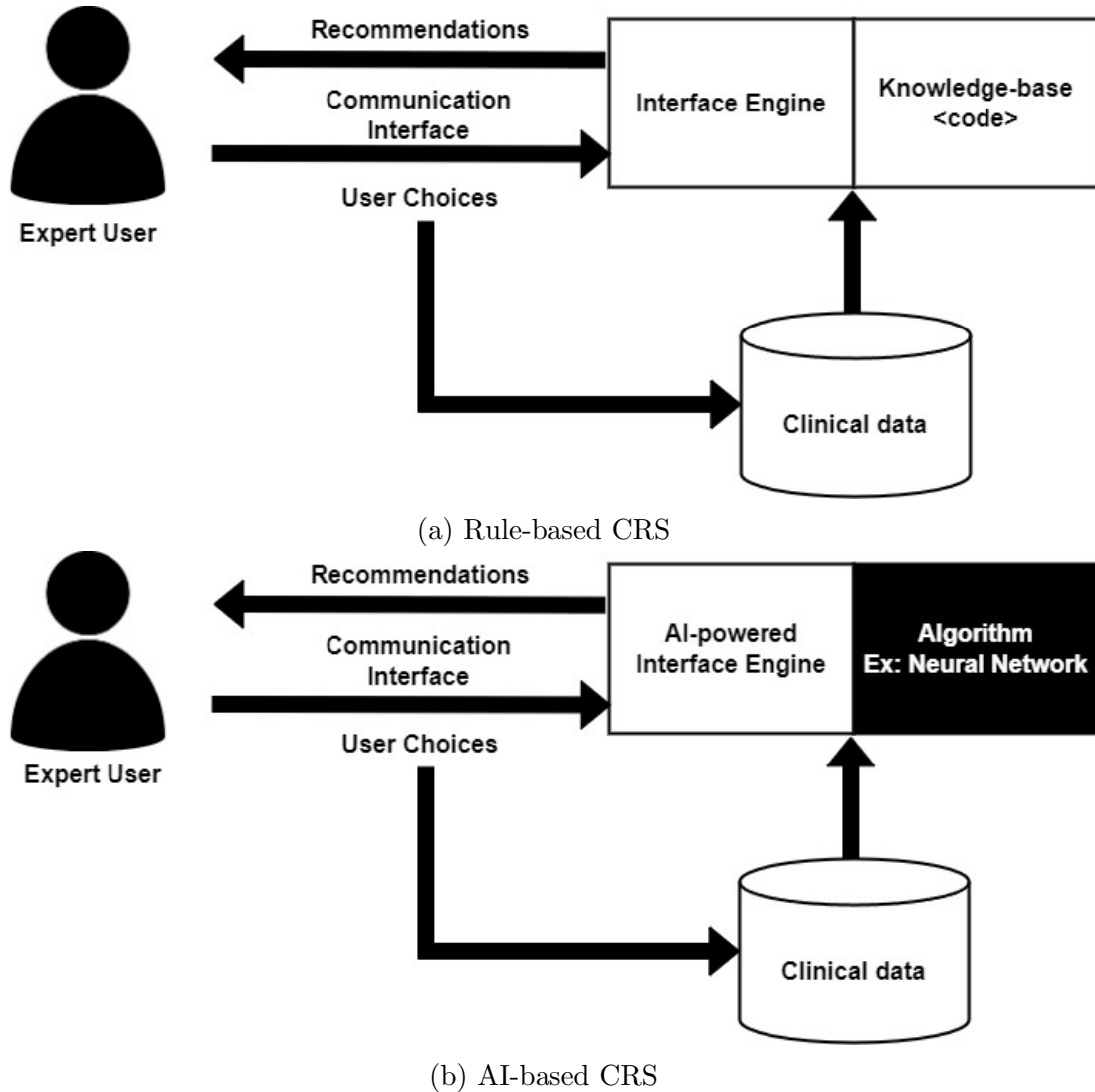


Figure 1.1: Categorization of Clinical Recommendation Systems

with diagnostic images, administrative and medical claim data, as well as various regulatory requirements. The adoption of EHR systems in the United States was encouraged by the enactment of the Health Information Technology for Economic and Clinical Health (HITECH) Act in 2009. This legislation provided incentives totaling \$30 billion, as reported by (Rouse, 2018). According to the source (ONC, 2022), there has been a significant increase in the adoption of EHRs by office-based clinicians. Specifically, the adoption rate has risen from 21% in 2004 to 87% in 2022. Furthermore, the proportion of clinicians who have adopted an essential EHR has tripled from 11% in 2006 to 54% in 2022.

EHR comprises a plethora of structured data such as (1) *numerical quantities*: patient demographics, clinical laboratory results such as height, weight, and blood



type; 2) *categorical values*: current Procedural Terminology (CPT) procedures or ICD codes; (3) *date/time objects*: temporal events of birth or admission; as well as unstructured data such as (4) *natural language free-text*, e.g., medical reports containing patient profiles, current health status, patient disease history, and discharge summaries; (5) *medical images* such as X-ray, CT, MRI, etc. (Gehrmann *et al.*, 2018). Structured EHR data does not require complex processing prior to performing statistical or machine learning tasks. However, it should be noted that most of the data present in EHRs today is unstructured and may require more complex processing before it can be used for these tasks (Joseph *et al.*, 2021). Researchers have endeavored to develop data-oriented models due to the vast amount of valuable information contained in EHR (Alqahtani *et al.*, 2022). The extensive collection of clinical data in diverse formats presents multiple challenges, including missing data and increased uncertainty. Utilizing EHRs containing unstructured data creates an opportunity to develop advanced techniques, such as predictive analysis frameworks or CRS, which can provide clinicians with valuable and improved diagnostic information. Predictive analysis is “*an advanced technique that uses powerful algorithms to identify patterns in historical data, which are then analyzed to make accurate predictions about future events or outcomes*” (Sundararaman *et al.* (2018); Ramesh and Santhi (2020)). Predictive analysis has played a crucial role in improving several healthcare trends by aiding clinicians and patients to enhance their medical activities, such as diagnosis (Sinaga and Putra, 2022). EHRs are utilized to extract disease diagnoses (Comito *et al.*, 2022) and medication information (Chen *et al.*, 2020) with increased precision and reduced costs. This thesis extensively studies the design and development of an effective AI-based clinical recommendation system that focuses on prognosis aid tasks using multimodal unstructured medical data.

## 1.1 Unstructured Medical Text Analysis

Unstructured medical text analysis involves the examination of unstructured medical text data, which can include various types of free-text medical information such as clinical notes, radiology reports, pathology reports, and discharge summaries (Spasic and Nenadic, 2020). The sample unstructured medical text data is shown in Figure 1.2. The figure shows that the unstructured clinical notes are unlike any regular text. They contain extended sentences with medical terms, punctuation, abbreviations, acronyms, misspellings, and incomplete sentences. Therefore, it is crucial to preprocess the notes and use sophisticated word embedding techniques

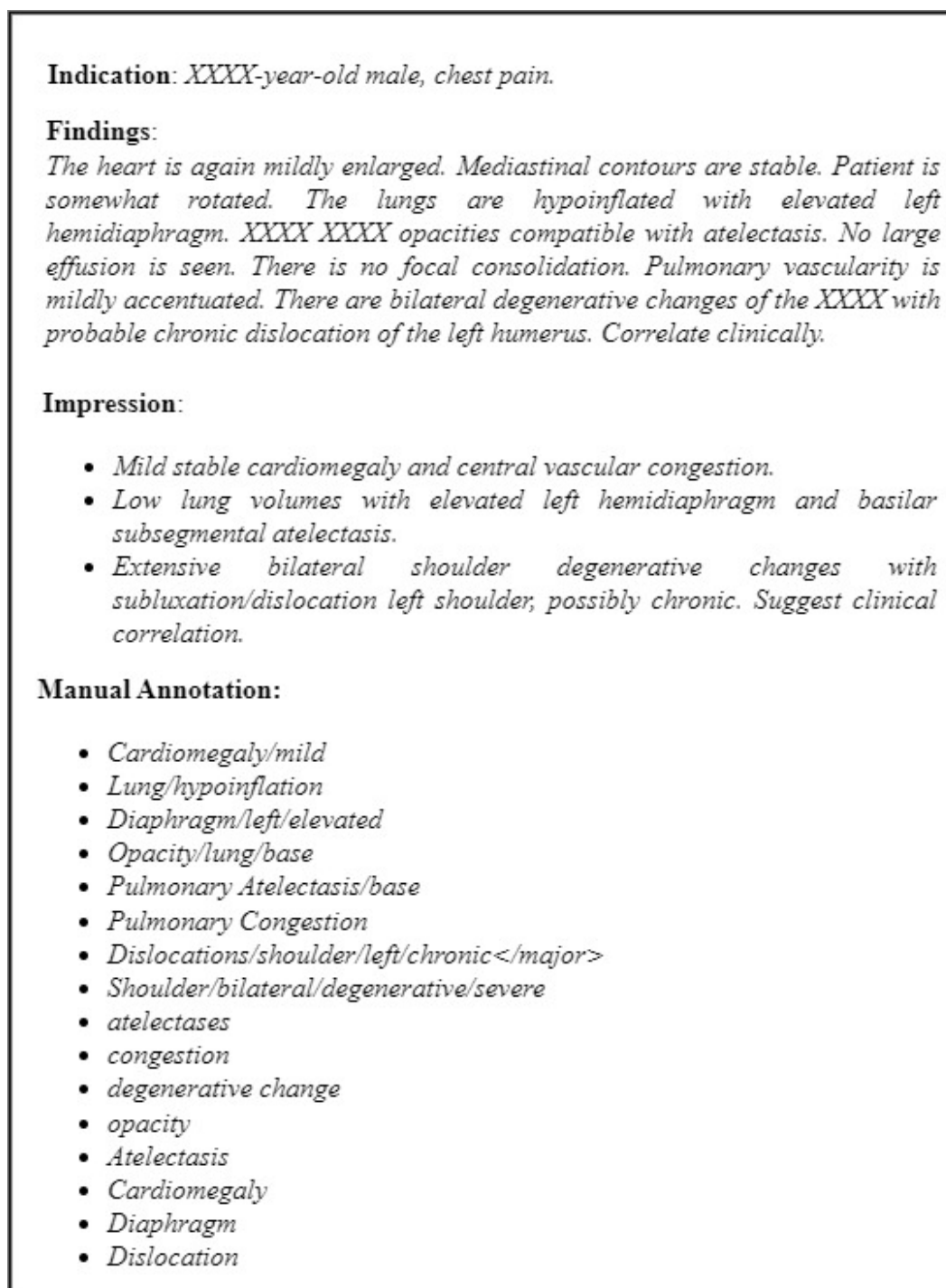


Figure 1.2: Sample radiology report ([Demner-Fushman et al., 2016](#))

to represent medical vocabulary accurately before applying statistical methods. Unstructured medical text analysis entails the extraction of significant information from unstructured text data and transforming it into organized data that can be utilized for decision-making and analysis. It commonly employs natural

language processing (NLP) methods, which can include techniques like named entity recognition, relationship extraction, and sentiment analysis (Kreimeyer *et al.*, 2017). Through repeated demonstrations, it has been shown that it is possible to extract hidden evidence from clinical narratives, which can then be utilized for extensive analysis at a later stage (Spasić *et al.*, 2020). The objective is to automatically recognize and extract crucial details such as patient demographics, medical diagnoses, treatments, and results (Mahbub *et al.*, 2022). This information can then be utilized to enhance clinical decision-making, quality of care, and research activities. The use cases of Unstructured text analysis consist of detecting adverse events (Henriksson *et al.*, 2015), pharmacovigilance (Lependu *et al.*, 2013), recruiting participants for clinical trials (Meystre *et al.*, 2019), and monitoring the occurrence and spread of diseases (Chen *et al.*, 2018).

## 1.2 Unstructured Medical Image Analysis

The analysis of medical images is of great importance in the identification and treatment of different medical ailments (Parmar *et al.*, 2018). The analysis of unstructured medical images entails the examination of images that lack a clearly defined structure, such as those obtained through X-rays, MRIs, or CT scans (Willemink *et al.*, 2020). Usually, unstructured medical image analysis involves utilizing a blend of techniques like image processing, computer vision, and machine learning to obtain relevant information from the images (Sarker, 2021). The process may involve tasks such as detecting significant areas, separating the image into distinct segments, and obtaining characteristics from the image that can be employed for additional investigation. Unstructured medical image analysis presents several significant obstacles, including managing image noise and artifacts (Sagheer and George, 2020), addressing the diversity of image acquisition and patient positioning (Dean and Scoggins, 2012), and managing the vast amounts of data generated by current medical imaging technology (Diaz *et al.*, 2021). Despite the difficulties, unstructured medical image analysis has the capacity to transform medical diagnosis and treatment by presenting precise and individualized perspectives on the health of the patient. Unstructured medical image analysis is an indispensable instrument for current medicine as it enables the identification of tumors, pulmonary diseases, and other anomalies as well as the monitoring of disease advancement over a period (Pandya *et al.* (2019); Bharati *et al.* (2020); Saeedi *et al.* (2023)). At present, disease diagnosis involves the manual examination and analysis of imaging data by skilled physicians and licensed professionals.

The diagnostic outcome obtained from manual image analysis by radiologists may not be consistent if the same images are re-examined after a certain period of time. This is a significant limitation of manual image analysis. Several factors could contribute to this inconsistency in the diagnostic outcome, including the following:

- *Inter-observer variability:* Radiologists may have varying levels of experience, knowledge, and expertise, which can result in differences in their interpretation of the same image (Obuchowicz *et al.*, 2020).
- *Intra-observer variability:* Inconsistencies in diagnostic outcomes can arise due to the fact that a radiologist may interpret an image differently at different points in time, even if it is the same radiologist analyzing the image (Hopper *et al.*, 1996).
- *Subjectivity:* The subjective nature of manual image analysis means that it can be impacted by a radiologist’s personal biases, level of experience, and expertise (Brady, 2016).
- *Fatigue and workload:* Radiologists may make errors or overlook important details in their analysis due to fatigue or a heavy workload, which can result in inconsistencies in the diagnostic outcomes (Hanna *et al.*, 2018).
- *Time elapsed:* Differences in diagnostic outcomes may occur when the same images are re-examined over time as the radiologist’s memory of the image may fade, or other factors may influence their interpretation of the image (Brady *et al.*, 2012).
- *Environmental factors:* A radiologist’s interpretation of an image can be influenced by external factors, such as lighting or distractions in the reading room (Woolen *et al.*, 2023).
- *Imaging artifacts:* Inconsistencies in diagnostic outcomes can occur due to the impact of image quality or artifacts, which can hinder a radiologist’s ability to accurately interpret the image (Bekiesińska-Figatowska, 2015).

Henceforth, the process of manual interpretation is often a difficult and time-consuming task, as diagnostic images for different diseases contain various patterns that can be challenging to identify. Fig. 1.3 displays some examples of X-ray images used for predicting pulmonary diseases.

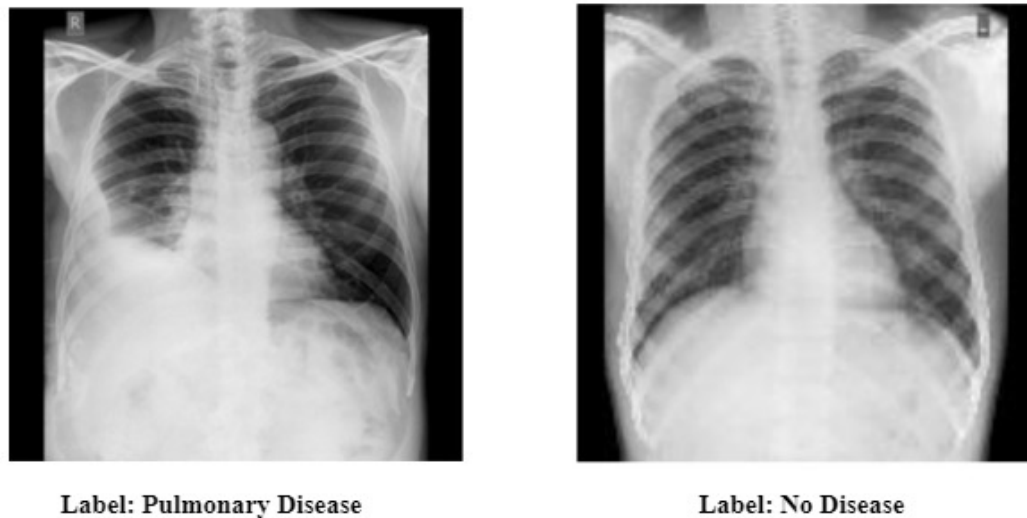


Figure 1.3: Sample chest X-ray with pulmonary abnormality and no disease classes (Demner-Fushman *et al.*, 2016)

### 1.3 Multimodal Medical Data Analysis

Data related to the same topics or objects can be obtained through multiple methods, with varying conditions or experiments, across different fields of study. The term “modality” pertains to these specific approaches of acquiring data (Lahat *et al.* (2015); Acosta *et al.* (2022)). Analyzing multiple modes of data together can lead to a more complete understanding of a specific task or topic and may offer novel insights that cannot be obtained by analyzing just one mode of data. While AI has demonstrated success in various areas such as speech recognition, natural image detection, and language translation, its application in healthcare has been limited by the intricate nature of the unique features or signals present within multimodal medical data (Acosta *et al.*, 2022). The use of wearable sensors has become more prevalent, and advancements in technology have made it easier to collect and combine data from different sources, resulting in an abundance of multimodal data. These data can be valuable in identifying, predicting, and preventing various diseases (Yang *et al.*, 2021). The majority of current AI research is centered on discovering, classifying, and predicting diseases based on data obtained from a single modality. However, clinicians utilize a diverse range of data from various sources to assess and plan treatment for patients (Nunes *et al.*, 2019). On the other hand, AI models that incorporate multimodal data available to clinicians for prognostic evaluation have displayed favorable outcomes in identifying and predicting diseases when compared to models that use only one modality of data Soenksen *et al.* (2022). The term “*Multimodal Medical Data*

*Analysis*” pertains to the analysis of medical data by utilizing different modes of information like images, text, and signals. This method can offer a more complete comprehension of a patient’s health status and enable precise diagnosis and treatment.

The utilization of multimodal data in intelligent healthcare systems was initially explored in the 1990s, and only a handful of studies were acknowledged during this early phase. Gradually, multimodal data became a crucial aspect of research in the healthcare system. To provide a brief overview of the past, [van der Putten \*et al.\* \(1995\)](#) created a transparent framework that enabled physicians to access multimodal data from various sources, including echocardiography, Cathlab databases, hospital information systems, and an Electrocardiogram (ECG) management system. The workstation was constructed using the C programming language on a UNIX platform and utilized an Interbase database and a CD-ROM for storage. During the initial phase, the idea of multimodal data analysis was a novel concept, and numerous challenges were faced while integrating and enhancing the use of multimodal data. The storage of multimodal data was cumbersome and costly, leading to significant storage difficulties. In order to overcome this issue, [Wood \*et al.\* \(1998\)](#) suggested a multimodal information system that could extract and generate information from various repositories based on specific requirements. Their objective was to simplify the coordination of data between different information sources from a wide range of domains. For annotating selected data, an object analyzer from Intext Inc. was employed, although the results of this annotation were unsatisfactory. Over time, the utilization of multimodal data in the healthcare system has steadily increased.

Subsequently, efforts were made to move beyond the mere storage of multimodal data and focus on its annotation. Medical experts typically rely on the comparison and correlation of data to achieve more precise clinical diagnosis and prediction. In their publication, [An \*et al.\* \(2008a\)](#) presented their research on visualizing multimodal data in EHRs. This represented a further development in the classification of electronic data into numeric texts and images. Additionally, the classified data were annotated in the study. In 2010, the concept of data fusion for retrieving multimodal data from electroencephalogram (EEG), MRI, and positron emission tomography (PET) was introduced by [Polikar \*et al.\* \(2010\)](#). This study introduced a new perspective on processing multimodal data in the healthcare system by proposing a diverse ensemble classifier solution that achieved 10% to 20% higher accuracy than previous methods. In 2013, several researchers brought a fresh perspective to the use of multimodal data in healthcare applications. For



instance, [Weibel \*et al.\* \(2013\)](#) presented an application that was designed for the analysis of multimodal EHR data. This work reduced the challenges of manual coding, and additional features such as audio tracks and gaze were integrated for various applications. By 2016, several researchers had started proposing their ideas for open-source software for medical imaging to address the curse of dimensionality. A shift towards the use of convolutional neural network (CNN) classifiers was also observed in 2016 ([Pinho and Costa, 2016](#)). As storage and processing capacities have developed and increased, healthcare data has also grown exponentially. From 2016 onwards, considerable research has been dedicated to analysing big data in the healthcare field, including data obtained from various sources ([Rehman \*et al.\* \(2021\)](#); [Amal \*et al.\* \(2022\)](#); [Kline \*et al.\* \(2022\)](#)). Over the past few decades, there has been a significant transformation in the use and growth of multimodal data in the healthcare industry. Initially, the focus was on storing such data, but with advancements in technology and machine learning, there has been a shift towards analysing this data.

Multimodal deep learning methods have revolutionized the way we utilize data from various sources. By combining data from multiple sources, such as images, videos, and LiDAR, these models can produce more accurate and valuable information than traditional single-modality approaches. Multimodal deep learning techniques have been successfully applied in various fields, including autonomous vehicles ([Person \*et al.\*, 2019](#)), social media video classification ([Trzcinski, 2018](#)), and emotion classification ([Pandeya and Lee, 2021](#)). For example, a fusion-based multimodal deep learning framework was proposed for safe navigation of autonomous vehicles, achieving 3.7% better performance compared to a uni-modal CNN classification architecture ([Person \*et al.\*, 2019](#)). Similarly, a multimodal model for social media video classification outperformed Google’s InceptionV3 model by approximately 12% in accuracy. In healthcare, multimodal deep learning models are being used to combine complementary contextual data to obtain more precise diagnostic results, overcoming the limitations of unimodal image-only approaches. This thesis showcases three different multimodal tasks that incorporate the analysis of both images and text: multimodal image-text analysis, cross-modal image-text analysis, and multimodal medical image analysis.

### 1.3.1 Multimodal Medical Image-text Data Analysis

Multimodal medical data analysis in radiology involves the use of both imaging techniques such as X-rays, CT scans, MRI, and ultrasound, as well as the accom-

panying textual information found in the report and patient history, to gain a comprehensive understanding of the patient’s condition (Zhang *et al.*, 2022). The multimodal clinical data is represented in Fig. 1.4 through a sample chest X-ray (CXR) along with its corresponding reports. Radiologists rely on various imaging techniques such as X-rays, CT scans, MRIs, and ultrasound to create images of internal organs and structures (Chanumolu *et al.*, 2022). These images are further analyzed and interpreted by the radiologists, and a report is generated that summarizes the findings and recommendations. Text and report case studies are a crucial aspect of multimodal medical data analysis in radiology. The report includes significant textual data like the patient’s medical history, clinical findings, and the radiologist’s analysis of the images. Medical professionals can obtain a more comprehensive understanding of the patient’s health and make well-informed decisions about their treatment by examining both the images and the associated text.

In summary, multimodal medical image-text analysis has the potential to revolutionize healthcare by jointly analyzing both image and text data. This approach can improve patient care by providing more precise and personalized treatments. However, the challenge lies in effectively combining data from different sources to maximize their unique features and generate more accurate diagnostic predictions. Since visual and textual features are distinct, there is a need to fuse them into a rich representation that can provide detailed information for better predictions.

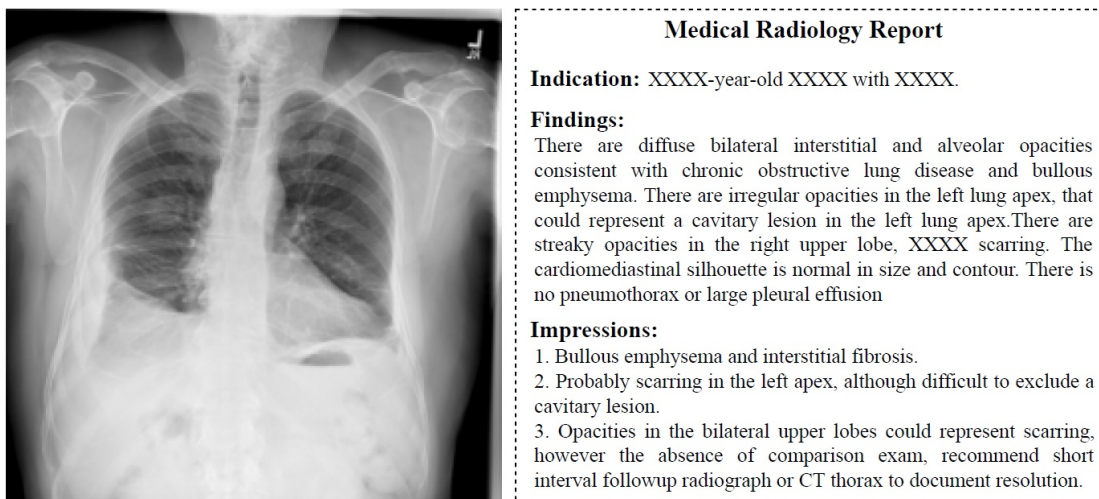


Figure 1.4: Sample CXR with associated clinical note (Demner-Fushman *et al.*, 2016)



### 1.3.2 Cross-Modal Medical Image-Text Analysis

The terms cross-modal and multimodal are often utilized interchangeably, but they have slightly different meanings in the context of data analysis. The term “*multi-modal*” pertains to the utilization of various modes or types of data for examining a phenomenon, such as a disease prediction or classification (Nasir *et al.*, 2023). On the other hand, “*cross-modal data analysis*” involves mapping data from different modalities onto a shared representation space, where they can be integrated and compared to gain a more comprehensive comprehension of the phenomenon. Examples of cross-modal applications include medical image captioning Singh and Parida (2022) and radiology report generation (Chen *et al.*, 2021). Cross-modal report generation from radiology images is the process of automatically generating a textual report from radiological images such as X-rays, CT scans, or MRIs (Gundogdu *et al.*, 2021). Usually, deep learning methods are employed to examine the images and produce a report that relies on the characteristics identified in the images. The typical procedure comprises multiple stages, such as preparing the image, extracting characteristics, and generating a report. Generating reports across different modalities holds the potential to enhance the precision and efficacy of radiology reporting by automating the task and easing the burden on radiologists (Alfarghaly *et al.*, 2021a). Nevertheless, there are still obstacles to surmount, such as the requirement for significant volumes of labelled data to train the deep learning models and the necessity to guarantee the precision and dependability of the produced reports (Ramirez-Alonso *et al.*, 2022).

### 1.3.3 Multimodal Medical Image Analysis

The area of multimodal medical image analysis concentrates on creating approaches and methodologies for examining and understanding medical images that are derived from diverse imaging modalities (Li *et al.*, 2021). Various technologies for imaging, such as X-ray, CT scan, MRI, ultrasound, and other modalities, can be used in medical settings (Huang *et al.*, 2020). The primary objective of multimodal medical image analysis is to integrate data from different imaging modalities to enhance the dependability and precision of medical diagnoses (Tan *et al.*, 2020). Researchers have investigated numerous potential uses of multimodal images to forecast different illnesses. In their study, Kabir *et al.* (2007) employed a Markov Random Field model to segment stroke lesions from a series of MRI images and developed an atlas of blood supply territories to differentiate between different stroke subtypes. Polikar *et al.* (2010) conducted a study com-

binning different biomarkers such as EEG, structured MRI, and PET to investigate how multiple modalities perform compared to using only one. They used ensemble classifiers that combined the results using sum and simple majority voting (SMV) fusion techniques. The results showed that the combined modalities increased classification performance by 10-20% when using the combined modalities. Nie *et al.* (2016) introduced a deep learning framework in 3D that extracts advanced features of brain tumours from different types of MRI images. The fig. 1.5 shows multiple images, including Diffusion Weighted Imaging (DWI), T2-weighted Fluid-Attenuated Inversion Recovery (T2-Flair), Apparent Diffusion Coefficient (ADC), and Susceptibility Weighted Imaging (SWI) MRI sequences of 10 patients data collected from the private medical institute. Medical professionals who specialize in radiology review all of these images in order to make a prediction about the presence of an acute infarct in a patient. Multimodal medical image analysis involves several challenges due to the complex and heterogeneous nature of the data. Some of the critical challenges include dealing with missing or inconsistent data across modalities, addressing variability in imaging protocols and quality across different datasets, managing a large amount of data, and devising efficient techniques to combine and merge information from multiple imaging modalities (Acosta *et al.*, 2022).

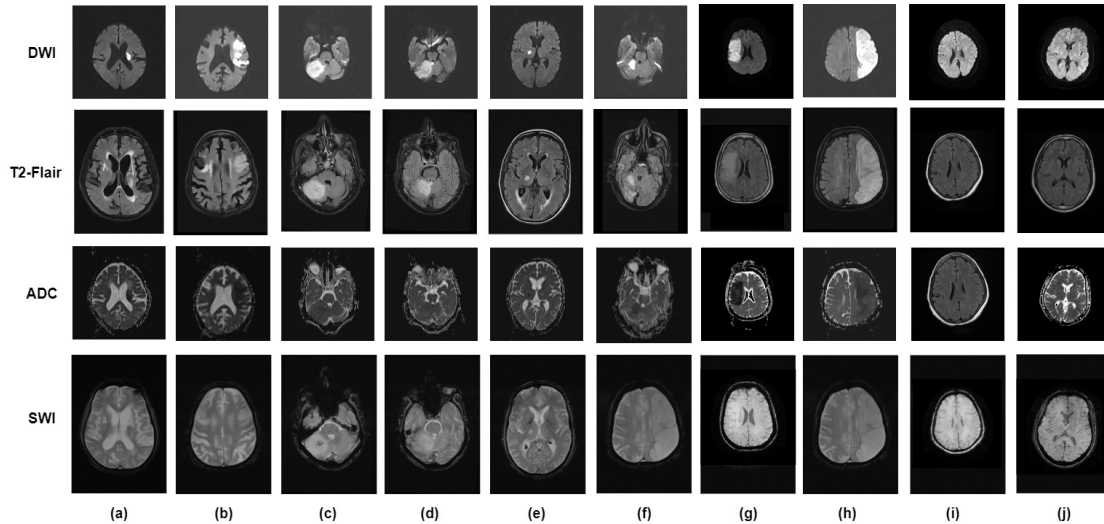


Figure 1.5: Top to bottom: The first row denotes the DWI MRI sequences, Second row indicates T2-Flair MRI Sequences, Third row represents ADC MRI Sequence and fourth row depicts the SWI MRI sequence of 10 patients data collected from private medical institute. Left to right: (a) to (h) represents the MRI sequences with Acute Infarct and (i) to (j) indicates the MRI sequences with no acute infarct.

Figure 1.6 provides an overview of the different components and concepts re-

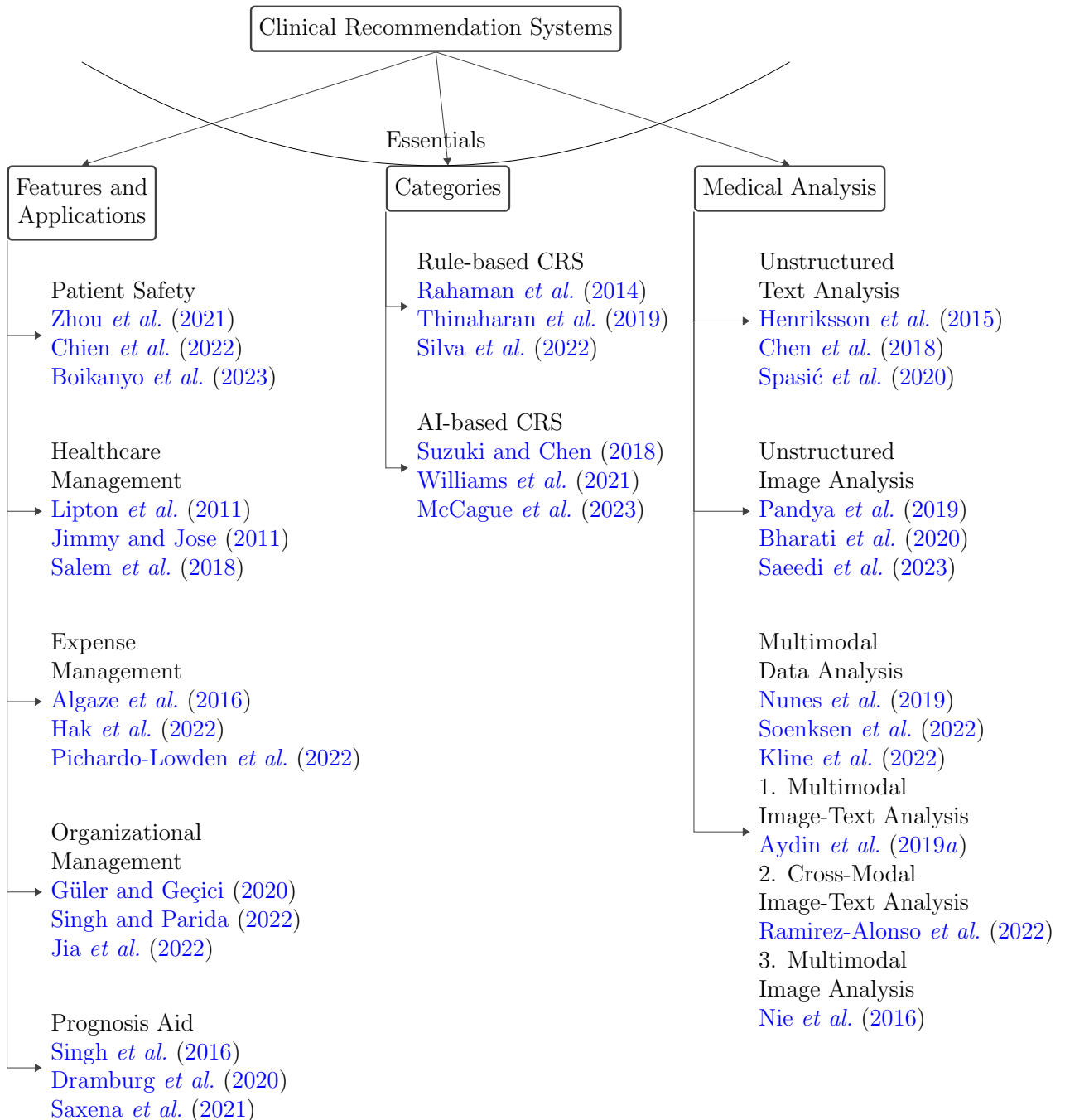


Figure 1.6: Essentials of Clinical Recommendation Systems

lated to clinical Recommendation Systems that are important for our research. Our thesis aims to develop an AI-based CRS framework that will assist in the prognosis of medical conditions. We will achieve this by analyzing unstructured data, such as medical text, images, and multimodal data, to provide diagnostic outcomes for patient-centered applications. Our research focuses on creating an effective framework that utilizes state-of-the-art technologies to improve medical decision-making and patient outcomes.

## 1.4 Prominent Obstacles and Concerns

The use of EHR-based CRS has the potential to revolutionize healthcare by improving patient outcomes and optimizing the use of resources. However, incorporating multimodal healthcare data into CRS presents notable challenges. The effectiveness of these systems hinges on how well they are able to synchronize with the current practices of healthcare providers. Therefore, it is essential to meticulously contemplate these difficulties during the development of CRSs, to guarantee their efficient implementation in clinical environments and to ensure that they offer maximum benefits to patients and healthcare professionals. The following are several obstacles that must be overcome when building an effective AI-based CRS that incorporates multimodal healthcare data.

1. *Robust data integration:* A significant obstacle is the amalgamation of information from different sources and formats, which may include electronic medical records, clinical notes, and medical images. It is essential to devise efficient strategies to combine this data into a consistent and interpretable format for the CRS to be successful.
2. *Data quality and standardization:* Since healthcare data is generated from various sources, it is prone to inconsistencies and errors. This may result in complications while integrating the data and can create challenges for the AI system to correctly understand the information. To tackle this obstacle, it is essential to adopt data quality assurance measures, including data cleansing and validation, to guarantee the precision and uniformity of the information. Moreover, unifying the data formats, coding methods, and vocabulary across various sources can simplify the process of integration and improve the efficiency of the AI system.
3. *Clinician acceptance and integration:* A key factor in the successful implementation and utilization of an AI-powered CRS in clinical settings is the

acceptance and integration of the system by clinicians. However, healthcare providers may have concerns about the accuracy, reliability, and potential impact of the system on their current workflows and decision-making processes. Including the insights and evaluations of healthcare providers in the development and testing stages can assist in resolving their apprehensions and improving their willingness to adopt the system.

4. *Dealing with unstructured medical data:* Dealing with unstructured medical data presents several challenges for healthcare organizations and AI systems. Extracting important information from unstructured medical data is one of the significant challenges faced in healthcare. This is mainly due to the fact that the data is in the form of text and contains complex medical terminologies that require advanced natural language processing techniques for proper analysis and interpretation. Moreover, unstructured medical data can be prone to errors, inconsistencies, and ambiguities that can negatively affect the performance and accuracy of the AI system.
5. *Dealing with diagnostic images:* Dealing with diagnostic images can be challenging due to the large size of the data files, which can make storage and processing of the images computationally intensive. An additional hurdle is the potential for different medical professionals to interpret the same image differently, which can result in inconsistencies and errors in the AI model's training and performance. The presence of inconsistencies in image quality can present a significant challenge in identifying any abnormalities or lesions, especially in early detection systems.
6. *Generalizability:* In the context of AI-based CRS, generalizability refers to the ability of the system to provide accurate and effective recommendations or decisions for patients with different medical conditions and in different healthcare settings. To ensure that an AI-based CRS can be widely adopted and used in clinical practice, it is essential to establish its generalizability. However, there are a number of difficulties related to achieving generalizability, including the wide variety and intricacy of medical conditions, the differences in treatment options and outcomes, and the variation in healthcare systems and policies across different regions and countries. To address these challenges, it is important for AI-based CRS to be trained on extensive and varied datasets that encompass a diverse set of medical conditions and patient populations. Moreover, the system should be tested and validated

in different clinical environments, and its effectiveness and safety should be compared against established clinical practice guidelines.

7. *Tolerance for predictive mistakes:* Tolerance for predictive mistakes refers to the acceptable margin of error in an AI-based CRS. While the ideal goal is to achieve zero predictive mistakes, this may not always be feasible, particularly in the early stages of implementation. In healthcare, where even a small mistake could have serious consequences for patients, it is critical to minimize the potential for predictive mistakes as much as possible. Therefore, the tolerance for predictive mistakes in an AI-based system should be very low, ideally approaching zero.
8. *Class Imbalance:* Class imbalance occurs when the distribution of classes or categories in a dataset is unequal, which can have a negative impact on the accuracy of the AI model's predictions or recommendations. In the field of healthcare, the occurrence of class imbalance is possible when there are limited cases of certain medical conditions or diseases. For example, in datasets of cancer patients, the count of patients with a less common type of cancer may be significantly lower compared to those with a more prevalent type of cancer. This can result in a situation of class imbalance, where the AI model might exhibit a preference for the larger class and could have reduced accuracy in predicting outcomes for the smaller class.
9. *Lack of transparency:* One challenge is the lack of transparency in the decision-making process of AI models. A major challenge in the interpretation of CRS is the use of black-box algorithms, which lack transparency in their decision-making process, making it challenging for healthcare providers to comprehend and rely on the system's recommendations. To overcome this challenge, efforts must be made to develop transparent AI systems that provide clear and concise explanations for their decision-making processes. One approach to addressing this issue is to utilize explainable AI techniques, such as Gradient-weighted Class Activation Mapping (Grad-CAM) ([Selvaraju et al., 2016](#)).
10. *Scarce data situation:* In the creation of AI-based CRS, a scarcity of data can be a significant challenge. This scenario may occur for different reasons, such as a scarcity of medical data that is exclusively available in private hospitals, or data being constrained to a particular medical specialty or geographical area. Consequently, it becomes challenging to obtain a comprehensive and

varied dataset to train the AI model. The limited availability of data can also have an impact on the precision and applicability of the AI model. Insufficient data can restrict the AI model's ability to learn and consequently lead to inaccurate predictions and subpar performance.

## 1.5 Summary

This chapter highlights the challenges present in current healthcare settings and the importance of implementing clinical recommendation systems to improve the quality of patient care. The chapter provides an overview of the different applications and categories of CRS. Additionally, it addresses the key issues related to medical data analysis, including analyzing unstructured text, images, and multi-modal medical data. Finally, the chapter identifies the prominent obstacles and concerns that must be considered when designing and developing an intelligent and effective AI-based CRS system to overcome these challenges. To assist healthcare providers in managing their workload and providing valuable information throughout the clinical process, it is essential to utilize methods that can effectively capture insights from various forms of medical data.

## 1.6 Thesis Organization

The rest of this thesis is organized as follows.

- In Chapter 2, a comprehensive review of the CRS in the medical field is provided along with an overview of the existing research gaps in the literature.
- In Chapter 3, formalizes the research problem and outlines the research objectives based on the literature presented in Chapter 2.
- Chapter 4 provides an in-depth analysis of proposed methods for analyzing unstructured medical text data.
- In Chapter 5, presents a detailed overview of the proposed framework for analyzing unstructured medical image data.
- Chapter 6 proposes various approaches for interpreting multimodal diagnostic images and their associated reports.
- Chapter 7 presents a technique for cross-modal diagnostic report generation through medical images.

- In Chapter 8, a framework is presented for the analysis of medical images that integrates multiple modalities.
- Chapter 9 offers a summary with conclusive remarks of the research work conducted and suggests potential avenues for future research in the field.



## Chapter 2

### Literature Review

EHR refers to the digital storage and management of patients' health information. This involves the electronic storage of various clinical data that pertains to a patient's medical history and treatment (Gold *et al.*, 2021). With the widespread adoption of EHRs in clinical settings, healthcare providers now have access to a vast amount of clinical data on the patients they serve. This information comprises medical records, prescribed drugs, laboratory test outcomes, medical imaging scans, and other relevant clinical details (Evans, 2016). The existence of a significant volume of clinical data has opened up opportunities for healthcare providers to develop CRSs that can assist them in making informed and evidence-based decisions (Dash *et al.*, 2019). Such systems can examine patient data to recognize possible health hazards, propose diagnostic and treatment alternatives, and keep track of patient progress (Tran *et al.*, 2020). It is crucial to introduce recommendation systems designed for medical purposes to address the gaps and provide support to patients and healthcare professionals in making more informed decisions concerning healthcare. To simplify the process of item selection for users, recommendation systems have been integrated into other applications like E-commerce platforms, digital content providers, and social network apps (Felfernig and Gula (2006); Tran *et al.* (2017)). In recent times, CRS has been designed to enhance medical recommendations and is extensively employed in the healthcare sector (Pincay *et al.* (2019); Sahoo *et al.* (2019)). The primary objective of this thesis is to concentrate on developing and designing a CRS capable of predicting diseases based on unstructured medical text data, unstructured medical image data, multimodal diagnostic image and text data, and multimodal medical image data. Additionally, we aim to create a CRS that can perform the cross-modal task of generating radiology reports from diagnostic images. Figure 2.1 showcases the classification of the CRS with respect to data utilization.

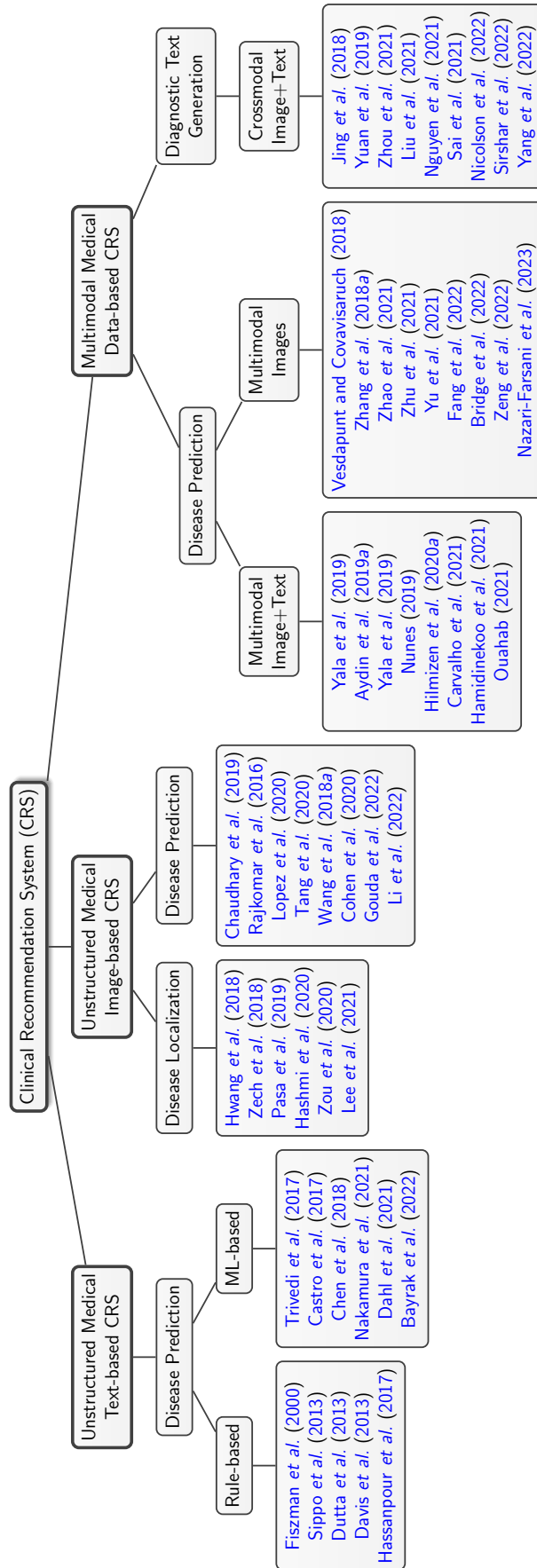


Figure 2.1: Classification of Clinical Recommendation Systems w.r.t data utilization

Our thesis extensively examines the different modules presented in Figure 2.1 and identifies research areas that require further exploration for each of these components.

## 2.1 CRs for Unstructured Medical Text Data Analysis

Analyzing unstructured medical text data in healthcare settings is more challenging than in other areas such as social media data or email classification (Mujtaba *et al.*, 2019). Unstructured medical texts can be difficult to understand because they often contain intricate medical terminologies, medical acronyms, words with spelling errors, and grammatical mistakes (Keselman and Smith, 2012). In order to deal with the complex vocabulary and irrelevant data in the corpus, it is necessary to perform a thorough pre-analysis of the corpus during the pre-processing stage (García *et al.*, 2016). This will help to address any lexical complications and noise in the data. In addition to the NLP challenges that are commonly encountered, mining radiology text reports presents a number of significant challenges. These include the difficulty of detecting and identifying normal or abnormal findings, as well as a lack of medical datasets that are available to the public (den Broeck *et al.*, 2005). Clinicians and researchers invest a significant amount of manual effort in generating, annotating, and benchmarking data for a specific decision-making task (Purushotham *et al.*, 2018). After just two years of being established, the Text REtrieval Conference (TREC) medical track was disbanded as a result of insufficient publicly available medical data (Voorhees, 2013).

### 2.1.1 Disease Prediction from Unstructured Radiology Reports

Radiology reports are a crucial component of the medical record that contain unstructured medical text data (Pandey *et al.*, 2020). They are typically generated by radiologists, who are trained to interpret medical images, such as X-rays, CT scans, and MRIs, and provide a written summary of their findings (Hartung *et al.*, 2020). Radiology reports are a rich source of information containing a patient's medical state, encompassing specifics regarding the location and intensity of any anomalies, in addition to any other pertinent clinical details (Pool and Goergen, 2010). Nevertheless, the textual data present in radiology reports is frequently

unorganized, implying that it lacks a uniform structure that can be conveniently scrutinized by computer systems (Krupinski *et al.*, 2011). In order to obtain valuable information from radiology reports, healthcare institutions and researchers employ techniques such as NLP to analyze and organize the unstructured data. There are two main categories of methods for classifying unstructured radiology reports: rule-based methods (Dutta *et al.* (2013); Fiszman *et al.* (2000); Hassanpour *et al.* (2017); Sippo *et al.* (2013)) and conventional ML-based methods (Castro *et al.* (2017); Johnson *et al.* (2014); Trivedi *et al.* (2017)). Rule-based methods for classifying data typically utilize traditional pattern matching techniques that depend on pre-established medical terminologies determined by radiologists or general medical terminologies derived from standard healthcare ontologies such as the Systematized Nomenclature of Medicine Clinical Terms (SNOMED) CT<sup>1</sup>. The principal drawback of rule-based methods is that the efficiency of the system depends exclusively on the precision of the pre-established patterns or medical keywords. On the other hand, ML-based techniques classify reports by utilizing medical features acquired from labeled reports. In their work, Castro *et al.* (2017) suggested a ML-based classifier for categorizing Breast Imaging reports. They employed Bag-of-Words (BoW) for extracting features and used Naive Bayes (NB) and Support Vector Machine (SVM) classifiers for the classification task. Both the term-document matrix technique (Trivedi *et al.*, 2017) and n-gram approach (Johnson *et al.*, 2014) have been utilized for the similar purpose of classifying radiology reports. DL techniques have demonstrated encouraging outcomes in general text classification tasks, such as sentiment analysis (Nedjah *et al.*, 2019) and extracting relationships from free-text (He *et al.*, 2018). The favorable results of DL techniques in numerous applications have motivated us to employ them in clinical decision making by predicting diseases from radiology reports.

The pre-processed texts need to be represented in vector space or word embeddings to be processed by the ML or DL Techniques. Word Embeddings such as Global Vector (GloVe) (Pennington *et al.* (2014)) and Word2Vec models (Mikolov *et al.* (2013b)) are feature modelling strategies in NLP, where every word is mapped to the dense and real-valued vector space that captures its meaning and syntactic properties of the words in raw corpus. Shin *et al.* (2017) applied Term Frequency-Inverse Document Frequency (tf-idf) for text encoding and CNN with an attention mechanism to classify the CT radiology reports obtained from the private medical institute. The proposed model was compared with logistic regression (LR), Random forest (RF) and SVM applied to 1400 reports. The proposed atten-

---

<sup>1</sup>SNOMED International-leading healthcare terminology. Online: <http://www.snomed.org/>

tion model achieved better performance compared to the three statistical models. Chen et al. [Chen et al. \(2018\)](#) proposed a deep learning framework to classify radiology reports from CT imaging reports extracted from two private institutions. The CNN model with the GloVe word embedding technique was utilized for classifying the reports and showcased the superiority of the proposed model compared to the traditional rule-based classifier PEfinder [Chapman et al. \(2011\)](#). Dahl et al. [Dahl et al. \(2021\)](#) proposed a CNN, Bidirectional Long Short-Term Memory (bi-LSTM), and SVM to classify and detect findings from the Norwegian radiology CT reports. The BoW and tf-idf word embedding techniques are utilized as an NLP strategy. The CNN and bi-LSTM models achieved slightly better results compared with the traditional SVM model. Nakamura et al. [Nakamura et al. \(2021\)](#) presented an automated detection and classification of actionable reports obtained from a Japanese private institution. The binary classification of CT reports is performed using four statistical methods: LR, gradient boosting decision tree (GBDT), bi-LSTM, and the Bidirectional Encoder Representations from Transformers (BERT) model. The BERT achieved a significantly higher area under the precision-recall curve (AUPRC) than the other three statistical models. Bayrak et al. [Bayrak et al. \(2022\)](#) proposed a MRI radiology report classification from the data acquired from a private medical institute in Turkey. The index-based word encoding strategy for word embedding conversion of a free-text and the long short-term memory (LSTM) network, bi-LSTM, and CNN for classifying the reports into epilepsy disease or not. The bi-LSTM showcased better performance compared to the other two deep learning strategies.

The above literature showcases that the selection of the NLP task significantly impacts the prediction or classification task on unstructured clinical notes. Most existing research utilizes radiology reports extracted from private medical hospitals. The radiology reports available today are scarce in number as they are restricted to private hospitals or are specific to a particular domain. We have found that the existing literature for disease prediction from unstructured free-text radiology reports is inadequate to be compared with any other prediction techniques ([Castro et al. \(2017\)](#), [Dutta et al. \(2013\)](#), [Fiszman et al. \(2000\)](#), [Hasanpour et al. \(2017\)](#), [Sippo et al. \(2013\)](#), [Trivedi et al. \(2017\)](#)). Due to insufficient benchmark studies on publicly accessible datasets, there is a need to establish the best prediction techniques for radiology reports.

## 2.2 CRSs for Unstructured Medical Image Data Analysis

Early diagnosis is essential for improving treatment results, decreasing the risks linked to disease prognosis, expanding the range of available treatment options, stopping the spread of contagious diseases, and reducing the overall impact of illnesses (Diogo *et al.*, 2022). Medical imaging modalities like X-rays, CT scans, MRI scans, and ultrasound are crucial in detecting diseases at an early stage by providing intricate images of the body’s internal structures. These images aid physicians in identifying irregularities, such as tumors or other abnormal growths, that might signify the occurrence of an illness (Hussain *et al.*, 2022). The timely detection of illnesses with medical imaging can substantially influence the outcomes for patients. For example, spotting cancer at an early stage using medical imaging can prompt timely intervention and treatment, potentially boosting the chances of survival (Miles, 2011). Furthermore, medical imaging can be utilized to track the advancement of diseases over time and assess the efficacy of treatments (Lao *et al.*, 2018). Medical imaging is employed in diagnosing various medical conditions, such as heart disease (Sharma *et al.*, 2021), lung disease (Kieu *et al.*, 2020), and neurological disorders (Zhang *et al.*, 2020).

Medical imaging is frequently capable of providing insights that are not accessible through alternative diagnostic methods, like blood tests or physical examinations (Puttagunta and Ravi, 2021). Currently, most medical imaging techniques require manual interpretation by trained clinicians and experts to make a medical prognosis. This implies that a skilled expert is required to visually inspect the images produced by the medical imaging technology in order to detect any possible abnormalities or anomalies (Hosny *et al.*, 2018). For example, medical professionals specializing in radiology, who have expertise in interpreting medical images, usually analyze X-rays, CT scans, MRI scans, and other imaging techniques to arrive at diagnostic conclusions. They must scrutinize the images meticulously and compare them with the typical anatomy and functioning of the body to identify any anomalies or abnormalities (Krupinski, 2010). However, this process of manual interpretation can be lengthy and susceptible to errors made by humans. Furthermore, the interpretation may be influenced by individual perspectives, leading to varying interpretations by different experts even for the same set of images (van Timmeren *et al.*, 2020). The rapidly increasing number of patients with chronic conditions is leading to a surge in demand for healthcare services. This, in turn, is placing a significant cognitive and diagnostic burden on

healthcare professionals who are tasked with manually inspecting and interpreting medical images (McPhail, 2016).

To overcome these difficulties, researchers are investigating the utilization of ML and DL algorithms to automate the interpretation of medical images (Yoon *et al.*, 2019). This has the potential to improve the speed and accuracy of diagnoses, particularly for diseases that are difficult to detect or have inconspicuous symptoms. In the fields of Radiology, Pathology, Cardiology, and Neurology, AI-based methods have been successfully employed to interpret imaging data. Radiology is a specialized field of medicine that employs diverse imaging techniques, including X-rays, CT, MRI, and ultrasound, to diagnose and treat medical conditions and injuries. The data obtained from imaging procedures during regular medical checkups is essential in the diagnosis and treatment of illnesses. Moreover, the process of collecting patient information utilized in radiology does not pose substantial risks or adverse effects. Therefore, in this research, we have utilized CXR imaging in the radiology field to predict pulmonary ailments.

### 2.2.1 Disease Prediction from Unstructured Chest X-ray Images

Medical image processing involves a collection of methods aimed at extracting significant clinical data from diverse imaging techniques, typically used for the purposes of diagnosis or prognosis (Varoquaux and Cheplygina, 2022). Due to the release of multiple, huge, publicly available diagnostic chest imaging datasets presented in Table 2.1, there has been a series of significant research explored in the field of disease diagnosis using deep learning techniques. The existing research focuses on various tasks involving detection or localization, classification, prediction, segmentation, and visualization of multiple diseases from the CXRs. The disease detection or localization task identifies the specific abnormalities within the CXR.

#### 2.2.1.1 Disease Detection and Localization Task

For a disease detection or localization task, Wang *et al.* (2019) presented the deep CNN model for localizing chest diseases from the Chest X-ray14 (Wang *et al.*, 2017c) dataset and compared it with the traditional CNN: ResNet-50 (He *et al.*, 2015), AlexNet (Krizhevsky *et al.*, 2012a), VGG16 (Simonyan and Zisserman, 2015a), and GoogleNet (Szegedy *et al.*, 2014). Rajpurkar *et al.* (2017) proposed a 121-layered Dense Convolutional Network named CheXNet to predict pneumo-



Table 2.1: List of some currently available diagnostic X-ray datasets for chest diseases.

Dataset	Dataset Description	Predictable Disease
NIH Chest X-ray14 (Wang <i>et al.</i> , 2017c)	112,120 images of 14 diseases gathered from 30,805 patient	Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Pneumonia, Nodule, Pneumothorax, Edema, Emphysema, Fibrosis, Pleural Thickening and Hernia
Pediatric CXR (Kermany <i>et al.</i> , 2018)	5856 CXR images in which 3883 are Pneumonia images	Pneumonia
CheXper (Irvin <i>et al.</i> , 2019)	224,316 CXR of 65,240 cases	14 Chest Diseases
MIMIC CXR (Johnson <i>et al.</i> , 2019b)	227,827 images with 14 chest disease images	14 Chest Diseases
Open-I (Demner-Fushman <i>et al.</i> , 2015)	7470 chest radiographs with frontal and lateral view	Pulmonary Edema, Cardiac Hypertrophy, Pleural effusion and Opacity
MC dataset (Jaeger <i>et al.</i> , 2014)	138 Chest images, 58 from Tuberculosis patient	Tuberculosis
Shenzhen (Jaeger <i>et al.</i> , 2014)	662 Chest images, 336 from Tuberculosis patient	Tuberculosis
KIT dataset (Ryoo and Kim, 2014)	10,848 chest images, 3828 from Tuberculosis patient	Tuberculosis

nia pathology from the Chest X-ray14 dataset (Wang *et al.*, 2017c). For the binary classification of pneumonia detection, the pretrained ImageNet weights (Deng *et al.*, 2009) were utilized. The authors demonstrated that CheXNet performs better for pneumonia detection from CXRs. Candemir *et al.* (2018) presented Deep CNN models such as AlexNet, VGG-16, VGG-19, and Inception V3 to detect Cardiomegaly from the Open-I CXR dataset. Hwang *et al.* (2018) proposed the ResNet-based model with 27 layers and 12 residual connections to detect active pulmonary tuberculosis in the large private CXR cohort. Likewise, as a detection task, Zech *et al.* (2018) incorporated the DenseNet121 model pretrained with ImageNet weights to detect pneumonia abnormality across NIH Chest X-ray14



and Open-I CXR datasets. The authors have utilized pooled datasets from various cohorts and trained the model on these datasets. Different radiologists will have different thresholds to detect diseases to report them. Hence, the pooling of datasets has significantly degraded the model's performance. [Pasa et al. \(2019\)](#) utilized the Convolution Neural Network-based model for faster diagnosis of tuberculosis diseases from two CXR cohorts and used the Grad-CAM technique to visualize the existence of tuberculosis in CXR. [Zou et al. \(2020\)](#) presented three deep learning models: ResNet50, Xception, and InceptionV3, for detecting and screening pulmonary hypertension from a private dataset collected from three institutes in China. [Hashmi et al. \(2020\)](#) used a weighted classifier that combines the weighted predictions of the state-of-the-art deep learning model to detect pneumonia in CXRs and also uses a heatmap to visualize the abnormalities. [Lee et al. \(2021\)](#) presents the ResNet101 and U-Net models pretrained on ImageNet to segment and detect the cardiomegaly diseases from the three medical cohorts.

#### 2.2.1.2 Disease Classification and Prediction Task

Correspondingly, the image-level prediction task involves analyzing the CXR image and predicting labels (classification) or continuous values (regression). We have grouped classification and prediction tasks as they use a similar type of architecture. [Rajkomar et al. \(2016\)](#) proposed the GoogleNet architecture to classify the CXRs into frontal and lateral. [Chaudhary et al. \(2019\)](#) uses the CNN-based deep learning model with three convolution layers, ReLU activation, pooling, and fully connected layers to diagnose pulmonary diseases from the NIH Chest X-ray14 dataset. [Tang et al. \(2020\)](#) identified the pulmonary abnormality using Deep CNN models and compared the performance with the radiologist's labels. [Cohen et al. \(2020\)](#), conducted an investigative study to find discrepancies while generalizing the classification models with five different CXR datasets. The DenseNet model has been used for this cross-domain study, and it has been found that the model with good performance does not agree on predictions, and the model with poor performance agrees on predictions. The authors have shown that the models trained on multiple datasets do not achieve true generalization. [Li et al. \(2022\)](#) proposed the U-Net and ResNet-based models to segment, classify, and predict pulmonary fibrosis from CXRs. [Aydin et al. \(2019a\)](#) proposed a pretrained Densenet121 model to classify the CXRs into normal and abnormal classes from the Open-I dataset and achieved 74% classification accuracy. [Lopez et al. \(2020\)](#) also applied the DenseNet121 model to classify the pulmonary abnormalities in CXRs from

the Open-I dataset. The authors achieved an AUROC of 0.61 and investigated reducing annotation burden by using the clinical report with CXR. Wang *et al.* (2018a) proposed a CNN-based network to extract the imaging features and classify the common thorax diseases from the three medical cohorts, including the Open-I dataset. The authors achieved an average AUROC of 0.741 and studied classifying the thorax diseases by jointly training the model with clinical reports.

Recent research on pulmonary diseases also focuses on detecting and classifying COVID-19 from CXRs. COVID-19 is a life-threatening infectious pulmonary disease that has caused a pandemic situation. Griner *et al.* (2021) used an ensemble of DenseNet-121 Networks to classify COVID-19 from the private CXR dataset. Kusakunniran *et al.* (2021) utilized the ResNet101 model to detect COVID-19 and produced a heatmap for segmenting lung areas from the private CXR dataset. Helal Uddin *et al.* (2022) proposed the CNN-based deep learning model named SymptomNet to detect COVID-19, and a heatmap was generated to visualize the disease. Giełczyk *et al.* (2022) presented the CNN-based deep learning method to classify COVID-19 and pneumonia from 6939 CXRs pooled from different Kaggle repositories. The authors also examined some preprocessing strategies such as blurring, thresholding, and histogram equalization. Gouda *et al.* (2022) proposed ResNet-50 based on two different deep learning models to detect COVID-19 from the 2790 CXRs pooled from various open-source repositories. A detailed summary of the literature review is shown in Table 2.2.

### 2.2.1.3 Data Augmentation vs. Synthetic Data Generation

Deep Learning is a variant of representation learning that uses a simple hierarchical structure obtained from a set of features extracted to define complex data representation. Deep Learning has become the cutting-edge technology in computer vision due to the development of GPU-based parallel computing hardware. In healthcare settings, deep learning has shown significant potential for analyzing structured (like clinical lab data) and unstructured data (like medical images, signals, etc.). Deep learning models applied to different modalities and organs have led to considerable advancements in medical image analysis. Radiology is a medical domain that utilizes imaging data like MRI, X-ray, and CT to monitor and diagnose illness. The medical imaging data are challenging to collect and annotate as the dataset is limited to a private institution. The annotation process requires an expert radiologist to manually label every image, which is a rigorous and time-consuming task. Deep learning algorithms necessitate a substantial

quantity of data to build a reliable model that can detect, segment, classify, and predict diseases from an image. The unbalanced and small dataset may not lead to an effective model and may cause overfitting of the majority class. The accuracy of the deep learning frameworks could be improved by utilizing the existing cohort more effectively.

Data augmentation is one such strategy that enlarges the size of the cohort through the random geometrical translation of the images. Conventional augmentation techniques like randomly rotating, flipping, shearing, or adding extra brightness or noise to the image are commonly used. The quantity of new data that may be produced by conventional data augmentation strategies is constrained by the limited number of simple-to-compute invariances like zooming, flipping, etc. Generative Adversarial Networks (GANs) (Goodfellow *et al.*, 2014a) were utilized to generate synthetic images to improve the training of deep learning models without having to use any pre-determined augmentation. The cycleGAN was proposed to produce synthetic non-contrast CT images to significantly improve the deep learning framework's generalizability for segmenting CT images (Sandfort *et al.*, 2019). Frid-Adar *et al.* (2018) presented a GAN-based model to enhance the performance of the convolutional neural network in classifying liver lesions from the CT images. Mondal *et al.* (2018) proposed a GAN-based model for semi-supervised segmentation of 3D multimodal medical images and showcased the performance increment compared to traditional segmentation tasks. The Conditional Generative Adversarial Networks (cGANs) were utilized to synthesize the MR images pertaining to Alzheimer's disease (AD) (Jung *et al.*, 2021). The Hierarchical Amortized GAN was proposed to generate high-resolution synthetic images from the 3D thorax CT and brain MR images (Sun *et al.*, 2022). The GAN-based data augmentation approaches have demonstrated remarkable outcomes in producing high quality images and improving the classifier's performance by enhancing its generalizing ability. In this study, we utilize data augmentation and synthetic data generation techniques on radiology images to empirically evaluate their effectiveness in improving radiology image classification.

Table 2.2: Summary of Literature Survey - Disease Prediction from Unstructured Chest X-ray Images

Author & year	Methodology	Task	Medical Domain	Abnormality	Imaging Data	Dataset
<a href="#">Rajkomar et al. (2016)</a>	The GoogleNet architecture is used to classify the CXRs into frontal and lateral.	Classification	Radiology	Pulmonary diseases	Chest X-ray	Private Dataset (909 Patients)
<a href="#">Rajpurkar et al. (2017)</a>	The 121-layered Dense Convolutional Network named CheXNet was used to predict Pneumonia pathology from CXRs, and for the binary classification of Pneumonia detection, pre-trained ImageNet weights were utilized.	Detection	Radiology	Pneumonia	Chest X-ray	NIH Chest X-Ray 14 (112,120 from 30,805 patients)
<a href="#">Candemir et al. (2018)</a>	Deep CNN models like AlexNet, VGG-16, VGG-19, and Inception V3 are utilized to detect Cardiomegaly from the CXRs.	Detection	Radiology	Cardiomegaly	Chest X-ray	Open-i (283 Cardiomegaly cases from 3683 patients)

Continued on next page

Table 2.2 – Continued from previous page

Author & year	Methodology	Task	Medical Domain	Abnormality	Imaging Data	Dataset
<a href="#">Hwang <i>et al.</i> (2018)</a>	The ResNet-based model with 27 layers and 12 residual connections is utilized to detect active pulmonary tuberculosis in the large private CXR cohort.	Detection	Radiology	Pulmonary Tuberculosis	Chest X-ray	Private Dataset (54,221 Normal CXRs and 6768 tuberculosis CXRs)
<a href="#">Zech <i>et al.</i> (2018)</a>	The DenseNet121 model, pre-trained with ImageNet weights, is trained and tested across different data cohorts to detect the pneumonia abnormality.	Detection	Radiology	Pneumonia	Chest X-ray	NIH Chest X-Ray 14 (112,120 from 30,805 patients), MSH (42,396 from 12,904 patients), Open-I (3,807 from 3,683 patients)
<a href="#">Pasa <i>et al.</i> (2019)</a>	A CNN-based model is proposed for faster diagnosis of tuberculosis diseases, and the Grad-CAM technique is incorporated for disease visualization.	Detection and Visualization	Radiology	Tuberculosis	Chest X-ray	NIH Tuberculosis CXR (138 and 662 patients), Belarus Tuberculosis Portal dataset (304 patients)

Continued on next page

Table 2.2 – Continued from previous page

Author & year	Methodology	Task	Medical Domain	Abnormality	Imaging Data	Dataset
<a href="#">Chaudhary et al. (2019)</a>	The CNN-based deep learning model with three convolutions, ReLU, pooling, and fully connected layers was proposed to diagnose chest diseases from CXRs.	Classification	Radiology	Pulmonary diseases	Chest X-ray	NIH Chest X-ray14 (1,12,120 CXRs)
<a href="#">Tang et al. (2020)</a>	Identifying abnormalities using Deep CNN models and comparison with the radiologist's labels.	Classification	Radiology	Pulmonary diseases	Chest X-ray	NIH ChestX-Ray14 (112,120 from 30,805 patients), Open-I (3,807 CXRs from 3683 patients), RSNA Dataset (21,152 patients)

Continued on next page

Table 2.2 – Continued from previous page

Author & year	Methodology	Task	Medical Domain	Abnormality	Imaging Data	Dataset
<a href="#">Cohen <i>et al.</i> (2020)</a>	Investigative study to find discrepancies while generalizing the models with multiple CXR datasets.	Classification	Radiology	Pulmonary diseases	Chest X-ray	NIH Chest X-ray14 (112,120 from 30,805 patients), PadChest (1,60,000 from 67,000 patients), MIMIC-CXR (227827 CXRs), Open-I (3,807 CXRs from 3683 patients), RSNA Dataset (21,152 patients)
<a href="#">Zou <i>et al.</i> (2020)</a>	Detection and screening of Pulmonary Hypertension using three deep learning models (Resnet50, Xception, and Inception V3)	Detection and Visualization	Radiology	Pulmonary hypertension	Chest X-ray	Private dataset (762 patients from three institute in China)

Continued on next page

Table 2.2 – Continued from previous page

Author & year	Methodology	Task	Medical Domain	Abnormality	Imaging Data	Dataset
<a href="#">Hashmi et al. (2020)</a>	A weighted classifier combining the weighted predictions of the state-of-the-art deep learning model is introduced to detect pneumonia in CXRs.	Detection and Visualization	Radiology	Pneumonia	Chest X-ray	Private dataset (7022 CXRs)
<a href="#">Griner et al. (2021)</a>	The classification of COVID-19 abnormalities is performed using an ensemble of DenseNet-121 Networks.	Classification	Radiology	COVID-19	Chest X-ray	Private dataset (12000 patients)
<a href="#">Lee et al. (2021)</a>	ResNet 101 and U-Net, pre-trained on ImageNet, are used to segment and detect the cardiomegaly diseases from the CXRs.	Segmentation and Detection	Radiology	Cardiomegaly	Chest X-ray	JSRT dataset (247 patients), Montgomery dataset (138 patients), Private dataset (408 patients).

Continued on next page



Table 2.2 – Continued from previous page

Author & year	Methodology	Task	Medical Domain	Abnormality	Imaging Data	Dataset
<a href="#">Kusakunniran <i>et al.</i> (2021)</a>	The ResNet101 model is utilized to detect COVID-19, and a heatmap is produced for the segmented lung area.	Detection and Visualization	Radiology	COVID-19	Chest X-ray	Private dataset (5743 CXRs)
<a href="#">Helal Uddin <i>et al.</i> (2022)</a>	The CNN-based deep learning model named SymptomNet is proposed to detect COVID-19, and a heatmap is generated to visualize the disease.	Detection and Visualization	Radiology	COVID-19	Chest X-ray	Private dataset (500 CXRs from Bangladesh)
<a href="#">Gielczyk <i>et al.</i> (2022)</a>	The CNN-based deep learning method is used to classify COVID-19 and Pneumonia. We also examined some pre-processing strategies like blurring, thresholding, and histogram equalization.	Classification	Radiology	Pneumonia and COVID-19	Chest X-ray	Pooled data from various cohorts (6939 CXRs)

Continued on next page

**Table 2.2 – Continued from previous page**

Author & year	Methodology	Task	Medical Domain	Abnormality	Imaging Data	Dataset
<a href="#">Gouda <i>et al.</i> (2022)</a>	The ResNet50-based two different Deep Learning approaches have been proposed to detect COVID-19.	Detection	Radiology	COVID-19	Chest X-ray	Pooled data from various cohorts (2790 CXRs)
<a href="#">Li <i>et al.</i> (2022)</a>	The U-Net and ResNet based models were proposed to segment, classify, and predict pulmonary fibrosis from CXRs.	Segmentation, Classification and Prediction	Radiology	Pulmonary Fibrosis	Chest X-ray	NIH Chest X-ray14 (Pulmonary fibrosis CXRs from 1,12,120 images)

## 2.3 CRSs for Multimodal Medical Data Analysis

EHRs are digital repositories that hold extensive patient medical information. These records are composed of multimodal data, meaning that they comprise diverse categories of information, such as textual and imaging data. EHR consists of a substantial volume of valuable information and, hence, provides researchers with the opportunity to establish data-driven models (Devarakonda and Tsou (2015); Jindal and Taneja (2015)). There has been a significant amount of research work carried out for predicting diseases from the Unimodal imaging data by utilizing only pixel-value information without leveraging the valuable clinical context from the structured or unstructured EHR (Gulshan *et al.* (2016); Hinton (2018); Dunnmon *et al.* (2018); Johnson *et al.* (2019a)). There has been a scarcity of research that leverages multimodal data containing both unstructured textual information and images. Multimodal medical image fusion can enhance the accuracy of diagnosis in radiology by combining different types of medical images (Hermessi *et al.*, 2021). Integrating different features from diverse modalities can furnish healthcare providers with a comprehensive understanding of a patient's medical condition, aiding them in making informed decisions about the most suitable treatment options.

Data analysis involves a structured approach that employs various methods such as data examination, refinement, conversion, and modelling. Figure 2.2 represents the general structure of multimodal medical data analysis. The various steps involved in multimodal medical data processing are as follows:

- *Feature Extraction:* The healthcare industry is experiencing a rise in high-precision, multimodal medical data due to the growing use of technology and mobility. The efficient utilization of this type of data can contribute significantly to the analysis and resolution of various healthcare challenges. Nevertheless, the diversified nature of multimodal data, such as text, images, and signals, poses a challenge in developing efficient data extraction algorithms. Feature extraction is the initial step in collecting raw data from various sources for further processing and storage (Chaudhury *et al.* (2016)). To extract features from medical data, various approaches have been proposed in the literature. For instance, Iakovidis and Smailis (2012) employed an unsupervised data mining approach to extract low-level data and their multiple features from a consolidated multimodal repository. In another work, Wang *et al.* (2018b) proposed a novel approach to represent complex medical data in a knowledge-based graph model. The graph similarity search

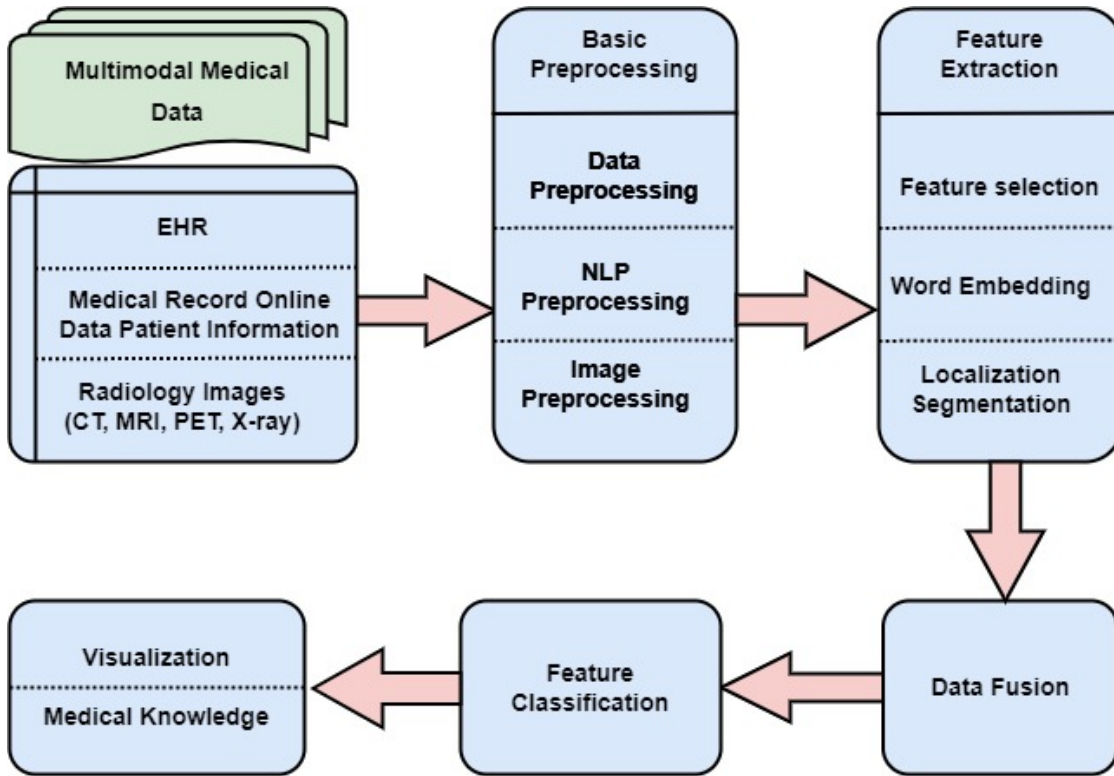


Figure 2.2: General Architecture of Multimodal Medical Data Analysis

was then applied to the knowledge graph, and lazy learning algorithms, such as dynamic time warping, were employed to find the similarity between the created graphs. The proposed method exhibited superior accuracy in comparison to the baseline models. In the healthcare industry, data is available in various formats and from diverse sources (Bleyer (1997)). To extract and process physiological data, such as galvanic skin response, heart rate, facial expression, text, and speech, a range of techniques are employed, including pattern matching, similarity search, feature extraction, automated annotation, classification, and clustering. In one study, Kurniawan and Pechenizkiy (2014) proposed a framework for stress analysis from multimodal affective data, such as physiological signals and external user data, including facial expression, speech, and text. Pattern mining techniques were utilized to extract features from various data models. In the cardiology domain, it is possible to extract data using various algorithms, and feature extraction can be accomplished by assigning the same label to similar data solutions (Syeda-Mahmood *et al.*, 2007). In sub-cancer pixel analysis of MRI and mammography images, decision tree models, chi-square, and automatic interaction detection methods are frequently utilized for feature extraction in

ML (Wu *et al.*, 2019).

Conventional ML techniques relied on human intervention to identify and extract particular features before forwarding them to the fusion or classification phase. With the emergence of deep learning, it has become possible to automatically extract features from medical data that contain multiple modes of information. Purwar *et al.* (2019) employed an AlexNet convolutional neural network model to extract features from both red blood cell imaging and organized blood reports, enabling the detection of microcytic hypochromia. Faris *et al.* (2021) presented a multimodal framework that utilizes both structured symptom data and unstructured medical questions, comprising a total of 263,867 instances, to support medical diagnosis via telemedicine. To extract features from the data, the authors employed tf-idf, hashing vectorizers, and doc2vec models. After the process of data extraction, the authors performed fusion and classification techniques to accurately predict disease diagnosis in telemedicine. It is noteworthy that CNN-based models are the preferred choice for deep learning feature extraction in numerous studies that extract features from diverse data modalities (Hilmizen *et al.* (2020b); Carvalho *et al.* (2021); Hamidinekoo *et al.* (2021)).

- *Data Fusion:* With the increasing availability of data from various sources, multimodality has become common in all fields, including the medical domain. Integration of multimodal data in medical decision support systems can significantly enhance their performance (Lahat *et al.*, 2015). Data fusion involves combining datasets from diverse sources and modalities, which poses significant challenges due to differences in frequencies and noise. Data fusion has been widely adopted by researchers as a technique for multimodal data analysis in various applications. Several data fusion approaches, such as data fusion for hybrid Brain-Computer Interface (BCI), rhythm-based BCI, and fusion of multiple heartbeat physiological signals, have been studied using comparative analysis by researchers (Chandra *et al.* (2019); Fazli *et al.* (2015)). Multimodal data typically contains multiple perspectives, and utilizing a multi-view approach to classify different subsets of the data has been found to be beneficial. In their study, Shachor *et al.* (2020) developed a new fusion framework that utilized a neural network and a mixture of views to process multimodal data, leading to a significant improvement in performance. Furthermore, in the field of multimodal medical data analysis, techniques such as anatomical structure identification, feature analysis,

and labeling approaches have been utilized to support 3D neuroanatomical database analysis, as demonstrated by Barillot *et al.* (1993). Reliable fusion techniques in the field of neuroimaging include the Markov-Penrose diagram of tensor network notation, Bayesian DAG, and coupled matrix tensor factorization, as suggested in the literature. A recent study has shown that a deep-gate convolutional neural network can be used to fuse multi-band images, with outstanding results obtained when fusing low and high-frequency components compared to existing systems (Lin *et al.* (2020)). According to recent research by Lin *et al.* (2020), combining low and high-frequency components through fusion yields remarkable results in comparison to current systems. A study by Adali *et al.* (2015) suggests that joint independent component analysis and transposed independent vector analysis models can effectively fuse MRI, EEG, and Structural MRI data. Furthermore, multi-band image fusion has a wide range of applications for enhancing image quality. Gaussian filters (Mohd *et al.* (2017)) and singular value decomposition (Nischitha and Padmavathi (2017)) are examples of filters that perform well and yield satisfactory results.

- *Classification:* The multimodal data fusion is followed by the classification and visualization tasks. In healthcare, the classification of diseases based on different medical data is crucial. This task can be achieved using machine learning (ML) and deep learning algorithms. For instance, breast cancer classification has been done using Ranklet transforms, LSP Ranklet transforms, and support vector machines (SVMs) Xi *et al.* (2017), while an encoder-decoder layer followed by a least-square algorithm has been used for ECG, MRI, and EEG signal compression classification (Zhang and Shen, 2011). Deep learning-based binary classification of chest diseases from CXR and associated radiology reports collected from the Indiana University (IU) dataset was performed by (Aydin *et al.*, 2019a) using a multimodal approach. The imaging features were obtained using a pre-trained CNN model, and the textual features were retrieved using a GloVe embedding model. The concatenated features were then passed through a fully connected network for classification. In another study, Lopez *et al.* (2020) compared the performance of a multimodal model and a unimodal model for CXR and associated radiology reports collected from the IU dataset. The fused features were passed through a fully connected deep neural network for classification, and the results showed a reduction of annotation burden through mul-

timodal learning. Data classification has diverse applications beyond visual and textual data, and it has been widely adopted in many fields. One such application is the design of support systems for Parkinson's patients based on their handwriting. In [Heidarvincheh et al. \(2021\)](#), a multimodal classification of Parkinson's disease (PD) in a home environment was proposed by extracting features from raw data obtained through a wrist-worn accelerometer and RGB-D camera. The silhouette images and accelerometer signals were preprocessed and classified as PD or healthy using an encoder-decoder CNN model. In [Ribeiro et al. \(2013\)](#), an approach was developed to classify chronic liver disease stages using clinical laboratory and ultrasound data, utilizing techniques such as SVM, Bayes, and K-means clustering. In [Yi et al. \(2022\)](#), a multimodal classification framework was proposed to categorize the severity of glaucoma from fundus and grayscale images collected from the Kunming Medical University. CNN-based classifiers were utilized for the classification task. Additionally, [Hilmizen et al. \(2020b\)](#) used a CNN-based classifier for classifying COVID-19 disease from multimodal CXR and CT features extracted using pre-trained VGG16 and ResNet models.

Multimodal classification has numerous applications in various medical domains, and ML and DL classifiers are used to categorize data for prognosis outcomes. Supervised learning is required for ML classifiers, which means that human intervention is necessary to manually pick features before passing them through the classifiers. However, deep learning classifiers do not require handcrafted features before feeding them into the fully connected layers for classification. Additionally, ML models do not learn incrementally, whereas DL classifiers can overcome this shortcoming by incrementally learning features.

- *Visualization*: Data visualization refers to the process of representing data and information in a visual and graphical format, such as charts, graphs, and maps, to make it easier to understand and analyze. The use of data visualization is crucial in the medical field to interpret and convey complex medical information, including medical images, clinical data, EHRs, and patient outcomes. By utilizing medical data visualization techniques, healthcare practitioners can identify patterns and trends in the data, make better-informed decisions, and ultimately improve patient outcomes. This section delves into data visualization techniques for multimodal medical data, which can be achieved through various algorithms, software, or hardware components.

For instance, [Levin \*et al.\* \(2005\)](#) designed hardware that used motion-based segmentation to visualize 4D cardiac data, which yielded significant performance improvements. Software-based data visualization methods, such as EHR data multimodal analysis using chromatogram plots, have also been explored. OpenCL, C++, and GUI toolkits are commonly used to visualize essential data features, and iterative visualization and MVC pattern algorithms have been widely adopted ([Manssour \*et al.\*, 2000](#)). When it comes to MRI, single-photon emission computerized tomography (SPECT), CT, and PET data visualization, 2D or 3D image visualization techniques are recommended, with inertial moment 3D visualization providing a better view. In addition, radar plots have been used to enhance saturation and transfer function to visualize multimodal data from image-guided neurosurgery ([Joshi \*et al.\*, 2010](#)). Trend charts, timelines, and data tables can also be utilized to visualize EHR and clinical data ([An \*et al.\*, 2008b](#)). Furthermore, [Song \*et al.\* \(2021\)](#) performed a quantitative analysis by integrating the Grad-CAM visualization technique into a multimodal fusion framework applied to MRI and PET images for diagnosing Alzheimer’s disease. Overall, data visualization is an essential tool for effectively representing complex data sets and aiding in the interpretation of multimodal medical data.

Multimodal information extracted from EHRs has been utilized in several tasks, including multimodal medical disease classification and prediction, generating diagnostic notes, and multimodal medical image fusion. This thesis explores several innovative approaches to multimodal analysis, including predicting pulmonary disease by fusing medical images and text, generating diagnostic reports from CXR, and detecting acute infarct by fusing MRI sequences.

### 2.3.1 Multimodal Diagnostic Image and Text Analysis

Data fusion specifies the integration of information from multiple modalities to retrieve complementary and more significant information for designing and developing effective, better-performing ML models than a model leveraging unimodal data. Our research has considered CXRs and associated radiology reports as multimodal data due to their existence in imaging and textual form. We can categorize medical data fusion techniques into early fusion, late fusion, and joint fusion methods ([Ramachandram and Taylor, 2017](#)). In early fusion techniques, also known as feature-level fusion, the features from heterogeneous sources or learned features retrieved from the neural networks, or manually extracted features are combined



and fed to a single ML or DL model to produce the final predictive decisions (Kharazmi *et al.* (2017); Li and Fan (2019); Purwar *et al.* (2020)). In late fusion techniques, also called a decision-level fusion, we fuse the predictions obtained from more than one ML or DL model to produce final predictive decisions (Qiu *et al.* (2018); Reda *et al.* (2018)). Whereas in Joint fusion techniques, also known as intermediate-level fusion, where the features from heterogeneous modalities are learned through the intermediate layers of neural networks, these learned features from multiple modalities are fused before being ingested into the final model to obtain predictive decisions (Yoo *et al.* (2017); Spasov *et al.* (2018); Yala *et al.* (2019); Aydin *et al.* (2019a)). The main distinction between the joint fusion approach and the early fusion strategy is that, with each training iteration, a better representation of learned features is obtained by back-propagating the loss to the feature-extracted neural network. As a result, joint fusion strategies are exclusively applied to neural networks due to their ability to back-propagate their loss to the feature retrieval network.

In the following literature review, we examine research that encompasses the fusion of medical image features with structured clinical measurements to predict various diseases. Kharazmi *et al.* (2017) in their research work, detected basal cell cancer from the multimodal dermoscopic images with structured clinical data like age, sex, size, and location of the lesion. They applied an early fusion strategy by concatenating the features extracted from both modalities using the CNN model. Li and Fan (2019) presented a multimodal framework for predicting Alzheimer's disease from the MRI and structured clinical information like assessments, questionnaire's and patient demographics. Here, the authors used concatenation for an early fusion of the imaging and clinical features obtained from CNN. Purwar *et al.* (2020) in their research work, detected microscopic hypochromia from the Red Blood Cell (RBC) images and structured clinical test reports, including blood count and other blood test parameters. The early fusion strategy was leveraged by concatenating CNN features obtained from imaging and clinical blood test reports. Qiu *et al.* (2018) implemented three CNN models to retrieve imaging features from three MRI images, and the late fusion strategy like mean, max, and majority voting is applied for fusing three images. Furthermore, in this study, two Multilayer Perceptron (MLP) models are employed to input non-imaging clinical assessment data, such as Mini-Mental State Examination (MMSE) and logical memory (LM) test results. The resulting features from these two MLP models are then combined with the imaging features through a majority voting strategy. Reda *et al.* (2018) proposed a meta-classifier based late fusion strategy, Stacked Nonnegativity Con-

straint Sparse Autoencoders (SNCSAE), for integrating features from MRI and structured PSA blood tests to predict the prostate cancer diagnosis. [Spasov \*et al.\* \(2018\)](#) applied CNNs to extract imaging features from the MRI and jointly fused them with structured clinical data, including demographics, genetic data, clinical assessments, and verbal learnings, before injecting it into a feed-forward Neural Network for predicting Alzheimer’s disease. [Yoo \*et al.\* \(2017\)](#) proposed joint and late fusion strategies using concatenating (joint) and Averaging (late) the imaging features extracted from MRI and clinical measurements using the CNN model to predict Multiple sclerosis. [Yala \*et al.\* \(2019\)](#) presented a joint fusion strategy by concatenating the CNN based pixel features and clinical features extracted from the mammograms and the clinical measurements, respectively. These features are further fed to the feed-forward neural network to predict breast cancer. [Carvalho \*et al.\* \(2021\)](#) proposed a multimodal framework to classify skin cancer from normal dermatoscopic images. The efficientNet-B3 model was utilized to extract features from the images, and a concatenation (joint fusion) strategy was applied to fuse ABCD features with the imaging features. So far, we’ve seen that structured clinical data combined with imaging features have considerably impacted disease prediction outcomes.

Likewise, we can leverage unstructured clinical reports with radiology imaging features to provide clinical context and improve the performance of disease predictions compared to unimodal text or image models ([Aydin \*et al.\* \(2019a\)](#); [Lopez \*et al.\* \(2020\)](#)). [Aydin \*et al.\* \(2019a\)](#) proposed a multimodal classifier with transfer learning to jointly fuse the medical report with the CXR images to classify patients into normal and abnormal classes. The authors considered a pre-trained DenseNet121 model to retrieve imaging features and a Glove embedding layer to produce textual features. Further, both features were concatenated before passing it to the dense feed-forward network. [Lopez \*et al.\* \(2020\)](#) presented the multimodal fusion strategy by applying DenseNet121 for image feature extraction from CXR and the word2vec model for text feature extraction from radiology reports. The authors have experimented with early, late, and joint fusion strategies by concatenating (early & joint) and averaging (late) imaging and textual modality features. [Nunes \(2019\)](#) proposed a multimodal framework for the classification of pulmonary diseases from the Indiana University dataset. To extract features from the radiology report, the LSTM based BioWordVec is applied, and EfficientNet-B5 is used to retrieve imaging features. The multimodal features obtained from single modal models are concatenated (joint fusion) and passed through a fully connected neural network to classify the diseases. [Huang \*et al.\* \(2020\)](#) presented

neural network-based text and image retrieval model to detect Pulmonary Embolism from CT images. The multimodal features obtained are fused (early & joint) using concatenation operations and late fused using averaging. [Hilmizen \*et al.\* \(2020a\)](#) and [Ouahab \(2021\)](#) proposed CNN-based feature extraction technique to detect COVID-19 from the two different modality images (CT + X-ray). The authors have applied joint fusion concatenation to integrate both imaging features. [Hamidinekoo \*et al.\* \(2021\)](#) applied Densely Connected Network (DCN) extract imaging features from the MRI and Whole Slide Imaging (WSI) images to detect Glioma disease. The major voting (late fusion) strategy was employed to ensemble the features obtained from two separate DCN models. All the above strategies for fusion use a straightforward concatenation strategy, ignoring inter-modal interactions between the two features.

The above literature shows that fusion approaches significantly improve performance compared with unimodal models when applied to medical cohorts. Incorporating clinical context by including structured or unstructured clinical data with medical images has provided better prognosis decisions. In most of the early fusion strategies ([Kharazmi \*et al.\* \(2017\)](#); [Li and Fan \(2019\)](#); [Purwar \*et al.\* \(2020\)](#)), the imaging features and clinical features retrieved from the neural networks are integrated or fused using concatenation approaches, forming a single plain vector, which does not always guarantee good results. In the late fusion techniques, the fusion strategies are basically focused on aggregating the results from the various unimodal models by using meta-classifiers, majority voting, mean, or max ([Reda \*et al.\* \(2018\)](#); [Qiu \*et al.\* \(2018\)](#)). The major limitation with the late fusion strategy is inter-modality dynamics; the interaction between the multimodal data is completely ignored. Most of the existing works ([Kharazmi \*et al.\* \(2017\)](#); [Li and Fan \(2019\)](#); [Purwar \*et al.\* \(2020\)](#); [Reda \*et al.\* \(2018\)](#); [Qiu \*et al.\* \(2018\)](#); [Spasov \*et al.\* \(2018\)](#); [Yoo \*et al.\* \(2017\)](#); [Yala \*et al.\* \(2019\)](#)), leverage the structured clinical data with the imaging features to predict the diseases from the multimodal medical cohort. After a thorough literature survey, it is observed that there has been a limited study carried out fusing unstructured radiology clinical free-text reports with the pixel features extracted from the radiology images to obtain valuable, meaningful prognosis information for the clinicians. DL and image captioning techniques have advanced significantly in recent years, allowing researchers to apply them to the cross-modal retrieval of generating radiology reports from the CXRs ([Wang \*et al.\* \(2018a\)](#); [Liu \*et al.\* \(2019b\)](#); [Gajbhiye \*et al.\* \(2020\)](#); [Alfarghaly \*et al.\* \(2021b\)](#); [Yuan \*et al.\* \(2019\)](#)). The overall summary of the literature review is presented in Table 2.3.

Table 2.3: Summary of Literature Survey - Multimodal Diagnostic Image and Text Analysis

Author & year	Methodology	Fusion Strategy	Multimodal Task	Disease	Imaging Data	Input non-imaging data	Dataset	# of Cases	Remarks
<a href="#">Wang et al. (2018a)</a>	CNN (image) + RNN (Text)	Joint Fusion - Concatenation	Classification & Cross-modal retrieval	Thorax Disease	CXR	Radiology Reports	IU dataset	3955	Intermodal Interaction Missing
<a href="#">Aydin et al. (2019a)</a>	Custom Glove (Text) + Pretrained Densenet121 (Image)	Joint Fusion - Concatenation	Classification	Pulmonary Diseases	CXR	Radiology Reports	IU dataset	3955	Intermodal Interaction Missing
<a href="#">Nunes (2019)</a>	LSTM based BioWord-Vec + EfficientNet-B5 (image)	Joint Fusion - Concatenation	Classification	Pulmonary Diseases	CXR	Radiology Reports	IU dataset	3955	Intermodal Interaction Missing

Continued on next page

Table 2.3 – Continued from previous page

Author & year	Methodology	Fusion Strategy	Multimodal Task	Disease	Imaging Data	Input non-imaging data	Dataset	# of Cases	Remarks
<a href="#">Yuan <i>et al.</i> (2019)</a>	Encoder (Images + text)	Joint Fusion - Concatenation, Early & Late Fusion Attention	Cross-modal retrieval	Pulmonary Diseases	CXR	Radiology Reports	IU dataset	3955	Intermodal Interaction Missing
<a href="#">Lopez <i>et al.</i> (2020)</a>	Word2Vec (Text) + DenseNet121 (Image)	Joint & early Fusion- Concatenation, Late Fusion-Averaging	Classification	Pulmonary Diseases	CXR	Radiology Reports	IU dataset	3955	Intermodal Interaction Missing
<a href="#">Huang <i>et al.</i> (2020)</a>	Neural Network (text + image)	Early & Joint Fusion - Concatenation, Late Fusion - Averaging	Detection	Pulmonary Embolism (PE)	Chest CT	EHR	Stanford University Medical Center	2500	Intermodal Interaction Missing

Continued on next page

Table 2.3 – Continued from previous page

Author & year	Methodology	Fusion Strategy	Multimodal Task	Disease	Imaging Data	Input non-imaging data	Dataset	# of Cases	Remarks
<a href="#">Hilmizen et al. (2020a)</a>	VGG16 and ResNet50 (for both images)	Joint Fusion - Concatenation	Classification	Covid-19	CXR and CT	-	Mixed dataset from kaggle Repository	1257	Intermodal interaction Missing and non-imaging data is not utilized.
<a href="#">Ouahab (2021)</a>	CNN (for both images)	Joint Fusion - Concatenation	Detection	Covid-19	CXR and CT	-	Mixed dataset from Kaggle Repository	1045	Intermodal interaction Missing and non-usage of non-imaging data.

Continued on next page

Table 2.3 – Continued from previous page

Author & year	Methodology	Fusion Strategy	Multimodal Task	Disease	Imaging Data	Input non-imaging data	Dataset	# of Cases	Remarks
<a href="#">Hamidinekoo et al. (2021)</a>	DCN (for both images)	Late Fusion - Major Voting	Classification	Glioma (Brain Tumor)	MRI and WSI	-	CPM-RadPath 2020 challenge dataset	329	Intermodal interaction Missing and non-usage of non-imaging data.
<a href="#">Carvalho et al. (2021)</a>	EfficientNet-B3 (images)	Joint Fusion - Concatenation	Classification	Skin Cancer	Dermoscopic images	ABCD Psuedo features	ISIC 2017 challenge dataset	2750	Intermodal Interaction Missing
<a href="#">Alfarghaly et al. (2021a)</a>	ChexNet (images) + Word2Vec (text)	Late Fusion-Attention	Cross-modal retrieval	Pulmonary Diseases	CXR	Radiology Reports	IU dataset	3955	Intermodal Interaction Missing

Our research work primarily focuses on predicting abnormalities from the multimodal CXR and its associated clinical reports by jointly fusing the pixel information with the radiology text feature using DL-based multimodal tensor fusion networks, considering inter-modality dynamics.

### 2.3.2 Cross-modal Medical Report Generation

Hospitals around the world heavily rely on medical imaging, which provides valuable insights for disease diagnosis and treatment planning. However, it is crucial for the radiologist to thoroughly examine the medical images in order to provide comprehensive findings and interpretations. In order to produce precise and reliable radiology reports, it is necessary for the radiologist to possess ample experience and devote a significant amount of time to scrutinizing the medical images (Jing *et al.*, 2018). A large number of radiology reports may end with inconclusive comments, resulting in patients undergoing additional tests, such as advanced imaging or pathology exams. The issue of the time required for a radiologist to create a detailed report is a significant concern, as on average, an experienced radiologist will need approximately 10–20 minutes to produce a thorough report. In situations such as overcrowded hospitals or during a pandemic, writing radiology reports can become challenging due to the ever-increasing number of cases (Yang *et al.*, 2022). These circumstances inspired our research into developing an automated radiology reporting system using a deep learning framework. Considerable progress has been made in the field of generating medical descriptions. Yuan *et al.* (2019) introduced an automatic report generation model that utilizes a multiview CNN encoder and a concept-enriched hierarchical LSTM. The model leverages multi-view information in radiology by employing visual attention in a late fusion manner and enriching the semantics in the hierarchical LSTM decoder with medical concepts. The authors Nguyen *et al.* (2021), presented a set of three modules consisting of classification, generation, and interpretation. For the classification module, a multi-view encoder is employed to extract visual features from chest X-rays, while a text encoder converts reports into embeddings. The generation module utilizes both visual and textual features to create text on a word-by-word basis. Finally, the interpretation module fine-tunes the text generated. Sai *et al.* (2021) showcased an automatic report generation model with the following stages: NLP Pipeline (Tokenization, Embedding, Removing Special Characters, etc.); CNN acts as an encoder in our model. A transfer Learning model, ChexNet is used to extract the features of the image. Hierarchical LSTMs



and Co-Attention Mechanism: Hierarchical LSTMs are designed to enrich the representation ability of the LSTM, and the co-attention mechanism provides the context. The sentence and word LSTMs then generate the final reports required. [Zhou \*et al.\* \(2021\)](#) presented a visual-textual attentive semantic model that uses DenseNet201 as a visual encoder model and BioSentVec as a text encoder. The LSTM model is utilized to generate the report. [Liu \*et al.\* \(2021\)](#) proposed an unsupervised Knowledge Graph Auto-Encoder (KGAE) model that utilizes independent sets of Chest X-ray images and their associated reports during the training phase. KGAE consists of a pre-constructed knowledge graph, a knowledge-driven encoder, and a knowledge-driven decoder. They have used the Knowledge-driven encoder to project medical images and reports to the corresponding coordinates in latent space and the Knowledge-driven decoder to generate a medical report on a given coordinate in that space. [Sirshar \*et al.\* \(2022\)](#) propose an encoder-decoder model with CNN used as a visual encoder and an RNN decoder with attention used to produce the radiology reports. [Nicolson \*et al.\* \(2022\)](#) presented the report generation framework, where the DenseNet pretrained on imageNet is used as an encoder for imaging feature extraction, and the BERT NLP encoder is utilized for textual feature extraction. The decoder model with attention is incorporated for report generation. The various general domain and domain-specific pre-trained checkpoints are evaluated, and the best checkpoints are chosen for warm starting the encoder-decoder of a CXR report generator. These warm starts help generate a diagnostically accurate report that can be used in a clinical setting. From the literature, it is observed that there is a significant need for improving performance and the quality of the reports generated. The summary of the literature is shown in [Table 2.4](#).

Table 2.4: Summary of Literature Survey.

Author & year	Methodology	Future Remarks	BLEU1	BLEU2	BLEU3	BLEU4	Dataset	# of images	#of reports
<a href="#">Yuan <i>et al.</i> (2019)</a>	Multiview CNN encoder and concept enriched hierarchical LSTM is used.	Low BLEU4 score	0.529	0.372	0.315	0.255	CheXpert	2,24,316	65,240
<a href="#">Sai <i>et al.</i> (2021)</a>	CNN: Encoder with transfer Learning Model (ChexNet) for image feature extraction; Hierarchical LSTMs and Co-Attention mechanism for report generation.	The BLEU1 and BLEU2 score is very low indicating the mismatch in the predicted report.	0.213	0.258	0.325	0.381	Open-I	7470	3955
<a href="#">Nguyen <i>et al.</i> (2021)</a>	Multiview image encoder and text encoder is used for visual and textual feature extraction; transfer encoder module for report generation	Some false positives are observed in generated reports.	0.515	0.378	0.293	0.235	Open-I	7470	3955

Continued on next page

Table 2.4 – Continued from previous page

Author & year	Methodology	Future Remarks	BLEU1	BLEU2	BLEU3	BLEU4	Dataset	# of images	#of reports
<a href="#">Liu <i>et al.</i> (2021)</a>	Knowledge Graph Auto-Encoder (KGAE) model is proposed consisting a pre-constructed knowledge graph, a knowledge-driven encoder and a knowledge-driven decoder.	Produces the superior performance compared to supervised model. BLEU score can be further improved.	0.417	0.263	0.181	0.126	Open-I	7470	3955
<a href="#">Zhou <i>et al.</i> (2021)</a>	The DenseNet201 and BioSentVec models with semantic attention for image and text feature extraction. The LSTM model is incorporated for report generation.	The final model achieves consistent improvements over all the evaluation metrics. The Bleu scores can be improved.	0.536	0.392	0.314	0.339	Open-I	7470	3955

Continued on next page

Table 2.4 – Continued from previous page

Author & year	Methodology	Future Remarks	BLEU1	BLEU2	BLEU3	BLEU4	Dataset	# of images	#of reports
<a href="#">Sirshar et al. (2022)</a>	The CNN-based feature extraction technique is used as an encoder, followed by an RNN decoder that generates reports.	Low BLEU4 score	0.58	0.342	0.263	0.155	Open-I	7470	3955
<a href="#">Nicolson et al. (2022)</a>	The DenseNet pretrained on imageNet weights and the BERT NLP is leveraged as encoder for visual and textual feature extraction. The decoder model with attention is incorporated for report generation.	Low BLEU4 score	0.4777	0.308	0.2274	0.1773	Open-I	7470	3955

### 2.3.3 Multimodal Medical Image Analysis

Medical image fusion is the process of combining information from multiple medical images into a single composite image, which can provide a more complete and accurate understanding of a patient's condition. In the case of acute infarct prediction from MRI sequences, multimodal medical image fusion can be used to integrate information from multiple MRI modalities (e.g., T1-weighted, T2-weighted, diffusion-weighted, and perfusion-weighted) to improve the accuracy of infarct prediction. Acute brain infarct is a prevalent cause of fatality and ailment globally, resulting in over 5.5 million deaths annually [Ovbiagele and Nguyen-Huynh \(2011\)](#). It is indicated by the abrupt appearance of clinical signs caused by focal or global brain dysfunction. These symptoms may persist for more than 24 hours or result in death, and there are no other identifiable factors other than the issues related to vascular origin. The stroke can be categorized as either an ischemic infarct or a hemorrhagic infarct. The occurrence of acute ischemic stroke is closely linked to the time elapsed since the stroke, which must not exceed 4.5 hours. Thrombolytic therapy is a diagnostic procedure to break up or dissolve the blood clots that should be initiated within 4.5 hours after the stroke ([Zhang \*et al.\*, 2021](#)).

[Lee \*et al.\* \(2020\)](#) proposed a ML technique to detect acute ischemic stroke within 4.5 hours from the DWI and T2-Flair MRI sequences of 355 patients collected from the South Korean medical centre. The image processing techniques were applied to infarct segmentation and image registration. For stroke onset time classification, three ML techniques, including SVM, RF, and LR, were utilized. The authors concluded that ML algorithms utilizing MRI scans were viable, and they exhibited even greater sensitivity than human interpretations in detecting cases within the time frame for acute thrombolysis. It was seen that the specificity achieved by the proposed framework is comparatively low, indicating increased false-positive cases, which may lead to haemorrhage upon thrombolysis, making it impractical for clinical practice. [Zhu \*et al.\* \(2021\)](#) proposed an automatic approach using a deep learning strategy to classify and identify the time of stroke onset from the DWI and T2-Flair MRI sequences gathered from the two different stroke centres in China. The sample of 268 de-identified patients is collected and classified into negative ( $\leq 4.5$  hours) and positive ( $>4.5$  hours). The MRI sequences were passed through atrous convolution in parallel to obtain the fine features. The segmentation of stroke ROI was performed using the efficientNet-B0-based U-Net model. Finally, the stroke onset time is classified through voting using five ML techniques. It was observed that the model presented by the au-

thors showcased significantly lower classification performance, indicating a higher misclassification rate of the onset time since stroke.

Vesdapunt and Covavisaruch (2018) presented semi-automated segmentation techniques using the Otsu method, Hill Climbing, Growent, and Fuzzy C-Means for stroke lesion segmentation from the 13 DWI sequences of six patients collected from a private hospital. The segmented lesions are compared with Flair MRI for testing as a gold standard. Zhang *et al.* (2018a) proposed a fully automatic and computationally efficient approach for the segmentation of brain stroke from DWI employing a 3D fully convolutional DenseNet. The proposed method includes data preprocessing, feature extraction, and segmentation steps from DWI MRI sequences collected from private and public dataset. Zhao *et al.* (2021) proposed a CNN-based method, the multi-feature map fusion network, for segmenting lesions resulting from acute ischemic stroke. This technique seeks to enhance segmentation precision by integrating various feature maps using CNNs. The DWI and ADC MRI sequences of 582 subjects were collected from Tianjin Huanhu Hospital, China. The proposed CNN-based multi-feature map fusion network offers a more accurate segmentation of acute brain infarct lesions than other standard techniques. Further investigation of the presented technique on more extensive and diverse datasets is needed to assess its generalizability and effectiveness in real-world clinical applications. In their study, Yu *et al.* (2021) introduced an attention-based CNN for predicting the affected tissue in cases of acute brain infarct. They discovered that DL techniques with fine-tuning were more effective than traditional thresholding approaches in predicting acute infarct tissue. However, the deep learning models may only generalize well to patients within the dataset used for training.

Fang *et al.* (2022) analyzed various ML techniques like deep forest (DF), SVM, RF and DL techniques like a Residual Neural Network (ResNet), CNN, and Long Short-Term Memory Network (LSTM) for predicting Ischemic stroke from the 16,403 structured medical data collected from the International Stroke Trial (IST) database. The authors concluded that DL techniques did not outperform ML models during the prediction task. The experiment was conducted on structured medical data, making it challenging to infer the lesion's exact location. Bridge *et al.* (2022) proposed an ML algorithm that evaluates the likelihood of infarct in each voxel present in 6,657 DWI and ADC MRI sequences collected from the Massachusetts General Hospital, USA. The probability above a given threshold point classifies the MRI sequence as positive. The authors concluded that the DWI and ADC images jointly enhance the ML models' performance compared to

a single MRI sequence as it provide complementary features. The MRI sequences were manually annotated before being input into the ML model, making it a labour-intensive task. The major limitation of the presented ML model is its capacity to apply to new data samples with different demographics, geography, and technical parameters, including scanner model and manufacturer.

A 3D-CNN was presented by [Zeng \*et al.\* \(2022\)](#) for assessing the extent of neurological damage caused by Ischemic stroke. To extract the feature map by utilizing spatial features from the 851 DWI images collected from Xiangtan Central Hospital (China), a CNN model with 17 layers, including convolution, max pooling, and fully connected layers, was suggested. The model achieved an AUC of 0.846 for DWI images with the size 256 X 256 X 64 volumetric pixels and obtained an AUC of 0.895 for DWI images with the size 128 X 128 X 32 volumetric pixels. The major limitation is the model's generalization ability, as the image parameters applied were not uniform. [Nazari-Farsani \*et al.\* \(2023\)](#) utilized a deep CNN with an attention-gated (AG) technique to predict the location and size of final infarct in patients suffering from an acute stroke from DWI and ADC MRI sequences. The model showcased the potential to make diagnostic decisions using MRI sequences with brain stroke. However, the threshold for infarct probability used to generate the probability map may not be optimal for all patients. Further studies are needed to validate the model's performance across different patient populations and imaging protocols and to investigate the impact of the model on clinical decision-making.

The above-reviewed literature focuses on the use of ML and DL algorithms to improve the accuracy of stroke diagnosis, lesion segmentation, classification and prediction from magnetic resonance imaging (MRI) sequences. Several studies have shown promising results in identifying the onset time of acute ischemic stroke within the 4.5-hour time window using ML techniques. The segmentation of stroke lesions from MRI sequences was also improved through semi-automated and fully automated approaches using various ML and DL models. However, the significant limitations of these models include low classification accuracy, increased false positive cases, labour-intensive annotation tasks, and limited generalizability to new data samples with different demographics, geography, and technical parameters. Overall, the literature suggests that ML and DL techniques have great potential to improve stroke diagnosis, lesion segmentation, acute infarct classification and prediction from MRI sequences. However, further investigation and improvement are needed to ensure their practicality and effectiveness in real-world clinical settings. The literature discussed above is summarized in [Table 2.5](#).

Table 2.5: Summary of Findings from Literature Review

Author & year	Method	Task	MRI Imaging Sequences				Text data	Dataset	Remarks
			DWI	T2-Flair	ADC	SWI			
<a href="#">Vesdapunt and Covavisaruch (2018)</a>	Otsu method, Hill Climbing, Growent and Fuzzy C-Means strategy is applied	Segmentation	✓	✓			Private Dataset (6 Patients)	Semi-automated process used making it difficult to apply to real-time applications.	
<a href="#">Zhang et al. (2018a)</a>	3D fully convolutional DenseNet	Segmentation	✓				Private Dataset (242 Patients) Public Dataset-ISLES2015 (28 Cases)	It can be enhanced by integrating additional imaging modalities or clinical information.	

Continued on next page



Table 2.5 – Continued from previous page

Author & year	Method	Task	MRI Imaging Sequences				Text data	Dataset	Remarks
			DWI	T2-Flair	ADC	SWI			
<a href="#">Lee et al. (2020)</a>	Image processing techniques followed by three ML techniques: LR, SVM, and RF were utilized for classification.	Classification	✓	✓			Private Dataset (355 Patients)	Extracting a set of handcrafted features is a time-intensive process. The specificity achieved is comparatively low.	
<a href="#">Zhu et al. (2021)</a>	Atrous convolution, efficientNetB0-based U-Net model and max voting from five ML techniques.	Classification and Segmentation	✓	✓			Private Dataset (268 Patients)	Low classification performance	
<a href="#">Zhao et al. (2021)</a>	CNN-based method, the multi-feature map fusion network, was introduced.	Segmentation	✓		✓		Private Dataset (582 Patients)	Generalizability of the model is restricted.	

Continued on next page

Table 2.5 – Continued from previous page

Author & year	Method	Task	MRI Imaging Sequences				Text data	Dataset	Remarks
			DWI	T2-Flair	ADC	SWI			
Yu <i>et al.</i> (2021)	An attention-gated CNN was presented.	Prediction	✓		✓		Private Dataset (237 Patients)	The DL models may not generalize well to patients outside of the cohort used for training.	
Fang <i>et al.</i> (2022)	ML techniques (SVM, RF and DF) and DL Techniques (CNN, LSTM and ResNet).	Prediction					✓	IST Dataset (16,403 patients)	The structured medical dataset used didn't show the lesion's position.
Bridge <i>et al.</i> (2022)	An ML algorithm was proposed.	Classification and Segmentation	✓		✓			Private Dataset (6,657 Patients)	Generalizability of the model is limited.

Continued on next page

Table 2.5 – Continued from previous page

Author & year	Method	Task	MRI Imaging Sequences				Text data	Dataset	Remarks
			DWI	T2-Flair	ADC	SWI			
<a href="#">Zeng et al. (2022)</a>	A CNN model with 17 layers was presented.	Prediction	✓				Private Dataset (851 Patients)	The model achieved an AUC of 0.846 and 0.895 for varied sized voxels. Generalisation ability was limited.	
<a href="#">Nazari-Farsani et al. (2023)</a>	DCNN with an AG mechanism was utilized.	Prediction	✓		✓		Private Dataset (445 Patients)	The threshold for infarction probability used to generate the probability map may not be optimal for all patients.	

## 2.4 Outcome of Literature Review

The comprehensive survey presented in the above section uncovered various shortcomings in the field of CRSs utilizing unstructured text, images, and multimodal data. Many developing countries lack a standard method for recording patient information, resulting in unstructured and unprocessed data. This creates difficulties in implementing text-based CRS, which depends on organized and processed data to provide healthcare providers with valuable insights and recommendations. The challenge of implementing text-based CRS in developing countries due to unstructured patient data has been highlighted in several studies ([Polnaszek \*et al.\*, 2016](#)). A study by [Khan and Banerji \(2014\)](#) notes that most hospitals in developing countries still use paper-based records or electronic systems with unstructured or semi-structured data. This makes it difficult to extract useful information from patient data and develop recommendation systems that can effectively support healthcare providers. Designing effective techniques to process unstructured clinical data and integrate it into CRS workflows can make a significant contribution to improving healthcare outcomes in developing countries with similar healthcare ecosystems, benefiting doctors and hospital personnel by providing valuable insights and recommendations.

Structured clinical data are widely used in decision support and recommendation systems due to their standardized format and ease of use. The structured data is typically organized into tables, fields, and codes that can be easily accessed and queried by computer programs, enabling efficient data processing and analysis ([Tayefi \*et al.\*, 2021](#)). Using structured clinical data has the benefit of offering a thorough perspective on a patient's medical background, encompassing aspects such as diagnosis, treatment, medications, and laboratory findings. This data is valuable to healthcare providers as they can use it to make well-informed decisions about patient care, such as determining the appropriate medication, ordering relevant tests, and formulating treatment strategies. Most of the existing works rely on only structured clinical data to provide prognostic outcomes. Nonetheless, solely depending on structured clinical data has its drawbacks. Structured data merely captures a fraction of the complete clinical information that is accessible, and a significant amount of critical clinical data is present in unstructured clinical notes. Such notes are written in natural language and could contain crucial clinical details, including patient history, symptoms, and treatment plans. Advanced NLP techniques are necessary to extract beneficial information from unstructured clinical notes by analyzing the text and retrieving pertinent information.

While there have been significant improvements in NLP over recent years, it remains a challenging task due to the intricate and variable nature of clinical notes. On the other hand, NLP and DL models have become more popular because of their ability to achieve high performance when trained with large amounts of data. However, medical datasets are currently limited to healthcare institutions, are domain-specific, and are small, making it difficult to train DL Models. Therefore, there is a requirement to develop NLP techniques and DL models capable of handling low data conditions for classifying and predicting diseases in radiology reports. In the last few years, several investigations (Trivedi *et al.* (2017); Shin *et al.* (2017); Dahl *et al.* (2021); Nakamura *et al.* (2021); Bayrak *et al.* (2022)) have suggested DL techniques to offer CRS for the prediction of diseases and vulnerabilities. Although some of these methods (Nakamura *et al.* (2021); Bayrak *et al.* (2022)) exhibited encouraging outcomes, there is undoubtedly potential for enhancement in the areas of data representations of the patient, neural network architectures, and explainability. Therefore, this thesis thoroughly investigates the avenue of developing more effective systems that provide precise predictive analytics tailored to individual patients, utilizing unstructured healthcare data in textual format.

The study conducted on CRS for pulmonary chest X-ray images indicated a significant requirement for improving the DL techniques employed. The existing deep learning strategies lack the ability to capture the more discriminative features of the receptive field. Medical CXRs come with varied-sized abnormalities; thus, most of the current techniques do not focus on multi-scale features (Rajpurkar *et al.* (2017); Rajkomar *et al.* (2016); Candemir *et al.* (2018); Zech *et al.* (2018); Lee *et al.* (2021); Li *et al.* (2022)). Most existing models utilize increased network parameters to detect pulmonary abnormalities, making them computationally expensive and challenging to use in mobile-vision applications. Moreover, the absence of transparency in current models poses a challenge for radiologists to comprehend the reasoning behind the model's diagnosis, thereby restricting its applicability in clinical settings. In this context, our primary goal is to build a better and more effective DL technique that is both lightweight and explainable for extracting multi-scale discriminative features by capturing a larger receptive field for predicting pulmonary abnormalities from chest X-rays. By developing an explainable and lightweight DL technique, we aim to mitigate the challenges posed by existing models and facilitate its adoption in clinical settings.

In medical imaging, unstructured medical images often contain intricate and irregular RoIs, making it difficult to isolate and extract meaningful information.

However, DL models, especially deep CNNs, have proven to be highly successful in automatically learning hierarchical and discriminative features directly from raw pixel data, eliminating the need for explicit feature engineering. This has led to remarkable achievements in medical image analysis. DL models offer numerous significant advantages over traditional machine learning models in this domain:

- *Automated Feature Learning:* DL models excel at automatically learning relevant features directly from raw medical images. This is particularly advantageous when dealing with complex and irregular RoIs in medical images. In contrast, traditional ML models often require manual, hand-crafted feature engineering, which can be time-consuming and may not fully capture the intricacies of the RoIs. DL models provide an end-to-end system that streamlines the feature extraction process (Iqbal *et al.*, 2023).
- *Hierarchical Representations:* With their deep architectures, DL models can learn hierarchical representations of the data, capturing both low-level and high-level features. This ability to learn multiple levels of abstraction is crucial when analyzing intricate RoIs and complex patterns within medical images. Traditional ML models may struggle to capture such hierarchies effectively (Torres-Velázquez *et al.*, 2021).
- *Scalability:* DL models are highly scalable and can handle large medical datasets effectively. Training DL models on powerful GPUs or TPUs allows for more extensive exploration of medical data, leading to potentially better predictive performance. Traditional ML models may face limitations when dealing with large-scale datasets (Vinod *et al.*, 2020).
- *State-of-the-art Performance:* In recent years, DL models, especially CNNs, have achieved state-of-the-art results in various medical imaging tasks. These tasks include disease prediction, lesion detection, and image segmentation. DL's ability to learn intricate patterns from vast amounts of data contributes to its superior performance compared to many traditional ML approaches (Rana and Bhushan, 2022).

In summary, DL models provide feature learning automation, hierarchical representation learning, scalability for large datasets, and top-notch performance, making them highly effective for medical image analysis tasks.

A thorough review of different prior studies is carried out with respect to medical data that involves multiple modes of imaging and non-imaging information.

The study shows that when used in conjunction, the clinical text and medical images improve the accuracy of classification due to their complementary features (Aydin *et al.* (2019a); Lopez *et al.* (2020); Huang *et al.* (2020); Alfarghaly *et al.* (2021b)). Many frameworks for predicting diseases from different sources of medical data use either early, late, or joint fusion techniques to handle the multimodal information. Early and joint fusion methods typically rely on simple concatenation to merge imaging and non-imaging features or two distinct imaging modalities. In contrast, late fusion approaches employ various meta-classifiers, majority voting, or averaging to combine the output from individual models focused on single modalities. However, neither concatenation nor late fusion methods account for the inter-modal dynamics between the heterogeneous features. Upon conducting a comprehensive review of the literature, it was discovered that only a few studies had explored the fusion of unstructured radiology clinical free-text reports with pixel features extracted from radiology images to provide clinicians with valuable and meaningful prognosis information. Our literature review indicates that the availability and accessibility of multimodal medical datasets are limited. Consequently, most existing models have only been tested on these small datasets, making it challenging to assess their generalizability to larger and more diverse datasets. To address these limitations, developing a deep learning framework that leverages an efficient multimodal feature fusion technique capable of combining low-level imaging and high-level textual features is necessary to generate accurate prognostic outcomes.

The accurate interpretation and summary of medical images, particularly those generated by radiology tests such as X-rays, CT scans, and MRIs, are crucial components of clinical diagnosis. Generating a diagnosis report from radiology images is an essential step in clinical diagnosis, and highly skilled radiologists are required for this task. However, the process can be time-consuming and mentally taxing for radiologists, especially in busy and overcrowded situations. To alleviate this burden and speed up the diagnosis process, there is a growing need for automated and reliable diagnostic report generation frameworks. Existing deep learning techniques for report generation have shown promise, but there is still room for improvement, particularly in terms of the BLEU score ((Jing *et al.*, 2018); (Yang *et al.*, 2022); Yuan *et al.* (2019); Nguyen *et al.* (2021); Sai *et al.* (2021); Nicolson *et al.* (2022)). One promising approach is to develop a cross-modal framework that combines textual and imaging features to assist radiologists in automatically generating accurate reports from medical images. By using such frameworks, healthcare providers can reduce the workload on radiologists, speed up the diag-

nosis process, and provide better patient care. Additionally, these frameworks can ensure consistency and accuracy in diagnosis reports, minimizing the risk of errors and improving the overall quality of patient care.

CRS for Multimodal image interpretation plays a crucial role in enhancing the accuracy of diagnostic outcomes by capturing complementary features from different diagnostic image modes or sequences. Radiologists rely on various MRI sequences, such as DWI, T2-flair, SWI, and ADC, to predict the presence of an Acute Infarct. In this medical condition, an accurate and timely diagnosis is crucial. Previous studies have employed ML models that extract handcrafted features from MRI data and feed them into the models for classification. However, this method is time-consuming and may not capture all the relevant information from the data. Therefore, it's crucial to develop improved approaches that incorporate all available MRI sequences to accurately predict Acute Infarct. Furthermore, we observed that radiologists use all four MRI sequences to identify acute infarct. However, to the best of our knowledge, no existing work has leveraged all the MRI sequences in predicting the disease ([Vesdapunt and Covavisaruch \(2018\)](#); [Zhang et al. \(2018b\)](#); [Lee et al. \(2020\)](#); [Zhu et al. \(2021\)](#); [Fang et al. \(2022\)](#)). This presents an opportunity to develop a more comprehensive and accurate approach that incorporates all four MRI sequences. Finally, we noted that the performance of most existing approaches is low, indicating the need for improved prediction accuracy. To achieve this objective, it is essential to design and develop a multimodal DL framework capable of extracting and fusing imaging features from all four MRI sequences. By doing so, we can improve prediction accuracy and surpass existing methods.

## 2.5 Summary

This chapter provides an extensive literature review of AI-based CRS, covering a range of techniques and frameworks. Our exploration of AI-based CRS focuses on three distinct categories of medical data: unstructured text data, unstructured image data, and multimodal data. Within each category, we investigate various approaches for constructing an effective AI-based CRS. Moreover, we examine three specific tasks associated with analyzing multimodal medical data, namely multimodal analysis using both images and text, cross-modal diagnostic text generation from images, and multimodal image analysis. After conducting an in-depth examination of the available literature, it became apparent that there is a clear requirement for diverse CRS implementations that can extract hidden knowledge



from a wide range of medical data, particularly for tasks like disease prediction. The literature review indicates that the current state-of-the-art CRS implementations primarily focus on a specific type of data or a specific disease, limiting their ability to handle diverse medical data and tasks. This finding underscores the need for more varied and adaptable CRS solutions that can extract insights from different healthcare data types, such as unstructured text, images, and multimodal data, to support diverse clinical applications. Through a comprehensive review of the literature, we have identified key areas for advancing the development of NLP and DL frameworks in clinical decision support systems. These areas include the extraction of textual features from unstructured diagnostic reports to aid in disease prediction, as well as the extraction of discriminative features from diagnostic chest X-ray images using DL techniques. Another critical area for development is the integration of image and text features for disease prediction, which requires the development of novel fusion strategies. Additionally, we have identified the need for automated report-generation methods for diagnostic images and the development of multimodal image analysis frameworks for disease prediction.

After considering these observations, the problem scope and statement addressed in this thesis were precisely defined and discussed in Chapter 3. In order to overcome the identified limitations and the research gaps, several methodologies have been proposed, which are briefly outlined in Chapter 3 and comprehensively discussed in subsequent chapters of this thesis.



# Chapter 3

## Problem Description

### 3.1 Background

The previous chapter provided an in-depth review of different approaches aimed at developing a viable AI-driven CRS framework capable of analyzing medical data across multiple modes. It also consolidated significant challenges and considerations that need to be addressed for constructing an improved and effective CRS. This chapter elucidates the specific research gaps that were identified and presents them as a problem statement. Additionally, it delineates the scope of the proposed research work featured in this thesis and provides a concise overview of the methodologies developed to address the formally defined issues.

### 3.2 Research Gaps

Chapter 2 of this thesis presents a detailed literature review of the current research on AI-based CRS that incorporates multiple types of medical data. It provides a thorough review of existing literature and highlights the outcomes and findings of the review. The following section briefly outlines the research gaps and limitations that have been identified:

- Clinical decision support or recommendation systems primarily rely on numerical or structured data, which is readily available in standard formats. However, there has been limited exploration of the potential of unstructured free-text data, such as radiology reports, which are a good source of valuable information for disease prediction.
- The performance of CRS that use unstructured medical data in text format for predicting diseases has demonstrated inadequate performance. Moreover,

the medical cohorts available in current practice are small, domain-specific, and limited to medical institutes. Henceforth, handling the unstructured radiology report in a low-data situation and predicting the disease outcome pose a challenge. This problem can be solved by developing advanced NLP techniques and DL models that can handle low data conditions for disease classification and prediction in radiology reports. Furthermore, improving the design and functionality of the existing CRS framework for unstructured medical text has the potential to significantly enhance their predictability.

- In medical image analysis, a significant obstacle lies in predicting diseases from unstructured medical images. The main difficulty is extracting the crucial and distinguishing features from the irregular RoI within the receptive field, which are essential for accurate diagnostic outcomes. These RoIs are often intricate and hard to isolate, making it challenging to gather the necessary information for precise disease prediction. Thus, the quest for effective, lightweight, and interpretable DL methods to accurately capture these discriminative RoI features remains an ongoing challenge. While traditional ML models have advantages in terms of being lightweight, the exceptional representation power of DL models, especially deep CNNs, makes them the preferred choice for medical image analysis.
- To enhance the accuracy of disease prediction from multimodal diagnostic data, such as chest x-rays and radiology reports, it is imperative to develop an effective multimodal fusion model that can seamlessly combine multiple radiology images and reports. However, there is a semantic gap between the low-level visual data and the high-level textual information, which poses a challenge. Thus, it is essential to create an effective multimodal system that can bridge this gap and enable an accurate disease prediction framework.
- Upon thorough review, it has been observed that cross-modal retrieval in report generation faces a significant challenge in terms of generating an overall description with low BLEU scores. The unstructured nature of radiology reports presents a severe obstacle to the performance of cross-modal retrieval in report generation. Addressing this challenge requires the development of more sophisticated techniques that can effectively process and extract relevant information from unstructured medical reports.
- Multimodal image fusion is a challenging task that requires capturing complementary features from multiple image sequences and representing them

in a common space for disease prediction. However, current techniques face considerable difficulties in accomplishing this effectively. To address this issue, an advanced method for fusing multiple imaging features is required to provide better diagnostic decisions.

- The current state of the art in CRS models for multimodal medical data exhibits poor generalizability, and their effectiveness on a broader range of datasets is uncertain. To address this issue, it is crucial to investigate the ability of these models to perform on augmented or synthetic data.

### 3.3 Scope of the Work

Upon reviewing the existing approaches for constructing CRSs, it becomes clear that the most effective ones utilize AI techniques such as ML or DL. The success of these recommendation systems is highly dependent on the accurate modeling and representation of multimodal medical data, as this is the foundation for prognostic models. To bridge the gaps and achieve the objective, the research presented in this thesis contributes significantly in five key areas, which are outlined below:

1. Designing and building an effective AI-based CRS for predicting diseases from the unstructured diagnostic text data.
2. To design and develop a lightweight and explainable AI-based CRS for predicting diseases from unstructured diagnostic images.
3. Design and development of an improved data fusion technique for combining complementary imaging and textual features for disease prediction from multimodal diagnostic data.
4. Designing a cross-modal framework for automatically generating the medical reports for a given diagnostic image.
5. To design and develop an enhanced multimodal image fusion network to integrate multiple diagnostic images for disease prediction.

#### 3.3.1 Problem Statement

Drawing from the limitations highlighted in the current literature regarding AI-based CRS that incorporate multiple types of medical data, the problem statement addressed by this thesis is formulated as follows:

*“Design and develop a framework for an effective Clinical Recommendation System on unstructured and multimodal data for the health sector through Machine Learning and Deep Learning Techniques”*

### 3.3.2 Research Objectives

After identifying gaps in previous research and formulating a problem statement, this thesis has established three main research objectives and four corresponding sub-objectives. The research work presented in the thesis is focused on addressing these objectives and sub-objectives:

1. To design and develop a technique for extracting medical concepts and predicting clinical outcomes from unstructured medical text data.
2. To design and build an effective mechanism for predicting clinical outcomes from unstructured medical images.
3. To design and develop a framework to integrate the complementary features from the Clinical multimodal data and predict disease outcomes.
  - To design and develop an improved data fusion technique to combine different features retrieved from Multimodal medical images and text for Predicting Diseases.
  - To design and develop an effective clinical recommendation system to perform cross-modal retrieval of diagnostic reports from medical images
  - To design and develop an effective multimodal image fusion network for integrating heterogeneous images for diagnostic disease prediction.
  - To assess and analyze the performance of the multimodal prediction compared to the single-modal prediction.

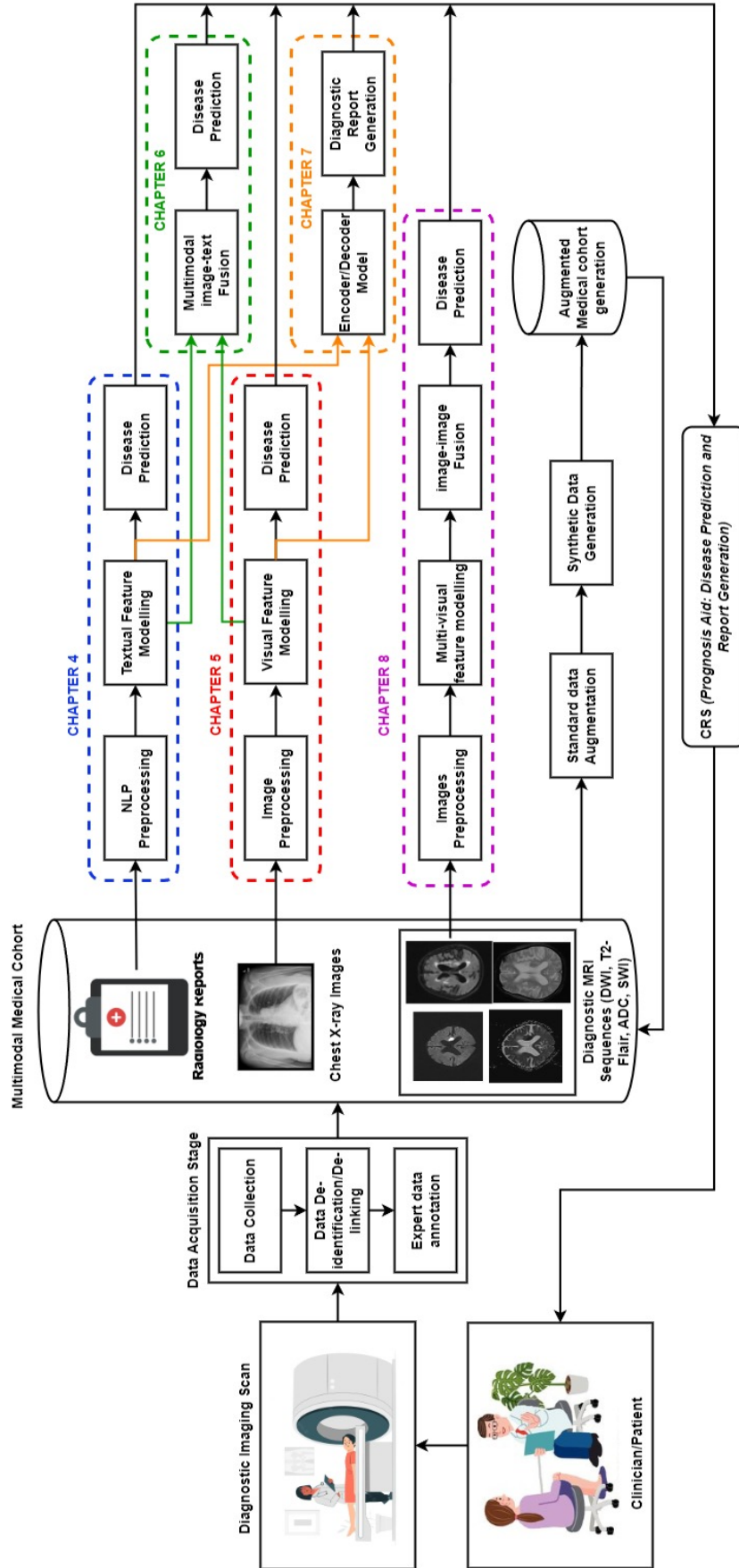


Figure 3.1: Systematic Overview of the Intelligent Framework for an Effective Clinical Recommendation System to Predict Diseases from Multimodal Medical Data.

## 3.4 Brief Overview of Proposed CRS Framework

We have presented a brief overview of our study that focuses on an AI-based CRS framework using multimodal medical data in Figure 3.1. The thesis presents a thorough examination of the contributions made in each chapter towards accomplishing our research objectives. This chapter gives a concise summary of the research work presented throughout the thesis, highlighting the key points and the overall scope of the study.

To build the necessary medical cohort, we obtained multimodal diagnostic data, such as CXRs with their reports and MRI sequences, from a private medical hospital. The collected data was subjected to de-identification/delinking procedures to maintain patient privacy and was then annotated by professional radiologists. Section 3.4.1 provides an explanation of the proposed methodology for extracting disease outcomes from unimodal, unstructured free-text reports using NLP pre-processing modules, text modelling, and classification stages. As discussed in Section 3.4.2, the methodology includes pre-processing unimodal diagnostic CXR images by resizing and removing noise before passing them through a visual feature modelling and disease classification/prediction module. Section 3.4.3 elaborates on the approach of predicting diagnostic abnormalities by creating a multimodal representation that fuses textual and imaging features obtained from reports and CXRs. In Section 3.4.4, we introduce an encoder-decoder-based deep learning module that enables cross-modal report generation from the input CXR. In Section 3.4.5, we demonstrate the prediction of prognostic outcomes through multimodal image fusion of four different MRI sequences. To increase the cohort size and generate high-resolution diagnostic images, we employ several data augmentation and synthesis techniques. The diagnostic outcomes of all the subnetworks mentioned above will aid high-level medical applications that are a primary requirement in the workflow of clinical recommendation systems.

### 3.4.1 Unimodal Medical Text Embedding Subnetwork for Disease Prediction from Unstructured Free-text Reports

Although healthcare providers often record clinical notes in an unstructured format, this type of documentation can provide a wealth of information about a patient’s medical condition, including their symptoms, disease progression, and treatment plans. However, despite its value, this information is frequently under-



utilized when predicting disease-specific conditions. Towards this end, there is a need to design and develop AI-based CRS that can directly analyze unstructured free-text data for disease prediction. Figure 3.2 presents the top-level process flow of the AI-based CRS with unstructured Free-text Reports for disease prediction. In Chapter 4, a practical text modelling approach is proposed that combines a Knowledge Base (KB) with deep learning to accurately mine text and predict pulmonary abnormalities from unstructured radiology free-text reports, even in a low-data condition. This method can potentially enhance the accuracy and efficiency of disease prediction, leading to better-informed patient decisions.

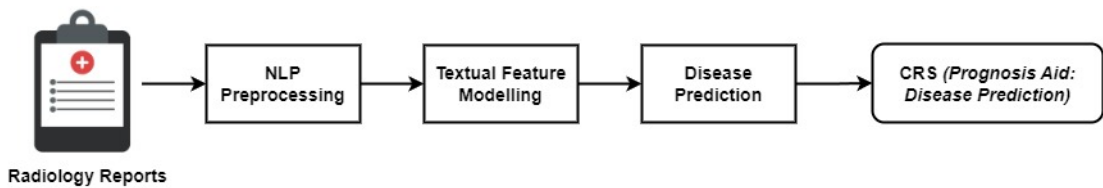


Figure 3.2: CRS with Unstructured Free-text Reports for Disease Prediction

As shown in Figure 3.2, the NLP preprocessing is applied to the unstructured radiology free-text reports to clean the data and make it ready to ingest into the NLP and DL models. We adopt Glove word embeddings with the Knowledge Base trained on a large corpus for effective text modelling. Further, we incorporate Convolutional Neural Network-based Discriminative Dimensionality Reduction to obtain the most discriminative feature vector. Finally, a fully connected Deep Neural Network is leveraged as the prediction model to detect the diseases.

### 3.4.2 Unimodal Medical Visual Encoding Subnetwork for Disease Prediction from Medical Images

As previously mentioned, this thesis focuses on using Chest X-ray medical images as a tool for predicting pulmonary disease. The texture and shape of the tissues in the diagnostic images are essential aspects of prognosis. Therefore, in the latest studies, the vast set of images with a larger resolution is paired with deep learning techniques to enhance the performance of the disease diagnosis in chest radiographs. Most of the attempts do not consider the computation overhead and lose the spatial details in an effort to capture the larger receptive field for obtaining the discriminative features from high-resolution chest X-rays. To address this, we propose a lightweight and explainable Unimodal Medical Visual Encoding Subnetwork (UM-VES) for predicting diseases from medical images. The architecture

diagram presented in Figure 3.3 illustrates the high-level design of an AI-powered CRS that utilizes unstructured diagnostic images for disease prediction. Specifically, the system pre-processes chest X-ray images by resizing them and removing noise before feeding them into the visual feature modelling and disease prediction module. Chapter 5 presents a detailed overview of the UM-VES framework. The UM-VES consists of the following four main subnetworks: (1) Multi-Scale Dilation Layer (MSDL), which includes multiple and stacked dilation convolution channels that consider the larger receptive field and capture the variable sizes of pulmonary diseases by obtaining more discriminative spatial features from the input chest X-rays; (2) Depthwise Separable Convolution Neural Network (DS-CNN) is used to learn imaging features by adjusting lesser parameters compared to the conventional CNN, making the overall network lightweight and computationally inexpensive, making it suitable for mobile vision tasks; (3) a fully connected Deep Neural Network module is used for predicting abnormalities from the chest X-rays; and (4) Gradient-weighted Class Activation Mapping (Grad-CAM) technique is employed to check the decision models' transparency and understand their ability to arrive at a decision.

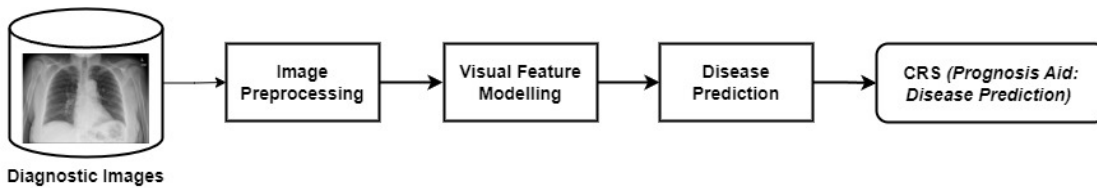


Figure 3.3: CRS with Unstructured Diagnostic Images for Disease Prediction

### 3.4.3 Deep Medical Multimodal Fusion Networks for Disease Prediction from Medical Text and Image Data

An AI-powered CRS that analyzes radiology images has the potential to assist medical practitioners in the diagnosis and prediction of pulmonary diseases. In order to improve our comprehension of pulmonary abnormalities, it is crucial to create a model that can efficiently make use of both radiology images and diagnostic reports. This will allow us to obtain a more comprehensive understanding of these conditions. While expert reports provide valuable information, they are often limited by the subjective interpretation of the individual writing them. Diagnostic scan data, on the other hand, offers a more objective view of the patient's condition. The merging of these two information sources will result in a more complete and precise depiction of pulmonary diseases, potentially resulting in improved

diagnosis, treatment, and, ultimately, better patient outcomes. In this direction, we proposed two Multimodal Medical Tensor Fusion Networks (i.e., the Compact Bilinear Pooling-based Medical Multimodal Fusion Network (CBP-MMFN) and the Deep Hadamard Product-based Medical Multimodal Fusion Network (DHP-MMFN) for predicting abnormalities from a radiology CXR and its associated reports. Chapter 6 provides a detailed overview of the multimodal fusion network, which is the foundation of an AI-powered CRS for disease prediction from multimodal clinical data. The process flow of this system is illustrated at a high level in Figure 3.4, showing how it leverages unstructured clinical data, including both text and images, to generate disease outcome predictions. To achieve this, the system uses a multimodal image-text fusion network to combine the extracted textual and imaging features, resulting in more accurate and comprehensive disease predictions.

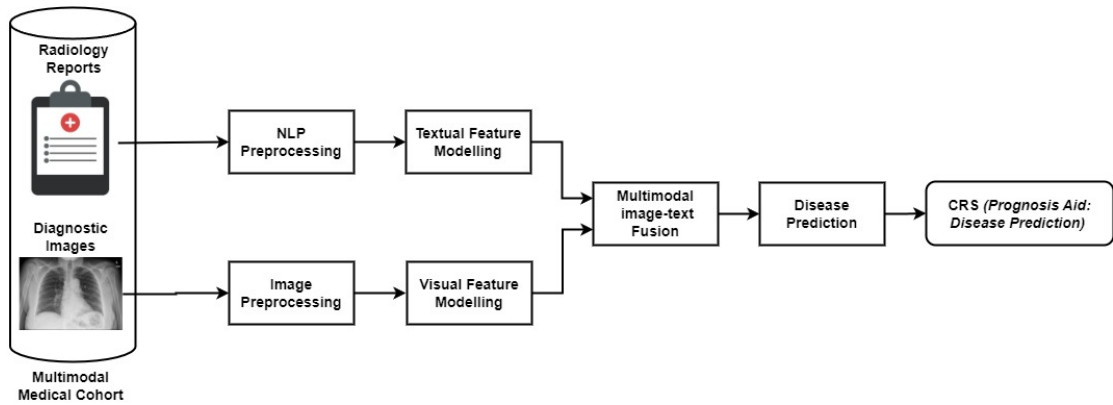


Figure 3.4: CRS with Multimodal Unstructured Clinical Data for Disease Prediction

### 3.4.4 Cross-modal Deep Learning Framework for Diagnostic Report Generation from Medical Images

Generating diagnostic reports for various medical conditions displayed in different types of medical scans, like X-rays, CT scans, and MRIs, is frequently required in the field of medical imaging. Typically, human experts examine the images and create detailed reports as part of this task. Nonetheless, this process can be prone to errors and is often a time-consuming endeavour. To address this challenge, we propose a deep encoder-decoder model to generate the reports from the CXR automatically. The high-level process flow of the AI-powered CRS for generating reports from CXR is shown in Figure 3.5. The proposed cross-modal framework

for automatic report generation takes CXRs, including frontal and lateral images, as input and produces the radiology report as an output. The detailed overview of the cross-modal framework is explained in Chapter 7.

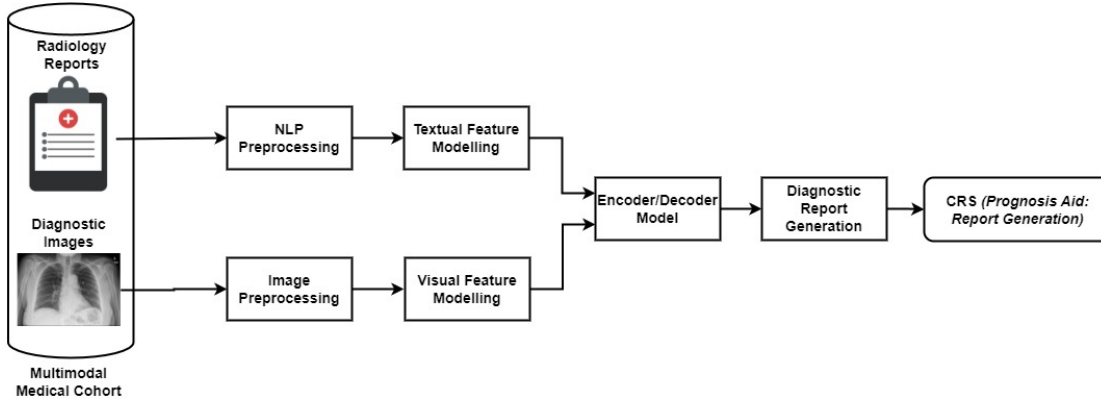


Figure 3.5: CRS with Multimodal Unstructured Clinical Data for Diagnostic Report Generation

### 3.4.5 Multimodal Image Fusion Network for Disease Prediction from Medical Images

As previously mentioned, we have chosen to utilize a multimodal fusion approach for predicting Acute Infarct, which involves combining data from multiple MRI sequences. Specifically, the selected sequences for fusion are DWI, T2-flair, SWI, and ADC. By merging information from these different sequences, we aim to enhance the accuracy and completeness of our prediction. This approach allows us to leverage the unique strengths of each sequence to gain a more comprehensive understanding of the patient's condition, which can assist in effective diagnosis and treatment planning. The diagram depicted in Figure 3.6 illustrates the high-level process flow of an AI-based CRS that predicts diseases from multimodal medical images. The system employs a multimodal feature modelling and fusion strategy to extract distinct visual features and merge them into a unified space. These fused features are then inputted into the disease prediction module, which generates a prognostic outcome. Towards this end, we propose two stacked multi-channel convolutional neural networks for predicting disease from multiple and individual MRI sequences. A detailed explanation of the various DL approaches undertaken to predict multimodal imaging features is presented in Chapter 8.

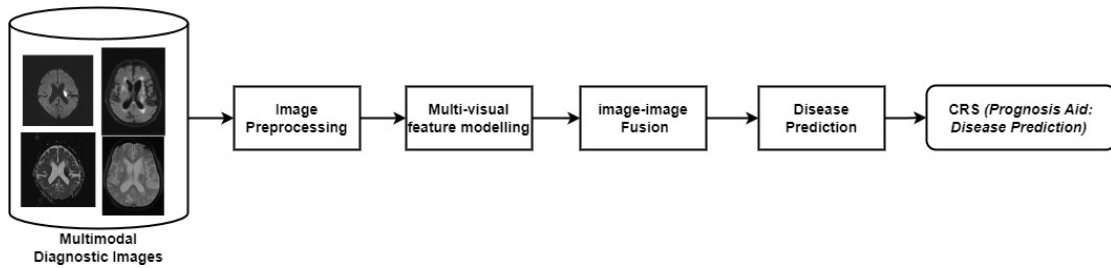


Figure 3.6: CRS with Multimodal Diagnostic Images for Disease Prediction

### 3.5 Research Contributions

This thesis aims to introduce a robust framework that can be employed to create a Computerized CRS that utilizes AI to predict diseases from various types of unstructured diagnostic data. This framework incorporates techniques from multiple modalities, such as diagnostic imaging and radiology reports, and leverages AI algorithms to analyze and extract meaningful information from this data. By utilizing this framework, healthcare providers can make well-informed decisions regarding the care and treatment of their patients, ultimately improving overall health outcomes of the individuals. Outlined below are the major contributions of this research work:

- Design of a practical text modelling approach that combines knowledge base and deep learning techniques, this study aims to extract latent features from radiology free-text reports, resulting in improved accuracy and efficiency of disease prediction.
- Development of an effective Multi-scale Deep Learning Network that can accurately detect abnormal chest conditions from radiographic images while providing clear and interpretable results.
- Designing of two effective Medical Multimodal Tensor Fusion Networks using Compact Bilinear Pooling and Deep Hadamard Product for predicting pulmonary abnormalities from the radiology CXR and text reports.
- Developing an advanced deep learning framework that combines textual and imaging features to create accurate and dependable radiology reports from CXR data, using encoder-decoder strategies.
- Development of an effective multimodal image fusion network that can extract multi-scale features from different MRI sequences to accurately predict the presence of Acute Infarct.

## 3.6 Summary

In this chapter, we highlight the research scope, followed by the identified shortcomings, which serve as the basis for defining the research problem. Three main research objectives and four corresponding sub-objectives are derived. The strategies for achieving these objectives are briefly introduced and will be elaborated on in the further chapters of this thesis.

## **PART II**

# **AI-based CRS for Unimodal Unstructured Medical Data Analysis**





## Chapter 4

# Unimodal Medical Text Embedding Subnetwork (UM-TES) for Disease Prediction from Unstructured Free-Text Reports

### 4.1 Introduction

Pulmonary diseases are the major cause of death worldwide due to tobacco smoking, air pollution, inhaling unwanted particles, radon gas, chemicals, etc. Pulmonary diseases involve various respiratory and lung disorders like pneumonia, chronic bronchitis, pleural effusion, and pulmonary fibrosis. The chances of risk involved in pulmonary diseases are high, and there is a need for timely treatment. The radiologist interprets the radiology imaging examination like a chest X-ray to diagnose the conditions affecting the lungs. The radiologists analyze the Chest X-ray and record their observations in the descriptive reports to validate the prognosis (Yasaka and Abe, 2018). These radiology reports contain rich information pertaining to patient demographics, disease findings, and conclusive remarks on the abnormalities. This essential information retrieved from the clinical narratives can be leveraged to improve the efficacy of clinical assessment, treatment, and research.

Usually, the radiologists manually categorize the clinical notes into normal (i.e., no diseases) and abnormal (i.e., pulmonary diseases). Manual classification of the radiology reports is labour intensive, time-consuming and prone to human error. Rapid and accurate identification of information contained in radiology narratives will minimize workloads, assist radiologists in decision-making, and prioritize patients with emergency care. Automating the task will benefit the radiologists with

less experience in predicting the abnormality when there is an increased number of patients at higher risk (Nakamura *et al.*, 2021). There has been significant growth in the usage of ML and DL strategies for automating disease prediction tasks from EHRs (Yasaka and Abe, 2018). The unstructured nature of the radiology free-text reports with complex vocabularies makes it difficult for the ML and DL models to extract features from the raw text. NLP plays a major role in extracting structured information from clinical text (Pons *et al.*, 2016).

The automated prediction of pulmonary abnormalities from the unstructured diagnostic notes can be further integrated into the existing medical diagnostic workflow to improve the health information system in the following ways:

- One potential approach would be to integrate the model as a decision support tool for radiologists. In this scenario, the model could be used to automatically flag radiology reports that are likely to contain pulmonary abnormalities, which could then be prioritized for review by the radiologist. The model could also be used to suggest possible diagnoses based on the irregularities detected in the report, which could help the radiologist arrive quickly at a more accurate diagnosis.
- Another approach would be to integrate the model into the EHR system used by the hospital or clinic. In this scenario, the model could be used to automatically populate the patient’s EHR with relevant diagnostic information based on the findings in the diagnostic notes. This could improve the accuracy and completeness of the patient’s medical record, improving the quality of care delivered to the patient.
- Finally, the model could also be used as a screening tool for large-scale population health studies. By analyzing radiology reports from a large cohort of patients, the model could identify patients at high risk of developing the pulmonary disease, which could inform targeted interventions and preventative care strategies.

The unstructured text in clinical practice is scarce in number, as most of the radiology reports are restricted to private institutions or domain-specific. The deep learning models produce better results when the cohort size is large. There is a need for a deep learning framework that accurately classifies the abnormalities from the radiology reports when the cohort size is small. In this research, we address the challenge of low data situation in the medical radiology report dataset by adapting the medical knowledge base at the text feature extraction stage. So,

we propose an NLP-based DL model that incorporates a knowledge base to convert the unstructured text into meaningful word embeddings. The framework proposed in this study employs the GloVe embedding technique in combination with a knowledge base to represent the features of the words. This approach is followed by the use of a deep neural network to predict the presence of pulmonary abnormalities. The proposed framework improves the accuracy and efficiency of NLP techniques by utilizing GloVe embeddings, which capture the semantic connections between words. Incorporating a knowledge base further enhances the model's predictive power by providing additional contextual information. The deep neural network component of the framework utilizes the feature representations generated by the GloVe embeddings and knowledge base to learn patterns and relationships between words, thereby enhancing the accuracy of the predictions.

#### 4.1.1 Problem Statement

Pulmonary diseases are a primary global health concern, causing significant morbidity and mortality. Timely detection and diagnosis of these diseases can enhance patient outcomes and increase their chances of survival. However, diagnosing pulmonary diseases using radiology imaging, such as chest X-rays, is often challenging due to the unstructured nature of radiology reports, which contain a large amount of text. Radiologists analyze the imaging results and prepare reports with findings and conclusive remarks, which can be difficult for healthcare providers to extract relevant information from, causing delays in treatment. Additionally, manually classifying radiology reports as normal or abnormal is a time-consuming and error-prone process. Also, the unstructured text in clinical practice is scarce in number, as most of the radiology reports are restricted to private institutions or are domain-specific.

The problem statement is defined as follows:

*Considering the challenges posed by lengthy diagnostic reports, clinical terminology, and a low data condition, devise and implement strategies for effective automated pulmonary disease prediction using unstructured free-text reports.*

In this chapter, we address this problem by developing the Unimodal Medical Text Embedding Subnetwork (UM-TES), which incorporates a knowledge base to learn the semantic pattern from the unstructured clinical notes. The main contribution of this research work is as follows:

- We propose an effective Unimodal Medical Text Embedding Subnetwork (UM-TES) to predict diseases from radiology reports. We point out the significance of incorporating Knowledge-based Medical Text modelling with Discriminative Dimensionality Reduction using CNN (DDR-CNN) and Deep Neural Networks for classifying and predicting diseases in Radiology Reports in a low-data environment.
- We carry out comprehensive analysis on two medical datasets (i.e., the publicly available Indiana University dataset (Demner-Fushman *et al.*, 2016) and a real-time corpus collected from a private medical institute) to illustrate the competency and rationality of the proposed DL Framework. A thorough investigation is conducted, and we benchmark the evaluation results of the Knowledge-based DL framework against the standard ML Techniques. To the best of our knowledge, this is the first work leveraging radiology data collected from an Indian Private Hospital.
- We examine the effect of dimensionality reduction using the DDR-CNN and the effectiveness of incorporating knowledge bases in enhancing the performance of the overall framework.
- We have conducted a systematic and comprehensive performance evaluation to check the efficacy of our proposed model with standard NLP techniques.

## 4.2 Methodology

The proposed UM-TES framework for Disease prediction from the unstructured radiology reports is shown in Figure 4.1. As an overview, the radiology findings are pre-processed to obtain the essential latent medical concepts. The word embeddings are learnt from the medical words by applying customized Clinical Knowledge-based Text modelling. The dense word embeddings obtained are mapped to the medical words from the findings in the Embedding Layer. Most Discriminative features are extracted by reducing the dimension using the Convolutional Neural Network. Finally, the flattened discriminative features are fed to the Deep Neural Network for prediction of the disease outcome.

### 4.2.1 Basic Pre-processing

To clean the data and make it ready to be ingested into the models, we pass the radiology report findings from both cohorts through the sequence of text pre-

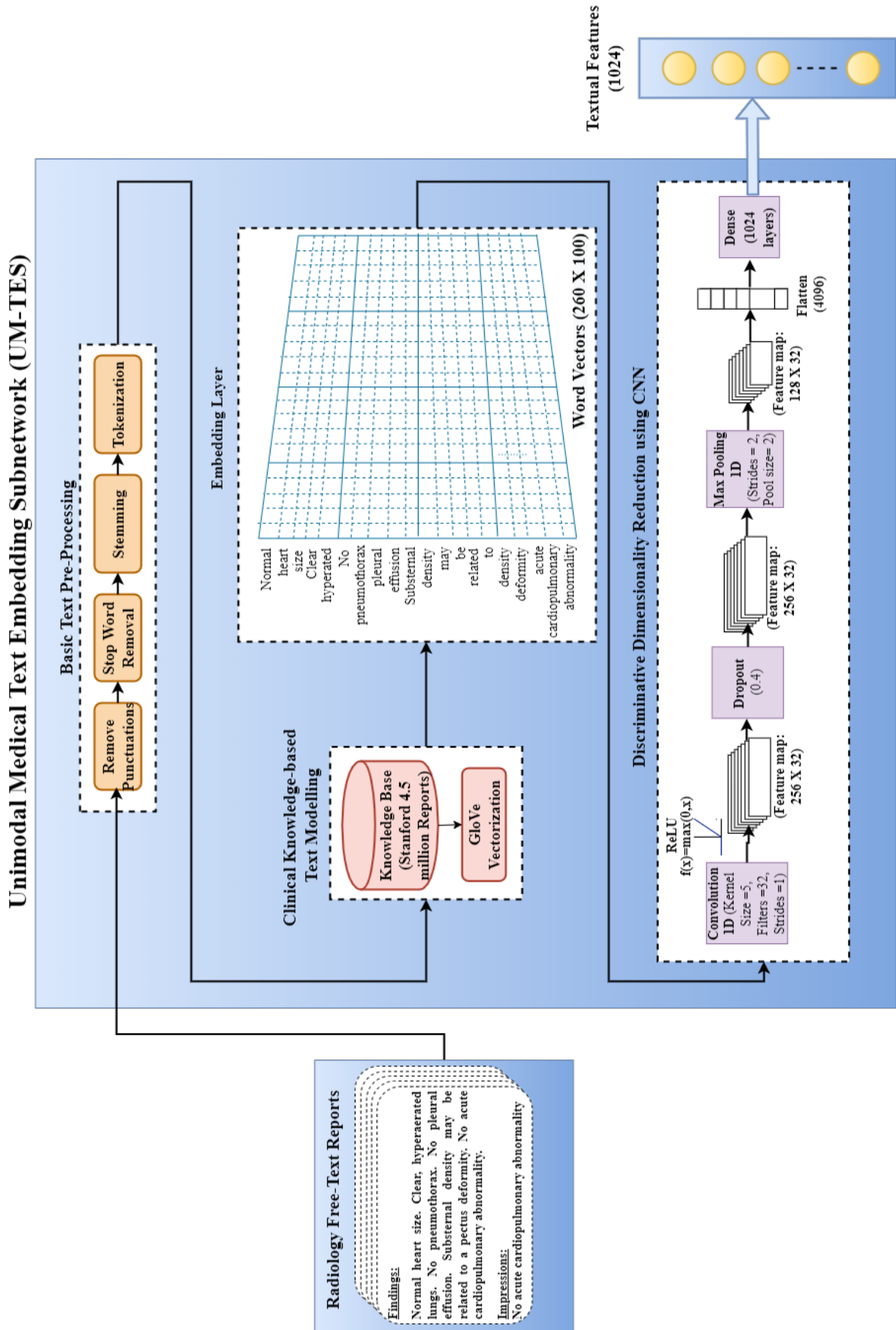


Figure 4.1: Proposed Unimodal Medical Text Embedding Subnetwork (UM-TES) for Text Feature Extraction

processing stages.

- *Punctuation Removal:* The presence of punctuation marks, such as periods, commas, colons, and semicolons, does not add any meaningful information to the text and can actually hinder the processing of textual data. To ensure that diagnostic conclusions are accurately processed when analyzing radiology reports, it is important to remove punctuation. For instance, the existence of unnecessary punctuation within the report may perplex the model in identifying the accurate disease condition.

For example, consider the following sentence from a pulmonary disease report: “Patient presented with shortness of breath, cough, and wheezing”. After the removal of punctuation, the sentence becomes: “Patient presented with shortness of breath, cough and wheezing”. Notice how the commas have been removed from the sentence, making it easier to process for machine learning models. Punctuation removal also helps to reduce the dimensionality of the text data, making it more manageable for downstream tasks like feature extraction and classification.

- *Stop word Removal:* Stopwords are commonly used words that do not provide much context to the text data and can be safely removed from the dataset without losing important information. In clinical notes like pulmonary disease reports, stop word removal can be an essential step to remove unnecessary words that do not contribute to the diagnosis or prediction of the disease. As an illustration, when analyzing a pulmonary disease report, certain common stop words like “a”, “an”, “the”, “and”, “in”, “on” and “of” may frequently occur. These words do not provide any specific information regarding the illness and may even introduce noise into the data. By eliminating these words, the data can become more concise and meaningful.

Consider the following sentence from a pulmonary disease report: “The patient has a history of smoking, which may have contributed to the development of chronic obstructive pulmonary disease (COPD)”. In this sentence, the stop words are “the”, “has”, “a”, “of”, “which”, “may”, “have”, “to”, “the”, “and”, “of”. Removing these words would result in the sentence: “patient history smoking contributed development chronic obstructive pulmonary disease (COPD)”. The remaining words convey the same meaning as the original sentence and can be more efficiently processed by NLP models.

- *Stemming*: The stemming process aids in the removal of suffixes to preserve only the base words. This method aids in normalizing the textual data by consolidating various forms of words, like plurals and verb tenses, into a solitary representation. In the context of radiology reports, stemming can help identify related words and reduce the feature space of the data, which in turn can enhance the proposed DL model’s performance.

To illustrate, in a diagnostic note on predicting pulmonary diseases, the term “effusions” could be simplified to “effusion”, making it possible to detect all occurrences of effusion in the text, regardless of whether the word is in its singular or plural form. Likewise, the term “fibrotic” could be reduced to “fibros”, making it possible to detect all instances of fibrosis in the text, whether the word is used as an adjective or a noun.

- *Tokenization*: The act of dividing a document or sentence into smaller units called tokens is known as tokenization. When dealing with radiology reports, tokens could signify either single words or phrases, and they can be beneficial in detecting patterns and connections within the textual information.

Let’s take an instance of a sentence from a radiology report: “A nodule in the right lung was detected on the chest X-ray”. After tokenization, the sentence may be broken down into the following tokens: [“A”, “nodule”, “in”, “right”, “lung”, “was”, “detected”, “on”, “the”, “chest”, “X-ray”]. In this sentence, each token corresponds to an individual word. Tokenization plays a crucial role in recognizing significant patterns and connections within the textual information, as well as extracting useful information that can be utilized in tasks like disease prognosis or image retrieval.

This research follows a series of steps to preprocess the text. The initial step involves removing punctuation, followed by the removal of frequently used words, known as stopwords, from the corpus. Next, a standardization process called stemming is applied to transform the words into their root or base form. Finally, the raw text is broken down into smaller units known as tokens.

## 4.2.2 Clinical Knowledge-based Text Modelling

We have proposed a Clinical knowledge-based Text modelling strategy for generating the discriminative word embeddings by jointly learning from the clinical reports and the clinical knowledge base. The main aim of incorporating the clinical knowledge base is to understand the infrequent clinical words, which pose a

significant challenge in capturing the semantics of the word. We have used an improved GloVe Model with the knowledge base to generate the most discriminative word embeddings.

- Global Vectors (GloVe) for Clinical Word Representations:** To derive the word vectors from the clinical radiology reports, the GloVe (Pennington *et al.*, 2014) model is based on the matrix factorization technique, leveraging statistical information obtained from the global word-word co-occurrence matrix. In particular, when provided with clinical radiology reports  $CR$ , the GloVe initializes by constructing the Word-word co-occurrence matrix  $M^c$ . In the co-occurrence matrix, the *target clinical word* is represented by the row, and the *context clinical word* is represented by the column. The word co-occurrence matrix obtained for a given clinical radiology reports with  $m$  words is  $m \times m$ . The  $M_{ij}^c$  indicates the tabular entries representing the total number of occurrences of the context clinical word  $\tilde{w}_j$  occurring in the target clinical word  $w_i$  in the radiology report cohort  $CR$ . Given the dimensionality  $d$  (i.e., a hyper-parameter set by the user), the GloVe embedding technique initializes by learning word vectors or embeddings  $cv_i, \tilde{c}v_i \in \mathbb{R}^d$  by conceding a given clinical context word  $cw_i$  as a target clinical word  $w_i$  or the context clinical word  $\tilde{w}_i$ , respectively.

The Co-occurrence probability of generating context clinical word  $\tilde{w}_j$  provided the target clinical word  $w_i$  is given by  $P(j | i) = M_{ij}^c / M_i^c$ . Here,  $M_i^c$  is the total occurrence of target clinical word  $w_i$  in the radiology report corpus. This probability provides how frequently the target clinical word is seen in the context clinical word in a large radiology report corpus. For example,  $cv_i^T \cdot \tilde{c}v_i$  provides the similarity between two clinical words  $w_i$  and  $\tilde{w}_j$ . The clinical word vectors for the radiology report cohort can be learnt from the global word-word co-occurrence matrix given by,

$$cv_i^T \cdot \tilde{c}v_j = \log P(j | i) = \log(M_{ij}^c) - \log(M_i^c) \quad (4.1)$$

Likewise,  $M_{ij}^c = M_{ji}^c$ ,

$$cv_j^T \cdot \tilde{c}v_i = \log P(i | j) = \log(M_{ij}^c) - \log(M_j^c) \quad (4.2)$$

Since,  $cv_i^T \cdot \tilde{c}v_j = cv_j^T \cdot \tilde{c}v_i$ , we can add the Eq. (4.1) and Eq. (4.2) to get,



$$2cv_i^T \cdot \tilde{c}v_j = 2\log(M_{ij}^c) - \log(M_i^c) - \log(M_j^c) \quad (4.3)$$

$$cv_i^T \cdot \tilde{c}v_j = \log(M_{ij}^c) - \frac{1}{2}\log(M_i^c) - \frac{1}{2}\log(M_j^c) \quad (4.4)$$

The right hand side of the Eq. (4.4) is the counts learnt from the radiology report cohort and the left hand side represents the learnable parameters. We contemplate  $\log(M_i^c)$  and  $\log(M_j^c)$  has the biases specific to the clinical words  $cv_i$  and  $\tilde{c}v_j$  to be learnt. To restore symmetry in Eq. (4.4), we add the scalar biased real-valued terms  $b_i$  and  $b_j$ , affiliated with clinical words  $cv_i$  and  $\tilde{c}v_j$  respectively.

$$cv_i^T \cdot \tilde{c}v_j = \log(M_{ij}^c) - b_i - b_j \quad (4.5)$$

$$cv_i^T \cdot \tilde{c}v_j + b_i + b_j = \log(M_{ij}^c) \quad (4.6)$$

The following Eq. (4.7) represents formulation of the optimization problem,

$$\min_{cv_i, \tilde{c}v_j, b_i, b_j} \sum_{i,j=1}^V (cv_i^T \cdot \tilde{c}v_j + b_i + \tilde{b}_j - \log(M_{ij}^c))^2 \quad (4.7)$$

The weighting function  $f(M_{ij}^c)$  is introduced to the eq. (4.7) to allocate lower weights for the frequently occurring words to avoid the objective function from skewing because of over-emphasizing the most commonly occurring word pairs in the medical cohort and is given by Eq. (4.9). Hence, the weighted least square errors are minimized to obtain the objective function of the GloVe Embedding model:

$$J_{CR} = \sum_{i,j=1}^V f(M_{ij}^c) (R^{CR} + b_i + \tilde{b}_j - \log(M_{ij}^c))^2 \quad (4.8)$$

Where,  $R^{CR} = cv_i^T \cdot \tilde{c}v_j$  is the scalar value produced by the inner dot product between the transpose of target word vector  $cv_i$  and the context word vector  $\tilde{c}v_j$ .

$$f(m) = \begin{cases} \left(\frac{m}{m_{max}}\right)^\alpha & \text{if } m < m_{max} \\ 1 & \text{otherwise} \end{cases} \quad (4.9)$$

The value of  $\alpha$  is set to 3/4 and  $m_{max}$  to 100 as the efficiency of the model

depends on the cutoff (Pennington *et al.*, 2014). As shown in the Eq. (4.8) obtained, the objective function defined from the clinical radiology reports aims to learn the co-occurrence between the two clinical words  $cw_i$  and  $c\tilde{w}_j$  by reducing the squared difference between the inner dot product and the logarithm of the co-occurrences between the context and target clinical words in the matrix  $M^c$ .

- **Incorporating Clinical Knowledge Base to GloVe Model:**

The GloVe is an unsupervised learning algorithm to obtain word embeddings or vector representation of words learnt from the given corpus and does not utilize any existing Knowledge Bases. As a result, GloVe cannot acquire reliable embeddings from infrequent words, posing a severe challenge in capturing semantics, which is essential in clinical text mining. Similarly, the clinical datasets available today are small, domain-specific and limited to private medical organizations, facing a significant challenge to learn word embeddings from the limited vocabulary. We utilize learned word embeddings as a knowledge base to address this problem while text modelling from the clinical radiology report corpus. Given a Clinical Knowledge Base  $CKB$ , we generate the objective function  $J_{CKB}$  to derive the semantic connection  $R(cw_i, c\tilde{w}_j)$  seen between the respective target clinical words  $cw_i$  and the context clinical words  $c\tilde{w}_j$ . The word embeddings trained on 4.5 million Stanford reports (Zhang *et al.* (2018b)) is considered in this experiment as a concrete case for a Clinical knowledge base. There are no specific guidelines to use any particular knowledge base; nevertheless, any knowledge base that creates a semantic relationship between the clinical words can be utilized as a Clinical Knowledge Base.

$$R^{CKB} = kv_i^T \cdot k\tilde{v}_j \quad (4.10)$$

Here,  $R^{CKB}$  represents the scalar value derived from the inner dot product between the knowledge vectors of target word  $kv_i$  and context word  $k\tilde{v}_j$  from Clinical Knowledge Base  $CKB$ , which corresponds to the clinical words  $cw_i$  and  $c\tilde{w}_j$  in the Clinical Radiology Reports  $CR$ . The objective function obtained after incorporating clinical knowledge base is as follows:

$$J_{CKB} = \sum_{i,j=1}^V f(M_{ij}^c) (R^{CKB} + b_i + \tilde{b}_j - \log(M_{ij}^c))^2 \quad (4.11)$$

Here,  $b_i$  and  $b_j$  represents the scalar biased real-valued terms associated with the clinical words  $cw_i$  and  $c\tilde{w}_i$  respectively. As shown in Eq. (4.11), the objective function obtained from the clinical knowledge base learns the co-occurrence between the two clinical words  $cw_i$  and  $c\tilde{w}_j$  by reducing the squared difference between the inner dot product of the knowledge vectors obtained from the clinical knowledge base and the logarithm of the co-occurrences between the clinical words  $cw_i$  and  $c\tilde{w}_j$  in the co-occurrence matrix  $M^c$ . Eq. (4.11) represents the weighted least squares loss, which measures the disparities between predicted word embeddings and the actual embeddings from the medical cohort. By minimizing this loss, we aim to make precise predictions that closely align with word co-occurrence probabilities in a large text corpus. It is essential to highlight that GloVe word embeddings are acquired through unsupervised learning, capturing statistical word relationships in the corpus. Consequently, our prediction task involves a regression-type problem, seeking to predict continuous values representing the word embeddings. In this context, the weighted least squares loss is a suitable choice for optimizing the model and refining the embeddings for our specific medical dataset.

The optimization algorithm employed for generating word embeddings was Stochastic Gradient Descent (SGD) (Ruder, 2016). We opted for SGD over other popular optimizers like Adam (Kingma and Ba, 2014) due to specific reasons that cater to our unique dataset and computational limitations. Firstly, our word embedding generation process is centered around a relatively small dataset, and our computational resources are limited. In consideration of these constraints, we found that SGD's simplicity and lower memory requirements make it a more practical and feasible choice. While Adam might demonstrate faster convergence on larger datasets, its advantages might not be as prominent when dealing with our specific dataset size. Secondly, we noticed that our word embedding generation process is relatively tolerant of noise. As a result, we found SGD's inherent noisiness to be advantageous in effectively navigating away from local minima. Since the word embedding optimization problem can involve non-convex loss surfaces, SGD's stochastic updates enable exploration of diverse regions in the loss landscape, holding the potential to produce improved word representations. Moreover, to ensure a comprehensive evaluation, we conducted multiple empirical comparisons between SGD and Adam on our specific task. Intriguingly, despite Adam's reputation for being less sensitive to hyperpa-

rameters, we noticed that SGD with a carefully tuned learning rate schedule outperformed Adam in terms of word embedding quality and downstream task performance. Notably, we employed a Greedy Search approach to select the hyperparameters, thereby ensuring a robust evaluation process.

### 4.2.3 Embedding Layer

The pre-trained word embeddings acquired from the 4.5 million Stanford Radiology reports are loaded into the embedding layer as a knowledge base, and word vectors for each clinical word from the clinical radiology reports are learnt. The radiology report findings are available in varied sizes; hence, they are padded to have the same length. Firstly, the unique words from the clinical corpus are extracted by tokenizing the report findings. Next, the hash of each word is generated by integer encoding for every clinical word. In other words, a lookup table or embedding matrix is created for each clinical word with the unique integer number as an index. Finally, the embedding layer emulates as a hidden layer by converting every integer input into one-hot vectors and performing matrix multiplication with an embedding weight matrix. The corresponding embedding matrix of the clinical words for a given radiology report corpus is produced as an output of the embedding layer. In our experiment, the pre-trained clinical knowledge base is ingested as a dictionary of 260 padded clinical words to generate the embedding weight matrix of word vectors with the output dimension of 100. The Algorithm 1 presents clinical knowledge-based Text modelling.

### 4.2.4 Discriminative Dimensionality Reduction using Convolution Neural Network (DDR-CNN)

Dimensionality Reduction is a crucial component of any Machine learning or Deep Learning task. The main objective of any Discriminative Dimensionality Reduction technique is to convert the embeddings from high-dimensional to low-dimensional space such that the essential discriminative information is preserved (Roweis and Saul, 2000). We employ Convolution Neural Network (CNN) to preserve the most discriminative features by learning the high-dimensional embeddings in lower-dimensional space and reducing storage and computing costs. The architecture proposed for DDR-CNN is a refined version of CNN's architecture (Collobert *et al.*, 2011). CNN was selected over other popular deep learning models, such as deep autoencoders (Siddique *et al.*, 2019), for dimensionality reduction

due to the following reasons:

- *Local Pattern Recognition:* CNNs are specifically designed to capture local patterns and spatial dependencies in data. In the context of word embeddings, CNNs can effectively identify and preserve important local contextual information, which is crucial for maintaining the semantic relationships between words.
- *Efficient Computation:* CNNs are computationally efficient compared to deep autoencoders, especially when dealing with high-dimensional data such as word embeddings. CNNs utilize shared weights and local receptive fields, making them more scalable for processing large datasets.
- *Translation Invariance:* CNNs possess a degree of translation invariance, meaning they can recognize patterns regardless of their location in the input. This property is beneficial for word embeddings, as it allows the model to capture similarities between words irrespective of their positions within sentences or documents.
- *Interpretability:* CNNs hierarchical structure makes it easier to interpret the learned features at different layers. This interpretability is valuable in understanding which linguistic features or attributes the model is focusing on during dimensionality reduction.
- *Non-linear Transformations:* CNNs inherently perform non-linear transformations, enabling them to capture complex relationships within the word embeddings. This flexibility is essential for accurately representing the nuances and complexities of natural language.
- *Domain Adaptation:* CNNs have shown good generalization abilities across domains, which can be advantageous when dealing with word embeddings obtained from different sources or datasets.

While (deep) autoencoders are indeed a popular choice for dimensionality reduction tasks, CNNs specific architectural characteristics and their ability to capture local patterns and semantic relationships in word embeddings make them a favorable choice in our study.

Let  $cx_i \in \mathbb{R}^d$  represent  $i$ -th clinical word in the sentence (i.e., findings from the radiology report) of  $d$ -dimensional word embedding. In our experiment, we have

considered 100 dimension word vectors for every clinical word (i.e.,  $d=100$ ). The clinical sentence of length  $k$  (i.e.,  $k=260$ ) with required padding is represented as:

$$cx_{1:k} = cx_1 \oplus cx_2 \oplus \dots \oplus cx_k \quad (4.12)$$

The  $\oplus$  in the above equation denotes the concatenation operator. The  $cx_{i:i+j}$  represents the concatenation of the clinical words  $cx_i, cx_{i+1}, \dots, cx_{i+j}$ . We generate the discriminative features by applying the convolution operation with the filter  $W \in \mathbb{R}^{md}$  to the window of  $m$  clinical words. In this experiment, we have utilized 32 filters with a window size of 5 (i.e.,  $m=5$ ). The new discriminative features  $cc_i = f(W \cdot cx_{i:i+m-1} + b)$  are obtained from the window of clinical words  $cx_{i:i+m-1}$ . Here,  $b \in \mathbb{R}$  represents the bias term and  $f$  indicates the Rectified Linear Unit (ReLU) (Agarap, 2018) activation function with kernel  $W$  applied on every available window of clinical words  $\{cx_{1:m}, cx_{2:m+1}, \dots, cx_{k-m+1:k}\}$  to produce the feature map,  $cc = [cc_1, cc_2, \dots, cc_{k-m+1}]$ , where  $cc \in \mathbb{R}^{k-m+1}$ .

Further, the dropout mechanism (Hinton *et al.*, 2012) is applied to regularize the network and address the overfitting problem. The dropout technique is used during the training process to prevent co-adaptation at the hidden layer by dropping out some number of layer outputs. Let  $z = [c\tilde{c}_1, c\tilde{c}_2, \dots, c\tilde{c}_n]$  denote the penultimate layer and  $n$  indicates the total number of kernels. The output of the forward propagation in the CNN is depicted by  $y = W \cdot z + b$ . The dropout mechanism is employed on the penultimate layer to obtain the following:

$$y = W \cdot (z \otimes \delta_i) + b, \quad (4.13)$$

Here,  $\otimes$  is the element-wise multiplication operation between the Bernoulli random variable or dropout rate represented by  $\delta_i \in \mathbb{R}^n$  with the feature maps in the penultimate layer. The  $\delta_i$  drops out the  $i^{th}$  neuron with the probability ( $p_i$ ) of becoming 1. In other words, using the dropout mechanism, we are dropping neurons and their connections with probability  $(1 - p)$ . Applying the dropout mechanism has regularized our convolution layer by providing maximum resilience against the overfitting problem. According to the outcome from the grid search approach (Bergstra and Bengio, 2012), the dropout probability of 0.4 is applied after the first hidden layer in our experiment. The feature map produced after passing through dropout layer is  $\tilde{y} = [\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_M]$ , where  $M = k - m + 1$ .

The feature map obtained is down-sampled (pooled) using the Max-pooling (Collobert *et al.*, 2011) strategy to decrease the dimensionality. The main objective is to learn the most discriminative features by computing the maximum or largest

value for every patch of the feature map. The max-pooling layer creates a lower resolution version of an input image and contains the essential structural elements. The pooled feature map containing summarized version of features  $y'$  is produced by applying strides  $s$  and pooling window size  $pw$ . Max-pooling layer operates on each available windows on the feature map  $\tilde{y}$  to obtain,

$$y' = \parallel_{i=1, i=i+s}^{L-pw+1} \max\{\tilde{y}_{i:i+pw-1}\}, \quad (4.14)$$

where  $y' \in \mathbb{R}^n$  and  $\parallel$  represents the customized concatenation operation. In our experiment, we have considered the stride  $s=2$  and pool window size,  $pw=2$  on the feature map of length,  $L=256$ . To begin with,  $i$  is initialized to 1 and incremented with the number of strides until it reaches  $L-pw+1$  to generate the discriminative feature map,  $y'_{1:128} = \max\{\tilde{y}_{1:2}\} \oplus \max\{\tilde{y}_{3:4}\} \oplus \dots \oplus \max\{\tilde{y}_{255:256}\}$ . The feature map obtained from the max-pooling layer is flattened and ingested into a dense layer to produce the features adjusting to the channel dimension.

#### 4.2.5 Network Structure of UM-TES:

In network architecture of UM-TES, the findings retrieved from the clinical radiology reports are preprocessed and are padded to make the input size 260. The proposed clinical knowledge-based text modelling is employed to generate the word embeddings of size 260x100. We use kernel size of 5, filter size of 32, strides of 1, pool size of 2, dropout of 0.4 and learning rate of 0.001 as hyperparameters for DDR-CNN. The discriminative clinical features of size 1024 are extracted by applying DDR-CNN on the word embeddings. We represent the final textual features obtained as  $M_t = \{t_1, t_2, t_3, \dots, t_{1024}\}$ .

**Algorithm 1:** Clinical Knowledge-based Text Modelling

**Input:** *Medical-Knowledge-Base*: Pre-Trained Word embeddings on 4.5 million Stanford Radiology Report, *Medical-Corpus*: Unstructured Radiology Reports

**Output:** A Text Model trained on Unstructured Radiology Report Corpus with Word Embedding Representation

```

1 initialization;
2 Function Clean( $S$ ):
3   Remove the Punctuation from  $S$ .
4   Remove Stopwords from  $S$ .
5   Stemming operation is applied on  $S$  to remove suffix.
6   return  $S$ ;
7 End Function
8 Function Tokenize(Medical-Corpus):
9   for each findings  $f_i \in$  Medical-Corpus do
10    |  $Cleaned-Findings \leftarrow Clean(f_i)$ ;
11    |  $tokens \leftarrow Cleaned-Findings$  are split into tokens;
12  end
13  return  $tokens$ ;
14 End Function
15 Function Convert-Word-to-Embeddings(Medical-Corpus):
    |
    | •  $Tokenized-Docs \leftarrow Tokenize(Medical-Corpus)$ 
    |
    | • Let  $v_1, v_2, \dots, v_n$  be the Unique medical words (i.e., Vocabulary) obtained
    |   from the  $Tokenized - Docs$ .
    |
    | • Generate one hot vector  $hv_{i:n}$  for each word in a Vocabulary  $v_{i:n}$ .
    |
    | • Let  $k$  be the input length of each  $Tokenized - Docs$  (Pad the documents if
    |   necessary).
    |
    | • Generate the knowledge-based word embeddings.
    |
    |   – The co-occurrence between the two medical words are learnt by the
    |     objective function defined in the Eq. (4.11) using Stochastic Gradient
    |     Descent by minimizing the  $kv_i, \tilde{kv}_j, b_i$  and  $\tilde{b}_j$  from the large medical
    |     corpus (refer Eq. (4.11) for term details).
    |
    |   – Load the Medical-Knowledge-Base as the Embedding Weight Matrix
    |     with  $d$ -dimension word knowledge vectors  $kv_{i:d}^1, kv_{i:d}^2, \dots, kv_{i:d}^k =$ 
    |      $kv_{i:d}^{j:k}$ , (where,  $i = 1, 2, \dots, d$ ) for all the medical words  $k$ .
    |
    | • The corresponding matrix is obtained by matrix multiplication
    |   between one hot vector of each word in a vocabulary and Embedding
    |   Weight Matrix through Embedding Layer with input size  $k$  and the
    |   output dimension  $d$  is,  $\tilde{kv}_{i:d}^{j:k} \leftarrow hv_{i:n} \times kv_{i:d}^{j:k}$ 
    |   return Word Vectors  $\tilde{kv}_{i:d}^{j:k}$  of size  $k \times d$ 
16 End Function

```



### 4.2.6 Fully connected Deep Neural Network (DNN) for Disease Prediction

Deep Neural Network (DNN) (Bengio (2009); Schmidhuber (2014)) is applied in the proposed framework for predicting whether the disease exists or not in the Radiology report. DNN is typically a multi-layer neural network, influenced by a biological neural network consisting of a collection of connected modules named neurons. DNNs comprise multiple such connected units between the input and output layers. The flattened discriminative features  $y' \in \mathbb{R}^n$  obtained from the CNN-DDR module are given as the input to a fully connected, four-layered Deep Neural Network for predicting whether disease exists or not in the radiology reports. The basic operation performed by DNN on the flattened discriminative features is forward propagation or inference, represented as,

$$y'_{i+1} = h(W_i \cdot y'_i + b_i), \quad (4.15)$$

Where  $W_{i=0,1,2}$  are the weight arrays and  $b_{i=0,1,2}$  are the bias of the DNN Layers. Here,  $h()$  is the non-linear function applied to every element of the feature vector. The non-linear function ReLU is applied at the penultimate layers, and the Sigmoid operation for the binary classification is applied at the output layers. For our experiment, based on a grid search approach Bergstra and Bengio (2012), dropout probability of 0.2 is applied after the first hidden layer to restrain the model from overfitting.

## 4.3 Comparison with State-of-the-art Text Modelling Strategies

We compare the proposed UM-TES model with the State-of-the-art text modelling strategies for predicting pulmonary diseases from radiology free-text reports, as shown in Figure 4.2. The findings section from the reports is extracted from the corpus and passed through the basic pre-processing stage to clean the data. The refined text is ingested into various NLP techniques, including the proposed framework to convert the text into meaningful word embeddings. The textual features retrieved are then given as input to a Fully connected Deep Neural Network (DNN) for pulmonary disease prediction. With the GloVe model, the following are the other state-of-the-art text modelling strategies considered for comparison:

1. **Bag of Words (BoW):** BoW (Sivic and Zisserman, 2009) is the primary

word embedding technique in which the sequence of words is converted into a bag of words. In the BoW strategy, the word occurrence count is calculated, disregarding the grammar. The vocabulary of unique words is generated from the radiology cohort, and the number of occurrences of each word is counted. The major drawback with the above technique is that the word count will provide information about the occurrences of the word in a cohort ignoring the context of the word.

2. **Term Frequency — Inverse Document Frequency (tf-idf)** tf-idf (Sammut and Webb, 2010) provides a statistical evaluation of how relevant a radiology word is to a cohort. The tf-idf comprises two steps: firstly, term frequency is calculated by counting the word occurrences in a radiology cohort. The inverse document frequency will diminish the weights provided on common words. tf-idf is based on BoW; subsequently, it doesn't capture the semantics of the medical word in a cohort.
3. **FastText:** FastText (Bojanowski *et al.*, 2016) is a type of word2vec model, where the n-gram (i.e., sequences of adjacent characters) of characters represents each radiology word. This will allow the model to learn the semantics of the shorter words and allows the model to understand the suffixes and prefixes of the radiology words. After the n-gram representation of the words, the skip-gram strategy is applied to learn the word embeddings. The major drawback of this method is the higher memory requirement as the model deals with the character of words.
4. **Continuous bag of words (CBOW):** CBOW (Wang *et al.*, 2017b) The CBOW is an unsupervised word2vec-based model which takes radiology context words as input and predicts the radiology target word. The lambda and softmax layers are utilised to learn the word embeddings by backpropagation strategy. We update weights in the embedding layer with each epoch using the gradient descent technique.
5. **Skip-gram:** Skip-gram (Mikolov *et al.*, 2013a) model predicts the radiology context word given the target word. The positive and negative input sample is created and fed as input to the model. These samples allow the model to learn the context and generate the semantic embeddings for each radiology word. The radiology target and context word pair given as input is merged to compute the dot product of the word embeddings. These embeddings are then passed through the sigmoid layer that provides either 0 or 1 as an

output. The obtained output is compared with the original label, and the loss is calculated by backpropagating for every epoch.

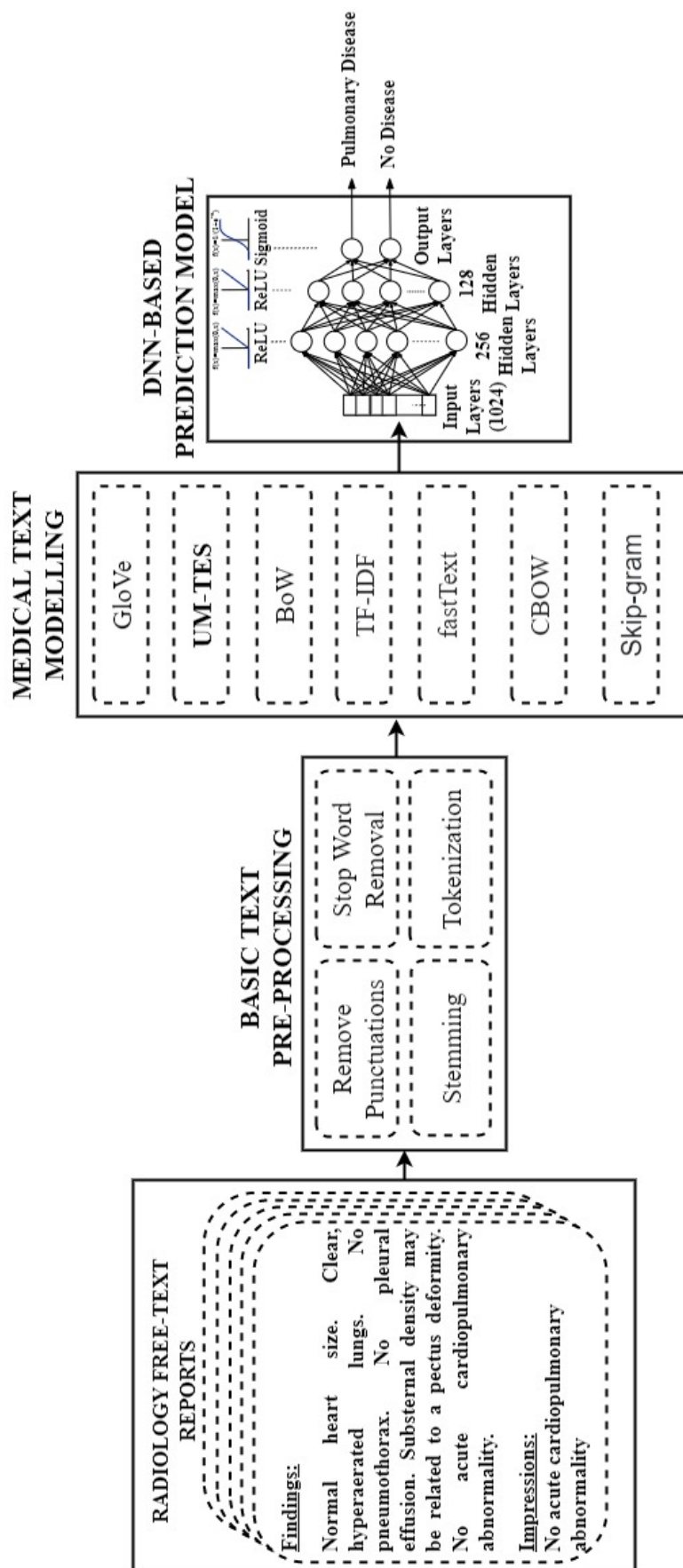


Figure 4.2: Overall workflow of various deep learning-based NLP framework for predicting pulmonary diseases from radiology free-text reports

## 4.4 Experimental Results and Discussion

This section explains the experimental evaluation of the proposed UM-TES Framework. First, the dataset and cohort selection are presented, followed by the evaluation metrics and benchmarking results.

### 4.4.1 Datasets and Cohort Selection

A limited dataset becomes a severe issue in the health domain when it happens to be multimodal data. In the case of images, there are some quality open source cohorts. Hence, there is a necessity to validate the effectiveness of the multimodal fusion models on the publicly available medical cohort and real-time data obtained from the private hospital. A comprehensive study was carried out on two clinical cohorts: the Indiana University chest X-ray dataset (Demner-Fushman *et al.*, 2016) and the real-time multimodal data acquired from a private medical hospital [KMC Hospital (Mangalore, India)]. For our investigation, the de-identified data is leveraged. The Institutional Ethics Committee (IEC) approval was granted by the Kasturba Medical College (KMC), Mangalore, for further research purposes. The two multimodal medical cohorts acquired consist of chest X-rays and associated radiology free-text reports. The two clinical cohorts are classified as *normal* (i.e., cases with no abnormal findings or any active diseases) and *abnormal* (i.e., cases with acute pulmonary and cardiopulmonary abnormalities like Pulmonary edema, Pleural effusion, Calcified granuloma, Pneumothorax, Cardiomegaly, Pulmonary atelectasis, Pneumonia, Opacity/lung base, etc.). Table 4.1 represents the summary of cases (chest X-ray with associated radiology reports) from the Indiana University and KMC Hospital dataset. A detailed benchmarking exercise is carried out on both clinical datasets to evaluate the proposed multimodal network.

- **IU Dataset:** The Indiana University dataset (Demner-Fushman *et al.* (2016)) is a publicly available multimodal dataset containing de-identified clinical radiology chest X-ray images and associated diagnostic reports. The majority of the existing work on the Indiana University dataset focuses on cross-modal retrieval of radiology reports given chest X-ray as input (Jing *et al.* (2017); Liu *et al.* (2019a); Xue *et al.* (2018)). To the best of our knowledge, limited research work has been carried out on this dataset in terms of multimodal disease prediction. The radiology reports associated with chest X-rays contain *findings*, *impression*, *indication*, and *Medical Subject Heading (MeSH)*

Table 4.1: Cohort Statistics: Chest X-Ray with associated Radiology reports from two Institutions

Characteristics	IU Dataset	KMC Dataset
Total No. of cases (Chest X-Ray with Radiology reports)	3996	502
Total No. of cases after removing missing cases	3638	502
Total No. of cases after data augmentation	6229	1498
Total No. of Sentences	17990	14537
Total No. of Words	143177	90221
Total No. of Vocabulary	1731	400
Percentage of Normal cases	38%	52%
Percentage of Abnormal Cases	62%	48%

indexing comprising encoded diseases and findings. We chose 3638 frontal chest X-rays with their respective radiology notes. The chest X-ray with missing reports is removed since we are conducting disease prediction from multimodal clinical data in our experiment. To obtain the ground truth annotation, we have used MeSH indexing to classify the multimodal clinical cohort into *normal* and *abnormal* classes. Finally, an experienced radiologist has carefully validated the annotated data to ensure the annotation is accurate. As discussed in Section 4.4.2, the selected cases are augmented to obtain 6229 chest X-rays with radiology reports for effective deep learning prediction.

- KMC Hospital Dataset:** The 502 chest X-rays with associated radiology reports were collected from the KMC Hospital (Mangalore, India). Both the Images and reports are de-identified to prevent any patients' personal information from being revealed. Each X-ray image and radiology report are manually reviewed by an experienced radiologist and are categorized into *normal* and *abnormal* classes. Equal data distribution is considered to avoid biasing the model into one category. The abnormal classes contain lung diseases such as pleural effusion, Tuberculosis, Pneumonia, Pneumothorax, Cardiomegaly, Consolidation, Edema, bronchiolitis, and Fibrosis. As the dataset collected was small in size, we applied various augmentation techniques as described in Section 4.4.2 to increase the cohort size to 1498 for effective multimodal disease prediction.

### 4.4.2 Data Preparation and Augmentation Stage

A large amount of high-quality data is required to develop a robust deep learning model with good performance (Chen and Lin (2014)). However, obtaining such data is challenging. One approach to addressing this issue is to enable practitioners to artificially expand the diversity of data available in the training set by augmenting the original dataset. Data augmentation also prevents overfitting problems and increases the model's ability to adapt to the new, unseen data derived from the same distribution as the one used to build the model (Dvornik *et al.* (2019)). As the size of the collected radiology medical cohort was too small for effective disease prediction, we applied data augmentation to produce a well-balanced and good-quality dataset. Data Augmentation must be carefully adapted, as the medical images are relatively sensitive to the various operations that can alter the original training set's actual distribution by introducing additional outliers.

As a part of data augmentation, we have used a series of geometric transformations on the training set. Following are the various data augmentation techniques applied to two medical cohorts: To begin with, we have applied rotation augmentation that randomly rotates the image by -5 to +5 degrees. Next, we have performed random zooming inside the X-ray image with a probability of 0.95 and a randomly chosen value between 1.1 and 1.5. Further, we have randomly changed the brightness of the image between 0.5 and 1.5. The value 0.0 produces the black image, the value 1.0 gives the original image, and a value greater than 1.0 generates a brighter image. Finally, a shear transformation is applied to tilt the image randomly between the range -5 and +5. The various data augmentation techniques applied to the medical cohort and the ranges of each method are shown in Table 4.2. For our experiment, we have used Augmentor (Bloice *et al.* (2017a)), a python-based image augmentation library, to perform the data augmentation of our medical cohort. Chapter 5 offers an extensive overview of the diverse data augmentation techniques applied to both the Indiana University and KMC hospital datasets.

### 4.4.3 Evaluation Metrics

The following six evaluation metrics were used to assess the effectiveness of the UM-TES framework: F1 score, precision, recall, accuracy, Matthews correlation coefficient (MCC) and Area Under the Receiver Operating Characteristics (AU-ROC). The number of abnormal cases in the medical corpus that are correctly predicted as abnormal cases by the prediction model is called as True Positive

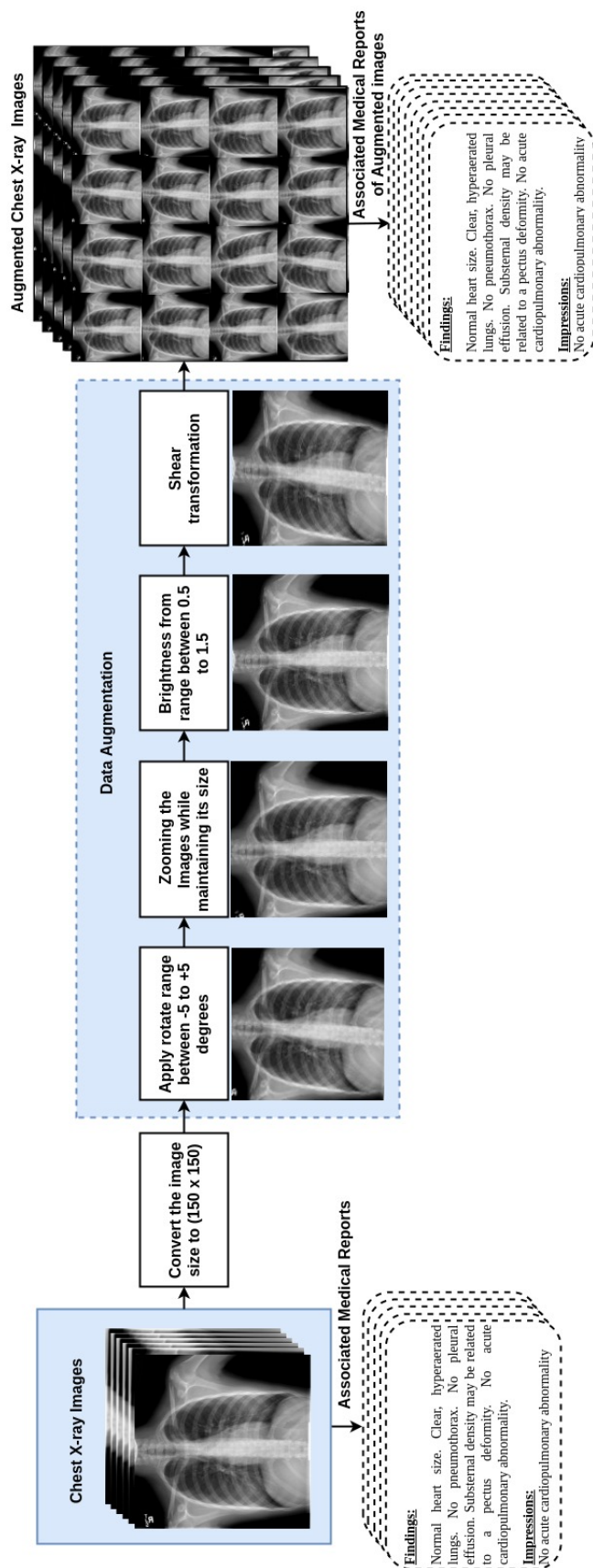


Figure 4.3: Data Augmentation of Radiology Data.



Table 4.2: The various data augmentation techniques applied on the medical cohort with the ranges of each techniques

Augmentation Techniques	Range
Rotation	[-5, +5]
Zooming	0.95
Brightness	[0.5, 1.5]
Shear Transformation	[-5,+5]

(TP). The number of normal cases in the medical corpus that are correctly predicted as normal cases by the prediction model is called True Negative (TN). The number of abnormal cases that are predicted as the normal case is called False Negative (FN). The number of normal cases predicted as the abnormal case is called as False Positive (FP). The accuracy is referred to as the number of exact predictions (i.e., TP+TN) from the given training set divided by the total number of predictions made, as shown in Eq. (4.16). Precision, as defined in Eq. (4.17) is the ratio of cases that are tested abnormal and are abnormal (i.e., TP) by the total number of correct predictions (i.e., TP+FP). Precision indicates the fraction of cases that actually have the diseases from the cases that are predicted to be diseased or abnormal. The recall, as defined in Eq. (4.18) is the ratio of cases that are tested abnormal and are abnormal (i.e., TP) by the cases that are actually abnormal (i.e., TP+FN). Recall shows the fraction of correctly predicted diseases from all the cases in the set that actually have the diseases. The higher the recall, the fewer actual instances of diseases go unpredicted.

We consider the F1-score to investigate the symmetry between precision and recall and also due to the uneven data distribution in the IU dataset. The F1-score, as depicted in Eq. (4.19) is the harmonic mean between precision and recall. We also consider MCC, which considers all four values of confusion metrics to check the balance between the two classes of different sizes. MCC, as defined in Eq. (4.20) is the correlation coefficient between the true and the predicted binary classification. MCC returns the value between +1 and -1, where +1 indicates an accurate prediction, 0 shows a random prediction, and -1 indicates a completely wrong prediction. The higher the correlation between the observed and predicted classes, the better the prediction. AUROC shows the performance measure of binary classification at different threshold settings. The AUROC curve is plotted between the True Positive Rate and the False Positive Rate. AUROC near 1 indicates good separability between the two classes (i.e., normal and abnormal cases), and near 0 indicates bad separation between the classes.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.16)$$

$$Precision(P) = \frac{TP}{TP + FP} \quad (4.17)$$

$$Recall(R) = \frac{TP}{TP + FN} \quad (4.18)$$

$$F1 - Score = 2 \cdot \frac{P \cdot R}{P + R} \quad (4.19)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4.20)$$

When dealing with almost balanced classes in the KMC dataset, it is essential to use a variety of evaluation metrics to gain a comprehensive understanding of the model's performance. While accuracy is commonly used, it may not be the most informative metric, as it can be sensitive to minor changes in predictions and doesn't account for varying levels of impact from misclassifications. Instead, we should consider using multiple evaluation metrics such as precision, recall, F1-score, MCC, and the AUROC curve. Each of these metrics provides unique insights into different aspects of the model's behavior and performance. Precision and recall are particularly important when misclassifying certain instances, which can have different consequences. For instance, in medical diagnoses, false negatives can lead to severe health issues, while false positives may cause unnecessary stress for patients. By focusing on precision and recall, we can better assess the model's ability to correctly identify positive instances among the predicted positive instances and all actual positive instances, respectively. In real-world scenarios, the ground truth labels may have inaccuracies themselves. By using multiple evaluation metrics, we can understand how the model performs under different evaluation assumptions and gain insights into the quality of the ground truth data. Considering a range of evaluation metrics allows us to have a more nuanced view of the model's performance, considering different aspects and consequences of misclassifications. This information becomes especially valuable in cases where the classes are nearly balanced and misclassifications may have a significant impact. In summary, utilizing a variety of evaluation metrics provides a more stable and insightful assessment of the model's performance, taking into account various real-world considerations and data characteristics. This approach helps in model selection, performance monitoring, and interpretation for practical applications.

#### 4.4.4 Results and Discussions

In order to comprehensively validate the proposed Medical knowledge-based Deep Learning framework, we leverage various traditional ML techniques such as SVM (Hearst (1998)), K-Nearest Neighbour (KNN) (Mucherino *et al.* (2009)), RF (Breiman (2001)), LR (Hosmer and Lemeshow (2000)) and AdaBoost classifier (AB) (Freund and Schapire (1996)). In order to comprehensively evaluate the proposed word embedding model, we have compared its performance with BoW, tf-idf, FastText, CBOW, Skip-gram, and GloVe embedding models. The extensive benchmarking investigations were conducted to verify the efficacy of the proposed model with respect to the state-of-the-art NLP and ML models. For our experiment, we have used the NVIDIA Tesla M40 server with a 24GB GPU, a 3TB Hard disk, 128 GB of RAM, and the Ubuntu server Operating System. The radiology corpora are divided into training and test sets (i.e., IU dataset: no. of training set = 5606, test set = 623, and KMC Hospital dataset: no. of training set = 1348, test set = 150). The UM-TES framework was trained for 100 epochs, and 10-fold stratified cross-validation was applied to examine the proposed model.

The findings extracted from the report are preprocessed and padded with an input size of 260. The proposed Knowledge-based Medical Text modelling is applied to the free-text to obtain word vectors of size  $260 \times 100 = 26000$ . The DDR-CNN is utilized to produce the most discriminative medical features of size 4096. Further, the DNN-based prediction model is employed on the low-dimensional features extracted from DDR-CNN to detect the pathology present in the reports. The grid search technique (Bergstra and Bengio, 2012) is used to choose the optimal hyperparameters for fine-tuning the model parameter setting. We implemented the proposed model using a well-known deep learning framework, Tensorflow (Abadi and *et. al.*, 2015). The following are the hyperparameters set for DDR-CNN and the DNN model: Kernel size: 5, filter size: 32, strides: 1, dropout probability: 0.4, pool size: 2, and learning rate: 0.001. We have also utilised the Adam optimizer and binary cross entropy as the loss function.

##### 4.4.4.1 Performance Analysis with the State-of-the-art NLP Techniques

The qualitative benchmarked results on the Indiana University and the KMC cohorts are shown in Table 4.3 and Table 4.4. The results show that the proposed deep learning framework with UM-TES outperforms state-of-the-art deep learning strategies. The proposed UM-TES has a staggering improvement in accuracy of 90.40% and 94.13% on the IU and KMC datasets, respectively, showcasing

the model’s prediction performance compared to the other state-of-the-art models. Our proposed UM-TES has achieved the 3% improved precision compared to CBOW and Skip-gram models, proving its ability to predict abnormal classes correctly. The increase in recall denotes the lesser chances of abnormal classes being unpredicted. The superior F1-Score and MCC of UM-TES indicate that the proposed model can accurately predict pulmonary disease despite a class imbalance problem. The proposed model has obtained an AUROC of 0.9555 and 0.9651 for the IU and KMC radiology cohorts, indicating that the model can accurately predict the normal and abnormal classes. The graphical visualization depicting the performance evaluation of the proposed UM-TES with state-of-the-art NLP techniques on the IU and KMC radiology cohorts is shown in Figure 4.4 and Figure 4.5. The analysis shows that the knowledge base incorporated for the proposed medical text modelling technique has significantly impacted performance by learning unseen or rare medical words. Henceforth, the proposed UM-TES can be incorporated when there is a low data condition while training the deep learning frameworks.

Table 4.3: Benchmarked performance analysis results of the proposed deep learning-based NLP technique with the state-of-the-art text modelling techniques on the diagnostic clinical free-text cohort collected from the publicly available Indiana University dataset

Models	Accuracy	Precision	Recall	F1-Score	MCC	AUROC
BoW	87.32%	0.8553	0.8150	0.8690	0.7253	0.8899
tf-idf	87.32%	0.8776	0.8026	0.8714	0.7336	0.8899
FastText	89.30%	0.8791	0.8608	0.8916	0.7931	0.9152
CBOW	89.37%	0.8956	0.8535	0.8916	0.7924	0.9221
Skip-gram	89.95%	0.8982	0.8601	0.8978	0.7924	0.9260
GloVe	87.27%	0.8741	0.8727	0.8729	0.7232	0.9333
<b>Proposed UM-TES</b>	<b>90.40%</b>	<b>0.9080</b>	<b>0.9040</b>	<b>0.9059</b>	<b>0.7939</b>	<b>0.9555</b>

#### 4.4.4.2 Performance Analysis with the State-of-the-Art ML techniques

The benchmark results are presented in Table 4.5 for the IU and KMC Hospital datasets. The proposed model with KB outperforms compared to the traditional ML models and DNN without KB on the publicly available dataset as well as on the real-time collected dataset. The proposed model with KB achieves high precision signifies that most cases belonging to the “abnormal” class are detected, which is the main objective of our disease prediction model. We have measured

Table 4.4: Benchmarked performance analysis results of the proposed deep learning-based NLP technique with the state-of-the-art text modelling techniques on the diagnostic clinical free-text cohort collected from KMC hospital

Models	Accuracy	Precision	Recall	F1-Score	MCC	AUROC
BoW	86.05%	0.9027	0.8358	0.8526	0.7418	0.8610
tf-idf	87.05%	0.9340	0.8396	0.8689	0.7768	0.8795
FastText	92.84%	0.9271	0.8974	0.9192	0.8701	0.9295
CBOW	92.72%	0.8764	0.8888	0.8822	0.8608	0.9294
Skip-gram	92.72%	0.9236	0.9214	0.9343	0.8608	0.9208
GloVe	92.53%	0.9270	0.9250	0.9290	0.8500	0.9630
<b>Proposed UM-TES</b>	<b>94.13%</b>	<b>0.9475</b>	<b>0.9413</b>	<b>0.9443</b>	<b>0.8827</b>	<b>0.9651</b>

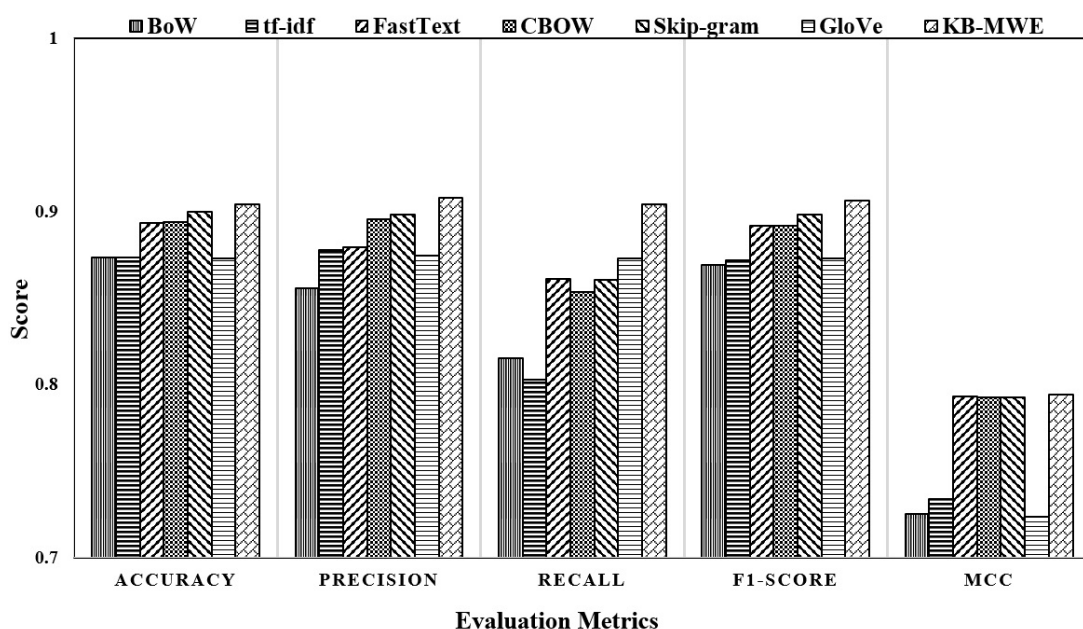


Figure 4.4: Performance analysis of UM-TES with state-of-the-art NLP models on IU cohort

the F1-Score and MCC, which are also essential evaluation metrics in our experiment, as our data exhibit a class imbalance problem (i.e., in the IU dataset, the pathology or abnormal class is in a higher number compared to the normal class; refer Table 4.1). The proposed model with KB has a staggering improvement of 11-31% in F1-Score and 9-40% in MCC for the IU dataset. The higher value of F1-Score and MCC of the proposed Knowledge-based Deep Learning model on both IU and KMC hospital data signifies that even if there was a class imbalance, the model was able to accurately classify according to the label “normal” and “abnormal”. The AUROC is plotted for the proposed model and the conventional ML

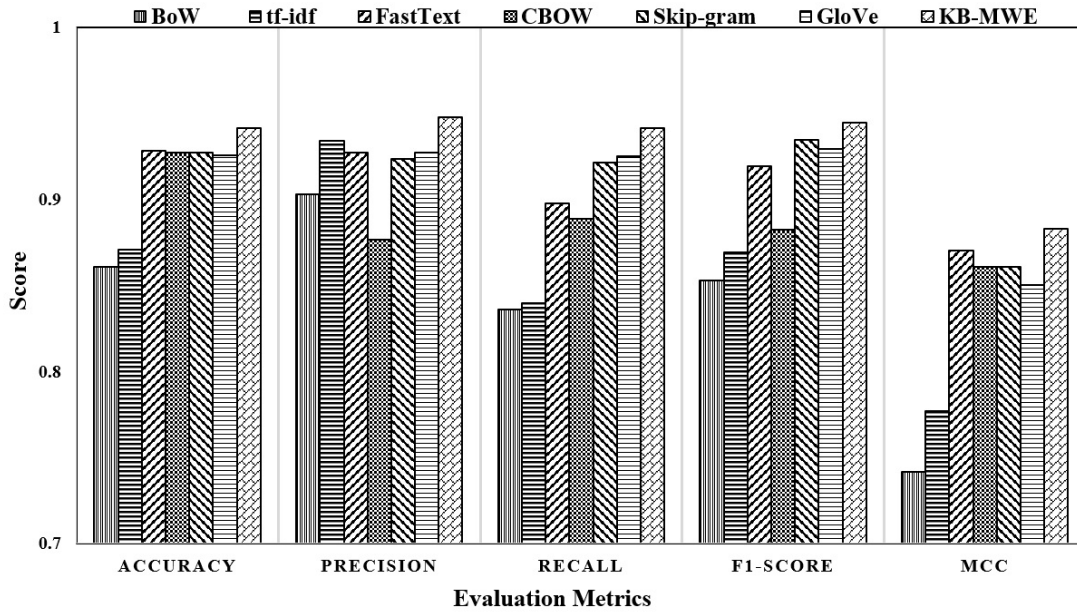


Figure 4.5: Performance analysis of UM-TES with state-of-the-art NLP models on KMC cohort

models for the IU and KMC hospital datasets, as shown in Figure 4.6a and Figure 4.6b respectively. It is apparent that the proposed model shows a considerable improvement of 4-15% and 0.29-8% for the IU dataset and the KMC hospital dataset in comparison with the standard ML techniques. The increased value of AUROC signifies that the model is accurate in distinguishing between reports with disease and those without disease. It is also shown that Random Forest has produced promising results with the proposed model, which can be attributed to its ensemble learning property. By leveraging ensemble learning, Random Forest combines multiple decision trees to make predictions, leading to improved performance in medical radiology report text classification. The effectiveness of Knowledge-based Medical Text modelling is explained in the following section.

#### 4.4.4.3 Effect of Clinical Knowledge-based Text Modelling

We have also examined the effect of customized Clinical Knowledge-based text modelling compared with the Glove word embedding, as shown in Figure 4.7a and Figure 4.7b respectively. There is a significant increase of 3% in terms of accuracy, precision, recall, F1 score and around 7% improvement in MCC for the IU dataset. There is a considerable gain of 2% in terms of accuracy, precision, recall, F1 score and around 3% improvement in MCC for the KMC dataset. The results depict that incorporating knowledge-base with the word-embedding models significantly



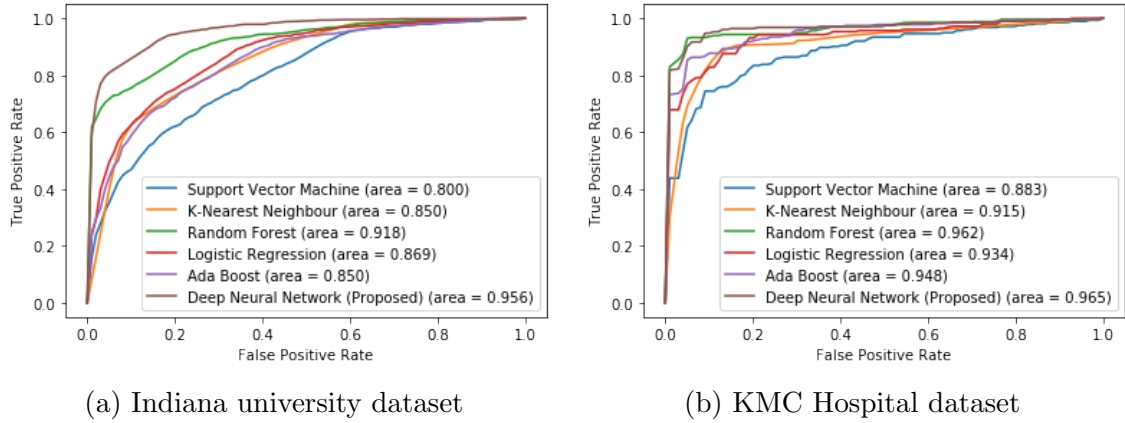


Figure 4.6: Comparing AUROC performance of proposed Deep learning model w.r.t. State-of-art Machine Learning techniques.

Table 4.5: Benchmarking the proposed DNN model with and without Knowledge Base (KB) against the State-of-the-Art Machine Learning Model w.r.t. Indiana University and KMC Hospital Dataset

Models	Indiana University Dataset					
	Acc.	Pre.	Recall	F-Sc.	MCC	AUROC
SVM	73.75%	0.7319	0.7374	0.5969	0.4149	0.7997
KNN	77.16%	0.7799	0.7716	0.7054	0.5214	0.8501
RF	86.36%	0.8729	0.8637	0.7866	0.7076	0.9175
LR	79.74%	0.7951	0.7974	0.7037	0.5550	0.8686
AB	76.55%	0.7787	0.7655	0.5913	0.4823	0.8499
<b>DNN-KB (Proposed)</b>	<b>87.27%</b>	<b>0.8741</b>	<b>0.8727</b>	<b>0.8129</b>	<b>0.7232</b>	<b>0.9333</b>
<b>DNN+KB (Proposed)</b>	<b>90.40%</b>	<b>0.9080</b>	<b>0.9040</b>	<b>0.8579</b>	<b>0.7939</b>	<b>0.9555</b>
Models	KMC Hospital Dataset					
	Acc.	Pre.	Recall	F-Sc.	MCC	AUROC
SVM	76.16%	0.7855	0.7619	0.7972	0.5286	0.8827
KNN	88.88%	0.8915	0.8888	0.8952	0.7769	0.9147
RF	93.53%	0.9417	0.9353	0.9370	0.8731	0.9622
LR	88.88%	0.8924	0.8890	0.8960	0.7767	0.9336
AB	88.28%	0.8886	0.8829	0.8909	0.7646	0.9482
<b>DNN-KB (Proposed)</b>	<b>92.53%</b>	<b>0.9271</b>	<b>0.9251</b>	<b>0.9289</b>	<b>0.8502</b>	<b>0.9630</b>
<b>DNN+KB (Proposed)</b>	<b>94.13%</b>	<b>0.9475</b>	<b>0.9413</b>	<b>0.9443</b>	<b>0.8827</b>	<b>0.9651</b>

Note: Acc.= Accuracy, Pre.=Precision, F-Sc.=F1-Score, DNN-KB=DNN without KB, DNN+KB=DNN with KB

increases the performance of the disease prediction due to the knowledge gained from the word embeddings trained on the large corpora.

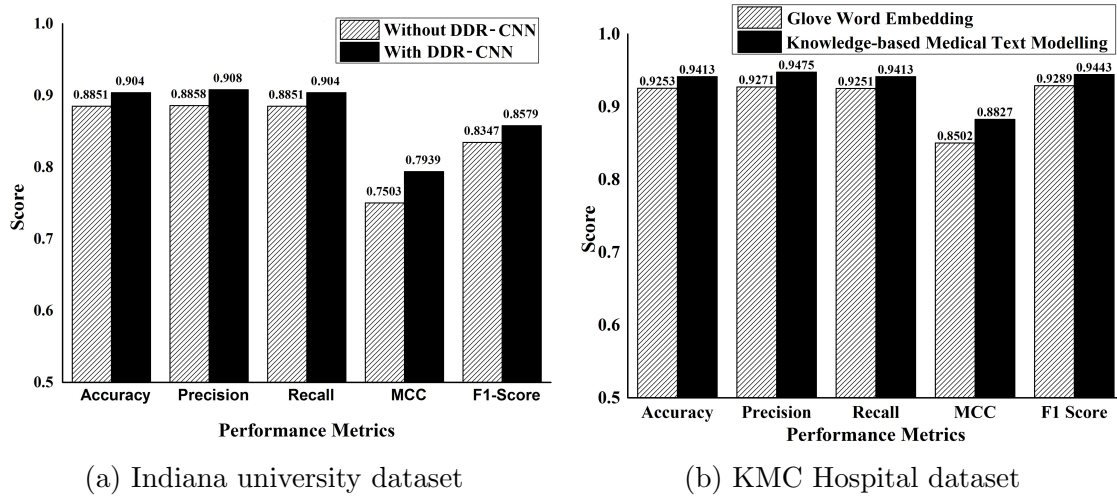


Figure 4.7: Effectiveness of Customized Clinical Knowledge-based Text Modelling compared to the GloVe Embeddings

#### 4.4.4.4 Effect of CNN-based Discriminate Dimensionality Reduction

We assessed the efficacy of our proposed model with and without DDR-CNN on the IU and KMC medical cohorts, as shown in Figure 4.8a and Figure 4.8b respectively. There is a substantial improvement of 2% in terms of accuracy, precision, recall, and F1 score when the DDR-CNN model is applied for the IU and KMC hospital datasets. There is an increase of 4% and 3% in terms of MCC metrics for the IU and KMC Hospital datasets, respectively. The results indicate that the DDR-CNN model obtains the most discriminative features that can predict the abnormality from the reports. The DDR-CNN module reduces storage and computational costs by transforming high-dimensional features into low-dimensional features while retaining the most discriminative features.



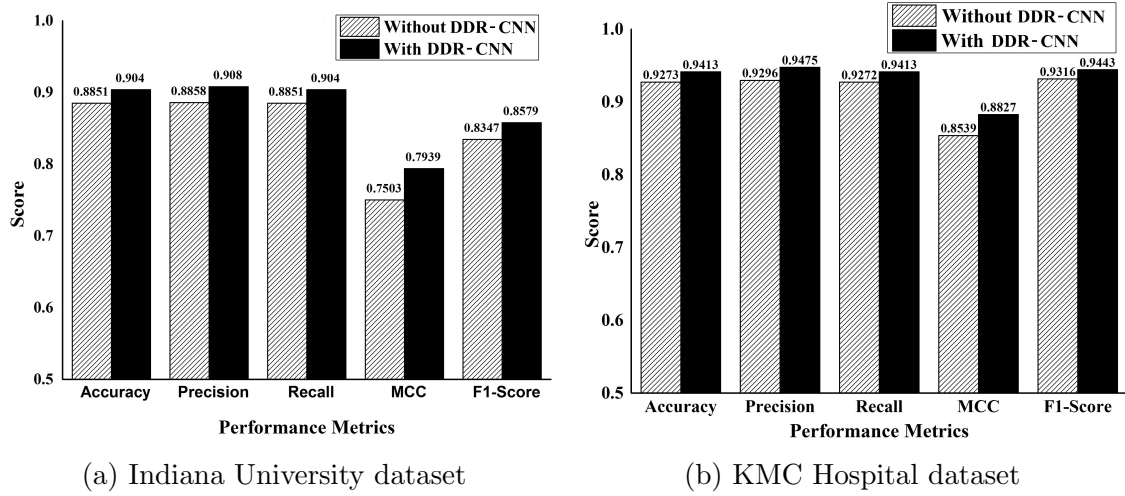


Figure 4.8: Comparison of performance metrics with and without DDR-CNN

## 4.5 Summary

In this chapter, we have proposed a UM-TES framework comprising clinical knowledge-based text modelling techniques with a deep learning framework to predict pulmonary diseases in radiology free-text reports. To model the text in the diagnostic reports, the GloVe Embedding model was used in conjunction with a knowledge base. The textual features were then processed using the DDR-CNN model to reduce their dimensionality. The final step was to apply a DNN to predict any abnormalities in the reports. Through our experimentation, we observed that the proposed UM-TES word embedding technique yielded superior performance when compared to state-of-the-art NLP models. Additionally, we evaluated the performance of the DNN classifier against that of a standard machine learning-based classifier and determined that the former achieved better results. Our observation revealed that the improved performance of UM-TES is attributed to the integration of a radiology knowledge base, which enhances prediction accuracy even when the training cohort is small in size. Consequently, the proposed model can be implemented in scenarios where data are scarce, which is often the case in the medical domain, where cohorts are institution-specific or restricted to specific domains. While deep learning models have the capability to learn representations directly from raw data, our UM-TES model employs a two-step approach that presents several noteworthy benefits. Firstly, it excels in producing semantically meaningful representations of medical words, leading to enhanced representation quality. Secondly, the model’s modularity, achieved by separating feature extraction (word embeddings) and classification (DNN), introduces a clear distinction of

responsibilities, making it more manageable and maintainable. Lastly, this design enhances the adaptability of the UM-TES model for NLP-based text classification tasks.

## Publications

*(based on study proposed in this chapter)*

1. Shashank Shetty, Ananthanarayana V S., and Ajit Mahale. (2020) Medical Knowledge-Based Deep Learning Framework for Disease Prediction on Unstructured Radiology Free-Text Reports Under Low Data Condition, *21st EANN (Engineering Applications of Neural Networks) 2020 Conference. EANN 2020., vol 2. Springer, Cham, Halkidiki, Greece. [https://doi.org/10.1007/978-3-030-48791-1\\_27](https://doi.org/10.1007/978-3-030-48791-1_27) [Core Ranked Conference - Springer Proceedings] (Status: Published Online)*
2. Shashank Shetty, Ananthanarayana V S., and Ajit Mahale. (2023). Diagnostic Performance Evaluation of Deep Learning-Based Medical Text Modelling to Predict the Pulmonary Diseases from the Unstructured Radiology Free-Text Reports. *Acta Informatica Pragensia*. Volume 12, Issue 2, <https://doi.org/10.18267/j.aip.214>, <https://aip.vse.cz/corproof.php?tartkey=aip-000000-0483> [Indexed: Scopus, IF: 1.15] (Status: Published Online)

## Chapter 5

# Unimodal Medical Visual Encoding Subnetwork (UM-VES) for Disease Prediction from Radiology Chest X-ray Image

### 5.1 Introduction

For decades, chest diseases have been one of the most prominent causes of anguish, fatality and use of health services worldwide. According to the World Health Organization, nearly 235 million people suffer from chronic respiratory disease every year. Yearly, there are two million rises in the number of chronic respiratory disease cases<sup>1</sup>. The impact of these diseases varies and rapidly spreads depending on geographic features, lifestyle, etc. Modern medical science relies on various radiological imaging data like CT, X-ray, MRI for disease diagnosis. X-ray is a technique used for decades by experts to visualize abnormalities in the acute and internal organs. CXR is considered the primary tool for diagnosing chest diseases, which may be due to factors such as accessibility, minimal radiation exposure, and reasonable commercial pricing, along with the diagnostic capability to identify a wide variety of pathologies. Annually, it was estimated that around 238 erect view CXR for every 1000 population was reported in developed countries<sup>2</sup>. Chest disease is analyzed from the CXR image in the form of blunted costophrenic angles, cavitations, infiltrates, consolidation, and broadly distributed nodules (Abiyev and Ma'aitah, 2018). By inspecting the CXR images, the radiologist can analyze the diseases and

---

<sup>1</sup>WHO Chronic respiratory diseases. Online: <https://www.who.int/health-topics/chronic-respiratory-diseases>

<sup>2</sup>United nations scientific committee on the effects of atomic radiation (UNSCEAR). Online: [http://www.unscear.org/docs/publications/2008/UNSCEAR\\_2008\\_Annex-A-CORR.pdf](http://www.unscear.org/docs/publications/2008/UNSCEAR_2008_Annex-A-CORR.pdf)

note the valuable findings in the reports. With the tremendous growth in diagnostic images, screening diseases with CXR becomes a tedious and time-consuming task for a radiologist. The computer-assisted clinical recommendation system can aid radiologists by minimizing their workload by providing primary screening (Zhang *et al.*, 2021). The advancement of CNN (Krizhevsky *et al.*, 2012a) has provided remarkable progress in various computer vision applications, including computer-assisted clinical recommendation systems. The possible benefits of automated clinical systems will be high sensitivity to minute findings, automating the tedious daily process, and providing analysis during the unavailability of the experts.

Furthermore, the abnormalities in CXR images come in various shapes and sizes. Also, every single abnormality in pulmonary diseases occurs in variable sizes. For example, different cases of a single pathology like pulmonary infiltrate exist in various forms and sizes. In CXR, there is a possibility of overlapping with the anatomical part and abnormalities, making it challenging to interpret from the CXR. In the case of frontal CXR, there are chances that the nodule posterior is likely to overlap with the heart. Henceforth, there is a need to learn multi-scale features from the CXR to accurately predict the varied sizes of disorders. Deep Learning has been a preferred approach for medical image processing tasks due to its significant impact in this field (Litjens *et al.*, 2017). Deep learning approaches usually require a massive amount of training data as there is a need to fine-tune a large number of parameters during the learning process. This has encouraged the research community to publish many diagnostic CXR cohorts with expert annotations for research purposes (refer Table. 2.1). As the size of the input images increases, there is a requirement to use a deeper network to assure that the receptive field of the network is wide enough. Several existing studies have used ResNet-50 (He *et al.*, 2015) and DenseNet-121 (Huang *et al.*, 2016) for capturing imaging features. Even though there is an improvement in performance, the computation cost and network parameters significantly increase due to the enlarged inputs integrated with the deep networks, further increasing the time taken to train and optimize the model. Consequently, this makes further deployment on mobile and embedded devices challenging.

### 5.1.1 Problem Statement

To enhance the accuracy of identifying abnormalities in diagnostic images such as X-rays, CT, and MRI scans, it is essential to create automated techniques capable

of dealing with the variety of internal organs illustrated in these images. The multidimensional nature and abundance of information present in medical images necessitate the development of efficient techniques for extracting optimal features from diagnostic cohorts. For CXR images, detecting abnormalities in various shapes and sizes necessitates learning multi-scale features. As the complexity and size of the image data increase, the deep learning models required to extract optimal features become more intricate, with higher computational demands and network parameters. This leads to longer training and optimization time, making the process time-consuming and resource-intensive. Additionally, these challenges become more pronounced when deploying these models on mobile and embedded devices, where processing power and memory are often limited. Consequently, developing techniques to mitigate these challenges is an important consideration when designing automated methods for medical image analysis. The problem statement is defined as follows:

*“Considering an unstructured diagnostic imaging cohort with a varied-sized pulmonary abnormality, design and build a lightweight and explainable multi-scale deep learning framework for predicting chest diseases to facilitate an intelligent clinical recommendation system.”*

In this chapter, we aim to expand the networks receptive field and learn multi-scale discriminative features by maintaining the model parameters effectively. The major contribution of this study is summarized as follows:

- With the focus of designing an effective deep learning network suitable to employ in cloud computing, mobile vision, and embedding system applications, we present an explainable and lightweight UM-VES framework to predict abnormal diseases from chest radiographs.
- To enlarge the receptive field and capture the discriminative multi-scale feature without increasing convolution parameters, we propose an effective Multi-Scale Dilation Layer (MSDL), which is conducive to learning varied sized pulmonary abnormalities and boosts the prediction performance.
- We adopt a lightweight Depthwise Separable Convolution Neural Network (DS-CNN) to learn the dense imaging features by adjusting lesser network parameters than the conventional CNN. We employed a fully connected Deep Neural Network to predict the abnormalities from the Chest Radiographs.

- We incorporated the gradient-weighted Class Activation Mapping (Grad-CAM) technique to visualize and localize the abnormalities in the chest region. This makes our network explainable by checking the decision model’s transparency and understanding their ability to arrive at a decision.
- We compared the proposed UM-VES with the existing state-of-the-art Deep Learning strategies. We assessed our model’s competence by applying it on two datasets: the publicly available Open-I dataset (IU) and Real-time diagnostic data collected from a private hospital.
- We propose Radiology Deep Convolutional GAN (RAD-DCGAN) inspired by DCGAN (Radford *et al.*, 2015) for performing data augmentation tasks that mainly enhance the performance of deep CNN classifiers. Following are the main contributions of this study:
  - We conducted a thorough quantitative analysis of the proposed RAD-DCGAN method and compared its performance with basic augmentation techniques, including rotation, zooming, brightness, and shearing. We also evaluated the combined images obtained from all the basic augmentation strategies to determine their impact on the model’s performance.
  - We utilize state-of-the-art deep learning models such as MobileNet, VGG16, EfficientNetB1, VGG19, ResNet50, Xception, InceptionV3, and DenseNet to classify diseases from X-ray and MR images generated by RAD-DCGAN and traditional augmentation techniques.

## 5.2 Methodology

We aim to design an effective deep learning network that is lightweight and explainable to predict abnormalities from the Chest X-ray. The general architecture of the proposed UM-VES is presented in Figure 5.1. The overall architecture of the proposed UM-VES with filter shape, stride, input size, and output size is shown in Table 5.1. We propose an MSDL subnetwork that incorporates three dilation convolution layers with varied dilation rates on the input CXR to obtain multi-scale features. The discriminative features obtained are passed through a series of DS-CNN to learn dense imaging features with lesser network parameters than conventional convolution networks. Finally, a fully connected DNN is applied to the extracted features for predicting the abnormalities from the CXR, and the

Grad-CAM strategy is employed to visualize the abnormalities by superimposing a heatmap on the CXR.

Table 5.1: Overall architecture of the proposed UM-VES: Multi-Scale Dilated Network with depthwise Separable convolution

Type	Filter Shape	Stride	Input Size	Output Size	
Dilated Convolution ( $d_r=1$ )	3 x 3 x 1	1	150 x 150 x 3	150 x 150 x 1	
Dilated Convolution ( $d_r=2$ )	3 x 3 x 1	1	150 x 150 x 3	150 x 150 x 1	
Dilated Convolution ( $d_r=3$ )	3 x 3 x 1	1	150 x 150 x 3	150 x 150 x 1	
Concatenation (Merge Layer)	-	-	150 x 150 x 1 ( $d_r=1$ ) 150 x 150 x 1 ( $d_r=2$ ) 150 x 150 x 1 ( $d_r=3$ )	150 x 150 x 3	
Convolution	3 x 3 x 32	2	150 x 150 x 3	75 x 75 x 32	
Depthwise Convolution	3 x 3 x 32	1	75 x 75 x 32	75 x 75 x 32	
Seperable Convolution	1 x 1 x 64	1	75 x 75 x 32	75 x 75 x 64	
Zero Padding	-	-	75 x 75 x 64	76 x 76 x 64	
Depthwise Convolution	3 x 3 x 64	2	76 x 76 x 64	37 x 37 x 64	
Seperable Convolution	1 x 1 x 128	1	37 x 37 x 64	37 x 37 x 128	
Depthwise Convolution	3 x 3 x 128	1	37 x 37 x 128	37 x 37 x 128	
Seperable Convolution	1 x 1 x 128	1	37 x 37 x 128	37 x 37 x 128	
Zero Padding	-	-	37 x 37 x 128	38 x 38 x 128	
Depthwise Convolution	3 x 3 x 128	2	38 x 38 x 128	18 x 18 x 128	
Seperable Convolution	1 x 1 x 256	1	18 x 18 x 128	18 x 18 x 256	
Depthwise Convolution	3 x 3 x 256	1	18 x 18 x 256	18 x 18 x 256	
Seperable Convolution	1 x 1 x 256	1	18 x 18 x 256	18 x 18 x 256	
Zero Padding	-	-	18 x 18 x 256	19 x 19 x 256	
Depthwise Convolution	3 x 3 x 256	2	19 x 19 x 256	9 x 9 x 256	
Seperable Convolution	1 x 1 x 512	1	9 x 9 x 256	9 x 9 s 512	
5 x	Depthwise Convolution	3 x 3 x 512	1	9 x 9 x 512	9 x 9 x 512
	Seperable Convolution	1 x 1 x 512	1	9 x 9 x 512	9 x 9 x 512
Zero Padding	-	-	9 x 9 x 512	10 x 10 x 512	
Depthwise Convolution	3 x 3 x 512	2	10 x 10 x 512	4 x 4 x 512	
Seperable Convolution	1 x 1 x 1024	1	4 x 4 x 512	4 x 4 x 1024	
Depthwise Convolution	3 x 3 x 1024	2	4 x 4 x 1024	4 x 4 x 1024	
Seperable Convolution	1 x 1 x 1024	1	4 x 4 x 1024	4 x 4 x 1024	
Global Average Pooling	Pool 4 x 4	1	4 x 4 x 1024	1 x 1 x1024	

### 5.2.1 Multi-Scale Dilation Layer

We propose a Multi-Scale Dilation Layer to obtain a broad receptive field using three-channel dilation convolution layers with varied dilation rates to capture the multi-scale discriminative features from the CXR images, as shown in Figure 5.2. The MSDL enlarges the receptive field using varied convolution kernels and captures the wider context from the input CXR with less cost. The complete region that an eye can see in the human visual system is called the field of view. The human visual system consists of millions of neurons that collect various pieces of





information. The receptive field can be defined as a small part of the total field of view in a biological neuron. In short, it's a portion of the information that is available to a single neuron. Correspondingly, the receptive field in deep learning is the part of the input region that produces the output feature (Araujo *et al.*, 2019).

Dilated or Atrous convolution was initially developed as an algorithm for the wavelet transformation (Holschneider *et al.*, 1989). The primary goal of dilation convolution is to enhance the image resolution by inserting “holes” (zeroes) in between every pixel in convolution filters, allowing the deep learning model to capture the dense features. Here, the zeros are viewed as the “gaps” between the pixels, and these gaps can be varied into different widths referred to as dilation rates (Wang *et al.*, 2017a). CNN is the widely applied deep learning model that includes various layers like input and output, convolution, pooling, and fully connected layers. The image features are captured by passing them through multiple layers at different levels. Out of all the layers, convolution and pooling are considered the crucial layers for learning features from the images. The convolution layer detects multiple spatial features from the input image through the receptive field, and the pooling layer progressively down-samples the size of these spatial patterns to decrease the computation cost and the number of parameters utilized (Yamashita *et al.*, 2018). The pooling layer in CNN provides a wider receptive field; however, the increased usage of the pooling layer results in the loss of information (Yu and Koltun, 2016). Hence, we have leveraged dilation convolution to capture the widened features without increasing the number of parameters to extract the discriminative features from the CXR. The standard 3D-convolution procedure can be mathematically shown as follows:

$$Z(t_h, t_w, t_c) = \sum_{l=1}^{T_H-1} \sum_{m=1}^{T_W-1} \sum_{n=1}^{T_C-1} Y(t_h + l, t_w + m, n) \cdot F(l, m, n) \quad (5.1)$$

In the above Eq. 5.1, the standard convolution operation is applied on the image  $Y(t_h, t_w, t_c)$  with the convolution filter  $F(l, m, n)$  to generate the output feature map  $Z(t_h, t_w, t_c)$ , where  $T_H$ ,  $T_W$  and  $T_C$  indicates the height, width and channel of the input chest X-ray image. The dilated convolution operation is the variant of the convolution operation, where filter parameters are varied differently. The same filter in the dilation convolution is applied at different ranges using varied dilation rates. This allows dilation convolution to have a broader receptive field than the traditional convolution operation. For example, in a standard convolution filter  $4 \times 4$ , the receptive field of  $4 \times 4$  is created with 16 parameters. In contrast,

the dilation convolution filter with  $4 \times 4$  and the dilation factor of 4 will create a receptive field of  $13 \times 13$  with 16 parameters. Henceforth, the broader coverage of the CXR image is obtained with the wider receptive field by linearly incrementing the parameter. Mathematically, the dilation convolution with the dilation rate  $d_r$  is represented as follows:

$$Z(t_h, t_w, t_c) = \sum_{l=1}^{T_H-1} \sum_{m=1}^{T_W-1} \sum_{n=1}^{T_C-1} Y(t_h + d_r \times l, t_w + d_r \times m, n) \cdot F(l, m, n) \quad (5.2)$$

As shown in the Eq. 5.2, when the  $d_r = 1$ , the dilation convolution operation acts similar to a normal convolution operation. Using the Atrous convolution operation, we propose a MSDL with a three-channel dilation operation. MSDL is obtained by stacking three atrous convolution operation with three different dilation factors to effectively capture the wider receptive field (refer Figure 5.2). The features obtained from three parallel dilation convolutions are concatenated to obtain the feature maps that are further forwarded to DS-CNN. As shown in Figure 5.2, all three atrous convolution operations maintain the same number of parameters:  $3 \times 3$  ( $d_r = 1$ ),  $3 \times 3$  ( $d_r = 2$ ) and  $3 \times 3$  ( $d_r = 3$ ). However, there is a broader coverage of the receptive field, capturing multi-scale features from CXR by varying the dilation rates. Let  $I_h \times I_w \times R$  be the dimension of the input CXR image ingested into three-channel atrous convolution in parallel and concatenated to obtain the activation map of dimension  $I_h \times I_w \times R$ . Here,  $I_h$  represents the height and  $I_w$  indicates the height and width of the input CXR, and  $R$  denotes the number of channels. To preserve the output size of MSDL to  $I_h \times I_w \times R$ , we have used three dilation convolutions (i.e.,  $R/3$ ). The MSDL adopts three dilation convolutions to broaden the receptive field without increasing the number of parameters and captures multi-scale features from the input diagnostic CXR image. The concatenated features from MSDL are further given input to DS-CNN to learn the dense imaging features.

### 5.2.2 Depthwise Separable Convolution Neural Network (DS-CNN)

We have used DS-CNN to learn in-depth imaging features from the multi-scale features extracted from the MSDL. The DS-CNN is a class of CNN that is generally used for two critical reasons: 1) It leverages a lesser number of parameters than the conventional CNN, 2) It is computationally inexpensive and can be utilized

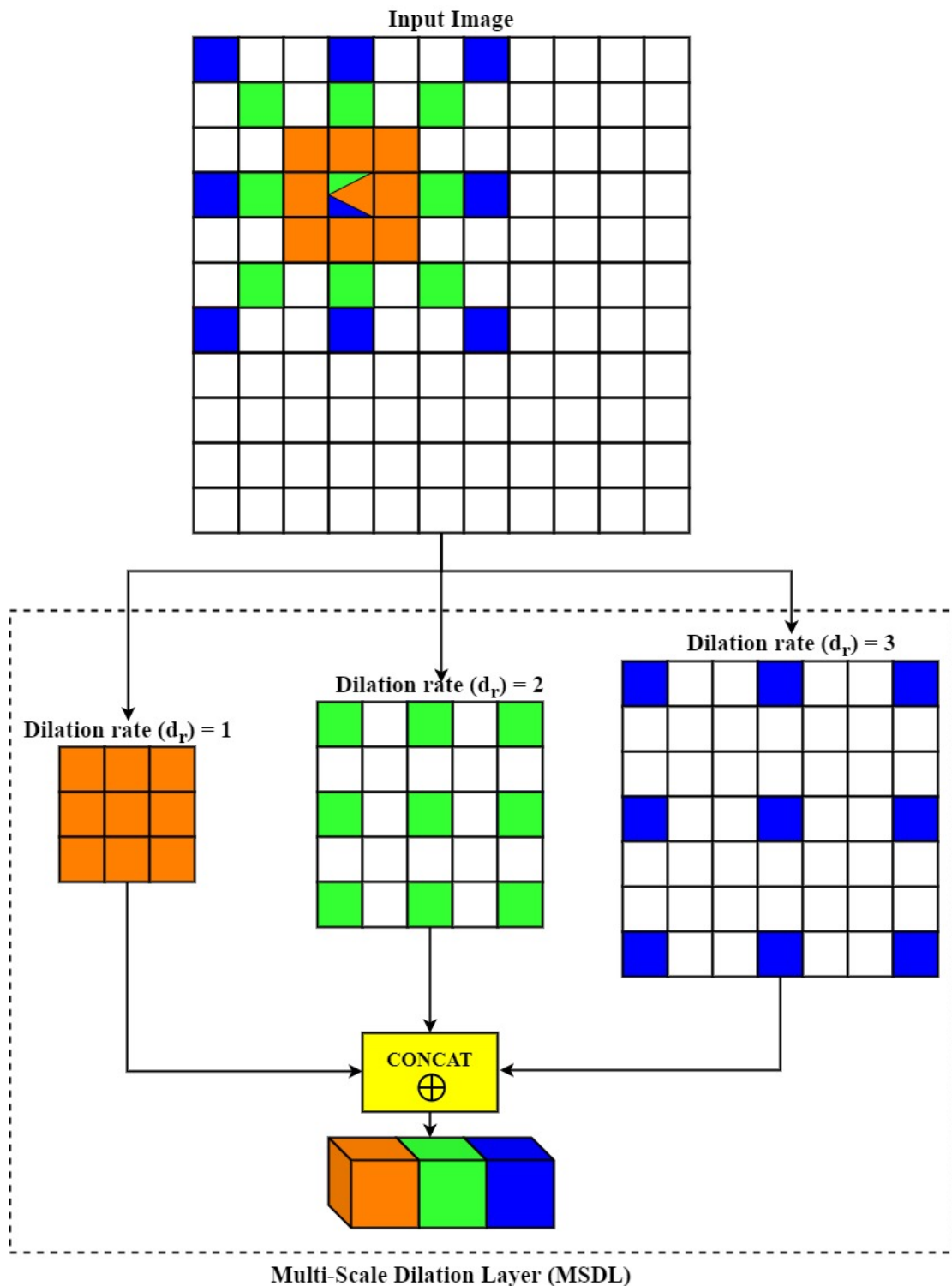


Figure 5.2: The Proposed Multi-Scale Dilation Layer. The three-channel Atrous convolution layer with dilation factors  $d_r = 1, 2, 3$  are stacked together to capture the wider receptive field. The resulting outcome from the three layers are concatenated to obtain the Multi-scale feature.

in mobile-based applications. DS-CNN has been utilized in some of the deep learning models like Xception (Chollet, 2016), and MobileNets (Howard *et al.*, 2017a). The DS-CNN can be further divided into Depthwise convolutions and pointwise convolutions. Figure 5.3 shows the difference between the traditional convolution filters and the Depthwise Separable filters. During the Depthwise convolution operation, the convolution is applied on one channel at a time using the  $S$  depthwise convolution filters (i.e.,  $C_j \times C_j \times 1$ ). Whereas in traditional convolution operation, the convolution is applied to all the  $R$  channels using the  $S$  filters (i.e.,  $C_j \times C_j \times R$ ). After the depthwise convolution operation, the pointwise convolution is applied on all the  $R$  channels with the  $S$  pointwise convolution filters (i.e.,  $1 \times 1 \times R$ ).

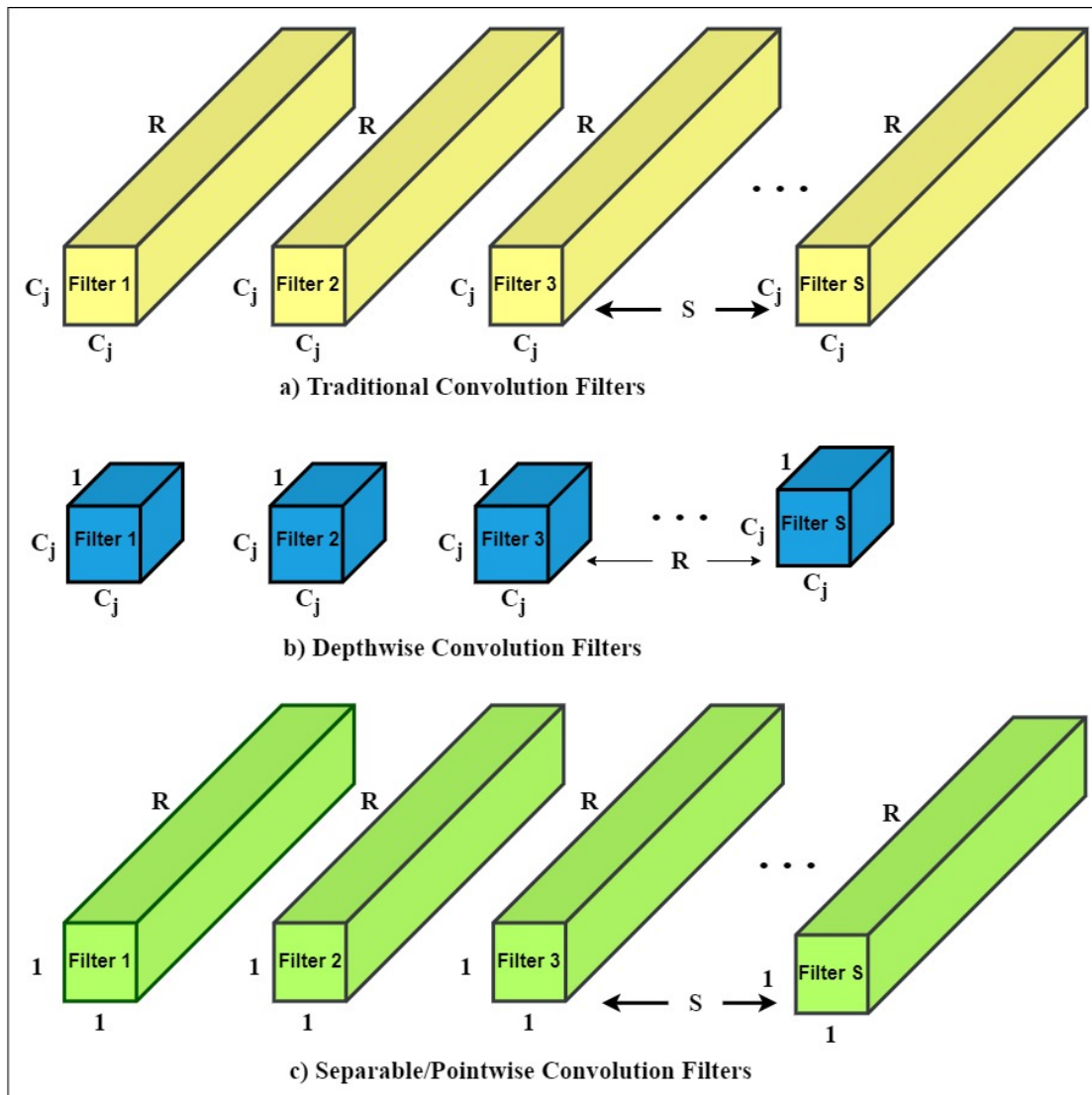


Figure 5.3: Conventional convolution filters and Depthwise Separable filters

The overall operation of the DS-CNN with depthwise and pointwise convolution operations is shown in Figure 5.4. Let us assume that the input feature map obtained from the MSDL layer applied on input CXR is  $Y$  with dimension  $I_h \times I_w \times R$ . If a multi-scale feature map obtained from the MSDL is ingested into the traditional convolution layer with kernels of size  $C_j \times C_j \times R$  then this convolution operation can be mathematically represented as follows:

$$Z_i = \sum_{k=1}^R Y_k \cdot C_i^j + b_k, \quad i = 1, 2, \dots, S \quad (5.3)$$

In the Eq. 5.3, the  $R$  and  $S$  indicate the input and output channels of the feature maps, respectively. Here,  $\cdot$  indicates the traditional convolution operator and the  $b_k$  represents the bias value. The output feature map generated from the standard convolution operation is represented by  $Z$  with size  $C_p \times C_p \times S$ . In the conventional convolution operation, the total number of multiplications in one convolution ( $T_{CNN}$ ) is equal to the size of the kernel and is denoted as follows:

$$T_{CNN} = C_j \times C_j \times R \quad (5.4)$$

As there are  $S$  kernels, the convolution operation is performed by striding every kernel vertically and horizontally  $C_p$  times. Hence, in the standard convolution operation, the total number of multiplications ( $Tot_{CNN}$ ) can be represented as follows:

$$Tot_{CNN} = S \times C_p \times C_p \times T_{sc} \quad (5.5)$$

Substituting the Eq. 5.4 in Eq. 5.5, we get Eq. 5.6,

$$Tot_{CNN} = S \times C_p \times C_p \times C_j \times C_j \times R \quad (5.6)$$

Unlike traditional convolution, in depthwise convolution, every kernel of size  $C_j \times C_j \times 1$  is applied on the single channel of the input activation map represented by,

$$Z_i = Y_k \cdot C_j + b_i, \quad k, i = 1, 2, \dots, R. \quad (5.7)$$

In the Eq. 5.7, the  $C_j$  represents the  $j^{th}$  depthwise filter, and the  $b_i$  indicates the bias value. The output feature map produced from the depthwise convolution operation is denoted by  $Z$  with size  $C_p \times C_p \times R$ . So, the number of multiplications

for a single depthwise convolution operation ( $T_{dc}$ ) can be depicted as follows:

$$T_{dc} = C_j \times C_j \quad (5.8)$$

The depthwise convolution operation is performed by sliding the kernel by  $C_p \times C_p$  times over  $R$  channels. So, the total number of multiplications by the depthwise convolution can be represented as follows:

$$Tot_{dc} = R \times C_p \times C_p \times T_{dc} \quad (5.9)$$

Substituting the Eq. 5.8 in Eq. 5.9, we get Eq. 5.10,

$$Tot_{dc} = R \times C_p \times C_p \times C_j \times C_j \quad (5.10)$$

The feature maps obtained from the depthwise convolution are passed through the pointwise convolution operation, where the  $1 \times 1 \times R$  kernel is applied on the input feature map to generate the final map of size  $I_h \times I_w \times S$ . Here, a single pointwise convolution operation needs  $1 \times R$  multiplications. The pointwise kernel is slided by  $C_p \times C_p$  times and hence, the total number of multiplications ( $Tot_{pc}$ ) can be formally represented as follows:

$$Tot_{pc} = R \times C_p \times C_p \times S \quad (5.11)$$

Therefore, the overall multiplication required for depthwise separable convolution operations is equal to the total number of multiplications needed in depthwise convolution ( $Tot_{dc}$ ) and pointwise convolution ( $Tot_{pc}$ ). The total multiplication of depthwise separable convolution operations ( $Tot_{DS-CNN}$ ) is given as follows:

$$Tot_{DS-CNN} = R \times C_p \times C_p \times C_j \times C_j + R \times C_p \times C_p \times S \quad (5.12)$$

So, to compare the complexity of DS-CNN with standard CNN, the ratio of Eq. 5.12 to Eq. 5.6 is computed as follows,

$$\frac{Tot_{DS-CNN}}{Tot_{CNN}} = \frac{R \times C_p \times C_p \times C_j \times C_j + R \times C_p \times C_p \times S}{S \times C_p \times C_p \times C_j \times C_j \times R} \quad (5.13)$$

Solve the Eq. 5.13 to obtain the Eq. 5.14,

$$\frac{Tot_{DS-CNN}}{Tot_{CNN}} = \frac{1}{S} + \frac{1}{C_j^2} \quad (5.14)$$

Here, Eq. 5.14 shows that the DS-CNN performs  $\frac{1}{S} + \frac{1}{C_j^2}$  times faster than the

standard CNN. Hence, dividing DS-CNN into two separate tasks (i.e., depthwise and pointwise operations) has significantly improved the computation speed and is lightweight compared to traditional CNN.

Figure 5.5 shows the general process flow of DS-CNN, followed by Batch Normalization and ReLU. To establish a larger gradient, we have utilized Batch Normalization and ReLU after every depthwise and pointwise convolution operation (Ioffe and Szegedy, 2015). Gradient represents the measure of the steepness of the slope. The higher the gradient, the steeper the slope, and the lower the gradient, the shallower the slope. Also, there is a need to learn in-depth features from the diagnostic CXR, and, hence, the use of the general process flow of DS-CNN (Figure 5.5) will make the deep learning network shallow. Therefore, in our proposed UM-VES, we have utilized 27 Batch Normalization and ReLU operations, 13 Depthwise and pointwise convolution operations, and a global average pooling layer to learn the discriminative features from the input CXR. Table 5.1 depicts the overall architecture with the network parameter details of the proposed UM-VES. The extracted features are further passed through the fully connected Deep Neural Network for abnormality prediction from the input CXR.

### 5.2.3 Network Structure and Training of UM-VES

In our network architecture of UM-VES, the input chest X-Ray is ingested into the UM-VES subnetwork we use Batch Normalization (BN) layers and ReLU layers after each depthwise convolution and pointwise convolution layer to make the gradient larger (Ioffe and Szegedy, 2015). The basic structure of depthwise separable convolution followed by batch normalization and ReLU is shown in Figure 5.5. We cannot obtain the deep imaging information from the medical chest X-ray images if the network is shallow; therefore, the use of one basic structure is insufficient to create an effective neural network. So, we have constructed a lightweight neural network using the basic structure as shown in Figure 5.5. Many such basic structures are joined together with global average pooling to form the UM-VES network structure. Overall in total, the UM-VES network structure includes a multichannel dilation layer, a concatenation layer, a convolution layer, 13 depthwise separable convolution layers, 27 BN and ReLU layers, and a global average pooling layer.

The multichannel dilation layer, including three dilated convolution layers with varied dilation rates (i.e.,  $r=1, 2$  and  $3$ ), captures a larger receptive field with valuable imaging features. The concatenated features from the three dilated



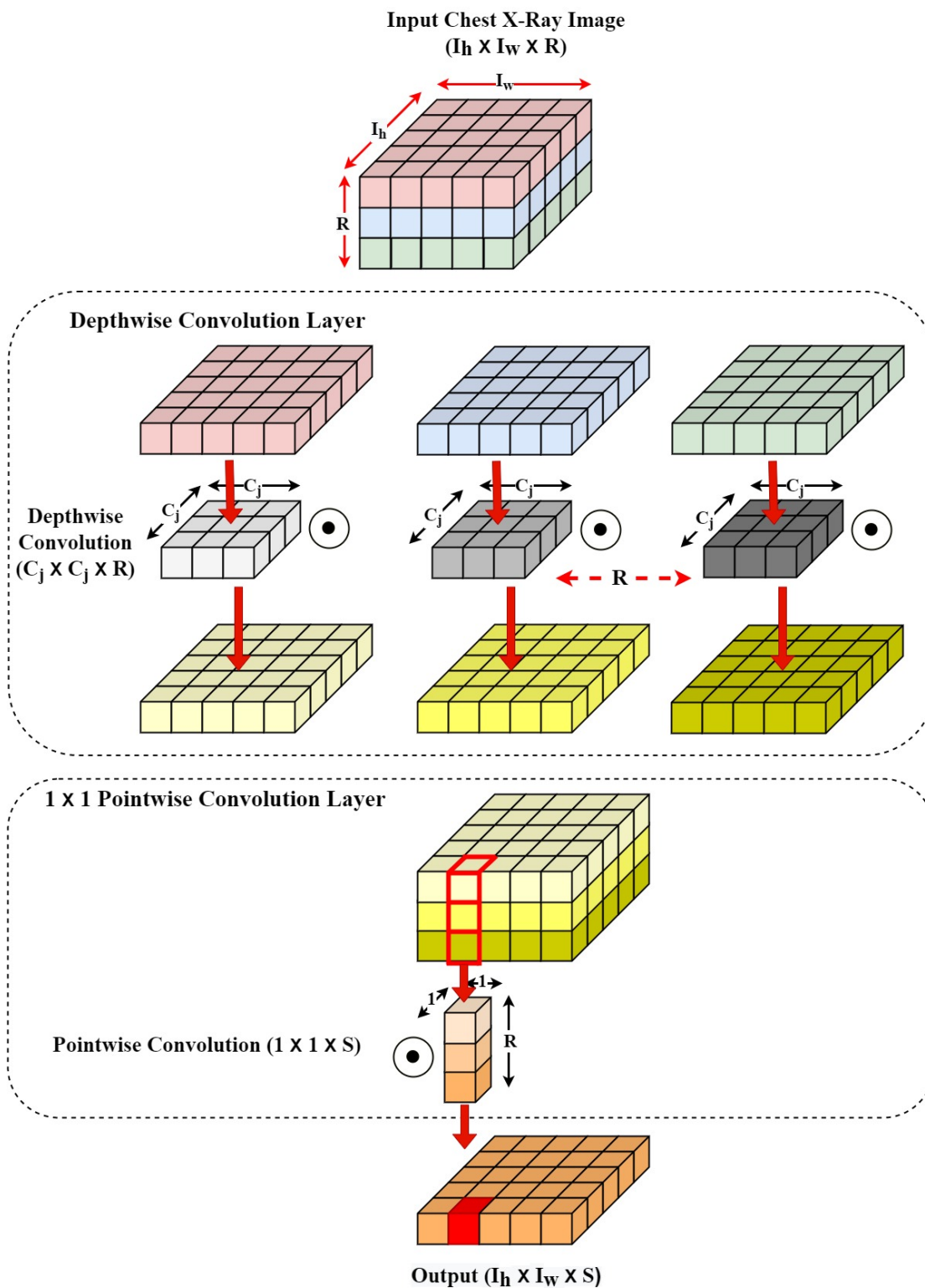


Figure 5.4: Overall operation of Depthwise Separable Convolution Neural Network (DS-CNN)



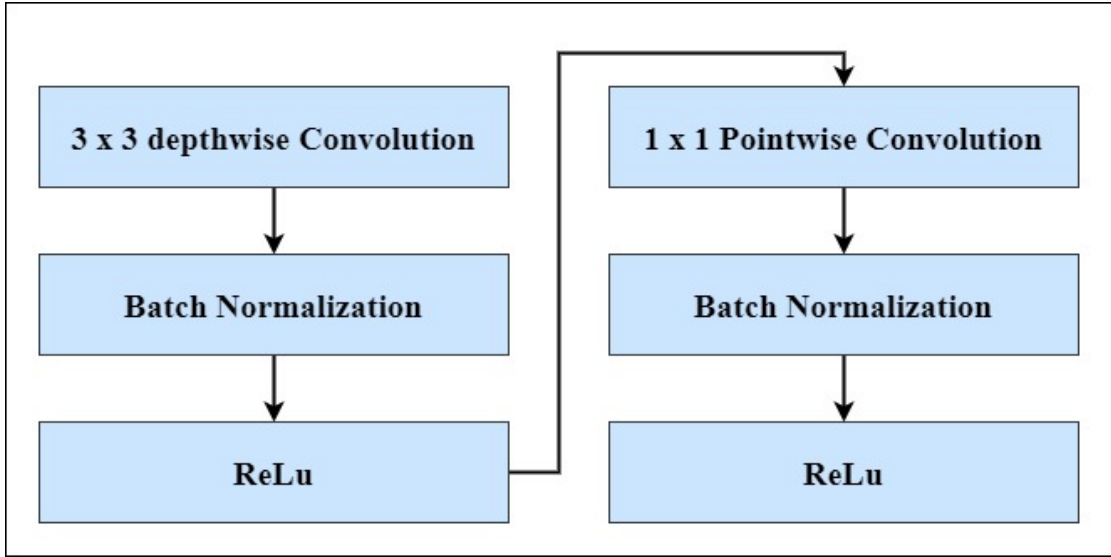


Figure 5.5: General process flow of the DS-CNN followed by Batch Normalization and ReLU

layers are further passed through the first full convolution layer to produce a new feature representation. The obtained feature is then passed to a series of depthwise separable convolution where  $3 \times 3$  depthwise convolution kernel is applied to each channel of the feature map, and further  $1 \times 1$  is used to combine the features obtained from the previous depthwise convolution operation. Usually, the standard convolution layer filters and combines the input into a new feature map in one step. In contrast, the depthwise separable convolution splits it into two stages of filtering and integrating the input into feature maps. This factorization process has the effect of substantially reducing the computation and the model size. After every  $1 \times 1$  pointwise convolution layer, we have added BN and ReLU layers to speed up training and enhance the network's generalization capability (Gu *et al.*, 2018). Finally, we apply global average pooling to obtain the feature map of size 1024. We represent the final imaging feature obtained as  $M_x = \{x_1, x_2, x_3, \dots, x_{1024}\}$ .

#### 5.2.4 Fully Connected Deep Neural Network for Abnormality Prediction

The Multi-scale in-depth features obtained from DS-CNN are flattened into a single dimension and ingested into fully connected DNN or dense layers to predict abnormalities from the Input CXR. In a fully connected DNN, every node or neuron in one layer is connected to every other neuron in the previous layer. The

main functionality of a fully connected DNN is to take flattened features obtained from the MSDL and DS-CNN as input and predict whether pulmonary disease exists or not in a diagnostic CXR. Every value from the flattened set of features obtained from MSDL and DS-CNN indicates the probability of that feature fitting into a particular category (i.e., disease or no disease). Hence, the fully connected DNN predicts and decides whether the diseases exist or not wholly based on the probabilities in the feature set. In our experiment, we used a three-layered DNN with two hidden layers of 256 and 128 units of neurons, followed by the output layer for binary predictions. Pictorially, the fully connected DNN for abnormality prediction is presented in Figure 5.6.

Let  $M_x = x_1, x_2, x_3, \dots, x_n \in \mathbb{R}^n$  be the flattened medical features obtained from the DS-CNN and input to the fully connected DNN. Let  $Z_j$  be the  $j^{th}$  output obtained from each layer and hence,  $Z_j$  can be calculated as follows:

$$Z_j = \phi(W_1 \cdot x_1 + W_2 \cdot x_2 + \dots + W_n \cdot x_n) \quad (5.15)$$

In the Eq. 5.15, the  $\phi$  represents the non-linear activation function, and  $W_1, W_2, \dots, W_n$  indicates the weight parameters. We have used the ReLU (Agarap, 2018) activation function for the first two hidden layers and the Sigmoid (Narayan, 1997) activation function for the final binary output layer. We have applied dropout = 0.2 to eliminate any overfitting problems during the network training.

### 5.2.5 Disease Visualization using Grad-CAM Technique

The MSDL and DS-CNN layers combined extract the multi-scale features from the input CXR. The features retrieved are given as an input to the fully connected DNN to convert these discriminative features into the probability score pertaining to both classes at the Softmax Layer. The class with the highest probability score will lead to the final prediction outcome (i.e., pulmonary disease present or not). Gradient Class Activation Map (Grad-CAM) is a mechanism used to generate the heatmap related to a particular class (Selvaraju *et al.*, 2016). The Grad-CAM provides a mechanism to check the decision model's transparency by localizing the abnormal image regions and makes our proposed model explainable by allowing us to understand the model's ability to arrive at a particular decision. Grad-CAM takes the gradients (or weights) from the final layers of DS-CNN and uses a heatmap to highlight the critical regions in the CXR for prediction. The areas with the highest gradient weights significantly impact the prediction result. Back

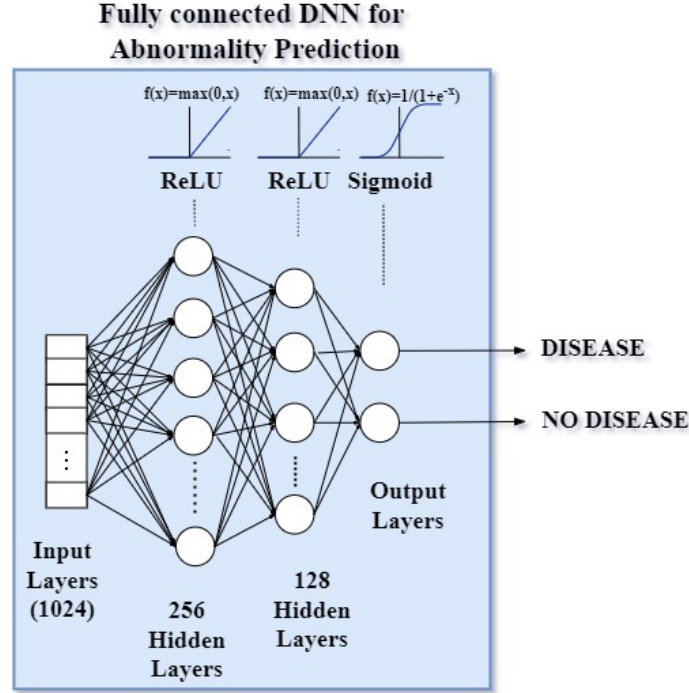


Figure 5.6: Fully Connected Deep Neural Network for abnormality prediction

propagating is computed with pulmonary disease = 1 and no pulmonary disease = 0, and the Global Average Pooling (GAP) (Lin *et al.*, 2014) of the gradient for every possible channel is calculated as follows:

$$Y_d = \frac{1}{f_H \times f_W} \sum_{l=1}^{f_H} \sum_{m=1}^{f_W} w_{i(l,m)} \quad (5.16)$$

In Eq. 5.16,  $Y_d$  represents the  $d^{th}$  one-dimensional feature after performing the GAP operation,  $f_H$  and  $f_W$  denotes the height and width of the two-dimensional activation map, respectively, and  $w_i$  is the  $i^{th}$  feature map at position (l,m) obtained from the DS-CNN. The updated weights are multiplied and added to the activation map. The output scores of both classes (i.e., disease and no disease) are computed as follows:

$$Score_C = \frac{1}{f_H \times f_W} \sum_j W_j^C F_j \quad (5.17)$$

Where,  $Score_C$  denotes the score of the proposed network in class  $C$ ;  $f_H$  and  $f_W$  denotes the height and width of the two-dimension activation map, respectively;  $W_j^C$  is the weight of the  $j^{th}$  activation map in class  $C$ , and  $F_j$  is the  $j^{th}$  activation map. The class discrimination positioning map is produced by comput-

ing the gradient between the score of the proposed network in class  $Score_C$  and the activation map  $F_j$  as follows:

$$\nabla_j^C = \frac{\partial Score_C}{\partial F_j} \quad (5.18)$$

Here,  $\nabla_j^C$  represents the gradient of the  $j^{th}$  activation map. The final sum produced is passed to ReLU to generate the Grad-CAM image.

$$HM^C = ReLU\left(\sum_j \nabla_j^C F_j\right) \quad (5.19)$$

Where,  $HM^C$  denotes the normalized heat map of class  $C$ . The detailed visual explanation of the proposed UM-VES for abnormality prediction in diagnostic CXR images using Grad-CAM is depicted in Figure 5.7.

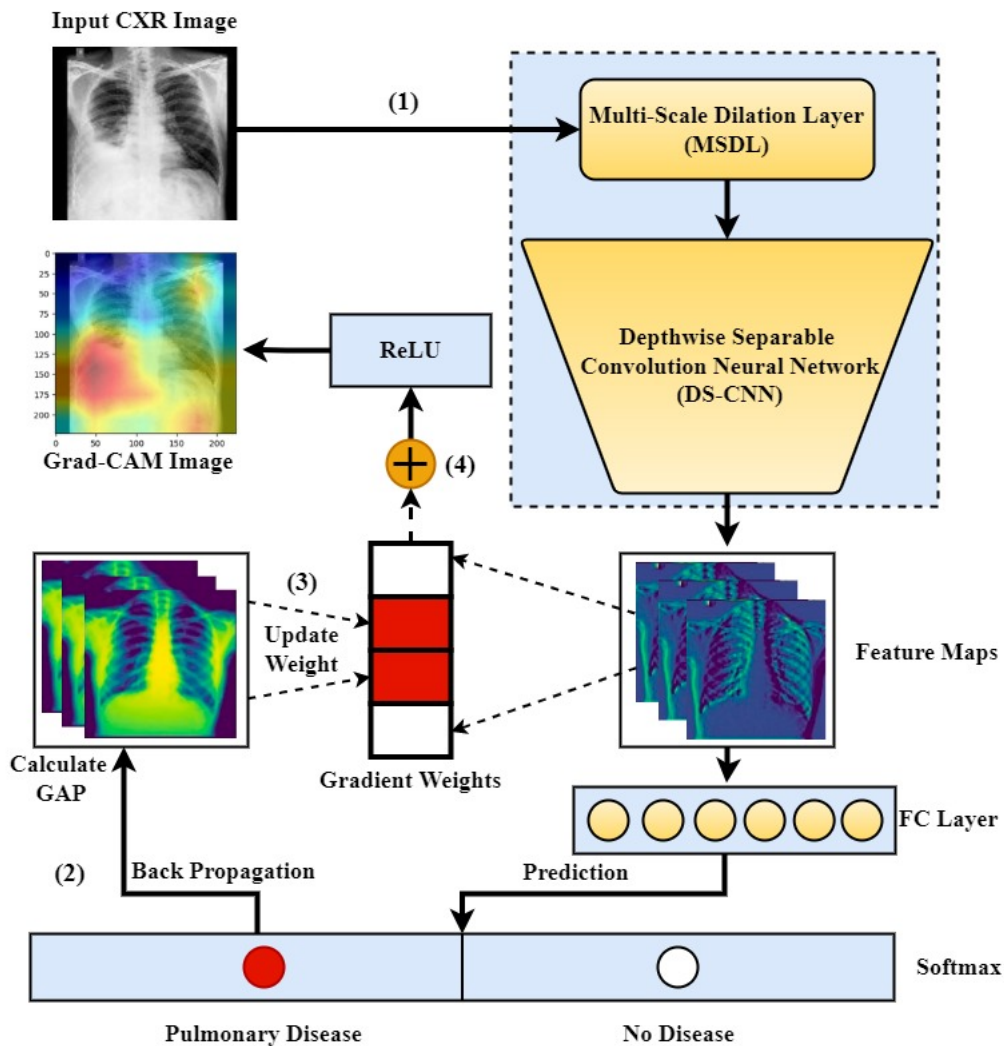


Figure 5.7: A visual explanation of the proposed UM-VES for abnormality prediction in diagnostic CXR images using Gradient-weighted Class Activation Mapping (Grad-CAM). (1) The CXR image is given as input to the network, and then prediction output is obtained by passing through the proposed deep learning network. (2) Back propagation is computed with Pulmonary Disease = 1 and No Pulmonary Disease = 0. (3) Calculating the GAP of the gradient for every possible channel and the gradient weights are updated for the proposed network. (4) The Grad-CAM is generated by multiplication and addition of weights to the activation map and ingesting the sum to the ReLU.

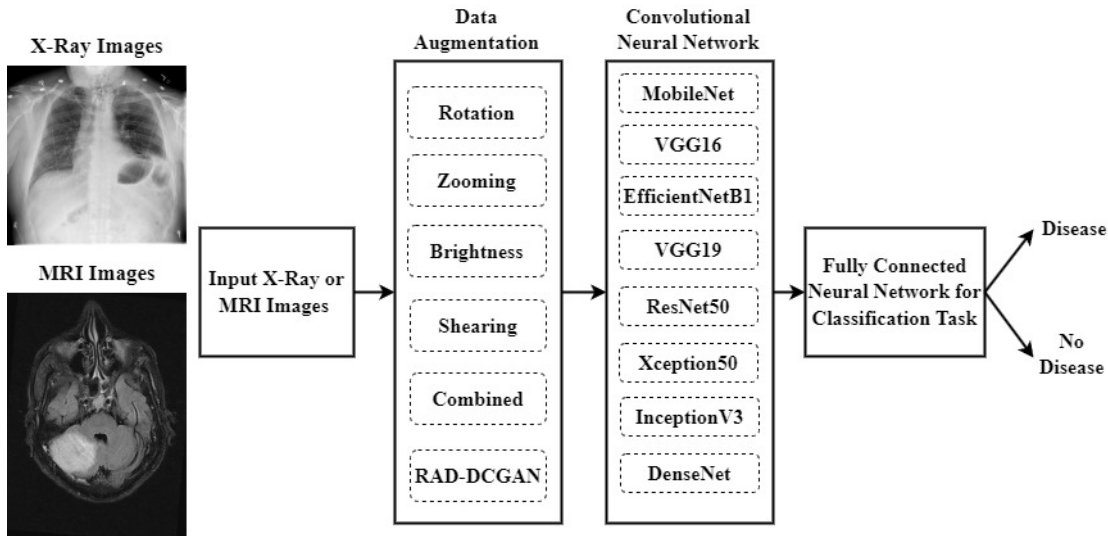


Figure 5.8: Schematic representation of the architecture used in this study for disease classification of radiology images using RAD-DCGAN and traditional data augmentation techniques.

### 5.3 Data Augmentation vs. Synthetic Data Generation: An Empirical Evaluation for Enhancing Radiology Image Classification

In this section, we propose Radiology Deep Convolutional GAN (RAD-DCGAN) inspired by DCGAN (Radford *et al.*, 2015) for performing data augmentation tasks that mainly enhance the performance of deep CNN classifiers. The schematic representation of the architecture used in this study for disease classification of radiology images using RAD-DCGAN and traditional data augmentation techniques is presented in Figure 5.8. We comprehensively analyze the proposed RAD-DCGAN with the basic data augmentation strategies for radiology X-ray and MR images. To begin with, the X-Ray or MR images are given as a separate input to the system, where they go through a series of data augmentation and synthesis processes. In addition, we aim to examine the effectiveness of the RAD-DCGAN method in contrast to traditional data augmentation methods by implementing different conventional deep learning approaches. Subsequently, we utilize a fully connected deep neural network to classify the diagnostic images into two categories (i.e., disease and no disease).

### 5.3.1 Basic Data Augmentation

We perform random (yet realistic) geometrical translations to the original image to enhance the diversity of the training samples. This process of transformation, called data augmentation, is employed to enhance the effectiveness of machine learning or deep learning models and prevent any overfitting problems. In the medical domain, gathering huge samples of medical data is not viable as manual data annotation needs an expert clinician's opinion and is time-consuming. Data augmentation comes to the rescue in a data-scarce situation by enhancing the cohort size through random transformation and introducing variability in the cohort. We have employed basic data augmentation techniques, including zooming, brightness, rotation, and shearing, for the spatial transformation of the X-rays and MRIs. The various data augmentation techniques applied are depicted in Figure 5.9.

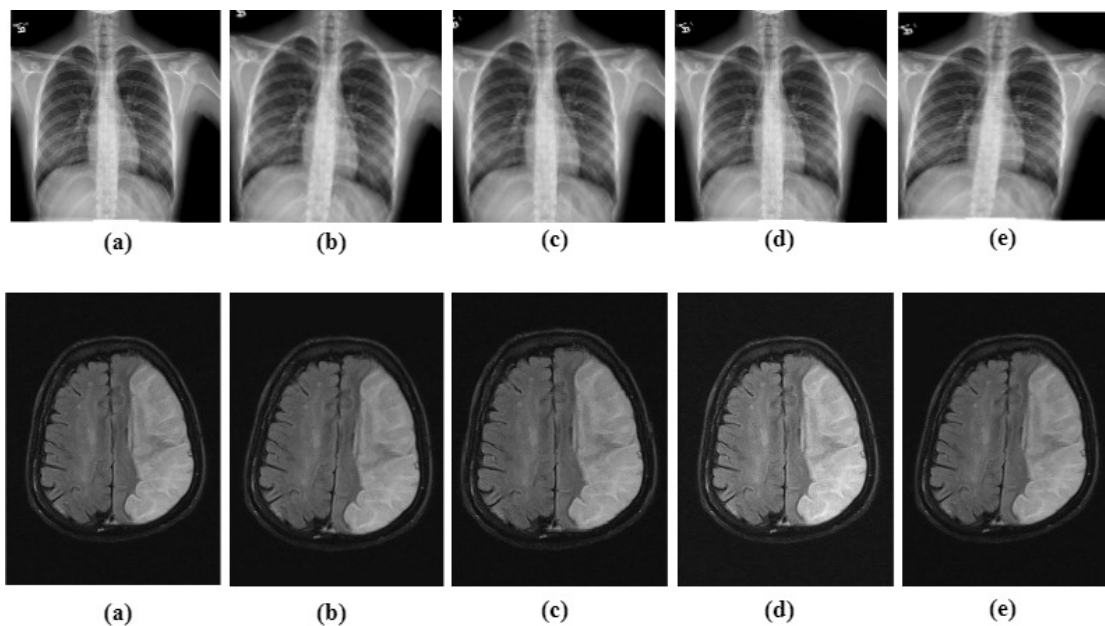


Figure 5.9: Basic data augmentation techniques applied on X-ray and MR images: (a) original X-ray and MR images, (b) rotated images, (c) zoomed images, (d) after increase in brightness and (e) sheared images

1. **Rotation:** The rotation technique allows us to rotate the MR and X-ray images by a certain degree. The degree of rotation should be carefully applied; otherwise, there is a possibility of obtaining upside-down images, which is unlikely to be seen in healthcare settings. In this study, We have used rotation degrees between  $-5$  and  $+5$ .

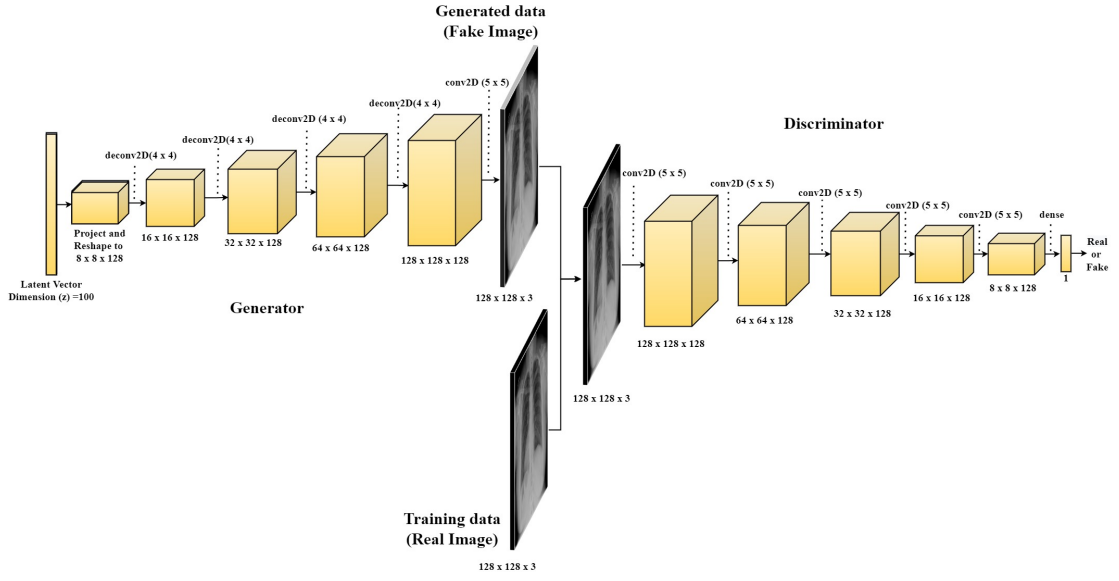


2. **Zooming:** The zooming technique is utilized to produce images with varying zoom levels. It either allows to zoom in on the image or it will enable adding extra pixels around the image to enlarge it. We have used the zoom range of 0.95, meaning the zoom-in of 5% is used.
3. **Brightness:** The brightness technique allows us to either increase the pixel value to result in a brighter image or reduce the pixel value to obtain darker images. For our experiment, we utilized the brightness range between 0.5 and 1.5.
4. **Shearing:** The primary purpose of the shearing is to give the model the ability to learn images from all angles and to give it the human perspective of viewing images from various angles. In this research, we have utilized the shear range between -5 and +5.
5. **Combined Augmentation:** We have incorporated a combination of data augmentation techniques, including rotation, brightness, zooming, and shearing. To assess their effectiveness, we trained a standard deep learning model on the data generated through the individual augmentation techniques and the combined augmentation strategy.

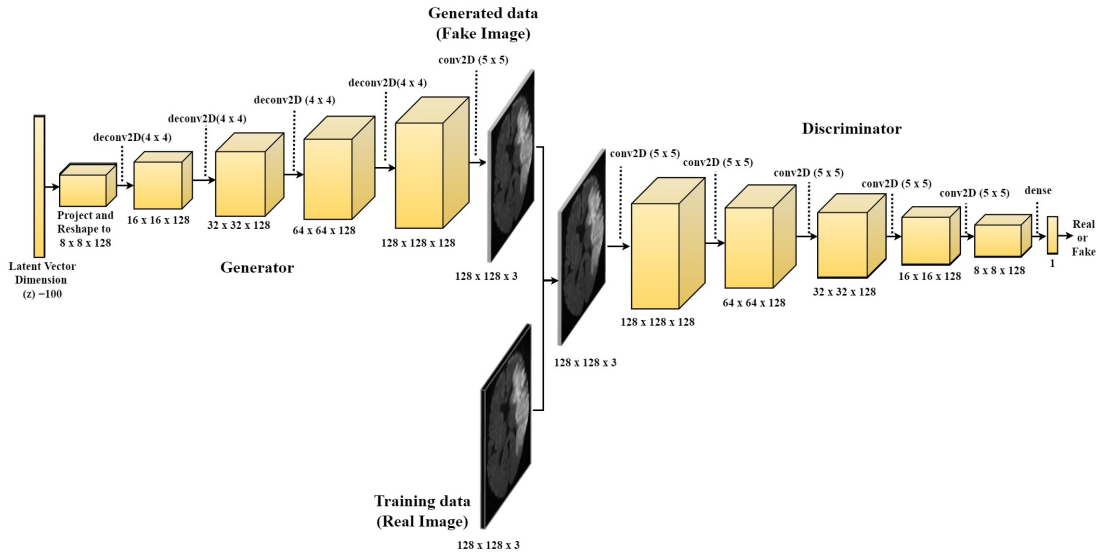
### 5.3.2 RAD-DCGAN for Synthetic Data Generation

Synthetic data can be defined as artificially generated data that mimics the original data. In the medical domain, synthetic data is very beneficial in solving data scarcity and would fast-track the time and energy required to collect/annotate the large cohort of medical data. [Goodfellow \*et al.\* \(2014a\)](#) presented GANs comprising two neural networks: generator and discriminator. The generator network creates a synthetic image mimicking the actual image by inputting random noise into the generator module. Whereas the discriminator network categorizes images into real images (i.e., the original image) and fake images (i.e., a synthetic image produced by a generator module). The GAN model generates realistic images by capturing the distribution of real images from the training set. It is hard to differentiate the synthetic image produced from the actual image. The RAD-DCGAN is a variation of the GAN model that can generate synthetic images from the radiology cohort. The proposed RAD-DCGAN for synthetic image generation from radiology images is shown in [Figure 5.10](#). The RAD-DCGAN contains two main components: the generator module and the discriminator module.





(a) for radiological X-Rays



(b) for radiological MRIs

Figure 5.10: The proposed RAD-DCGAN for synthetic image generation from radiology images.

1. **Generator Network** To begin with,  $100 \times 1$  random noise vector is ingested into a generator network and is reshaped into  $8 \times 8 \times 128$  by feeding it to a dense layer. Additionally, the outcome from the dense layer is sent through a sequence of four de-convolution layers (also known as convolution-transpose layers) to create the upsampled feature maps, which results in the synthetic image of size  $128 \times 128 \times 3$ . We apply the Leaky ReLU activation function to all four de-convolution layers, and the hyperbolic tangent (tanh) is employed for the final convolutional layer. It is seen that using bounded activation

functions like leaky ReLU and tanh allows the RAD-DCGAN to saturate swiftly while learning features and allowing it to cover the color space of the training distribution. To stabilize the training process, we have applied batch normalization to all four de-convolution layers. Finally, a synthetic image with size  $128 \times 128 \times 3$  is obtained as an output from the generator module. Figure 5.11a displays the comprehensive framework of the generator module.

- Discriminator Network** The detailed framework of the discriminator network is depicted in Figure 5.11b. The primary motto of the discriminator is to categorize the generated MR or X-ray image as real or fake. The original MR or X-ray image with a size of  $128 \times 128 \times 3$  is ingested into the discriminator module along with the synthetic image produced from the generator network. In the discriminator network, the four convolution operations are performed, and finally, the sigmoid activation function is employed to categorize the radiology image as real or fake. The leaky ReLU activation function, followed by batch normalization, is utilized with four convolution operations of the discriminator network.

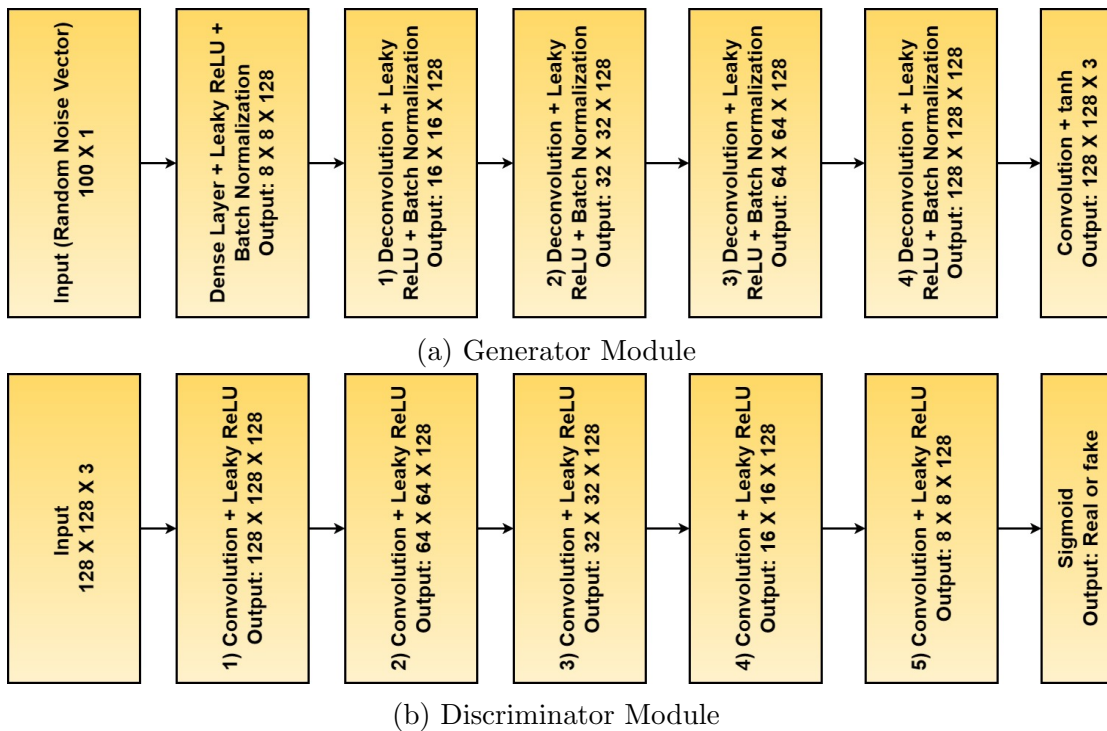


Figure 5.11: General architecture of generator and discriminator module of RAD-DCGAN

### 5.3.3 Objective Function of RAD-DCGAN

The objective function of RAD-DCGAN is to narrow the gap between the probability distribution of original and synthetic radiology images. In this research, we have utilized minimax loss (Goodfellow *et al.*, 2014a) as depicted in Eq. 5.20. The minimax loss allows the loss function to be reduced for the generator module in the proposed RAD-DCGAN, whereas the same loss is maximized in the discriminator module. In the RAD-DCGAN, the generator and discriminator modules are trained simultaneously, similar to an analogy of a min-max game, where the two players play opposing roles (i.e., the generator and discriminator module) with the value function  $V_R(Df, Gf)$ .

$$\min_{Gf} \max_{Df} V_R(Df, Gf) = \mathbb{E}_{y \sim p_{rad}(y)} [\log(Df(y))] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - (Df(Gf(z))))] \quad (5.20)$$

Here,  $y$  represents the original radiology image and  $Df(y)$  denotes the probability that the  $y$  originated from the initial data distribution and not from the generated data distribution.  $\mathbb{E}_{y \sim p_{rad}(y)}$  is the expected value over the actual samples of radiology images  $y$  and  $\mathbb{E}_{z \sim p_z(z)}$  represents the expected value over all the generated synthetic samples. The  $p_z(z)$  denotes the random noise variable that is given as the input to the generator module and  $Gf(z)$  is a differentiable generator function used to map to the data space.

The generators distribution  $P_g$  is learnt over the actual radiology data  $y$  by establishing an input noise variable  $P_z(z)$  and mapping it to to data space  $Gf(z)$ . The discriminator function  $Df(y)$  denotes the likelihood of  $y$  coming from actual data rather than  $P_g$ . The discriminator is trained so that the correct labels are allocated to training cases and the synthetic cases produced by  $G$ . The generator is simultaneously trained so that the  $\log(1 - (Df(Gf(z))))$  is minimized.

### 5.3.4 Loss Function of RAD-DCGAN

The main function of the discriminator network is to differentiate between the artificial radiology images (i.e., the synthetic image produced by the generator network) and the real radiology images (i.e., actual images from the training sample). The basic job of the discriminator is binary classification, so as a loss function, we have incorporated binary cross-entropy. Eq. 5.21 represents the binary cross-entropy:

$$J_{BCE}(w) = \frac{1}{R} \sum_{r=1}^R [l_r \times \log(h_w(y_r)) + (1 - l_r) \times \log(1 - h_w(y_r))] \quad (5.21)$$

Where  $R$  is the total count of radiology image samples (i.e., X-ray or MR images) for training in mini-batch (i.e., splitting the training set into small batches),  $l_r$  represents the target label for the training sample  $r$ . The target label for the actual sample is 1, and for the synthetic sample is 0. The  $y_r$  denotes the input training sample  $r$ , and  $h_w$  represents the neural network model with the weights  $w$ . In the Eq. 5.21, the summation indicates the average cost of overall samples in an entire batch  $R$  in the radiology cohort. The  $l_r \times \log(h_w(y_r))$  represents the multiplication of the actual label  $l_r$  and the logarithm of the prediction obtained. The predicted features obtained by the RAD-DCGAN are denoted by  $h_w(y_r)$ , and for instance, the loss is 0 (i.e.,  $-\log(1)$ ), when the training model produces output 1, which is the ideal case for predicting the radiology image sample to be real by penalizing false negatives. Whereas,  $(1 - l_r) \times \log(1 - h_w(y_r))$  penalizes false positive cases in the model output.

## 5.4 Experimental setup

This section offers a comprehensive overview of various aspects of our study, including parameter configurations, the selection of the radiology cohort, techniques used for data augmentation, and the evaluation metrics employed.

### 5.4.1 Parameter Configurations of Proposed UM-VES and State-of-the-Art Deep Learning Models

For our experimental analysis, we have utilized the NVIDIA Tesla M40 server with the following hardware specifications: 128GB RAM, 24GB GPU, 3TB HD, and Linux server OS. We have used Python 3.6 with open-source software Keras and the Tensorflow library (Abadi and et. al., 2015). The open-I and data collected from KMC private hospitals are divided into training/validation, and test sets as given in the Table. 5.3. The proposed UM-VES is trained for 20 epochs for 10-cross fold validations. The overall layer-wise hyperparameter information of the UM-VES is presented in Table 5.1. The UM-VES consists of MSDL with three-channel parallel dilation convolution layers with a dilation factor,  $d_r = 1, 2, 3$ . We have employed the grid search approach (Bergstra and Bengio, 2012) to select the

optimum hyperparameters for our proposed model and the state-of-the-art deep learning models.

Table 5.2: Parameter details of all the state-of-the-art Deep Learning Models and the proposed UM-VES

<b>Models</b>	<b>Total Parameters (in Millions)</b>
MobileNet	3.2289
VGG16	14.7147
EfficientNetB1	6.5752
VGG19	20.0244
ResNet50	23.5877
Xception	20.8615
InceptionV3	21.8028
DenseNet121	25.1283
<b>Proposed UM-VES</b>	<b>4.8105</b>

After fine-tuning the hyperparameters, the learning rate of 0.001 has been used, and the stochastic gradient descent-based Adam optimizer is leveraged. In the proposed UM-VES, the CXR image of size  $150 \times 150$  is passed as an input to the network, and the multi-scale feature of size 1024 is produced through the global average pooling layer. Further, the output clinical features are ingested into a fully connected DNN, where two hidden layers of 256 and 128 units are used with the ReLU activation function. Finally, the softmax activation function is applied in the third dense layer with two units for binary abnormality prediction from CXR. The dropout probability (Srivastava *et al.*, 2014) of 0.2 and the early-stopping strategy (Yao *et al.*, 2007) are employed to avoid the overfitting of the proposed UM-VES. The proposed UM-VES and the state-of-the-art deep learning models are initialized with the ImageNet pre-trained weights (Deng *et al.*, 2009) and later retrained on the Open-I and KMC cohorts. Usage of ImageNet pre-trained weights addresses the problem of the enormous dataset needed for deep learning training. The parameter details of all the state-of-the-art Deep Learning Models and the proposed UM-VES are shown in Table 5.2. The proposed model is lightweight, like MobileNet and EfficientNetB1, which have mobile-centric applications.

### 5.4.2 Radiology Cohort Selection

For our experiment, we have utilized two radiology cohorts: 1) Publicly available Open-I or IU dataset (Demner-Fushman *et al.*, 2015), 2) Data collected

Table 5.3: Dataset Statistics: Detailed description of the CXR diagnostic images from two medical repositories

Dataset Description	Open-I Cohort	KMC Cohort
Tot. # of CXR images	3996	502
Tot. # of CXR images after removal of missing reports	3638	502
Tot. # of CXR after standard data augmentation	6229	1498
Tot. # of Training/Validation Set	5606	1348
Tot. # of Test Set	623	150
Tot. % of Normal cases (i.e., No Pulmonary diseases)	38%	52%
Tot. % of Abnormal cases (i.e., Pulmonary diseases)	62%	48%

from the KMC private hospital (Mangalore, India). The data collected from the KMC private hospital was de-identified, and approval from the Institutional Ethics Committee (IEC) was granted to use the dataset for research purposes. The detailed statistics and descriptions of the two medical repositories are presented in the Table. 5.3. Both the radiology cohorts are categorized into “normal” (i.e., CXR images with no pulmonary or chest diseases) and “abnormal” (i.e., CXR images with pulmonary diseases like Pulmonary Atelectasis, pulmonary fibrosis, pulmonary edema, etc.). Most of the existing research on the Open-i dataset deals with cross-modal retrieval tasks to generate a radiology report from CXR images (Jing *et al.* (2017); Liu *et al.* (2019a); Xue *et al.* (2018)). After a thorough survey, it is observed that limited study is carried out on classification and prediction tasks. In this regard, we have refined the dataset according to the classification and prediction tasks. The CXR images in the Open-I cohort consist of associated radiology reports with findings, impressions, indications, and Medical Subject Heading (MeSH). MeSH comprises the specific details pertaining to the diseases, and we have extracted the ground-truth annotations from the MeSH. The annotations are validated to check their correctness by experienced radiologists. Also, to evaluate the performance of the proposed UM-VES model, comprehensive benchmarking is performed and compared with various state-of-the-art deep learning models. The experienced radiologists manually annotated the dataset collected from KMC Hospital as per the gold standards (Wissler *et al.*, 2014).

### 5.4.3 Data Augmentation Settings

For our experiment, we have utilized two radiology cohorts: 1) Publicly available Open-i (IU) dataset (Demner-Fushman *et al.*, 2015), 2) Data collected from the KMC private hospital (Mangalore, India). The detailed statistics of the dataset are presented in the Table. 5.3. Considering the limited dataset, which may lead to an overfitting problem when passed through the proposed deep learning framework with multiple iterative layers. Data augmentation strategies are applied to resolve these shortcomings. The data augmentation process is applied to the CXRs just before the training process to improve the performance of the proposed model by preventing overfitting. Chest X-rays are relatively sensitive to the different geometric transformation operations as they might introduce new outliers; hence, careful adoption of data augmentation techniques is needed. We have applied a series of data augmentation techniques like rotation, zooming, brightness, and shearing for image augmentation. The process flow of the various data augmentation pipeline is shown in Figure 5.12. The detailed image augmentation settings of various augmentation strategies applied to diagnostic CXRs are presented in Figure 5.4. In this study, we have incorporated augmentor (Bloice *et al.*, 2017a), a python toolkit for image augmentation to increase the size of the medical cohort.

Table 5.4: Image augmentation settings

Augmentation Strategies	Value
Rotation range	[-5, +5]
Zoom range	0.95
Shear range	[-5, +5]
Brightness range	[0.5, 1.5]

### 5.4.4 Evaluation Criteria

We have used six standard evaluation criteria: Accuracy (Acc.), Precision (Pre), Recall (Rec.), F1-Score(F1), MCC and AUROC to examine the performance of the proposed UM-VES on the two medical CXR cohorts. Section 4.4.3 provides a comprehensive explanation of the evaluation metrics chosen along with the rationale behind their selection.



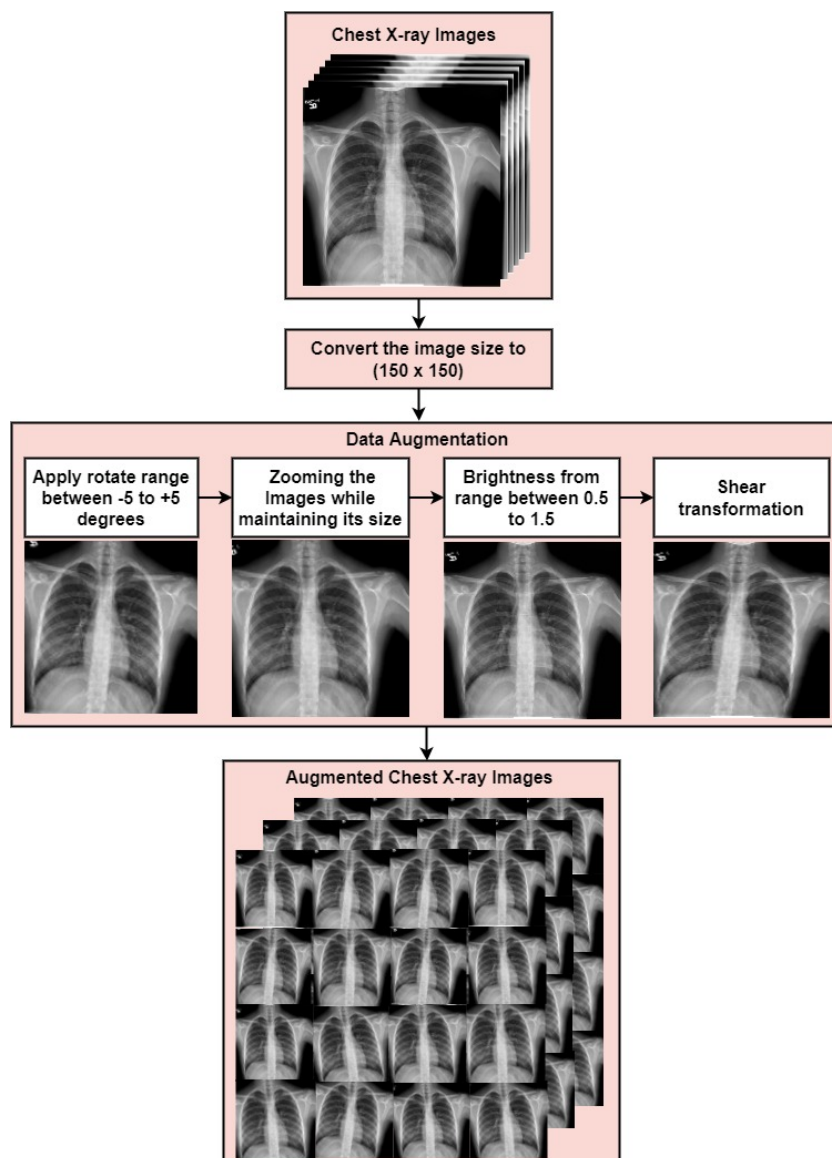


Figure 5.12: Systematic data augmentation process flow of diagnostic CXRs

## 5.5 Results and Discussions

This section highlights the experimental analysis of the proposed UM-VES. We have compared the proposed model with state-of-the-art Deep Learning models. Also, we have compared the result of the proposed model with the existing work on the Open-I dataset. We have also showcased the qualitative analysis of the proposed UM-VES model by visualizing and localizing the abnormalities in the chest regions.



### 5.5.1 Quantitative Analysis of Proposed UM-VES with the Fine-tuned Pre-trained Deep Learning Models

The detailed quantitative analysis of the proposed UM-VES model is performed, and the results are compared with the State-of-the-art Deep Learning frameworks for the publicly available Open-I Dataset and the real-time diagnostic data collected from KMC Hospital (refer Table 5.5 and Table 5.6). The graphical representation depicting the performance analysis of the proposed UM-VES with the different baseline deep learning models for Open-I and KMC CXR datasets is shown in Figure 5.15 and Figure 5.16. The proposed model has achieved consistent performance for accuracy, precision, recall, F1-score, MCC, and AUROC. For both Open-I and KMC hospital cohorts, the model performs better than the existing pre-trained state-of-the-art deep learning models like MobileNet, VGG16, EfficientNetB1, VGG19, ResNet50, Xception, InceptionV3, and DenseNet121. It is evident from Table 5.5 and Table 5.6 that the MSDL layer considerably impacts performance by obtaining a broad receptive field and capturing multi-scale features. The proposed UM-VES model achieves significantly higher precision and recall compared to the other baseline models. This shows that our proposed model is able to decrease false positive and false negative predictions. The F1-score and MCC of the proposed model are high compared to other State-of-the-art models, indicating that our proposed model can effectively classify even though there is a class imbalance. The proposed UM-VES model has achieved a higher AUROC of 0.8572 and 0.8793 for Open-I and KMC datasets compared to existing state-of-the-art deep learning models, indicating that the model can better distinguish between pulmonary disease and no disease from the CXRs. Other lightweight deep learning networks like MobileNet and EfficientNetB1 have also achieved promising results for both Open-I and KMC hospital datasets.

It is also seen in Table 5.2 that the proposed UM-VES requires only 4.8105 million training parameters. The UM-VES is lightweight and five times smaller compared to the extensively utilized DenseNet121 model (25.1283 million parameters) on the Open-I dataset for pulmonary disease classification (Zech *et al.* (2018); Aydin *et al.* (2019b); Wang *et al.* (2018a)). As a result, the training of the UM-VES is faster than most of the existing deep learning strategies like VGG16, VGG19, EfficientNetB1, ResNet50, Xception, InceptionV3, and DenseNet121. The proposed UM-VES model utilizes comparatively shallow architecture, consisting of fewer layers than other baseline deep learning models. However, the proposed model outperforms the existing state-of-the-art models, which have deeper archi-

Table 5.5: Benchmarked Experimental results of proposed UM-VES Model with the state-of-the-art Deep Learning Model on Open-I Dataset.

<b>Models</b>	<b>Acc.</b>	<b>Pre.</b>	<b>Rec.</b>	<b>F1</b>	<b>MCC</b>	<b>AUROC</b>
MobileNet	0.7675	0.7670	0.767	0.7668	0.5339	0.8108
VGG16	0.6357	0.6361	0.64	0.64	0.5605	0.8418
EfficientNetB1	0.7805	0.7803	0.7801	0.7802	0.5605	0.8418
VGG19	0.6357	0.6361	0.64	0.647	0.2722	0.6357
ResNet50	0.7465	0.7436	0.745	0.746	0.492	0.7901
Xception	0.77	0.776	0.77	0.76	0.573	0.8109
InceptionV3	0.7473	0.7471	0.748	0.746	0.4993	0.8004
DenseNet121	0.7336	0.74	0.7354	0.7346	0.4688	0.8003
<b>Proposed UM-VES</b>	<b>0.7922</b>	<b>0.7926</b>	<b>0.7928</b>	<b>0.7927</b>	<b>0.5855</b>	<b>0.8572</b>

Table 5.6: Benchmarked Experimental results of proposed UM-VES Model with the state-of-the-art Deep Learning Model on KMC hospital Dataset.

<b>Models</b>	<b>Acc.</b>	<b>Pre.</b>	<b>Rec.</b>	<b>F1</b>	<b>MCC</b>	<b>AUROC</b>
MobileNet	0.7804	0.7801	0.7801	0.7803	0.5604	0.8228
VGG16	0.6623	0.6621	0.6623	0.6622	0.5731	0.8314
EfficientNet	0.7945	0.7943	0.7942	0.7941	0.5858	0.8330
VGG19	0.6642	0.6641	0.6641	0.6653	0.3822	0.6642
ResNet50	0.7657	0.7656	0.7656	0.7654	0.5102	0.8012
Xception	0.7821	0.7823	0.7822	0.7821	0.5168	0.8351
InceptionV3	0.7741	0.7743	0.7743	0.7741	0.4963	0.8103
DenseNet121	0.7511	0.7513	0.7513	0.7511	0.4826	0.8099
<b>Proposed UM-VES</b>	<b>0.8225</b>	<b>0.8201</b>	<b>0.8200</b>	<b>0.8200</b>	<b>0.6401</b>	<b>0.8793</b>

tectures, making our model less computationally expensive with reduced training time. Figure 5.13 and Figure 5.14 represents the experimental observation of the loss and accuracy vs. the total number of epochs w.r.t 10-fold cross-validation for the open-I and KMC hospital datasets. It is observed that the loss gradually drops after every epoch for all the folds, and the accuracy remains stable after a few initial variations. We have saved the model weights with the highest performance for every fold.

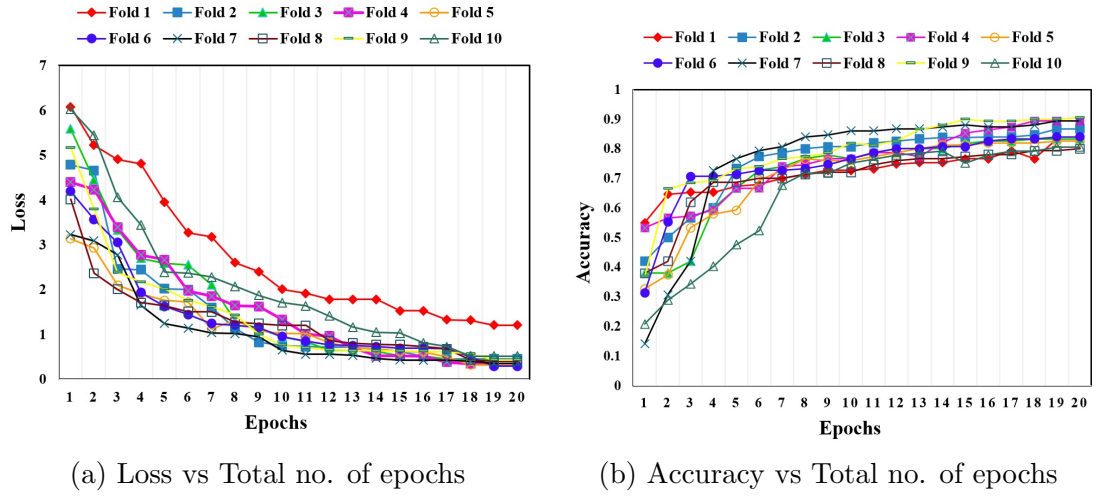


Figure 5.13: Experimental observation of the loss and accuracy vs total number of epochs w.r.t 10-fold cross-validation for Open-I X-ray dataset

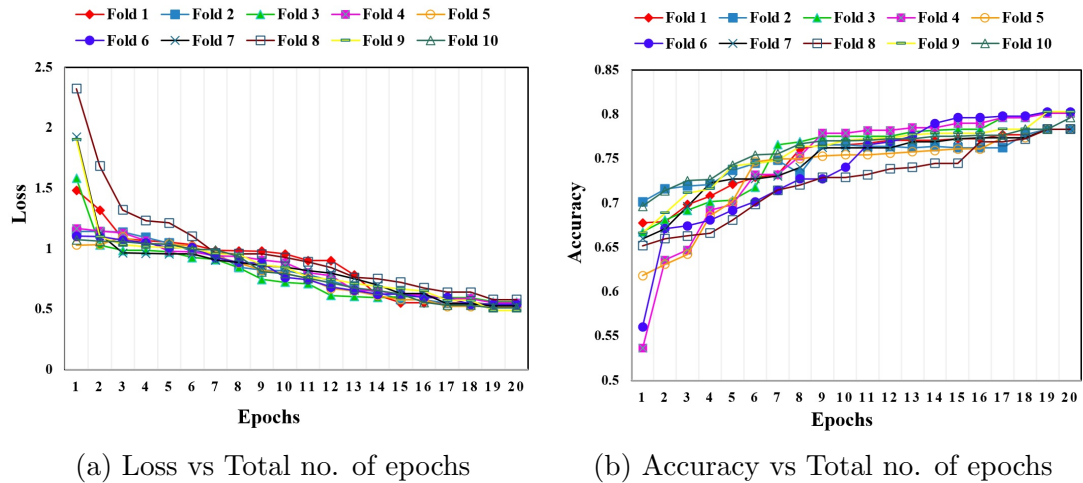


Figure 5.14: Experimental observation of the loss and accuracy vs total number of epochs w.r.t 10-fold cross-validation for KMC Chest X-ray dataset

### 5.5.2 Performance Analysis of Proposed UM-VES with the Existing State-of-the-art Deep Learning Strategies on Open-I Dataset

We have also compared the performance of the proposed UM-VES model with the existing benchmarked deep learning models on the Open-I dataset. After a comprehensive survey, we found four research papers using the Open-I dataset for the classification task. Table 5.7 presents the details of the evaluation metrics obtained from the existing research articles on the Open-I dataset compared with the proposed UM-VES. [Zech et al. \(2018\)](#), [Aydin et al. \(2019b\)](#), and [Lopez et al. \(2020\)](#)

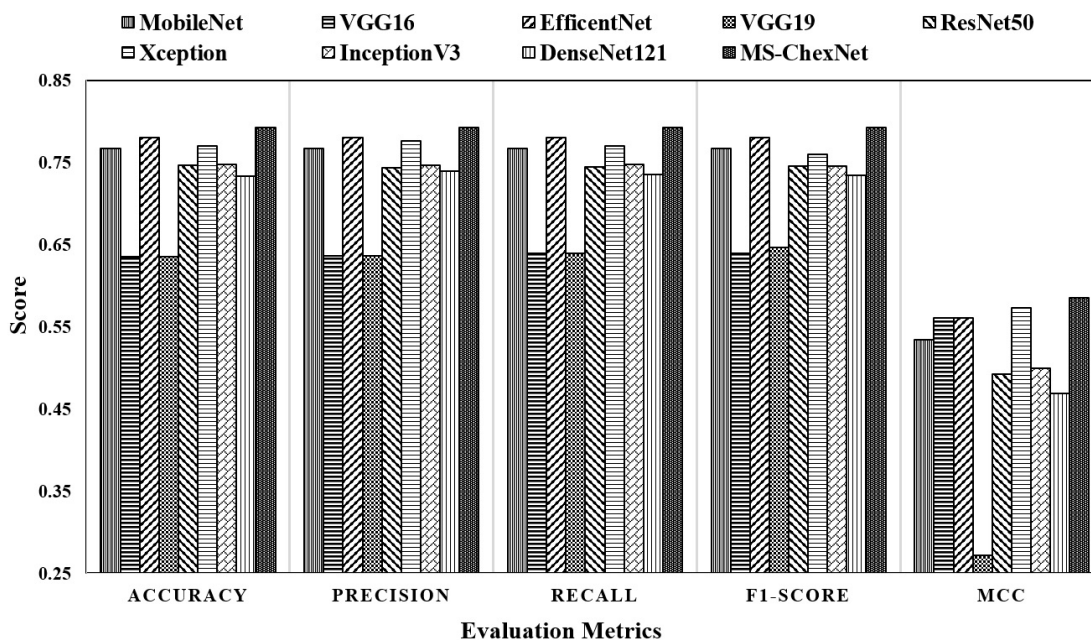


Figure 5.15: Performance analysis of proposed UM-VES with the different baseline deep learning model for Open-I CXR dataset

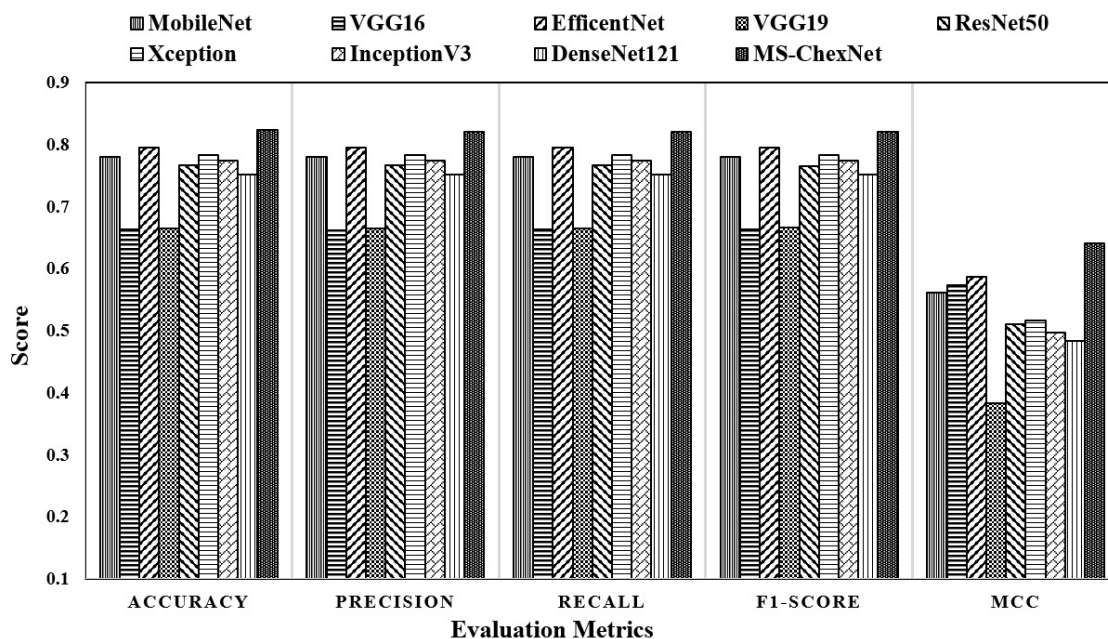


Figure 5.16: Performance analysis of proposed UM-VES with the different baseline deep learning model for KMC Hospital CXR dataset

presented a variation of the denseNet121 model, and it is observed that our proposed UM-VES model has achieved better performance with respect to accuracy, precision, recall, F1-Score, and AUROC. It is also observed that the existing works

have not considered all the standard evaluation metrics, which are essential while performing the prediction task on the Open-I dataset. Wang *et al.* (2018a) proposed a CNN-based model to predict pulmonary disease from the Open-I dataset and attained an AUROC of 0.741. It is found that the proposed UM-VES model has produced a higher AUROC of 0.8572, showcasing the impact of the MSDL layer on the performance of the model by obtaining a broader receptive field and capturing the multi-scale features for efficient prediction of pulmonary diseases.

Table 5.7: Performance analysis of the proposed UM-VES with the existing state-of-the-art deep learning strategies on Open-I Dataset

Reference	Acc.	Prec.	Rec.	F1	MCC	AUROC
Zech <i>et al.</i> (2018)	-	-	-	-	-	0.725
Aydin <i>et al.</i> (2019b)	0.74	-	-	-	-	-
Wang <i>et al.</i> (2018a)	-	-	-	-	-	0.741
Lopez <i>et al.</i> (2020)	-	0.52	0.42	0.46	-	0.61
<b>Proposed UM-VES</b>	<b>0.7922</b>	<b>0.7926</b>	<b>0.7928</b>	<b>0.7927</b>	<b>0.5855</b>	<b>0.8572</b>

### 5.5.3 Qualitative Analysis of Proposed UM-VES

Figure 5.17 depicts some sample qualitative results of disease visualization from CXR with the grad-CAM technique with its ground-truth label, and the radiologist highlighted CXR. The visualization techniques allow our proposed model to be explainable by iterating back and understanding the model’s ability to arrive at a decision. The Grad-CAM method (Selvaraju *et al.*, 2017) uses the gradient of the interesting concept in a given convolution layer. The main goal is to highlight the significant regions and generate a coarse localization map. The area with a red colour indicates the part of the model where attention is strong, and blue represents the part where attention is weak. The first four rows indicate the CXR with pulmonary abnormalities, and the last row shows the CXR with no abnormalities. For comparison purposes, we received localized and labelled CXRs from expert radiologists and compared them with the predicted CXRs from the proposed UM-VES Model. It is observed from the findings that the proposed UM-

VES model can reach a performance level similar to that of expert radiologists. We can suggest that the lightweight and explainable UM-VES model has the potential for preliminary examination of CXRs in radiology workflows to assist radiologists when resources are scarce and improve the overall prediction accuracy.

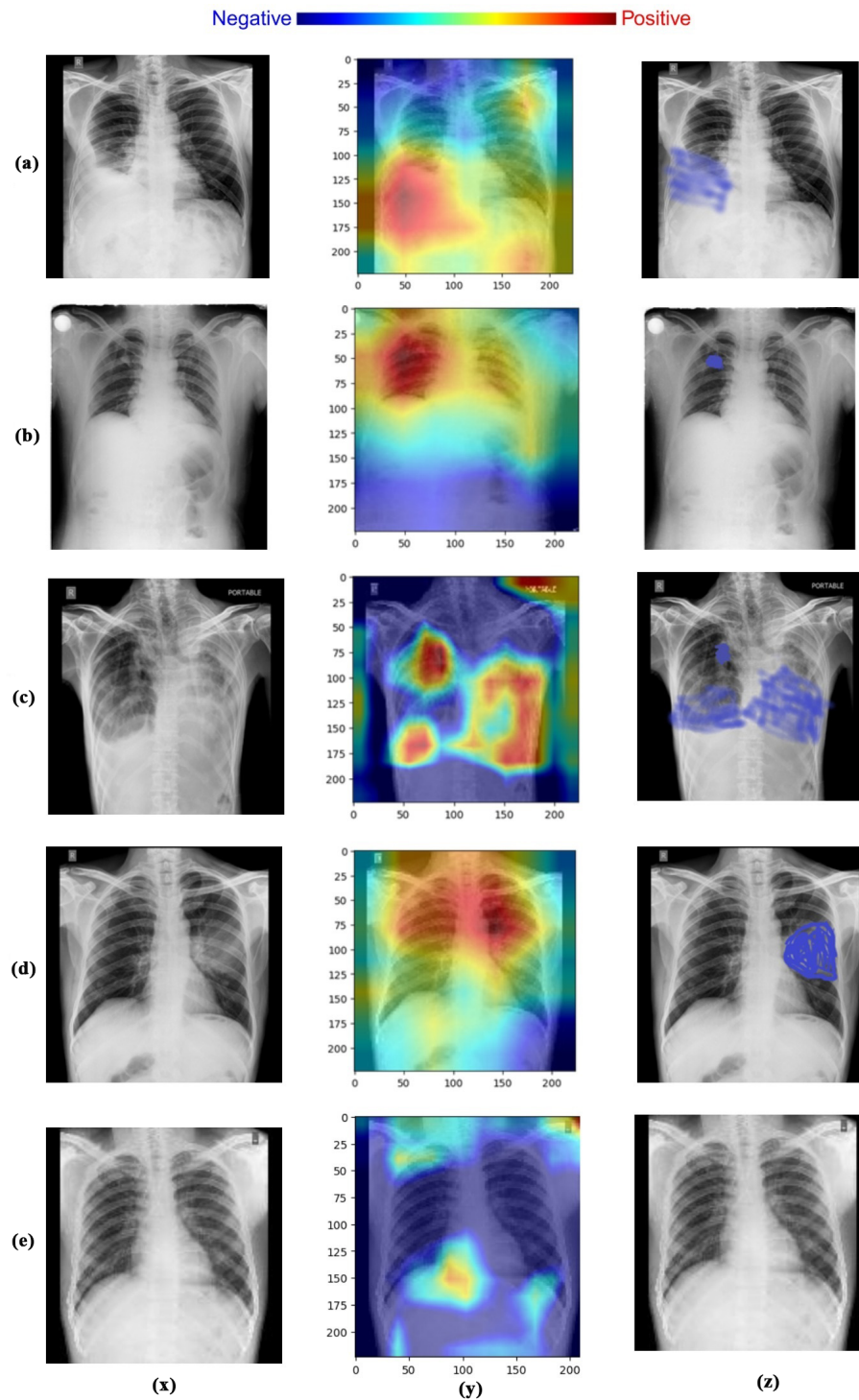


Figure 5.17: Disease Visualization with Grad-CAM Technique with its ground-truth label and the radiologist's highlighted radiographs. From left to right: (x) are the original Chest radiographs; (y) are the heatmap overlaid on the radiographs, where the areas marked with a peak (red) in the heatmap indicate abnormalities with high probabilities; (z) are the same chest X-rays with abnormalities highlighted (blue) by the experienced radiologist, From top to bottom: (a) to (d) are the chest radiographs with pulmonary abnormalities, and (e) are the chest radiographs with no abnormalities.



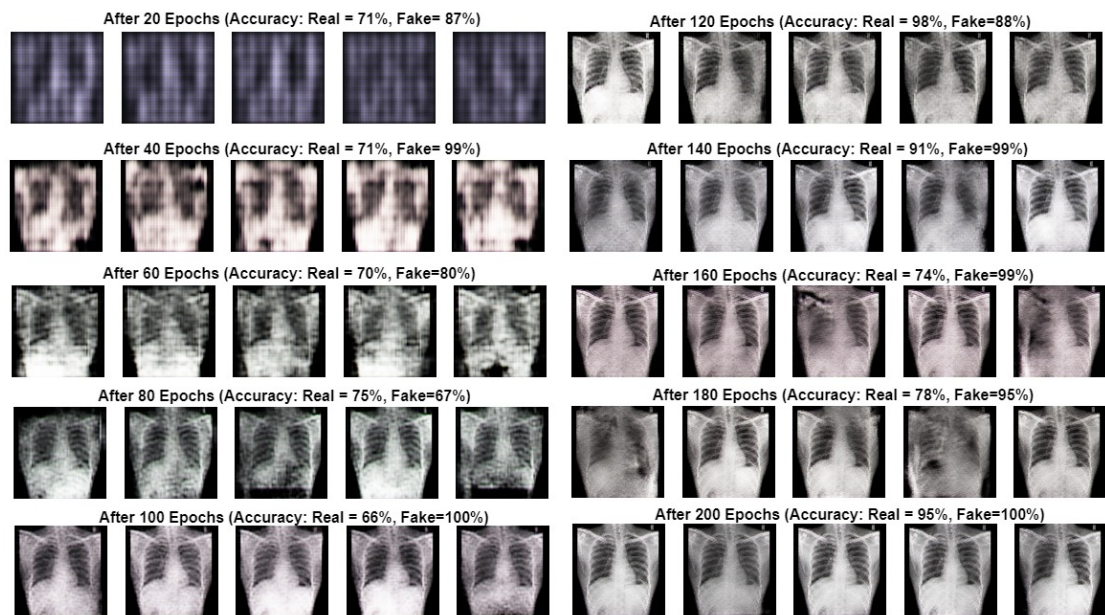
## 5.6 Data Augmentation vs. Synthetic Data Generation

In this section, we utilize data augmentation and synthetic data generation techniques on radiology images to empirically evaluate their effectiveness in improving radiology image classification.

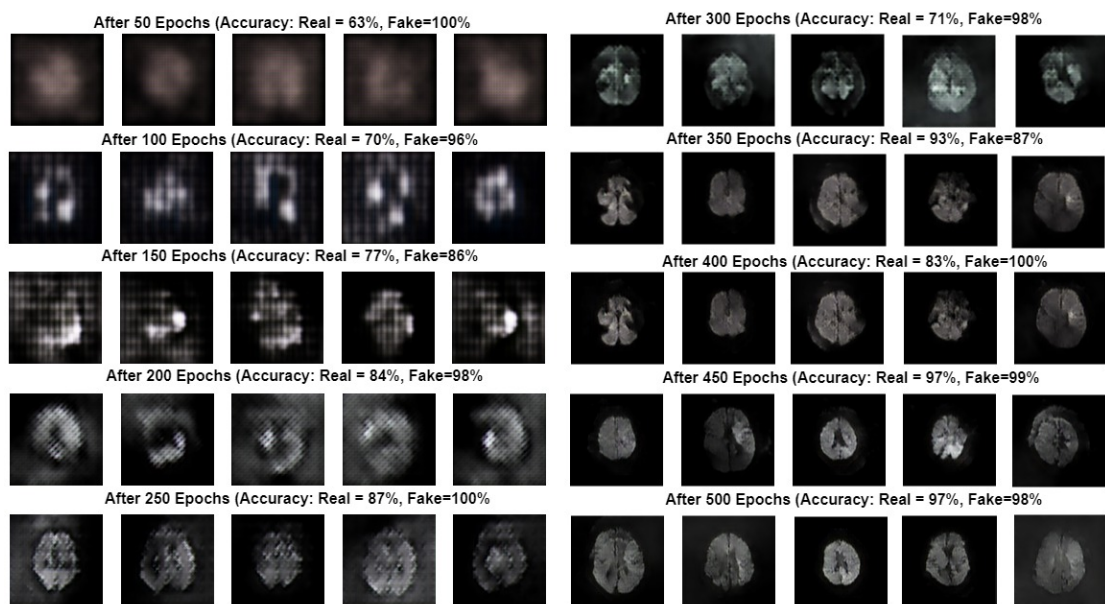
### 5.6.1 An Empirical Evaluation for Enhancing Radiology Image Classification

In this section, we discuss the experimental setup of the overall study, followed by the cohort selection and detailed results. The hardware and software required for the comprehensive experiment are as follows: The NVIDIA Tesla M40 server with 128 gigabytes of RAM, 3 terabytes of HD, 24 gigabytes of CPU, and Python 3.6 with Keras, TensorFlow library was utilized. For X-ray images, we trained RAD-DCGAN for 200 epochs, and we could observe that it could generate X-ray images that resembled the original image in around 100 epochs. Likewise, for MR images, we trained RD-DCGAN for 500 epochs, and after 400 epochs, MR images similar to the original image were obtained. The generation of synthetic images after every 20 and 50 epochs in X-ray and MR images is presented with the accuracy of real and fake classification in Figure 5.18. For the chest X-ray dataset collected from KMC hospital, the training time was approximately 59.91 minutes, and this was conducted over 200 epochs. Additionally, on the MRI sequence dataset from the same hospital, the training time for the GAN model was approximately 111.33 minutes, spanning 500 epochs. We have used the Python toolkit augmentor (Bloice *et al.*, 2017b), which provides a library for performing various augmentation operations to enlarge the radiology cohort obtained. To check the efficacy of the proposed RAD-DCGAN compared to basic augmentation strategies, we have obtained the classification accuracy by applying various state-of-the-art convolution neural network models like MobileNet (Howard *et al.*, 2017a), VGG16 (Liu and Deng, 2015), EfficientNetB1 (Tan and Le, 2019), VGG19 (Liu and Deng, 2015), ResNet50 (He *et al.*, 2016a), Xception (Chollet, 2017a), InceptionV3 (Szegedy *et al.*, 2016a) and DenseNet (Huang *et al.*, 2017a). We have fine-tuned the hyperparameters of the pre-trained models by tweaking them so that they can adapt to the disease classification task. The pre-trained deep learning frameworks were initiated with the imageNet weights and later retrained on the radiology images obtained from the KMC hospital.





(a) X-ray images - 200 Epochs



(b) MR Images - 500 Epochs

Figure 5.18: The generation of synthetic data after every 20 and 50 epochs in X-ray and MR images, respectively

### 5.6.2 Cohort Selection

We have acquired 502 X-ray Images (Normal Case =240, Abnormal Case=262) and 991 MR Images (Normal Cases=497, Abnormal cases=494) from Kasturba Medical College (KMC), Mangalore (India). The Institutional Ethics Committee (IEC) approval was taken to utilize the de-identified data for research purposes.

We applied basic augmentation techniques to increase the cohort size to 1498 X-ray images ( $train_{set}=1198$ ,  $test_{set}=150$  and  $validation_{set}=150$ ) and 3962 MR Images ( $train_{set}=3169$ ,  $test_{set}=397$  and  $validation_{set}=396$ ), and we have used the proposed RAD-DCGAN for synthetic image generation to enlarge the size of the cohort to 1498 X-ray Images ( $train_{set}=1198$ ,  $test_{set}=150$  and  $validation_{set}=150$ ) and 2154 MR Images ( $train_{set}=1723$ ,  $test_{set}=216$  and  $validation_{set}=215$ ). While splitting the dataset into train, test, and validation, we ensured that no samples from the test or validation set were present in the training set to avoid the data leakage problem.

### 5.6.3 Results and Discussions

The proposed RAD-DCGAN comprises of two main stages: 1) Discriminator module training; and 2) Generator module training. While training the generator network, the fake or synthetic radiology images are generated, and they are categorized with the real radiology images while training the discriminator network. Firstly, we train the discriminator network with the batch of actual radiology image samples to calculate  $\log(D(y))$ . Later, we train the discriminator network with the batch of synthetic samples produced by the generator network to calculate  $\log(1 - (D(G(z))))$ . For the proposed RAD-DCGAN, the generator and discriminator modules are trained simultaneously to calculate the loss and accuracy as presented in Figure 5.19 and Figure 5.20 for X-ray and MR images, respectively. The discriminator network loss is stabilized when it reaches 0.5, which is when the discriminator is made to categorize the images into real and fake. Parallely, the training accuracy of the discriminator network must be greater than 50% while training on real and fake images. The loss of the proposed RAD-DCGAN's discriminator and generator modules for both actual and synthetic radiology samples is around 0.5, and the discriminator module's accuracy is around 80-90%, showcasing that the proposed model achieves stable equilibrium.

Further, we perform an empirical evaluation of the proposed RAD-DCGAN compared with basic augmentation strategies by training eight separate standard convolutional neural network models and the proposed UM-VES model. We trained the models on the data generated from the various data augmentation techniques shown in Figure 5.8. The detailed classification performance of the various augmentation methods on the radiology X-ray and MR images is presented in Table 5.8 and Table 5.9. We have resized the radiology images to  $150 \times 150 \times 3$  and passed them as input to the various state-of-the-art deep learning networks.

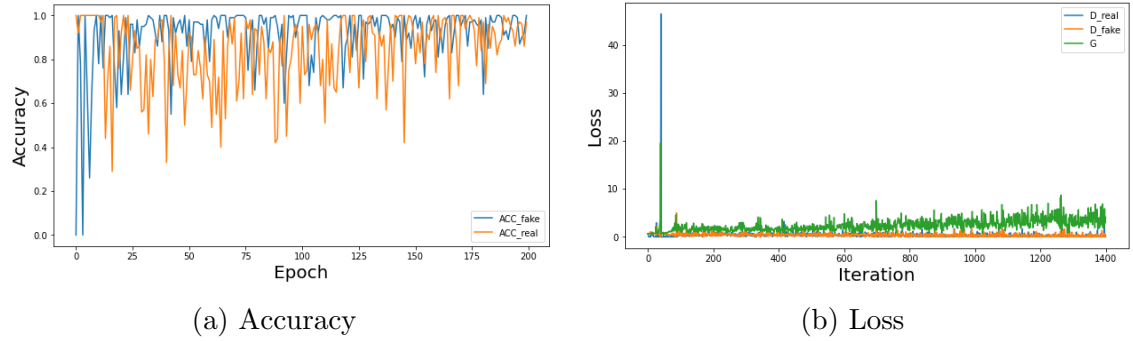


Figure 5.19: Accuracy and loss during the training of discriminator and generator component in RAD-DCGAN on X-ray images

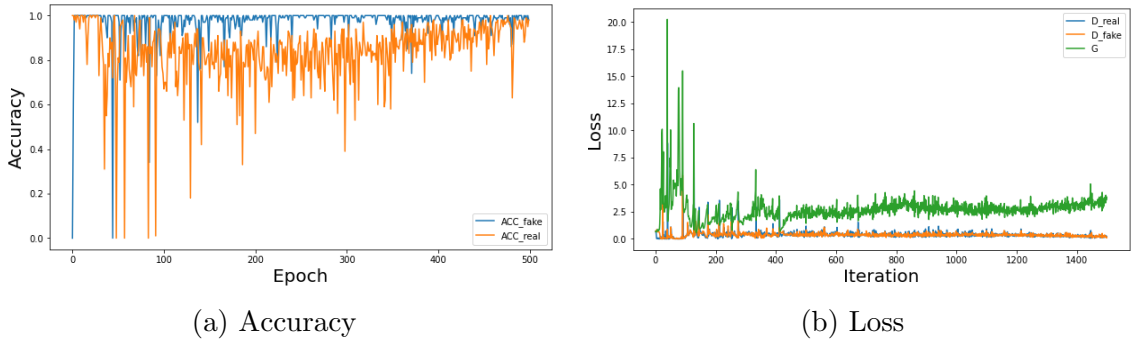


Figure 5.20: Accuracy and loss during the training of discriminator and generator component in RAD-DCGAN on MR images

Table 5.8: Classification performance metrics for Chest X-Ray Images

Methods/ Models	Accuracy (%)								
	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
Rotation	77.83	66.10	79.24	66.11	76.37	77.97	77.16	74.80	<b>82.28</b>
Zooming	77.7	66.06	79.12	65.91	76.02	77.66	76.87	74.59	<b>84.32</b>
Brightness	76.91	64.11	76.35	64.31	74.71	75.63	75.58	73.78	<b>78.41</b>
Shearing	77.62	65.83	79.16	65.65	76.01	77.55	76.82	74.37	<b>83.12</b>
Combined Augmenta- tion	78.04	66.23	79.45	66.42	76.57	78.21	77.41	75.11	<b>82.25</b>
<b>RAD- DCGAN</b>	<b>82.15</b>	<b>70.53</b>	<b>82.94</b>	<b>70.88</b>	<b>79.42</b>	<b>82.62</b>	<b>81.62</b>	<b>79.88</b>	<b>85.16</b>

Note: a) MobileNet, b) VGG16, c) EfficientNetB1, d) VGG19, e) ResNet50, f) Xception, g) InceptionV3 (h)DenseNet (i)UM-VES

The accuracy for each model is calculated while training and testing the models on the data generated using the proposed RAD-DCGAN and various basic augmentation techniques like rotation, zooming, brightness, shearing, and combined augmentation. It is observed that the models trained on the traditional augmentation data produced accuracy of about 76.91% to 78.04% for MobileNet, 64.11%

Table 5.9: Classification performance metrics for MRI T2-Flair sequences

Methods/ Models	Accuracy (%)								
	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
Rotation	92.21	87.57	45.43	91.69	66.58	92.34	90.16	88.87	<b>94.41</b>
Zooming	91.58	86.42	44.25	90.47	65.14	91.09	89.24	87.12	<b>93.10</b>
Brightness	91.14	85.77	43.01	89.92	64.61	90.05	88.47	86.13	<b>92.87</b>
Shearing	92.10	87.11	44.29	91.38	66.14	91.77	89.54	87.89	<b>93.24</b>
Combined Augmenta- tion	93.44	88.88	47.97	92.42	67.67	93.43	91.41	90.00	<b>95.72</b>
<b>RAD- DCGAN</b>	<b>97.20</b>	<b>97.67</b>	<b>54.88</b>	<b>96.27</b>	<b>91.62</b>	<b>95.81</b>	<b>97.20</b>	<b>95.81</b>	<b>98.24</b>

Note: a) MobileNet, b) VGG16, c) EfficientNetB1, d) VGG19, e) ResNet50, f) Xception, g) InceptionV3 (h)DenseNet (i)UM-VES

to 66.23% for VGG16, 76.35% to 79.45% for efficientNetB1, 64.31% to 66.42% for VGG19, 74.71% to 76.57% for ResNet50, 75.63% to 78.21% for Xception, 75.58% to 77.41% for InceptionV3 and 73.78% to 75.11% on DenseNet for X-ray images (refer to Table. 5.8). For MR images, it is seen that the accuracy obtained is about 91.14% to 93.44% for MobileNet, 85.77% to 88.88% for VGG16, 43.01% to 47.91% for efficientNetB1, 89.92% to 92.42% for ResNet50, 64.61% to 67.67% for Xception, 90.05% to 93.43% for InceptionV3 and 86.13% to 90.00% for DenseNet model trained on various basic augmentation techniques (refer to Table. 5.9). The effectiveness of the UM-VES model in accurately classifying images has been demonstrated through its superior performance when compared to state-of-the-art CNN models using augmented or synthetic images.

When trained on radiology images generated using rotation augmentation, the deep learning models achieved superior performance compared to other traditional augmentation strategies. The models trained on radiology images produced from brightness augmentation comparatively gave lesser accuracy than other basic augmentation techniques for both the X-ray and MRI cohorts. Empirically, it is also observed that the images generated from the combined augmentation achieved a better performance than the individual augmentation techniques. With respect to, the deep learning classifiers trained on the proposed RAD-DCGAN model have achieved significantly higher accuracy of at least 3-4% more than the models trained on the images generated using traditional augmentation strategies for both the radiology images. The superior performance achieved by the proposed RAD-DCGAN compared to conventional augmentation strategies indicates that the synthetic image generated has some additional information required for deep learning classifiers, which also prevents the models from overfitting.

## 5.7 Summary

In this chapter, we propose a lightweight and explainable deep learning network named UMVES, a Multi-Scale Chest X-ray Network that consists of MSDL and DS-CNN layers, to predict pulmonary diseases from the CXR obtained from the publicly available Open-I dataset and the CXR data collected from the private medical hospital. The MSDL layer captures the multi-scale features with the help of a broader receptive field, and the DS-CNN layer learns the imaging features by adjusting lesser parameters. The quantitative and qualitative analyses of the proposed UM-VES model are performed on both CXR datasets. The experimental validation was observed through evaluation metrics like accuracy, precision, recall, F1-score, MCC, and AUROC. The experimental results show that the proposed model outperformed baseline deep learning techniques and existing state-of-the-art approaches. The MSDL layer in the proposed model has significantly impacted the prediction outcome by capturing the Multi-scale features from the CXR. The grad-CAM method is employed to visualize the pulmonary abnormalities from the CXR and to check the model's ability to arrive at a decision. The obtained grad-CAM CXR samples are compared with the CXRs labelled by expert radiologists. It is observed that the UM-VES can reach a performance level similar to that of the radiologists. This study also presents RAD-DCGAN for generating synthetic images from the radiology X-ray and MRI cohorts collected from a private medical hospital (KMC Hospital, India). We have conducted a comprehensive qualitative analysis of the proposed RAD-DCGAN compared with conventional data augmentation techniques like rotation, zooming, brightness, and shearing. The eight state-of-the-art deep learning classifiers and the proposed UM-VES are used to check the efficacy of the data generated from the proposed RAD-DCGAN and the traditional data augmentation techniques. The detailed investigation shows that the synthetic data generated through the proposed RAD-DCGAN has achieved a significantly higher classification accuracy of 3-4% compared to the data generated through basic data augmentation strategies. This superior performance is due to the higher resolution synthetic images generated with additional information, which aids the classifier's performance.

## Publications

*(based on study proposed in this chapter)*

1. Shashank Shetty, Ananthanarayana V S., and Ajit Mahale (2022). "MS-

CheXNet: An Explainable and Lightweight Multi-Scale Dilated Network with Depthwise Separable Convolution for Prediction of Pulmonary Abnormalities in Chest Radiographs”, *Multidisciplinary Digital Publishing Institute (MDPI) Mathematics*, 10, no. 19: 3646. <https://doi.org/10.3390/math10193646> [Indexed: SCIE & Scopus, IF: 2.592] (*Status: Published Online*)

2. Shashank Shetty, Ananthanarayana V S., and Ajit Mahale. ”Data Augmentation vs. Synthetic Data Generation: An Empirical Evaluation for Enhancing Radiology Image Classification”, *IEEE 17th International Conference on Industrial and Information Systems (ICIIS'23)*. [Core Ranked Conference] (*Status: Accepted For Presentation*)

# **PART III**

## **AI-based CRS for Multimodal Unstructured Medical Data Analysis**





## Chapter 6

# Deep Medical Multimodal Fusion Networks (DMMFN) for Disease Prediction from Radiology Chest X-ray Images and its Associated Reports

### 6.1 Introduction

Pulmonary diseases are the most commonly found infections caused due to air pollution, tobacco smoking, breathing radioactive chemical elements, asbestos, or any other unwanted particles. The various Pulmonary diseases are Tuberculosis, Pneumothorax, Cardiomegaly, Pulmonary atelectasis, Pneumonia, etc. Some symptoms of pulmonary diseases include wheezing, shortness of breath, chronic cough, weight loss, etc. The risk factor involved in pulmonary diseases is high, and hence, timely prediction of the abnormality is vital. Modern medicine is significantly reliant on synthesizing data and information from various modalities, including diagnostic imaging data (i.e., CT, X-Ray, MRI, etc.), structured laboratory data, unstructured narrative text data (i.e., medical reports), and, in certain situations, audio, video, signals (i.e., ECG signals), or any other observational data. Radiology is one such critical medical discipline involving medical imaging like X-Ray, CT, MRI, etc. to investigate the internal structure of the body and detect any abnormality. One of the most frequently available diagnostic procedures for detecting and assessing abnormalities in the chest and lung area is a CXR. The grid of normal (No diseases) and abnormal (pulmonary diseases) CXR from the Indiana University dataset is presented in Figure 6.1. Radiologists will analyze these CXRs and prepare clinical notes detailing the conditions visualized from the medical image.

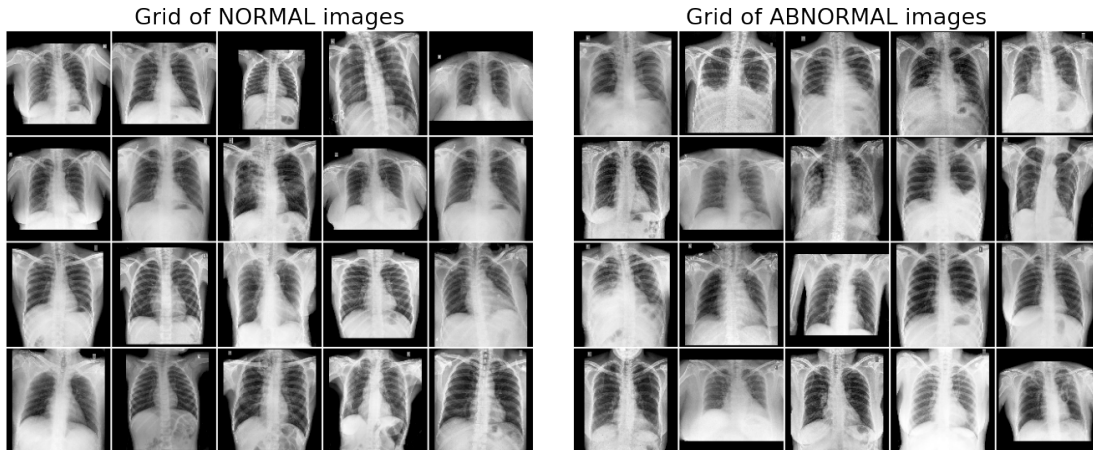


Figure 6.1: The grid of normal (No diseases) and abnormal (pulmonary diseases) CXR from Indiana University dataset.

It is especially true in the area of diagnostic image analytics, where a thorough understanding of clinical context is necessary to make accurate therapeutic recommendations. For example, on numerous occasions, the absence of laboratory test information while investigating medical images resulted in poor correlation and reduced clinical utility for the radiologists (Leslie *et al.* (2000); Cohen (2007)). A large percentage of radiologists (i.e., 87%) said clinical information substantially impacted image analysis during the survey of radiologists (Boonn and Langlotz, 2009). Radiology is not the only medical field where the significance of using medical context for precise interpretation of imaging data is recognised. But several other image-based medical areas like dermatology, ophthalmology and pathology also rely upon clinical information to assist visual analysis (Wong *et al.* (2015); Comfere *et al.* (2013); Jonas *et al.* (2017)).

Physicians can evaluate the imaging findings in the proper clinical context if related and accurate details pertaining to the current medical conditions and previous clinical history are available during the analysis. This information also leads to a more pertinent differential diagnostic process that provides useful reports for clinicians, which will help improve patients' prognoses. The number of radiological imaging tests is increasing in this digital era. Hence, to fulfil this high workload requirement, a physician, on average, needs to evaluate diagnostic imaging every 10 minutes over eight hours of a day, contributing to weariness, stress, and higher inaccuracies (McDonald *et al.*, 2015). The DL approach has recently exhibited promising outcomes in various research domains. Also, in the medical field, it is rapidly growing because of its ability to provide an automated system by complementing or augmenting the cognitive tasks of overburdened clinicians

(Dean *et al.* (2015); Banerjee *et al.* (2019); Kumar and Jayadev (2020)). CNN, one of the DL models, has provided significant performance in applications involving the analysis and categorization of images and is usually adapted to radiological imaging. Initially, the CNNs were widely applicable in medical image analysis, including CXRs, diabetic retinopathy, and skin cancer (Hinton (2018); Dunnmon *et al.* (2018); Johnson *et al.* (2019a); Gulshan *et al.* (2016)). However, these proposed models considered only pixel data from the single input modality and could not derive a context from the other medical information as it is done in common medical practice, hampering clinical translations.

We can illustrate with an elementary exemplar like the detection of pneumonia from a chest radiograph, where researchers have previously worked on building DL models to automate the process of identifying and classifying pathologies from chest radiographs (Rajpurkar *et al.* (2018); Majkowska *et al.* (2020)). But, implementing these models has minimal impact on clinical procedures due to the non-usage of clinical context and diagnostic values. Despite having the visual features from the chest radiographs of patients with pneumonia, in general, they cannot possibly identify any other differential diagnoses, meaning they could be non-specific and ambiguous. An accurate diagnosis requires laboratory data, clinical reports, and values. To rephrase it, the chest radiograph results suggest pneumonia in a patient with fever, shortness of breath, and chest pain; however, in another case, it might represent other causes of chest conditions such as pleural effusion, cardiomegaly, pulmonary edema, or even cancer of the lungs. There are many such indefinite instances over multiple dimensions of the healthcare domain where the data fed with the clinical context, including structured or unstructured data, have significantly impacted the medical imaging interpretations. EHRs that may be structured clinical data or unstructured clinical reports are of paramount importance for the precise and clinically apt understanding of medical imaging. Henceforth, the automated analysis of visual features from medical imaging alongside the data extracted from the EHR, like patient demographics, history of illness, and laboratory values after testing, will definitely give more clinically consistent and high-efficacy models.

Multimodal DL approaches take input from different data modalities and fuse them to produce more consistent and valuable information obtained from various single data sources. These fusion-based DL models have been giving promising results and are also successfully applied outside the field of healthcare, like object detections for autonomous vehicles (Person *et al.*, 2019), classifying social media videos Trzcinski (2018) and emotion classifications (Pandeya and Lee, 2021). As an

exemplar, for safe navigation of autonomous vehicles, a fusion-based multimodal DL framework was proposed by the authors that takes the input from the images and LiDAR data points to detect the objects on the road effectively. The multimodal DL model achieved 3.7% more performance compared to the uni-modal CNN classification architecture (Person *et al.*, 2019). Likewise, while performing social media video classification from text and video sources, the proposed multimodal model gave approximately 12% higher accuracy compared to the uni-modal Google’s InceptionV3 model. The increase in performance or accuracy is not the only criteria being considered for justifying the use of deep multimodal learning in clinical imaging. But the main motto is to combine the complementary contextual data to obtain more precise diagnostic results by limiting the challenges of unimodal image-only approaches.

The core research challenges in multimodal learning are the fusion of multimodal data to utilize the benefits of the complementary features from the various heterogeneous sources and cater to more effective diagnostic predictions. Since the visual and textual features are mutually exclusive, there is a need for rich fusion representation to provide fine-grained knowledge for further prediction.

### 6.1.1 Problem Statement

Currently, relying solely on image-only approaches presents multiple challenges, and to overcome these challenges, it is necessary to merge visual features from medical imaging with data extracted from Electronic Health Records (EHR). This EHR data includes patient demographics, medical history, and laboratory test results, which provide relevant and precise information about the patient’s current medical condition and past medical history during the analysis. However, the key research challenge in multimodal learning is creating a comprehensive fused representation that offers detailed knowledge for further prediction. The main issue is that visual and textual features are mutually exclusive, making it difficult to blend the data from different sources and make more precise diagnostic predictions. Thus, the challenge is to develop a methodology that effectively merges these two types of data to create a more comprehensive representation that can be used to provide accurate therapeutic recommendations in diagnostic image analytics. The problem statement is defined as follows:

*“Considering the set of multimodal unstructured medical images and its associated clinical text data, design and develop an effective deep learning model for fusing complementary visual and textual features to*

*create a unified representation in a shared space for disease prediction to support an intelligent clinical recommendation system.”*

The primary objective of this chapter is to provide a comprehensive overview of the research study that was conducted to address this significant problem. We propose an effective fusion strategy using the DL framework to fuse the features extracted from diagnostic images and text to predict pulmonary abnormalities. The following are the critical contributions of our proposed research work:

- We propose two effective Medical Multimodal Tensor Fusion Networks: Compact Bilinear Pooling (CBP) and Deep Hadamard Product (DHP), for predicting abnormalities from the radiology CXR and text reports.
- We conducted a thorough investigation and compared the unimodal vs multimodal models for disease abnormality predictions from the multimodal radiology cohorts.
- We have analyzed the performance of the proposed models by applying them to standard augmented data and the synthetic data generated to understand their ability to predict from the new and unseen data.
- The Proposed unimodal and multimodal models are assessed and analyzed in two heterogeneous diagnostic cohorts: Publicly available multimodal medical cohort containing CXRs with diagnostic reports from Indiana University and data acquired from the private medical hospital.
- We benchmarked the performance of the proposed multimodal prediction model with respect to state-of-the-art medical fusion techniques.

## 6.2 Methodology

Our research proposes a new deep learning framework for predicting abnormalities in medical images using a heterogeneous radiology cohort. Figure 6.2 illustrates the architecture of our proposed Multimodal Medical Tensor Fusion Network. To obtain the disease outcome, our system takes chest X-rays along with their corresponding radiology reports as input. We propose two types of Unimodal models: UM-TES (discussed in Chapter 4) and UM-VES (discussed in Chapter 5), which process the radiology report text and the CXR image separately. We also propose two Multimodal models: CBP-MMFN and DHP-MMFN, which combine

the features extracted by the Unimodal models. Finally, the combined features are passed to a fully connected Deep Neural Network (DNN) to predict the disease outcome.

This section describes the proposed Deep Medical Multimodal Fusion Networks (DMFN), where we explain two proposed strategies: Compact Bilinear and Deep Hadamard Product. The main aim of any fusion strategy is to integrate multiple unimodal representations into a multimodal representation. Most of the previous multimodal research focused on concatenating visual and textual features without considering inter-modal interaction. Both of our proposed multimodal fusion strategies effectively explore inter-modal interactions by explicitly aggregating visual CXR features and textual clinical report features. Let  $M^t = \{t_1, t_2, t_3, \dots, t_p\}$  be the medical textual features in the form of tensors that are obtained from the UM-TES as discussed in Chapter 4 and  $M^x = \{x_1, x_2, x_3, \dots, x_q\}$  be the medical imaging features in the form of tensors that are generated from the UM-VES as explained in Chapter 5. Following are the two multimodal strategies applied to these tensors obtained from the two unimodal models:

### 6.2.1 Compact Bilinear Pooling-based Medical Multimodal Fusion Network (CBP-MMFN)

Bilinear pooling provides local pairwise interaction between every tensor of both the visual features extracted from the CXR and the textual features generated from the associated radiology reports in a multiplicative fashion. However, bilinear pooling results in a high dimensional representation, resulting in an infeasible set of parameters to be learnt. We propose a novel Compact Bilinear pooling to aggregate the heterogeneous visual and textual features by projecting a higher dimensional outer product representation in a low-dimensional space, effectively capturing the high-order correlation between the imaging and textual tensors. In our proposed CBP-MMFN network, we use the two-fold Cartesian product to combine two unimodal tensors  $M^t$  and  $M^x$  on the following vector field:

$$\left\{ (M^t, M^x) \mid M^t \in \begin{bmatrix} M^t \\ 1 \end{bmatrix}, M^x \in \begin{bmatrix} M^x \\ 1 \end{bmatrix} \right\} \quad (6.1)$$

Every neural coordinate  $(M^t, M^x)$  can be visualized as a 2-D point in the two-fold cartesian space characterized by the textual and imaging embedding dimensions. Mathematically, it is equivalent to the Kronecker product or outer product between the medical textual feature  $M^t$  and the imaging feature  $M^x$  and

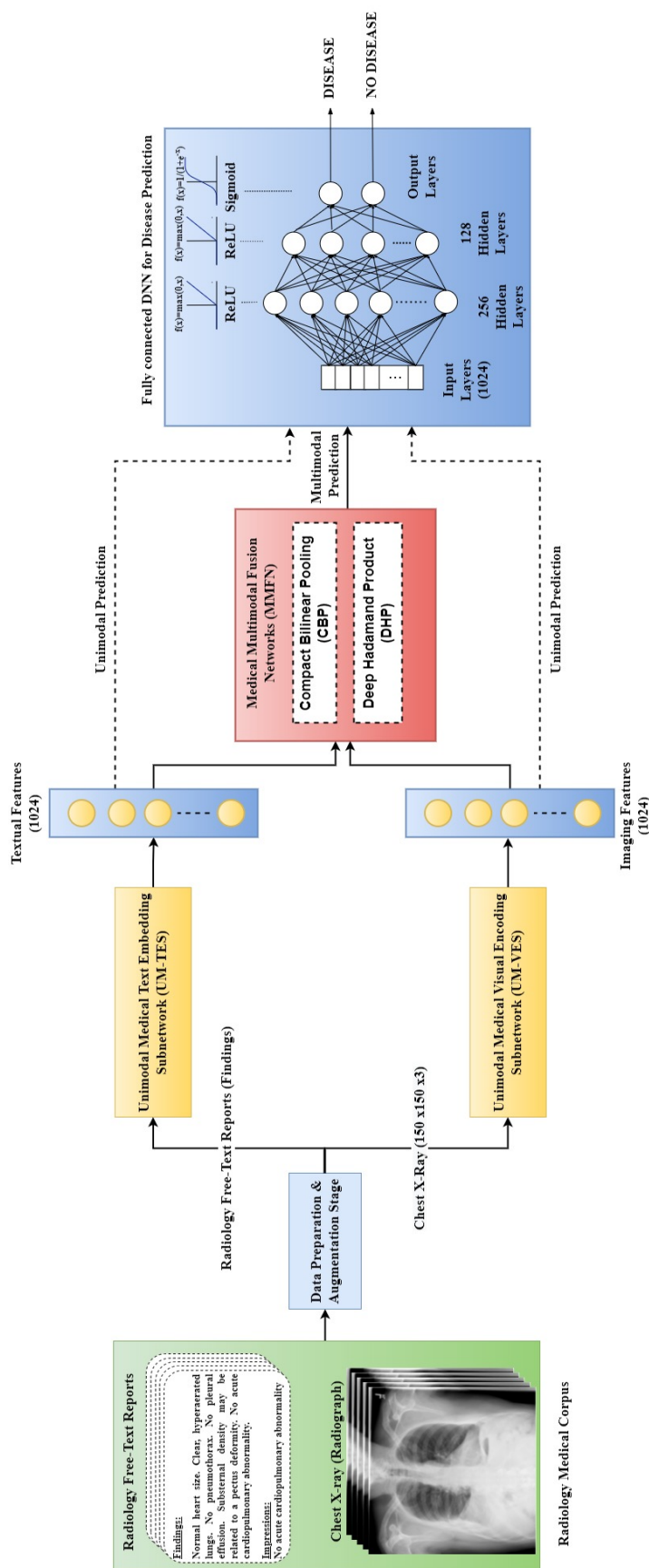


Figure 6.2: Proposed Multimodal Medical Tensor Fusion Network-based DL Framework for predicting Abnormality from the heterogeneous radiology CXR and text reports.



the multimodal bilinear pooling obtained is defined as follows:

$$M^{CBP} = \begin{bmatrix} M^t \\ 1 \end{bmatrix} \otimes \begin{bmatrix} M^x \\ 1 \end{bmatrix} = \sum_{i=1}^p \sum_{j=1}^q M_i^t (M_j^x)^T \in \mathbb{R}^{N_H \times N_W} \quad (6.2)$$

Here,  $\otimes$  represents the Kronekar product or the outer product between the two vectors, and  $M^{CBP} \in \mathbb{R}^{N_H \times N_W}$  (i.e.,  $1024 \times 1024$ ) is the bilinear interaction map consisting of all possible combinations of the multimodal representation capturing inter-modal interactions between textual and imaging features as shown in Figure 6.3. For instance, if  $M^t$  and  $M^x$  are defined as,

$$M^t = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_p \end{bmatrix}, M^x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_q \end{bmatrix} \quad (6.3)$$

then the kronekar product or the outer product between the two vectors can be obtained as follows:

$$M^t \otimes M^x = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_p \end{bmatrix} \otimes \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_q \end{bmatrix} = \begin{bmatrix} t_1 x_1 & t_1 x_2 & \cdots & t_1 x_q \\ t_2 x_1 & t_2 x_2 & \cdots & t_2 x_q \\ \vdots & \vdots & \ddots & \vdots \\ t_p x_1 & t_p x_2 & \cdots & t_p x_q \end{bmatrix} \quad (6.4)$$

The  $ij^{th}$  component of the features obtained can be depicted as:

$$(M^t \otimes M^x)_{ij} = t_i x_j, \text{ for all } 1 \leq i \leq p \text{ and } 1 \leq j \leq q \quad (6.5)$$

Mathematically, the bilinear pooling is generated by the outer product of the unimodal feature, and hence, this is the core design component of our fusion model. We argue that employing outer products is beneficial in three-folds: 1) It encodes more tensor correlation between the two different modalities, capturing inter-modal interactions. 2) It has more significance than the straightforward concatenation operation, which keeps the basic information obtained from different modalities without modelling any correlation. 3) Finally, the promising benefit lies in the bilinear interaction map obtained from the outer product, lies in a 2D format similar to that of an image. In this view, the inter-modal correlation encoded in the bilinear interaction map can be viewed as the local feature of an image. DL models have demonstrated remarkable success in computer vision.



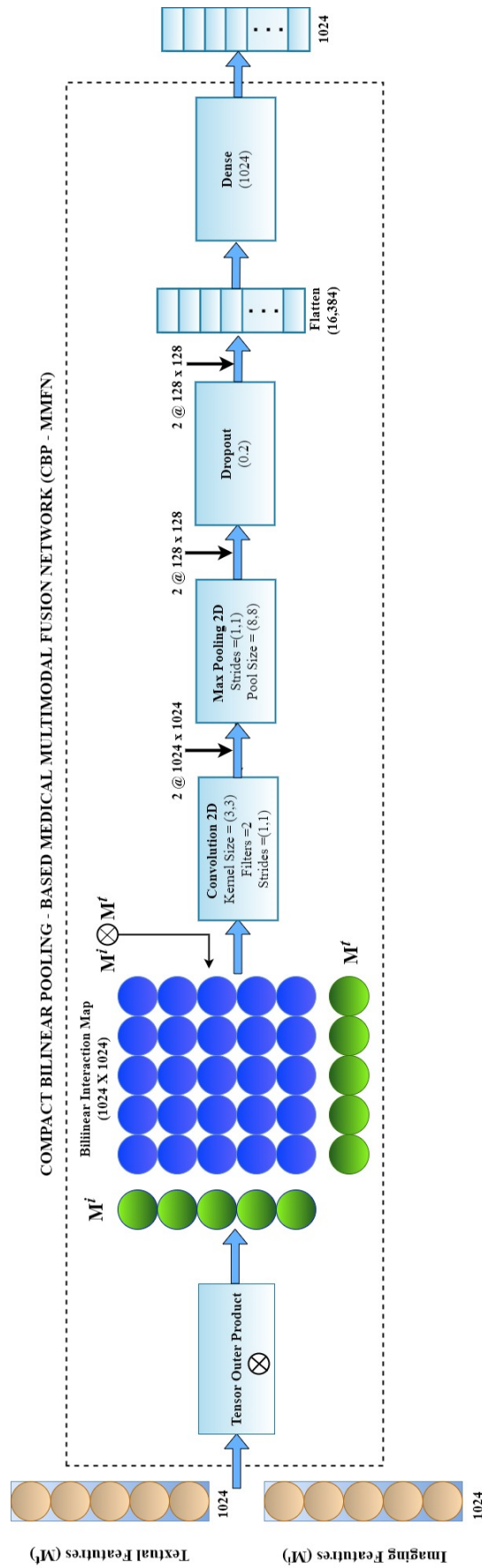


Figure 6.3: Proposed Compact Bilinear Pooling-based Medical Multimodal Fusion Network (CBP-MMFN)

Consequently, constructing bilinear 2D maps allows us to leverage sophisticated DL techniques for learning bilinear interaction functions in multimodal disease prediction tasks. In applications such as visual question answering (Fukui *et al.*, 2016), the concept of integrating visual and textual features in the outer product has produced competent outcomes.

Empirically, we noticed that the multimodal representation obtained after the fusion is of high dimension, but the possibilities of overfitting are negligible. We believe this is mainly because the output neurons obtained from the CBP-MMFN are easy to understand and semantically significant (i.e., the manifold of the multimodal representation is not complex but has a relatively high dimension). Henceforth, the following layers in the network find it easy to interpret the meaningful information. However, the bilinear interaction maps obtained are usually of high dimension, which increases the computation cost and storage space. Hence, We propose compact bilinear pooling employing CNN over the bilinear interaction map to learn the high order correlation between the multimodal pairwise embeddings and reduce the high dimensional representation to the lower-dimensional space.

$$z = conv(M^{CBP}, W) = f\left(\sum_{i=1}^{N_H} \sum_{j=1}^{N_W} M_{p+i-1, q+j-1}^{CBP} \cdot W_{i,j} + b\right) \quad (6.6)$$

Here,  $M^{CBP}$  represents the bilinear interaction map with the size  $(N_H, N_W)$ , which is fed to the convolution module. The ReLU activation function with  $W$  convolution filters is leveraged on every accessible window of the interaction map to generate a discriminative activation map  $z$  of size,  $dim(conv(M^{CBP}, W)) = (N_H, N_W, N_C)$ . Here,  $N_C$  represents the number of channels generated from the convolution filter  $W$ . The pooled feature obtained from the convolution layer is downsampled using the Max pooling method. This mechanism will reduce the high-dimensional feature map to low dimensional space while retaining valuable information. The down-sampled feature map obtained by applying max pooling is asserted as follows:

$$z' = pool(z)_{p,q,r} = max(z_{p+i-1, q+j-1, r})_{(i,j) \in [1,2,\dots,P]^2} \quad (6.7)$$

Where,  $Z' \in \mathbb{R}^{N_H \times N_W}$  represents the activation map obtained by performing the max pooling operation with the pooling size  $P \times P$ . We adopt the dropout mechanism on the feature map  $Z'$  to prevent overfitting. The discriminative compact features are flattened and ingested into a fully connected prediction module

to predict the diseases.

The bilinear pooling adopted for fusing multimodal clinical reports and the CXR results in a richer representation of pairwise interaction between the visual and textual features. However, the quadratic expansion obtained from the outer product operation yields a high-dimensional interaction map with expressive multimodal features. This significantly increases the computation cost and adversely affects practical applications. Henceforth, we propose a compact bilinear pooling to boost the training process, where we adopt CNN to attain second-order interaction between multimodal features and reduce the bilinear interaction map from high dimension to lower dimension space. Algorithm 2 presents the proposed Compact Bilinear Pooling-based Medical Multimodal Fusion Network for predicting pulmonary abnormalities from multimodal clinical data.

---

**Algorithm 2:** Proposed Compact Bilinear Pooling-based Medical Multimodal Fusion Network (CBP-MMFN)

---

**Input:** Textual feature extracted from the clinical report using UM-TES,  $M^t \in \mathbb{R}^p$  and Visual feature extracted from the CXR using UM-VES,  $M^x \in \mathbb{R}^q$

**Output:** Compact multimodal representation,  $\Phi(M^t, M^x) \in \mathbb{R}^d$

```

1 for  $i \leftarrow 1, 2, \dots, p$  do
2   for  $j \leftarrow 1, 2, \dots, q$  do
3     /* Outer product between the textual features and the visual
       features to obtain the bilinear interaction map */
4      $M^{CBP} \leftarrow \begin{bmatrix} M^t \\ 1 \end{bmatrix} \otimes \begin{bmatrix} M^x \\ 1 \end{bmatrix} \leftarrow M_i^t (M_j^x)^T, M^{CBP} \in \mathbb{R}^{N_H \times N_W}$ 
5     /* Applying Convolution and Pooling operation to reduce the high
       dimensional interaction map into lower dimensional space */
6      $z \leftarrow conv(M^{CBP}, W) \leftarrow ReLU(\sum_{i=1}^{N_H} \sum_{j=1}^{N_W} M_{p+i-1, q+j-1}^{CBP} \cdot W_{i,j} + b)$ 
7      $z' \leftarrow pool(z)_{p,q,r} \leftarrow max(z_{p+i-1, q+j-1, r})_{(i,j) \in [1,2, \dots, P]^2}$ 
8      $\phi \leftarrow z^* \leftarrow Dropout(z', 0.2)$ 
9     /* The compact multimodal features is then flattened and passed through
       fully connected network for disease prediction */

```

---

## 6.2.2 Deep Hadamard Product-based Medical Multimodal Fusion Network (DHP-MMFN)

We also propose another tensor fusion strategy, the Deep Hadamard Product-based Medical Multimodal Fusion Network (DHP-MMFN), which provides local

element-wise interaction between every tensor of imaging features extracted from CXR and the textual features obtained from the radiology reports. This allows tensor elements in the same position as both imaging and visual features to interact multiplicatively and results in an interactive map of the same dimension as input vectors. In our proposed DHP-MMFN network, we use the two-fold element-wise product (or Hadamard product) to combine the two unimodal tensors  $M^t \in \mathbb{R}^d$  and  $M^x \in \mathbb{R}^d$  of same size  $d$ . The multimodal element-wise representation is defined as follows:

$$M^{DHP} = \begin{bmatrix} M^t \\ 1 \end{bmatrix} \odot \begin{bmatrix} M^x \\ 1 \end{bmatrix} = \sum_{i=1}^d M_i^t \odot M_i^x \in \mathbb{R}^d \quad (6.8)$$

Here,  $\odot$  indicates the hadamard product or the element-wise multiplication between the two vectors, and  $M^{DHP} \in \mathbb{R}^d$  (i.e., 1024) is the pairwise interaction map with the same dimension  $d$  as that of two input vectors, consisting of multimodal representation capturing inter-modal interactions between textual and imaging features as shown in Figure 6.4. For instance, if  $M^t$  and  $M^x$  are defined as,

$$M^t = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_d \end{bmatrix}, M^x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \quad (6.9)$$

then the hadamard product or the element-wise multiplication between the two vectors can be obtained as follows:

$$M^t \odot M^x = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_d \end{bmatrix} \odot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} = \begin{bmatrix} t_1 x_1 \\ t_2 x_2 \\ \vdots \\ t_d x_d \end{bmatrix} \quad (6.10)$$

The  $i^{th}$  component of the features obtained can be denoted as:

$$(M^t \odot M^x)_{ii} = t_i x_i, \text{ for all } 1 \leq i \leq d \quad (6.11)$$

The features obtained from DHP-MMFN are flattened and fed to the fully connected module for disease prediction. A deep Hadamard product-based fusion strategy allows intermodal information flow between medical imaging and textual features by robustly capturing the high-level interaction between the multi-

modal features. The Deep Hadamard product provides bilinear interaction with a rich representation of imaging and textual features combined. DHP-MMFN integrates data from two different data distributions into a global space in which intermodality dynamics are obtained. Hence, the expressive features generated by DHP-MMFN significantly improve the overall DL framework's performance and provide supreme results compared to standard concatenation techniques. This is due to the fact that the DHP-MMFN generates a better correlation between the visual and textual features. Algorithm 3 presents the proposed Deep Hadamard Product-based Medical Multimodal Fusion Network for predicting pulmonary abnormalities from multimodal clinical data.

---

**Algorithm 3:** Proposed Deep Hadamard Product-based Medical Multimodal Fusion Network (DHP-MMFN)

---

**Input:** Textual feature extracted from the clinical report using UM-TES,  $M^t \in \mathbb{R}^p$  and Visual feature extracted from the CXR using UM-VES,  $M^x \in \mathbb{R}^q$

**Output:** Pairwise Interaction map,  $\Phi(M^t, M^x) \in \mathbb{R}^d$

1 **for**  $i \leftarrow 1, 2, \dots, d$  **do**

/\* Hadamard product or the element-wise multiplication between the two vectors to obtain the pairwise interaction map \*/

2  $M^{DHP} \leftarrow \begin{bmatrix} M^t \\ 1 \end{bmatrix} \odot \begin{bmatrix} M^x \\ 1 \end{bmatrix} \leftarrow M_i^t \odot M_i^x \in \mathbb{R}^d$

/\* The multimodal pairwise interaction features is then flattened and passed through fully connected network for disease prediction \*/

---

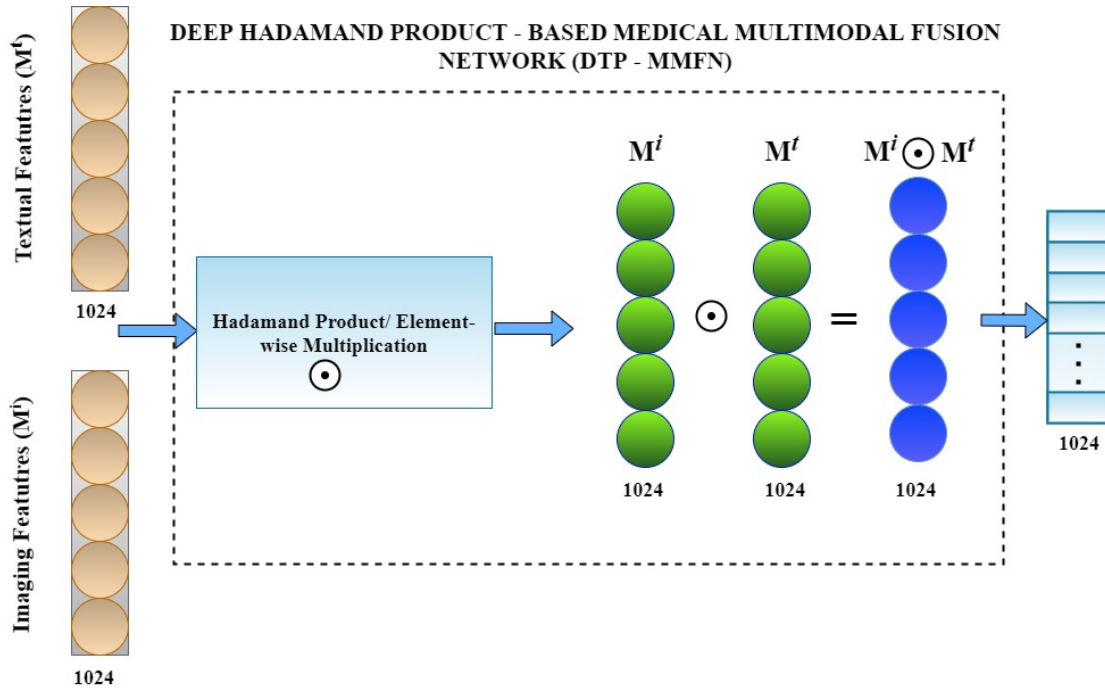


Figure 6.4: Proposed Deep Hadamard Product-based Medical Multimodal Fusion Network (DHP-MMFN)

### 6.3 Experimental Setup and Evaluation

In this section, we will provide an overview of the materials used in our research, followed by an explanation of the data augmentation techniques that were applied to improve the medical cohort size. Finally, we will present a detailed analysis of the experimental results obtained from our study.

#### 6.3.1 Datasets and Cohort Selection

A limited dataset becomes a severe issue in the health domain when it happens to be multimodal data. In the case of images, there exists some quality open source cohorts. Hence, there is a necessity to validate the effectiveness of the multimodal fusion models on the publicly available medical cohort and real-time data obtained from the private hospital. A comprehensive study was carried out on two clinical cohorts: the Indiana University CXR dataset (Demner-Fushman *et al.*, 2016) and the real-time multimodal data acquired from a private medical hospital [KMC Hospital (Mangalore, India)]. For our investigation, the de-identified data is leveraged. The IEC approval was granted by the Kasturba Medical College (KMC), Mangalore, for further research purposes. The two multimodal medical cohorts acquired consist of chest X-rays and associated radiology free-text reports. The

two clinical cohorts are classified as “normal” (i.e., cases with no abnormal findings or any active diseases) and “abnormal” (i.e., cases with acute pulmonary and cardiopulmonary diseases like Tuberculosis, Pneumothorax, Cardiomegaly, Pulmonary atelectasis, Pneumonia, Opacity/lung base, etc.). Table 6.1 represents the summary of cases (CXR with associated radiology reports) from the Indiana University and KMC Hospital datasets. A detailed benchmarking exercise is carried out on both clinical datasets to evaluate the proposed multimodal network.

Table 6.1: Cohort Statistics: CXR with associated clinical diagnostic notes from two clinical cohorts

Characteristics	IU Dataset	KMC Dataset
Total No. of cases (CXR with Radiology reports)	3996	502
Total No. of cases after removing missing cases	3638	502
Total No. of cases after standard data augmentation	6229	1498
Total No. of cases after synthetic data generation	6229	1498
Total No. of Sentences	17990	14537
Total No. of Words	143177	90221
Total No. of Vocabulary	1731	400
Total No. of Training/Validation Set	5606	1348
Total No. of Test Set	623	150
Percentage of Normal cases	38%	52%
Percentage of Abnormal Cases	62%	48%

### 6.3.2 Evaluation Criteria

We have used six standard evaluation criteria: Accuracy, Precision, Recall, F1-Score, MCC, and AUROC to examine the performance of the proposed multimodal fusion models on the two medical CXR cohorts. Section 4.4.3 provides a comprehensive explanation of the evaluation metrics chosen along with the rationale behind their selection.

### 6.3.3 Data Preparation and Augmentation Stage

#### 6.3.3.1 Standard Data Augmentation

A huge amount of high-quality data is required to develop a robust DL model with good performance [Chen and Lin \(2014\)](#). However, obtaining such data is challenging. One solution is to enable practitioners to artificially expand the

diversity of data available in training set by augmenting the original dataset. Data augmentation also prevents overfitting problems and increases the capacity of the model to adjust to the new, unseen data, which is derived from a similar distribution as that of the one used to build the model [Dvornik \*et al.\* \(2019\)](#). As the size of the collected radiology medical cohort was small for effective disease prediction, we applied data augmentation to produce a good amount of class balanced dataset. Data Augmentation must be carefully adapted as the medical images are relatively sensitive to the various operations that can alter the original training set's actual distribution by introducing additional outliers. [Chapter 5](#) offers an extensive overview of the diverse data augmentation techniques applied to both the Indiana University and KMC hospital datasets.

### 6.3.3.2 Generation of Synthetic CXRs using DCGAN

Generative Adversarial Network (GAN) ([Goodfellow \*et al.\*, 2014b](#)) is a DL model used to generate a new set of data from the training set with a similar data distribution. The GAN model comprises two main modules: The generator module generates synthetic or fake images that resemble images in a training set. The discriminator module focuses on the classification or identification of generated images as real or fake. In this experiment, we have utilized a variant of the GAN model named Deep Convolution Generative Adversarial Network (DCGAN) ([Radford \*et al.\*, 2016](#)) to generate synthetic CXRs from the original set of images. DCGAN uses deep convolutional networks instead of fully connected networks as in GAN. In general, the convolution networks can capture better regions of spatial correlations in the images, and hence, the DCGAN is a better fit for image or video data. The architectural diagram of DCGAN is shown in [Figure 6.5](#).

The  $100 \times 1$  latent vector (i.e., random noise) is given as an input to the dense layer of a generator module. The random noise vector is reshaped and transformed into  $8 \times 8 \times 128$ . The output from the dense layer is upsampled by passing it through a series of four transposed convolution (or deconvolution) layers to produce the fake image of size  $128 \times 128 \times 3$ . Leaky Rectified Linear Unit (Leaky ReLU) ([Xu \*et al.\*, 2015](#)) is employed as an activation function for all the transposed convolution layers, and the tanh activation ([Xu \*et al.\*, 2016](#)) function is applied on the final output layer within the generator module. To stabilize the learning process and normalize the input, we have utilized Batch normalization after every transposed convolution layer. Each layer has the deconvolution (deconv2D) layer followed by Leaky ReLU and Batch Normalization. A detailed overview of



generator architecture is depicted in Figure 6.6.

The main aim of the discriminator module is to categorize the generated image as real or fake. The  $128 \times 128 \times 3$  CXR image generated from the generator is passed as an input to the discriminator module. The discriminator module consists of a set of convolution layers through which the CXR image is downsampled. Finally, the image is classified as real or fake when passed through the output layer. The Leaky ReLU activation function is utilized in the convolution layer, and the sigmoid activation function is used in the output layer. The detailed architecture of the discriminator architecture is presented in Figure 6.7. The task of the discriminator is considered a binary classification problem since the images are classified as real or fake.

There are two main steps for training DCGAN. First, the generator module is trained to produce the fake image, and later, the discriminator module is trained to accurately classify the image as real or fake. We have utilized binary cross-entropy as the loss function, and the Adam optimizer is used with the beta1 hyperparameter of 0.5. The DCGAN is trained for 200 epochs with a learning rate of 0.0002. The dropout probability of 0.4 is applied before the sigmoid output classification. The stride of (4,4) and (5,5) are used for the generator deconvolution layers and the discriminator convolution layers, respectively. After 200 epochs, the DCGAN could generate synthetic CXR images resembling the original images.

### 6.3.4 Network Configurations and Parameter Settings

The NVIDIA Tesla M40 server with 128 GB of RAM, a 24GB GPU, a 3TB Hard drive, and a Linux server operating system is used for our investigation. The multimodal clinical cohort is split into training/validation and testing sets (refer to Table 6.1). Both unimodal and multimodal networks were trained for 100 epochs and 10-cross fold validation to evaluate the performance of the proposed models. The Python 3.6 with Keras library and Tensorflow (Abadi and et. al., 2015), a popular DL platform, were utilized to implement the proposed multimodal medical tensor fusion network. The optimum hyperparameters are determined by exploiting the grid search strategy (Bergstra and Bengio, 2012) to fine-tune the model's parameter setting. We have utilized Adam (Kingma and Ba, 2015) as an optimizer, and the binary cross-entropy is leveraged as a loss function.

For the text encoding model UM-TES, we use the GloVe model with CKB, which takes findings from the radiology report with dimension 260 as an input and gives the word embeddings of size  $260 \times 100$  as an output. The output dimension of

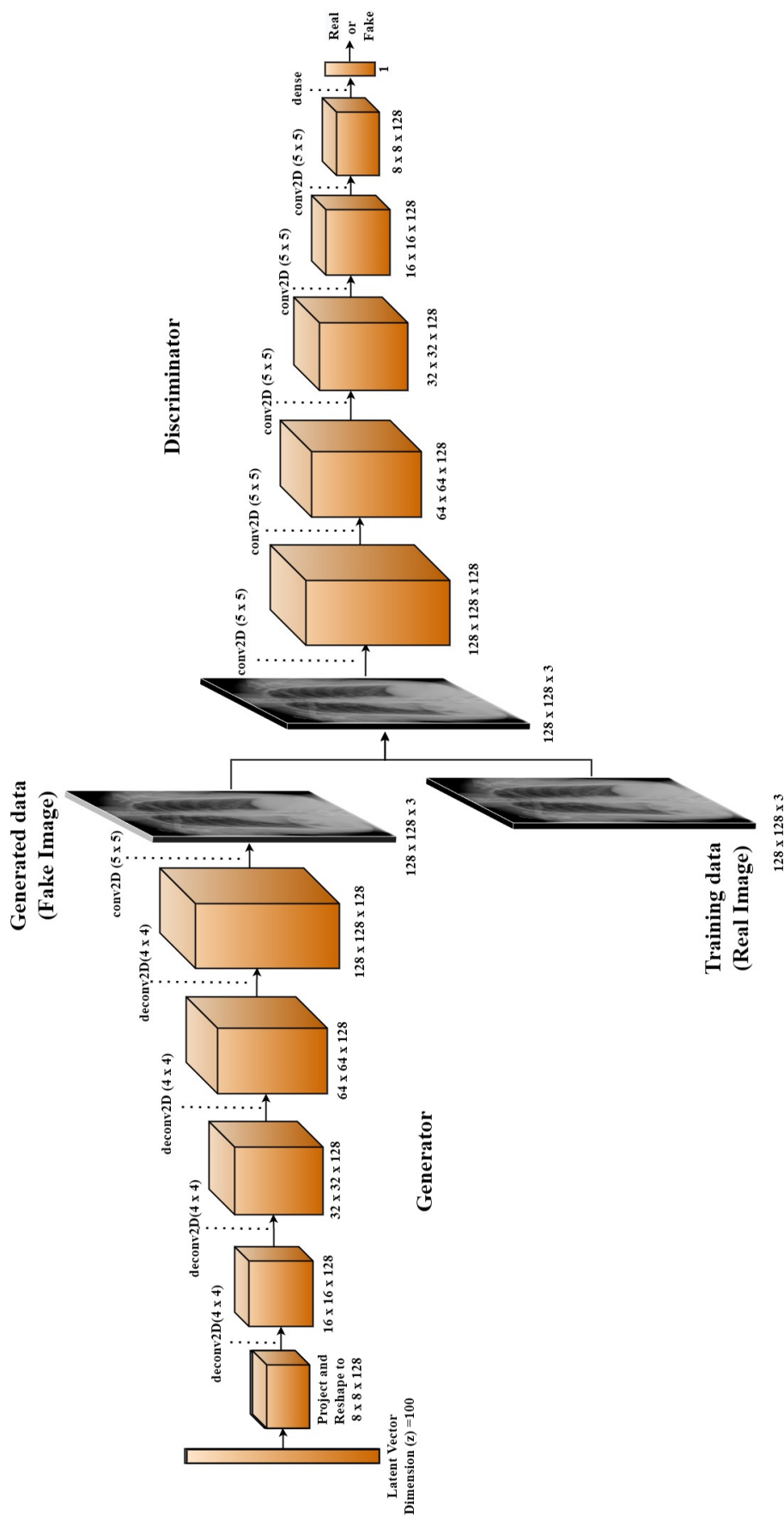


Figure 6.5: Architectural diagram of Deep Convolution Generative Adversarial Network

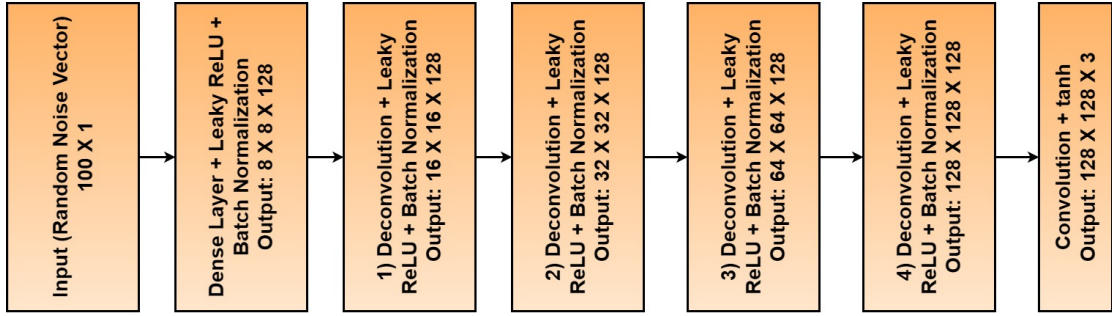


Figure 6.6: Generator Module architecture of DCGAN

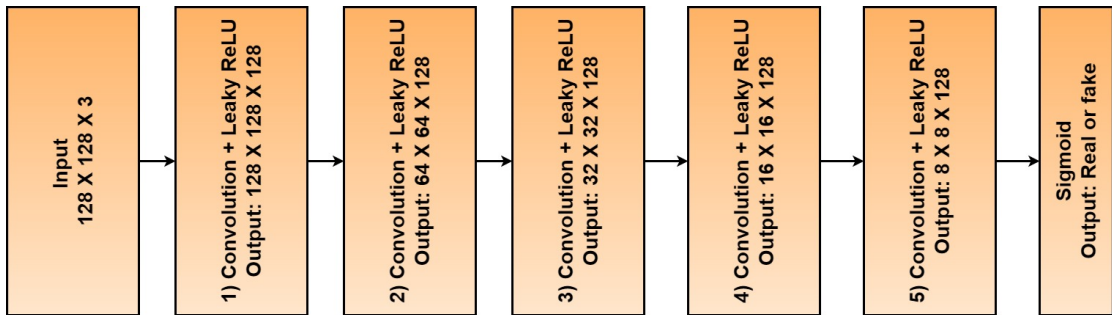


Figure 6.7: Discriminator Module architecture of DCGAN

the GloVe model is set to 100. The DDR-CNN is applied to the word embeddings of size  $260 \times 100$  to generate discriminative clinical features of size 1024. Based on the grid search strategy (Bergstra and Bengio, 2012), we use a kernel size of 5, a filter size of 32, strides of 1, a pool size of 2, a dropout of 0.4, and a learning rate of 0.001 as hyperparameters for DDR-CNN. For the image encoding model UM-VES, we utilize a multichannel dilation layer followed by concatenation, 13 depthwise separable convolution layers, 27 BN and ReLU layers, and a final global average pooling layer. We have used the optimal learning rate of 0.001 for UM-VES. The CXRs with an input dimension  $150 \times 150 \times 3$  is passed as an input to the UM-VES to obtain the imaging features of size 1024 from the final global average pooling layer.

The textual features with size 1024 obtained from UM-TES and the imaging features with size 1024 generated from UM-VES are passed through two Multimodal models, CBP-MMFN and DHP, separately to produce effective multimodal representations. In CBP-MMFN, a bilinear pooling map with a size of  $1024 \times 1024$  is produced by the outer product between the textual and imaging features. These features are further ingested into the CNN layer with kernel size: (3,3), Filters: 2, strides: (1,1), pool size: (8,8) and dropout: 0.2 to obtain the compact multimodal features of size 1024. In DHP-MMFN, the Hadamard product

between the imaging and textual features is generated to produce the multimodal features of size 1024. The multimodal feature of size 1024 is passed through a fully connected DNN with three hidden layers (i.e., 256, 128, and 2) for disease prediction. The ReLU non-Linear activation function (Krizhevsky *et al.*, 2012b) is applied at the initial two layers, and the Sigmoid logistic function is utilized at the final layers for binary classifications. The model regularization dropouts (Hinton *et al.*, 2012) with probability of 0.2 are used after the first layer to prevent the model from overfitting. The early stopping mechanism (Yao *et al.*, 2007) is incorporated to stop the training process when the performance of the validation set degrades during the training epochs. The early stopping avoids overtraining and regularizes the model by reducing the overfitting problem during training and enhancing the model’s generalization ability.

### 6.3.5 Ablation Study

In this section, we conduct an ablation study of the proposed Multimodal Medical Tensor Fusion Network to ablate from its complete form and check each module’s contribution in predicting diseases from heterogeneous clinical data. The ablation study is conducted on two multimodal clinical datasets (i.e., Indiana University and KMC hospital datasets), as outlined in Table 6.2. The multimodal network is divided into two unimodal subnetworks (i.e., UM-TES and UM-VES) and two multimodal subnetworks (i.e., CBP-MMFN and DHP-MMFN). In our ablation study, we have analyzed the performance of unimodal subnetworks by removing the multimodal subnetworks. The imaging and textual features obtained from UM-VES and UM-TES are separately passed to the fully connected prediction model. We have also investigated the performance of the two multimodal subnetworks by combining the visual and textual features obtained from the two unimodals. The result shows that including multimodal models to combine imaging and textual features has significantly increased the overall network’s performance. The CBP-MMFN is seen as the dominant multimodal tensor fusion scheme in our experiment. The reason is that the CBP-MMFN achieves intermodal interaction and captures more correlation between the textual and imaging features obtained from the unimodal models.

Table 6.2: Experimental results of the proposed unimodal and multimodal model for abnormality prediction from CXR images and its associated radiology reports collected with standard augmented data from Indiana University and KMC hospital dataset.

Model Type	Modality	Proposed Models	Indiana University Dataset					
			Accuracy	Precision	Recall	F1 Score	MCC	AUROC
Unimodal	Image Only	UM-VES	79.22%	0.7926	0.7928	0.7927	0.5855	0.8572
	Text Only	UM-TES	90.40%	0.9080	0.9040	0.9059	0.7939	0.9555
Multimodal	Image + Text	DHP-MMFN	96.27%	0.9629	0.9629	0.9634	0.9263	0.9844
		CBP-MMFN	97.35%	0.9735	0.9737	0.9736	0.9472	0.9876
Model Type	Modality	Proposed Models	KMC Hospital Dataset					
Unimodal	Image Only	UM-VES	82.25%	0.8201	0.8200	0.8200	0.6401	0.8793
	Text Only	UM-TES	94.13%	0.9475	0.9413	0.9443	0.8827	0.9651
Multimodal	Image + Text	DHP-MMFN	95.46%	0.9512	0.9506	0.9509	0.9018	0.9805
		CBP-MMFN	96.94%	0.9787	0.9635	0.9661	0.9385	0.9834

### 6.3.6 Performance Analysis of Unimodal and Multimodal Models

In order to validate the impact of the two proposed heterogeneous medical fusion networks, we conducted a comparative study with the two unimodal models. The benchmarked results of the proposed unimodal and multimodal models for predicting disease abnormalities from the Indiana University and KMC hospital datasets are presented in Table 6.2. Figure 6.8 and Figure 6.9 shows the graphical summary showcasing performance metrics of proposed unimodal vs multimodal models for the Indiana University and KMC Hospital datasets, respectively.

The proposed heterogeneous models DHP-MMFN and CBP-MMFN are proven as the supreme models with approximately 6-7% and 16-17% increase in the performance (Accuracy, Precision, recall, F1-score, and MCC) compared to the Unimodal strategies UM-TES and UM-VES, respectively, for the Indiana University corpus. The multimodal models have also shown a significantly higher AUROC curve when compared to unimodal models, proving that the multimodal models (DHP-MMFN and CBP-MMFN) have better capability in distinguishing between positive (i.e., normal) and negative (i.e., abnormal) classes. Also, the text-only (UM-TES) model has outperformed the image-only (UM-VES) model with an over 10-11% increase in the evaluation metrics score. For the KMC Hospital dataset, the multimodal models DHP-MMFN and CBP-MMFN have yielded a better performance of 13-16% than the unimodal image-only model (UM-VES). The text-only (UM-TES) model is the highest performing unimodal, with significantly good results. The multimodal models DHP-MMFN and CBP-MMFN have better competence in characterizing the dataset into positive and negative classes than the unimodal models, as reflected in the AUROC of multimodal models compared to the AUROC of unimodal models. The multimodal models have a higher F1-score and MCC than unimodal models, signifying that even if there is a class imbalance in the Indiana University dataset, the multimodal models are competent enough to classify the radiology exams into normal and abnormal accurately.

From the comprehensive analysis of two medical cohorts, it is seen that both multimodal models have given significantly superior results compared to the unimodal models. This is because, in multimodal strategies, there is an intermodal interaction that is missing in unimodal models. It is observed that for the Indiana University and the KMC hospital cohorts, the text modality has a considerable impact on the performance of the multimodal models than the imaging modality, as reflected by an AUROC of 0.9555 and 0.9651 for text-only models compared to

an AUROC of 0.8572 and 0.8793 for image-only models. We believe this is because of the annotation process of the Indiana University and KMC hospital datasets. The annotators have focused on text being assigned to the labels of the radiology reports, and the results reflect that the most discriminative features are found in the text modality. Also, it is seen that the knowledge base incorporated in the UM-TES has significantly impacted the performance of the text-only model. The extensive vocabulary of clinical words was obtained from the knowledge base that helps the model learn unseen and infrequent words. The CBP-MMFN has given the highest performance in two multimodal clinical cohorts out of two multimodal models. The intermodal correlation encoded in the bilinear interaction map of CBP-MMFN has more discriminative features than the DHP-MMFN interaction map.

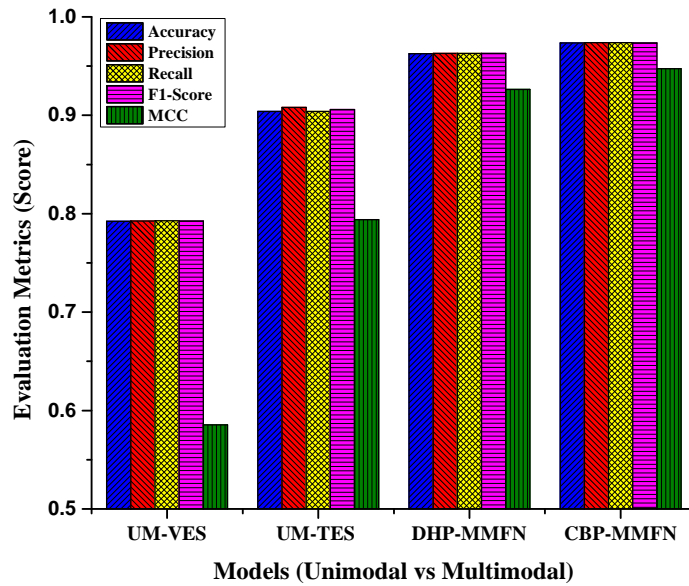


Figure 6.8: Comparison of performance metrics of proposed unimodal vs multimodal models for Indiana dataset

### 6.3.7 Performance Analysis on Synthetic Data Generated

Table 6.3 showcases the performance of the proposed unimodal and multimodal models for pulmonary abnormality prediction from the synthetic data generated using DCGAN. As described in Section 6.3.3.2, we have used DCGAN for generating synthetic CXRs. The discriminator network is made to guess between the

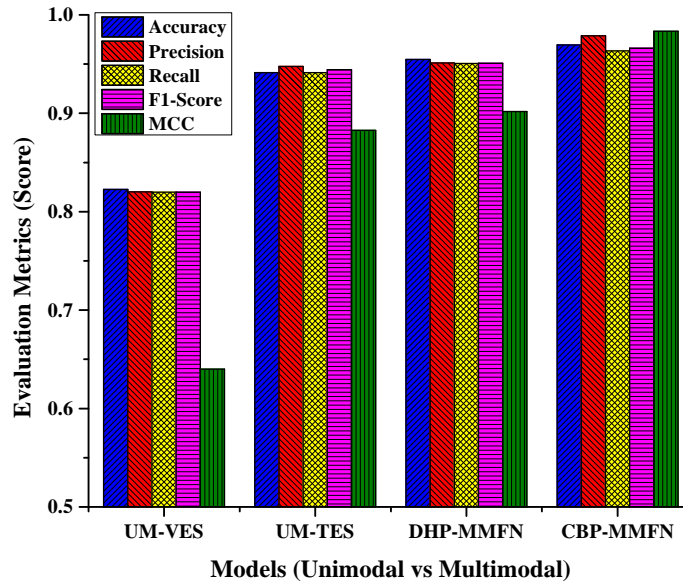


Figure 6.9: Comparison of performance metrics of proposed unimodal vs multimodal models for KMC dataset

fake and real CXR images. The discriminator accuracy in determining real and fake samples during training (200 epochs) of DCGAN for the Indiana University and KMC hospital datasets is shown in Figure 6.10. The generator and discriminator losses during the training (Batch size= 128 and 200 epochs) of DCGAN are depicted in Figure 6.11, where discriminator and generator losses are around 0.5 for both the datasets. The process of synthetic data generation from the latent noise after every 20 epochs for the KMC hospital dataset is presented in Figure 6.12. The total number of cases after synthetic data generation for the Indiana University and KMC hospital datasets is shown in the Table. 6.1.

It is observed that training our proposed unimodal and multimodal models with the actual CXR images and synthetic CXR images yields better performance than training the models with the actual CXR images with the standard augmented images (refer to Table Table. 6.2). There is a significant increase in accuracy of 1-2% in image only, text only, and image+text models for both medical cohorts. This suggests that the generated synthetic CXRs comprise more meaningful features that help enhance the model's performance. Also, there is an increase in precision and recall recorded for pulmonary abnormality prediction from both datasets. This shows that the features obtained are more discriminative to categorize between the disease and no disease. The higher MCC and F1-Score are



obtained for both unimodal and multimodal models, showcasing accurate disease prediction despite an uneven class distribution in the Indiana University dataset. The synthetic CXR data generated provides more variability to the two medical cohorts by increasing their size. The synthetic CXR generated comprises varied features compared to the original set of images. The proposed models provide good results on this synthetic data, proving that the proposed unimodal and multimodal models can predict and adapt to new and unseen data. The comparison of proposed unimodal vs. multimodal performance metrics on actual and synthetic multimodal data for Indiana University and KMC hospital datasets is shown in Figure 6.13 and Figure 6.14, respectively.

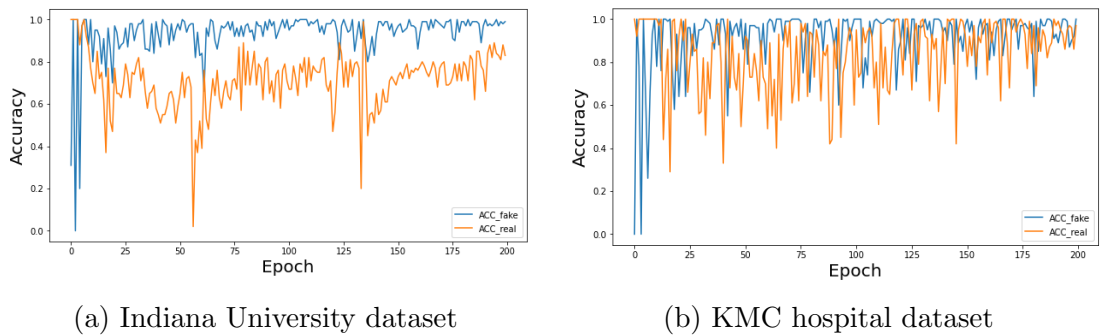


Figure 6.10: Discriminator Accuracy on real and fake samples during training of DCGAN

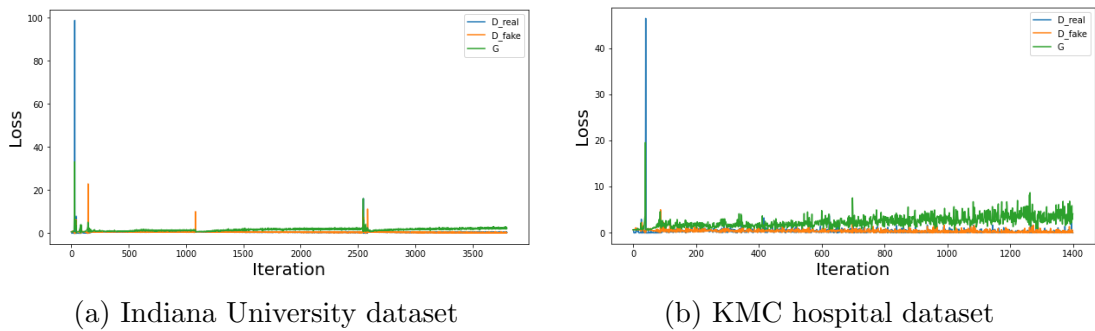


Figure 6.11: Generator and Discriminator loss during the training of DCGAN

### 6.3.8 Performance Comparison with the State-of-the-art Models

The experimental outcomes of the existing State-of-the-Art multimodal fusion strategies on Indiana University CXR and their associated clinical reports are outlined in Table 6.4. Our two multimodal fusion strategies, DHP-MMFN and

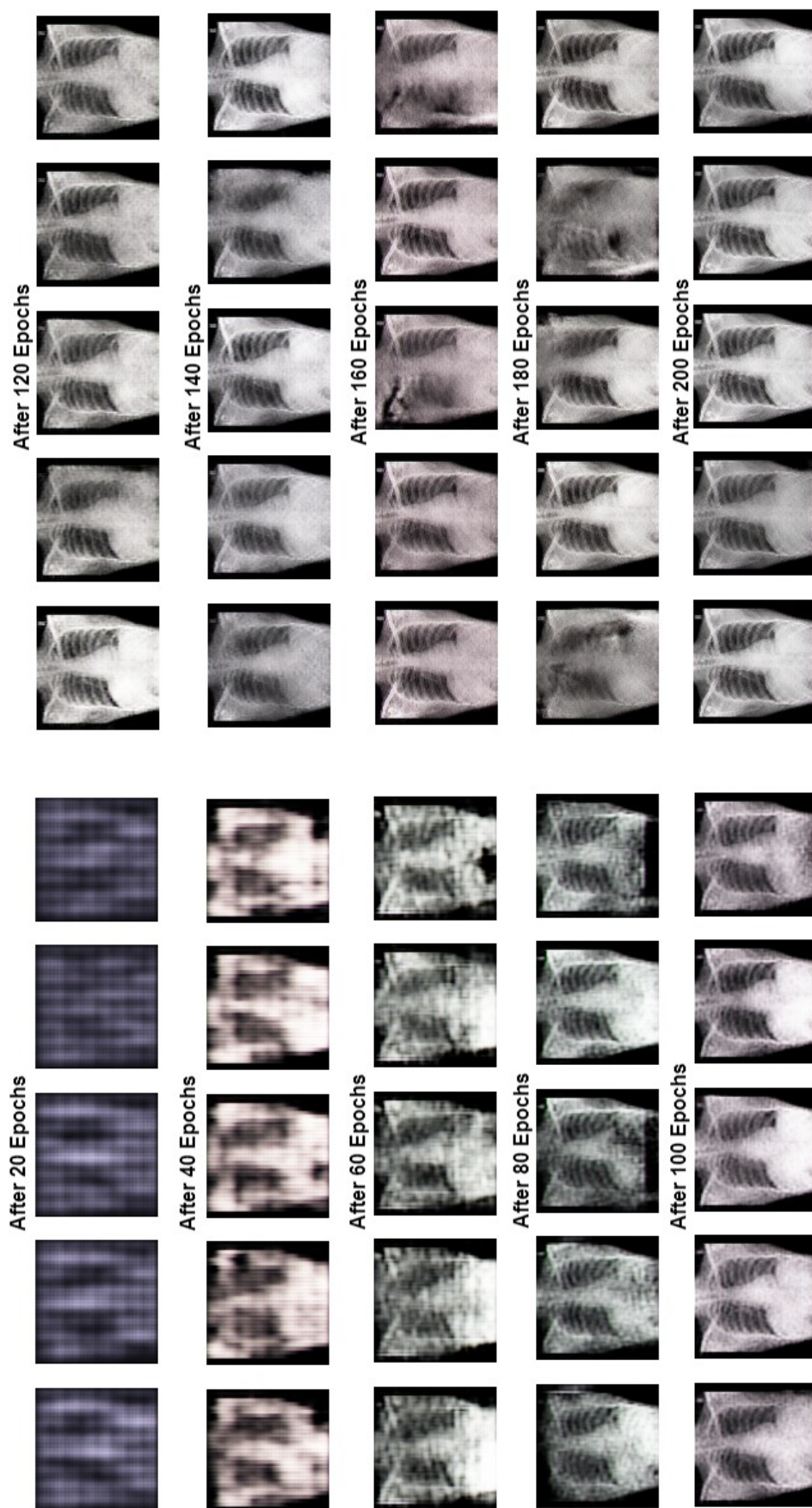


Figure 6.12: The synthetic CXR images generated after every 20 epochs for KMC hospital dataset

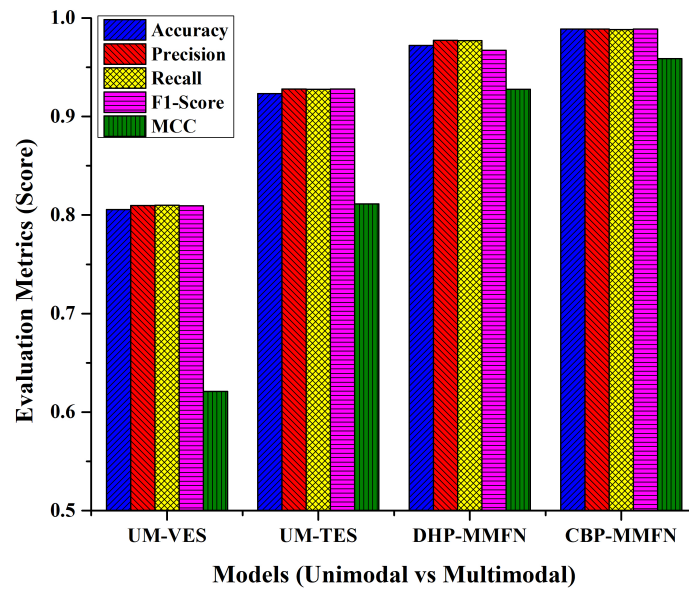


Figure 6.13: Comparison of performance metrics of proposed unimodal vs multimodal models on Actual data with synthetic data generated from Indiana University dataset

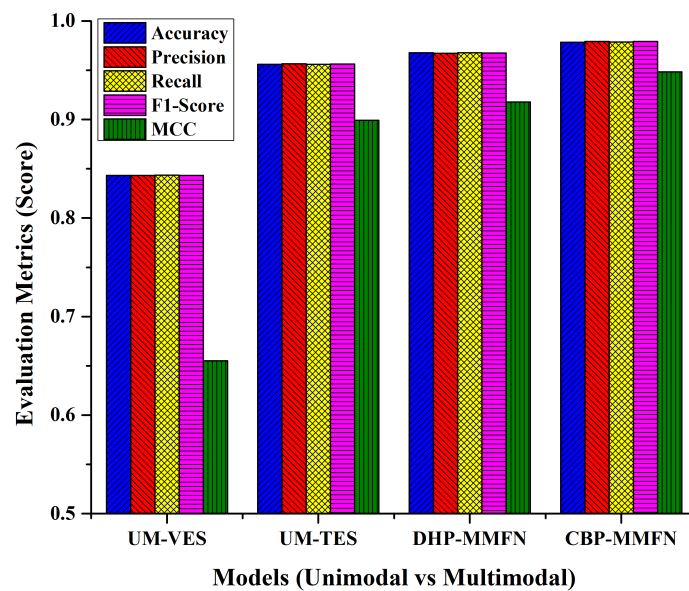


Figure 6.14: Comparison of performance metrics of proposed unimodal vs multimodal models on Actual data with synthetic data generated from KMC Hospital dataset

Table 6.3: Experimental results of the proposed unimodal and multimodal models for abnormality prediction on Actual data with Synthetic CXRs and radiology reports generated from Indiana University and KMC hospital datasets using DCGAN.

Model Type	Modality	Proposed Models	Indiana University Dataset					
			Accuracy	Precision	Recall	F1 Score	MCC	AUROC
Unimodal	Image Only	UM-VES	80.56%	0.8096	0.8098	0.8095	0.6211	0.8633
	Text Only	UM-TES	92.32%	0.9278	0.9275	0.9279	0.8112	0.9570
Multimodal	Image + Text	DHP-MMFN	97.21%	0.9774	0.9769	0.9673	0.9277	0.9877
		CBP-MMFN	98.88%	0.9888	0.9883	0.9887	0.9588	0.9889
Model Type	Modality	Proposed Models	KMC Hospital Dataset					
			Accuracy	Precision	Recall	F1 Score	MCC	AUROC
Unimodal	Image Only	UM-VES	84.33%	0.8432	0.8436	0.8432	0.6551	0.8829
	Text Only	UM-TES	95.61%	0.9565	0.9561	0.9563	0.8991	0.9679
Multimodal	Image + Text	DHP-MMFN	96.78%	0.9673	0.9677	0.9674	0.9178	0.9877
		CBP-MMFN	97.83%	0.9791	0.9787	0.9791	0.9483	0.9971

CBP-MMFN, remarkably surpass the six existing state-of-the-art multimodal fusion strategies. Both the proposed multimodal models have shown 14-15% more accuracy compared to the existing model fusion technique (Aydin *et al.*, 2019a), where they have used concatenation for fusing the tensors. The proposed multimodal models have achieved superiority over another multimodal model fusion strategy, MFT (Lopez *et al.*, 2020), where standard concatenation is applied to combine the visual and textual features. Our multimodal models yield better results due to the intermodal interaction between the text and imaging features, which is missing in standard concatenation, where features from two modalities are joined before passing through the prediction model. It is found that the state-of-the-art MFT, EFT, and LFT models, including the proposed DHP-MMFN and CBP-MMFN models, have competent recall rates, which proves that all the models effectively predict the true positives (i.e., disease abnormalities). That is, recall indicates how many were correctly identified as disease abnormality out of all the cases having disease abnormalities. As observed from Table 6.4, our proposed model yields 5-10% more precision compared to existing models. This proves that our proposed model correctly predicts the abnormalities from all the cases. The proposed models, CBP-MMFN and DHP-MMFN, have superior F1 scores compared to state-of-the-art models. The higher model F1-score signifies that the proposed model has accurately categorized the cohort as normal and abnormal, even though there exists a class imbalance in the Indiana University dataset. There is a significant improvement of 5-7% in the AUROC of the proposed fusion model compared to state-of-the-art models, indicating that there is excellent separability between two classes (i.e., normal and abnormal). From the experiment, it was found that the proposed multimodal models have significantly outperformed the state-of-the-art models. In the existing state-of-the-art approaches, either concatenation or late fusion techniques like averaging are used to combine the features, where the inter-modal interactions are completely ignored. In our proposed model, the inter-modal dynamics between the visual and textual modalities are considered, which is the major reason for performance gains.

After performing a comprehensive investigation on two multimodal clinical datasets, it is found that multimodal learning provides a benefit over unimodal learning when predicting diseases from the radiology CXR and associated clinical free-text notes. With regards to the two proposed multimodal fusion strategies, CBP-MMFN performs better than the DHP-MMFN model across publicly available Indiana University cohorts and data collected from the KMC hospital. The superior results in CBP-MMFN are obtained because of the intermodal dynamics

Table 6.4: Comparing the performance of the proposed multimodal models against the existing state-of-the-art medical multimodal fusion model for abnormality prediction from CXR and its associated radiology reports from the Indiana University dataset. The results of the state-of-the-art medical multimodal fusion model is taken from their published research work.

Medical Multimodal Fusion Model	Reference	Accuracy	F1 Score	Precision	Recall	AUROC
Wang et al. (2019) - RNN (Text) + CNN (Image) + Concatenation (Fusion) (2019)	Wang et al. (2018a)	-	-	-	-	0.965
Aydin et al. (2019) - Custom Glove (Text) + Pretrained Densenet121 (Image) + Concatenation (Fusion) (2019)	Aydin et al. (2019a)	81%	-	-	-	-
Nunes et al. (2019) - LSTM based BioWordVec (text) + EfficientNet-B5 (image) + Concatenation (Fusion) (2019)	Nunes (2019)	96.34	0.7919	0.8654	0.7299	-
Model Fusion Technique - Word2Vec (Text) + DenseNet121 (Image) + Concatenation (Fusion) (2020)	Lopez et al. (2020)	-	0.86	0.78	0.97	0.91
Early Fusion Technique (2020)	Lopez et al. (2020)	-	0.91	0.90	0.92	0.93
Late Fusion technique- Word2Vec (Text) + DenseNet121 (Image) + Averaging (2020)	Lopez et al. (2020)	-	0.89	0.83	0.96	0.93
DHP-MMFN	This Study	97.21%	0.9634	0.9629	0.9629	0.9844
CBP-MMFN	This Study	98.88%	0.9736	0.9735	0.9737	0.9876

between the textual and imaging modalities. The Bilinear interaction map generated from the outer product of visual and textual features in CBP-MMFN generates a far more expressive multimodal feature representation, encoding more tensor correlation than the simple concatenation operation and element-wise product. Hence, the discriminative features extracted from the CBP-MMFN model provide a significant performance gain over the uni-modal models and the DHP-MMFN model. The unimodal text-only model (UM-TES) has given more promising results than the proposed unimodal image-only (UM-VES) model. The two major reasons for it are as follows:

- Incorporating a *CKB* helps to jointly learn word vectors from the cohort and knowledge base, which increases the vocabulary size and allows the learning of infrequent clinical words.
- It has been found that radiology reports have more discriminative features than CXRs. This is because the annotators have focused on the text being assigned to the labels of the radiology reports.

The proposed models also recorded good performance on synthetic data generated using DCGAN, proving their ability to predict from new and unseen data. We also observed that the existing state-of-the-art multimodal fusion techniques applied to radiology images and their associated reports are either straightforward concatenation or late fusion techniques like averaging, which ignore inter-modal interaction among the two modalities. The proposed multimodal medical tensor fusion techniques have given supreme results compared to the existing state-of-the-art methods. The proposed models focus on inter-modal dynamics, which find the tensor correlation between the textual and imaging modalities. The experimental results prove that the multimodal representation obtained from the proposed model has more expressive features than the traditional concatenation strategy.

## 6.4 Summary

The chapter introduces two Multimodal Medical Tensor Fusion Networks, CBP-MMFN and DHP-MMFN, which are designed to predict abnormalities in radiology CXR scans and their associated reports. The proposed models are evaluated on two multimodal radiology datasets: a publicly available Indiana University dataset and real-time data collected from KMC Private Hospital. We found that the multimodal models perform better than unimodal models and show superior performance compared to state-of-the-art heterogeneous fusion techniques for predicting



abnormalities in the radiology cohort. Additionally, we evaluate the models on synthetic data generated using DCGAN and observe that both the unimodal and multimodal models perform better on the synthetic data generated, showcasing their ability to predict from new and unseen data distributions. Overall, the chapter highlights the potential of multimodal learning for predicting abnormalities in radiology CXR with associated reports. The proposed models show competitive performance and have the potential for applications in real-world healthcare settings.

## Publications

*(based on study proposed in this chapter)*

1. Shashank Shetty, Ananthanarayana V. S., and Ajit Mahale (2022), Comprehensive Review of Multimodal Medical Data Analysis: Open Issues and Future Research Directions, *Acta Informatica Pragensia (AIP)*, 11(3), 423-457, <https://doi.org/10.18267/j.aip.202> [Indexed: Scopus, IF: 1.15] (*Status: Published Online*)
2. Shashank Shetty, Ananthanarayana V. S., and Ajit Mahale (2023), Multimodal medical tensor fusion network-based DL framework for abnormality prediction from the radiology CXRs and clinical text reports. *Multimedia Tools and Applications*, Springer Publisher, <https://doi.org/10.1007/s11042-023-14940-x> [Indexed: SCIE & Scopus, IF: 2.577] (*Status: Published Online*)



## Chapter 7

# Cross-Modal Deep Learning Framework for Diagnostic Report Generation from Chest X-ray Images

### 7.1 Introduction

Medical imaging is an essential aspect of modern healthcare that offers doctors valuable insights into a patient's internal structures. Imaging techniques like X-rays, CT scans, MRIs, and ultrasounds allow doctors to diagnose illnesses and plan treatments with accuracy. Radiologists are healthcare specialists who specialize in analyzing medical images to create comprehensive radiology reports that contain detailed findings and impressions (Jing *et al.*, 2018). Examining medical images and creating diagnostic notes is a time-consuming process that demands a high degree of proficiency. Radiologists must thoroughly scrutinize the medical images, cross-reference them with the patient's medical history, and deliver precise and thorough interpretations. Despite their experience and proficiency, radiologists may come across indeterminate results that necessitate additional investigation. Consequently, patients may need to undergo further testing, such as advanced imaging or pathology, to determine the root cause of the issue. Furthermore, when hospitals become crowded, or there is a surge in patient numbers during pandemics, radiologists may find it challenging to meet the demand for detailed reports (Yang *et al.*, 2022). The increased workload may cause burnout, errors, or delays in providing care to patients. To overcome these difficulties, experts have turned to AI and DL technologies to automate the generation of radiology reports. Using this approach can lessen the burden on radiologists, enhance the speed and precision of diagnoses, and ultimately improve patient care.

### 7.1.1 Problem Statement

The increasing demand for accurate and timely radiology reports, coupled with the challenges radiologists face in examining medical images and creating diagnostic notes, has led to burnout, errors, and delays in providing care. While experts have turned to AI and DL technologies to automate the generation of radiology reports, implementing and adopting these technologies face several challenges. These include the need to address concerns regarding the accuracy and reliability of automated notes, integrate these technologies into existing clinical workflows, and ensure that they are accessible and affordable to all healthcare facilities. Addressing these challenges will be crucial in realizing the potential of AI and DL technologies to improve the speed, accuracy, and efficiency of radiology diagnoses, ultimately enhancing patient care outcomes. The problem statement is defined as follows:

*“Considering the multimodal medical cohort containing radiology images with associated diagnostic notes, design and develop an automatic diagnostic report generation by analyzing the visual features from the Chest X-ray scans.”*

We propose a solution to the challenge by developing a deep encoder-decoder model that can automatically generate reports from chest X-rays. To achieve this, we have utilized a Multi-channel dilation layer with Depthwise Separable Convolution Neural Network to extract imaging features and knowledge-based text modelling for textual feature extraction. Finally, the LSTM model is used to fine-tune the generated report. We summarize the contributions of this study as follows:

- We propose an encoder-decoder-based deep learning framework to generate diagnostic radiology reports for given chest x-ray images.
- We have developed a dynamic web portal that can efficiently take in chest X-ray images as input and generate radiology reports as output, thereby providing an accessible and user-friendly solution.
- We conduct a comprehensive analysis and compare the performance of the proposed model with state-of-the-art deep learning approaches.

## 7.2 Methodology

The proposed encoder-decoder framework aims to generate radiology reports from chest X-rays, which include both frontal and lateral images. During the training process, both the chest X-rays and the corresponding reports are provided as input to the encoder. The encoder consists of two components: the UM-VES (refer Chapter 5) for extracting visual features and the UM-TES (refer Chapter 4) for extracting textual features. These features are then used by the LSTM-based decoder to generate the reports. The encoder operates by processing each item in the input sequence and aggregating the captured information into a context vector. Once the entire input sequence has been processed, the encoder transfers the context vector to the decoder, which generates the output sequence item by item. This process allows the model to effectively combine the visual and textual information and generate contextually relevant reports. The detailed architecture of the proposed cross-modal retrieval is shown in Figure 7.1. During the training phase, the model aims to establish connections between the textual information in the reports and the visual features extracted from the chest X-ray images. The UM-TES approach is employed to encode the textual information, while the UM-VES technique is used to extract visual features. These modalities are then integrated into a joint representation, enabling the model to learn the correlations between the input chest X-ray images and the associated textual information in the reports. By iteratively optimizing the model's parameters, it gradually acquires the capability to generate coherent and contextually relevant reports. During the testing phase, only the chest X-ray images are provided as input to the trained model. Drawing upon the learned associations between the image and the textual information from the training phase, the model generates a report based solely on the input image. This process is achieved by utilizing the decoding mechanism of the trained model, such as a Long Short-Term Memory (LSTM), to generate the text-based output.

### 7.2.1 Unimodal Medical Visual Encoding Subnetwork (UM-VES)

The proposed UM-VES utilizes a multichannel dilation layer with a depthwise separable convolution neural network to extract the imaging features. Compared with the conventional convolution layer, the proposed multichannel dilation convolution layer yields complete imaging information by producing a larger receptive

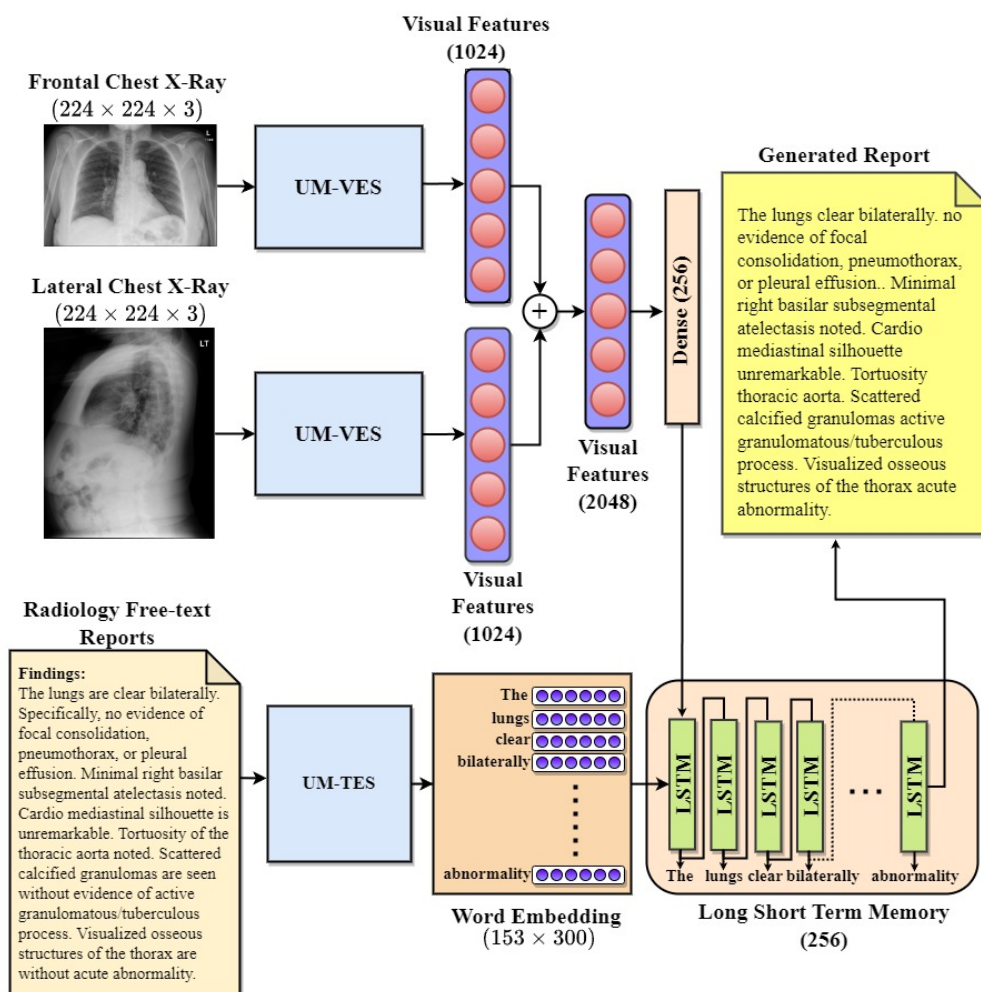


Figure 7.1: Overall architecture of the proposed cross-modal deep learning-based model for automatic report generation

field without increasing the network parameters. Further, the Depthwise Separable convolution network is applied instead of the traditional convolution network to minimize computation burden at each layer evenly. The UM-VES for extracting the textual features from the chest X-ray is presented in Chapter 5. The UM-VES framework is used to extract visual features from both the frontal and lateral CXR images independently, and the resulting features are combined by concatenation.

## 7.2.2 Unimodal Medical Text Embedding Subnetwork (UM-TES)

As an overview, the radiology findings are pre-processed to obtain the essential latent medical concepts. The word embeddings are learnt from the medical words by applying customized Clinical Knowledge-based Text Modelling. Dense word

embeddings obtained are mapped to the medical words from the findings in the Embedding Layer. The detailed explanation of UM-TES for extracting textual features is explained in Chapter 4.

### 7.2.3 Long Short-term Memory-based Report Generation

The fundamental concept of utilizing LSTM for report generation centers around the memory cell, denoted as  $c$ , which primarily stores the information on the input received at any given moment. The function of these cells is controlled by layers or gates that are inserted in a multiplicative manner and can maintain values of either 0 or 1, which are determined by the gates. Specifically, three gates are employed to monitor whether the present value of the cell should be disregarded, if the new cell value should be generated (output gate 0), or if it should be interpreted as input, as illustrated in Figure 7.2. The Equation 7.1, 7.2 and 7.3 depict the input, forget, and output layers, respectively.

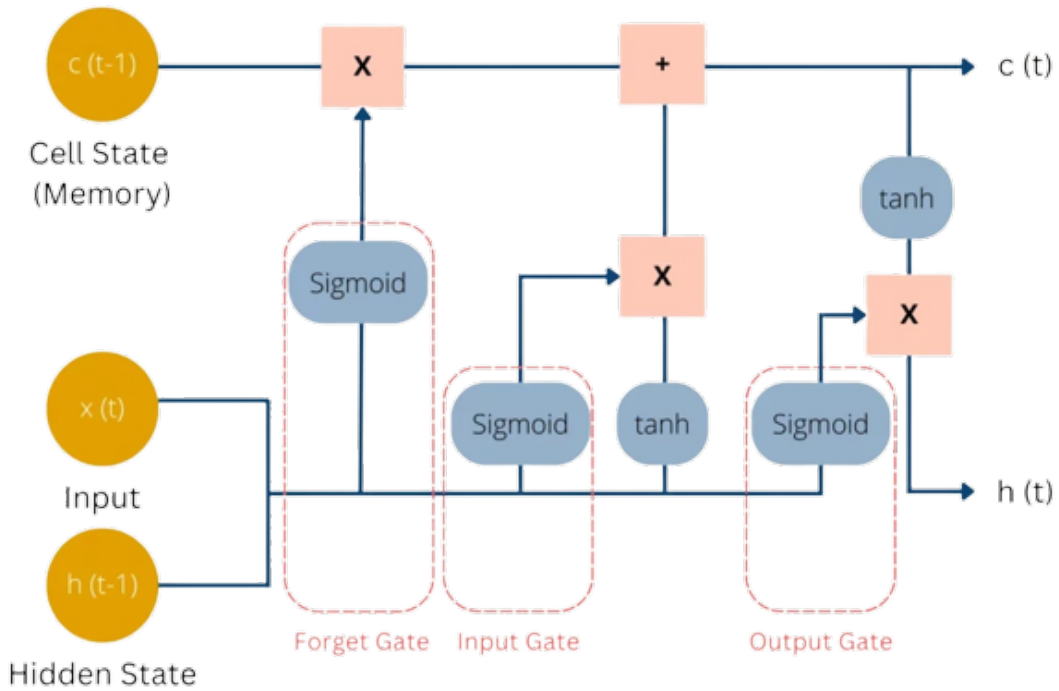


Figure 7.2: Long Short-term Memory Architecture

$$input_t = \sigma(W_{iy}y_t + W_{im}m_{t-1}) \quad (7.1)$$

$$forget_t = \sigma(W_{fy}y_t + W_{fm}m_{t-1}) \quad (7.2)$$

$$output_t = \sigma(W_{oy}y_t + W_{om}m_{t-1}) \quad (7.3)$$

The equation 7.4, 7.5 and 7.6 represent the other operation of the LSTM model.

$$cell_t = forget_t \odot cell_{t-1} + input_t \odot h(W_{cy}y_t + W_{cm}m_{t-1}) \quad (7.4)$$

$$cell_t = output_t \odot cell_t \quad (7.5)$$

$$P_{t+1} = Softmax(m_t) \quad (7.6)$$

Where,  $input_t$ ,  $forget_t$  and  $output_t$  denotes the output of the input, forget and output gates, respectively at time  $t$ ;  $y_t$  represents the input vector at time  $t$ ;  $m_{t-1}$  is the hidden state of the LSTM at time  $t-1$ ;  $W_{iy}$ ,  $W_{fy}$ ,  $W_{oy}$ ,  $W_{cy}$ ,  $W_{im}$ ,  $W_{fm}$ ,  $W_{om}$  and  $W_{cm}$  indicates the weight matrices that manage that manage the input and hidden connections between the input, forget and output gates and cells;  $cell_t$  represents the state of the cell at time  $t$  and  $P_{t+1}$  represents a probability distribution over a set of possible outcomes at time  $t+1$ .

### 7.3 Web-based Framework for Report Generation

We utilized the Flask web framework to create a user-friendly web interface for our model. By uploading both frontal and lateral X-ray images through this interface, users can obtain reports with ease. To streamline the user experience, we implemented Ajax, a technique that enables data to be sent and retrieved asynchronously in the background of the application without requiring the entire page to be reloaded. This approach is particularly useful when we want to update specific portions of an existing page without redirecting or reloading the page for the user. As depicted in Figure 7.3, in order to obtain a report, users are required to upload both frontal and lateral X-ray images. After clicking on the 'Generate Report' button, an Ajax request is sent to the Flask App hosted on the server. The Flask application utilizes the uploaded images to generate predictions for the report, which are then transmitted back to the client side. Upon receipt, the predicted report is displayed to the users.

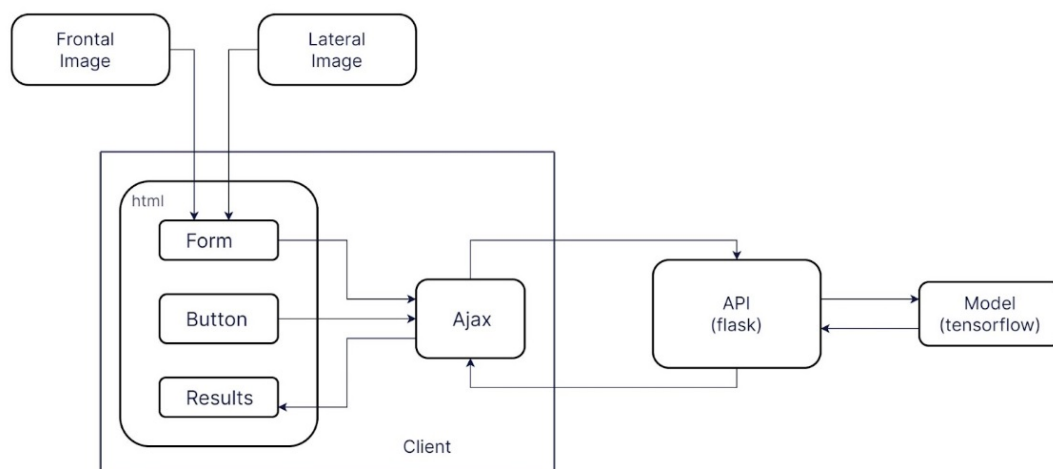


Figure 7.3: Client-Server interaction used for predicting reports

## 7.4 Experimental Setup and Evaluation

For model training, we utilized the IU Chest X-Ray Collection, which includes a comprehensive set of chest x-ray images accompanied by their corresponding diagnostic reports. Please see Section 4.4.1 for a detailed description of the data cohort employed in this study, which comprised 7,470 pairs of images and reports (Number of cases=3996). The reports contained two main sections, impressions and findings. In our investigation, we selected frontal and lateral images and the content of the findings section as the target captions to be generated. To conduct our experiment, we removed cases without reports and frontal/lateral images, ultimately working with 3,638 cases. Two methods were used to generate text reports: greedy search and beam search. Greedy search is an algorithmic approach that incrementally constructs a solution by selecting the next piece that seems to provide the most immediate benefit. In contrast, beam search expands on the greedy search technique by generating a list of the most likely output sequences, each with its own score. The sequence with the highest score is then chosen as the final result.

To evaluate the performance of the generated reports, we incorporated the BLEU score. The BLEU (Bilingual Evaluation Understudy) Score is a method used to evaluate the similarity between a generated sentence and a reference sentence. The score ranges from 0.0, indicating a total mismatch, to 1.0, indicating a perfect match. This approach involves tallying the number of matching n-grams in the candidate text with those in the reference text. For instance, a uni-gram or 1-gram would correspond to each token, whereas a bi-gram comparison would

correspond to each pair of words. Achieving a perfect score is not practical, as it necessitates an exact match with the reference, which even human translators cannot achieve. Furthermore, comparing scores across datasets can be difficult due to the number and quality of the references used to determine the BLEU score.

We compute the BLEU score for an automatic report generated using beam and greedy search. It is observed that beam search produces a superior BLEU score compared to the greedy search algorithm. The qualitative analysis of the proposed deep learning-based model using a beam and greedy search algorithm is shown in the Table. 7.1. The BLEU score of 0.5459, 0.4131, 0.386 and 0.3552 are obtained for different n-grams in the greedy search approach. The beam search approach produces a BLEU score of 0.5881, 0.4325, 0.4017 and 0.3860. We have also compared the results with the existing automatic diagnostic report generation work. Most of the existing work has shown lesser BLEU4 as it compares the four words together with the ground truth. Our proposed model outperforms the existing models while generating robust diagnostic reports. This may be due to the multi-channel visual features and knowledge-based discriminate text features extracted in the encoder of the proposed network. The detailed analysis of the various existing models is shown in Table. 7.2.

Table 7.1: Performance analysis of the proposed model

Method	Bleu1	Bleu2	Bleu3	Bleu4
Greedy Search	0.5459	0.4131	0.3864	0.3552
Beam Search	0.5881	0.4325	0.4017	0.3860

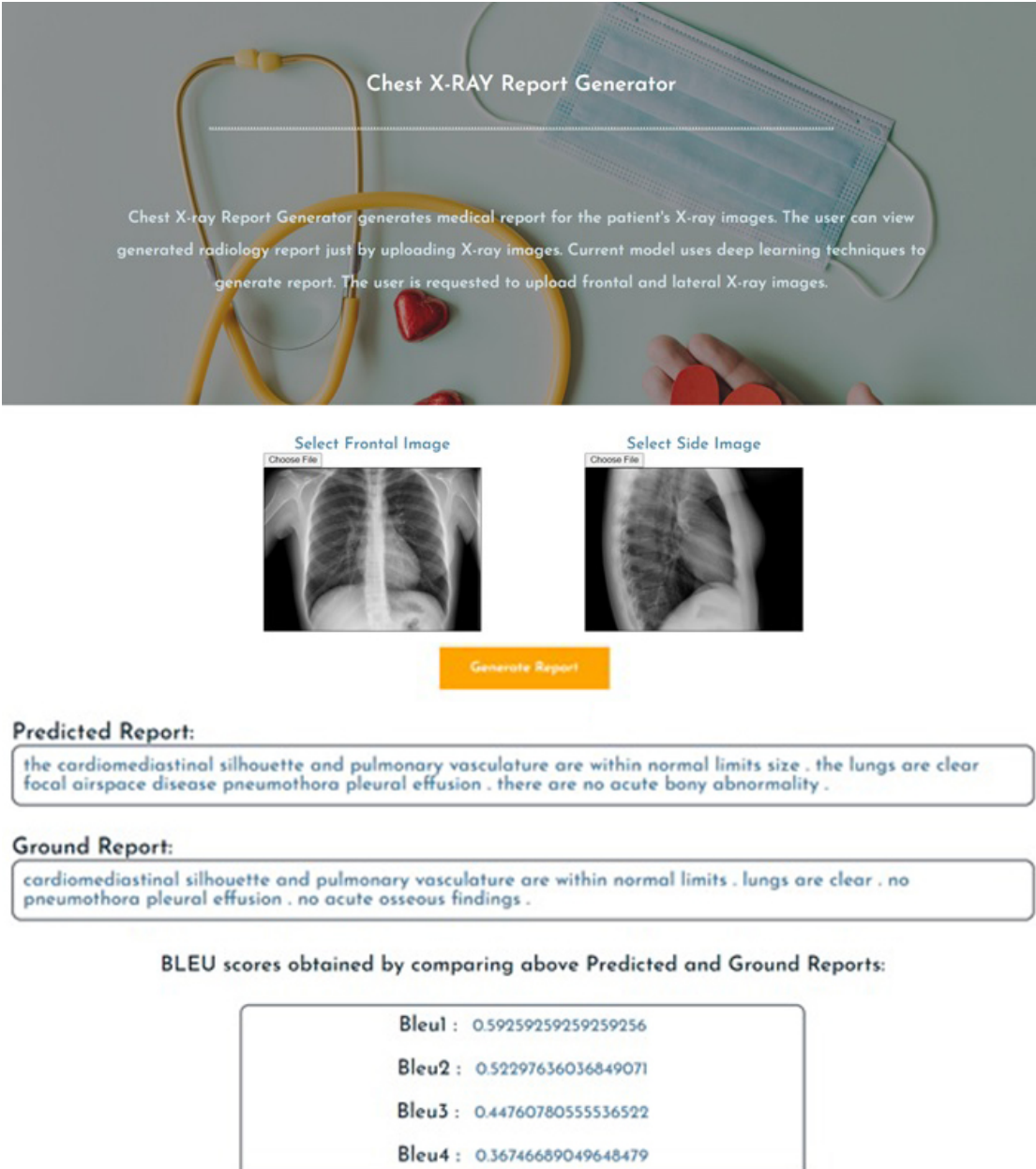
Table 7.2: Performance analysis compared with existing work of report generation

Method	Bleu1	Bleu2	Bleu3	Bleu4
Sai <i>et al.</i> (2021)	0.213	0.258	0.325	0.381
Nguyen <i>et al.</i> (2021)	0.515	0.378	0.293	0.235
Liu <i>et al.</i> (2021)	0.417	0.263	0.181	0.126
Zhou <i>et al.</i> (2021)	0.536	0.392	0.314	0.339
Sirshar <i>et al.</i> (2022)	0.58	0.342	0.263	0.155
Nicolson <i>et al.</i> (2022)	0.4777	0.308	0.2274	0.1773
<b>Proposed Model</b>	<b>0.5881</b>	<b>0.4325</b>	<b>0.4017</b>	<b>0.3860</b>

We designed and developed a flask web application interface for quantitative analysis of the model. Figure 7.4 shows the web interface to upload the chest x-ray images and produce the diagnostic report. The user has to input frontal



and lateral chest X-ray images into the web interface. When the user clicks the “generate report”, an Ajax request will be sent to the Flask App on the server, where the Flask application uses the uploaded images to predict reports. The predicted reports will be sent back to the client, where they are displayed to the users with the BLEU score.



**Chest X-RAY Report Generator**

Chest X-ray Report Generator generates medical report for the patient's X-ray images. The user can view generated radiology report just by uploading X-ray images. Current model uses deep learning techniques to generate report. The user is requested to upload frontal and lateral X-ray images.

Select Frontal Image

Select Side Image

**Predicted Report:**  
the cardiomeastinal silhouette and pulmonary vasculature are within normal limits size . the lungs are clear focal airspace disease pneumothora pleural effusion . there are no acute bony abnormality .

**Ground Report:**  
cardiomeastinal silhouette and pulmonary vasculature are within normal limits . lungs are clear . no pneumothora pleural effusion . no acute osseous findings .

**BLEU scores obtained by comparing above Predicted and Ground Reports:**

Bleu1 : 0.59259259259259256  
Bleu2 : 0.52297636036849071  
Bleu3 : 0.44760780555536522  
Bleu4 : 0.36746689049648479

Figure 7.4: The dynamic web portal for automatic diagnostic report generation.

## 7.5 Summary

This chapter describes an automated framework that employs a deep learning-based encoder-decoder approach to generate reports from chest X-ray scans. The modules used in the framework, such as UM-VES, UM-TES, and LSTM, are discussed in detail. In addition, a dynamic web framework was developed and implemented that accepts chest X-ray images as input and generates diagnostic reports as output. To evaluate the proposed framework, a comprehensive set of experiments was conducted, and the results were compared with those of state-of-the-art report generation frameworks. The proposed framework yielded better performance, as evidenced by an improved BLEU score compared to existing models.

## Publications

*(based on work presented in this chapter)*

1. Shashank Shetty, Ananthanarayana V. S., and Ajit Mahale, Cross-Modal Deep Learning-based Clinical Recommendation System for Radiology Report Generation from Chest X-rays, *International Journal of Engineering*, [Indexed: ESCI & Scopus, IF: 1.64] (*Status: Published Online*)

## Chapter 8

# Multimodal Image Fusion Network for Acute Infarct Prediction from MRI Image Sequences

### 8.1 Introduction

Acute brain infarct is a prevalent cause of fatality and ailment globally, resulting in over 5.5 million deaths annually (Ovbiagele and Nguyen-Huynh, 2011). It is indicated by the abrupt appearance of clinical signs led by focal or global brain dysfunction. These symptoms may persist for more than 24 hours or result in death, and there are no other identifiable factors other than the issues related to vascular origin. The stroke can be categorized as either an ischemic infarct or a hemorrhagic infarct. According to the regulations for early thrombolytic treatment of patients with Acute ischemic infarct, immediate medical diagnosis is the primary therapy for minimizing the brain tissue damage to the maximum extent (Powers *et al.* (2018); Jiang *et al.* (2021)). An acute ischemic infarct is caused due to the instant obstruction of blood flow to a specific area of the brain, leading to the death of brain tissue (French *et al.*, 2016). The leading cause of this occlusion is typically a blood clot that obstructs the arteries that supply blood to the brain. Usually, this kind of infarct is associated with an acute ischemic stroke, characterized by neurological symptoms resulting from a disruption in blood flow to a particular region of the brain. When the supply of blood to the brain is obstructed, the brain cells die within a matter of minutes due to insufficient oxygen and nutrients. This blockage may lead to rapid loss of brain function and result in life-threatening symptoms and complications. The specific symptoms experienced by the patients depend on which part of the brain is affected by the infarct. Such symptoms may include one-sided body weakness or paralysis, problems with speech or language

comprehension, vision loss, and issues with balance and coordination (Shi *et al.* (2022); Winder *et al.* (2022)).

Among the various imaging modalities available, MRI is the most accurate imaging modality for diagnosing acute brain infarct. MRI is a non-invasive imaging method that provides high-resolution images comprising the anatomy of the human body, including organs and tissues. MRI is regularly used to identify and diagnose different cardiovascular and cerebral disorders in the initial stages (Madai *et al.* (2014); Liang *et al.* (2021)). The most reliable MRI sequence for early acute infarct detection is DWI (Arenillas *et al.*, 2002). It can detect a brain infarct within 3 to 6 hours of the onset of the disease process, making it an essential tool for diagnosing and treating ischemic stroke. DWI works by gauging the random Brownian movement of water particles inside a particular tissue region. The ADC provides a quantitative estimate of the extent of limited diffusion within that area. In the DWI sequence, hyperintensity is observed for the ischemic infarct area, and hypointensity is seen in the ADC sequence (Ogbole *et al.*, 2015). In addition to DWI, T2 FLAIR MR imaging is also used to detect acute ischemic infarcts. This sequence involves nulling the signal of cerebrospinal fluid (CSF) using an inversion time ranging from 1500 to 2500 ms. The resulting image will show an area of hyperintensity on an acute ischemic infarct (Brant-Zawadzki *et al.*, 1996). Another MRI technique, SWI, is sensitive to compounds that distort local magnetic fields, such as deoxyhemoglobin, intracellular methemoglobin, and hemosiderin. This fully velocity-corrected, 3D gradient-echo high spatial resolution technique can detect hemorrhagic infarcts as the area of blooming on MRI (Hermier and Nighoghossian, 2004). Figure 1.5 showcases the DWI, T2-Flair, ADC, and SWI MRI sequences of 10 patients data collected from a private medical institute.

MRI captures blood flow dynamics that aid physicians in assessing the risks and benefits of reperfusion therapy (Kang *et al.* (2012); Kim *et al.* (2023)). However, determining the appropriate course of action is challenging due to the varying size, shape, and location of lesions, as well as the intricate cerebral hemodynamic process (Goyal *et al.*, 2021). Hence, there is a requirement for an automated approach utilizing Artificial Intelligence techniques to predict Acute Infarct from MRI images (Qiu *et al.* (2021); Ozkara *et al.* (2023)). Such a system can aid Radiologists in providing accurate and swift diagnoses, guiding treatment approaches, and ultimately leading to reduced mortality and morbidity rates (Werdiger *et al.* (2022); Shetty *et al.* (2022)). CNNs are a specific kind of neural network that have been devised to process data that is presented in grid-like structures, such as images (Albawi *et al.*, 2017). CNNs have exhibited superior performance com-

pared to traditional methods that rely on manually crafted features (Yamashita *et al.*, 2018). These networks have demonstrated remarkable effectiveness in various computer vision applications, that includes image recognition (Tajbakhsh *et al.*, 2016), identifying and recognizing objects (Ren *et al.* (2017); Hassanzadeh *et al.* (2020)).

To accurately analyze a single case, it's crucial to consider the valuable information contained in various MRI sequences such as DWI, ADC, T2-Flair, and SWI. However, the complex relationships between these sequences and the subtle changes they indicate may not be easily detectable when viewed as a single sequence. To overcome this challenge, we have used multiple channels to combine the information from these sequences and capture the complex interplay between them, improving the model's ability to detect and interpret fine changes. In this investigation, the stacked multi-channel CNN framework is proposed for acute infarct prediction, leading to more accurate and reliable detection of ischemic brain lesions from MRI sequences like DWI, ADC, T2-flair, and SWI. from 1500 to 2500 ms. The resulting image will show an area of hyperintensity on acute ischemic infarct (Brant-Zawadzki *et al.*, 1996). Another MRI technique, SWI, is sensitive to compounds that distort local magnetic fields such as deoxyhemoglobin, intracellular methemoglobin, and hemosiderin. This fully velocity-corrected 3D gradient echo high spatial resolution technique can detect hemorrhagic infarcts as the area of blooming on MRI (Hermier and Nighoghossian, 2004). Figure 1.5 showcases the DWI, T2-Flair, ADC and SWI MRI sequences of 10 patients data collected from private medical institute.

MRI captures blood flow dynamics that aid physicians assess the risks and benefits of reperfusion therapy (Kang *et al.* (2012); Kim *et al.* (2023)). However, determining the appropriate course of action is challenging due to the varying size, shape, and location of lesions, as well as the intricate cerebral hemodynamic process (Goyal *et al.*, 2021). Hence, there is a requirement for an automated approach utilizing Artificial Intelligence techniques to predict Acute Infarct from MRI images (Qiu *et al.* (2021); Ozkara *et al.* (2023)). Such a system can aid Radiologists in providing accurate and swift diagnoses, guiding treatment approaches, and ultimately leading to reduced mortality and morbidity rates (Werdiger *et al.* (2022); Shetty *et al.* (2022)). CNNs are a specific kind of neural network that have been devised to process data that is presented in grid-like structures, such as images (Albawi *et al.*, 2017). CNNs have exhibited superior performance compared to traditional methods that rely on manually crafted features (Yamashita *et al.*, 2018). These networks have demonstrated remarkable effectiveness in var-

ious computer vision applications, that includes image recognition (Tajbakhsh *et al.*, 2016), identifying and recognizing objects (Ren *et al.* (2017); Hassanzadeh *et al.* (2020)).

To accurately analyze a single case, it's crucial to consider the valuable information contained in various MRI sequences such as DWI, ADC, T2-Flair, and SWI. However, the complex relationships between these sequences and the subtle changes they indicate may not be easily detectable when viewed in single sequence. To overcome this challenge, we have used multiple channels to combine the information from these sequences and capture the complex interplay between them, improving the models ability to detect and interpret fine changes. In this investigation, the stacked multi-channel CNN framework is proposed for acute infarct prediction leading to more accurate and reliable detection of ischemic brain lesions from MRI sequences like DWI, ADC, T2-flair and SWI.

### 8.1.1 Problem Statement

Acute brain infarct is a severe global health issue that can cause significant harm, so diagnosing it quickly and accurately is crucial to prevent further damage to brain tissue. MRI is a highly accurate non-invasive method for detecting acute brain infarcts, and specific MRI sequences, such as DWI, T2-Flair, ADC and SWI, can help identify both ischemic and hemorrhagic infarcts. However, manually interpreting these MRI sequences can be time-consuming and error-prone. Therefore, using AI techniques, particularly CNNs, can provide an automated and more reliable approach to predicting acute infarcts from MRI images. Using AI-based strategies, radiologists can provide swift and precise diagnoses, improving treatment decisions and reducing morbidity and mortality rates. This study aims to develop an automated CNN-based approach to predict acute infarcts from MRI images, thus improving patient outcomes.

*“Considering the multimodal medical Image cohort with multiple MRI sequences including DWI, T2-Flair, ADC and SWI, design and develop an effective deep learning strategy to predict acute infarct accurately to facilitate an intelligent clinical recommendation system”*

The following are the key contributions of the proposed research addressing the above challenges:

- We present a novel framework for acute infarct prediction from MRI sequences using deep learning techniques. Our framework consists of a contour-

based brain segmentation technique to isolate the brain contours from the MRI data. We then propose stacked multi-channel convolutional neural networks (SMC-CNN-M and SMC-CNN-I) to predict the disease from multiple MRI sequences and individual MRI sequences. To visualize the disease in the MRI data and assess the model’s potential to predict acute infarct, we incorporate Gradient-weighted Class Activation Mapping (Grad-CAM).

- The proposed framework was evaluated on a medical cohort collected from a private medical hospital, and we benchmarked their classification performance against the baseline deep learning network. We performed an ablation study on different MRI sequences to assess the efficacy of each sequence. Furthermore, we generated synthetic data using DCGAN and compared the performance of our proposed models on the synthetic data.

## 8.2 Materials

In this section, we will delve into the intricacies of our data collection and cohort selection processes and then elaborate on the methods used for data augmentation and synthesis. To begin with, we will explain how we carefully chose the cohort for our study and the steps taken to collect the necessary data. Our focus was on ensuring that our data was representative of the population being studied, and that it was collected in a consistent and systematic manner. We will provide a detailed account of our cohort selection criteria and the measures we implemented to ensure data quality. After describing the data collection process, we will move on to the topic of data augmentation and synthesis. Here, we will explain how we used various techniques to enhance the amount and diversity of our data. We will discuss the various methods used to create synthetic data, and how these were combined with the real data to create a more comprehensive dataset.

### 8.2.1 Data Collection

In our experiment, we gathered MRI sequences (DWI, T2-flair, ADC, and SWI) from KMC Hospital (Mangalore, India), which were then de-identified or de-linked to protect patient privacy. We received approval from the IEC to use this cohort for research in the area of health informatics. The detailed cohort statistics providing information about the MRI sequences gathered from KMC Hospital are presented in Table 8.1. The MRI sequences were captured using several machines, including 1.5T Siemens Magnetom Avanto, 1.5T Signa Exite, and 1.5T Siemens Symphony,



Table 8.1: Cohort Statistics: Detailed description of the MRI sequences collected from KMC private hospital.

<b>Dataset Description</b>	<b>KMC Hospital Cohort</b>
Total # of cases included with four MRI sequences	1267
Total # of cases after pre-processing	991
Total # of cases with Acute Infarct	494
Total # of cases with no disease	497
Total # of training set	793
Total # of validation set	99
Total # of test set	99
Total # of training set (after standard augmentation)	3169
Total # of validation set (after standard augmentation)	397
Total # of test set (after standard augmentation)	396
Total # of training set (after synthetic data generation)	1532
Total # of validation set (after synthetic data generation)	192
Total # of test set (after synthetic data generation)	192

and a total of 1267 cases with all four MRI sequences were collected. These MRI sequences were passed through the basic preprocessing stage to exclude irrelevant cases. The following cases are excluded: (a) cases with Hemorrhagic stroke (i.e., 68 cases), (b) cases with metallic implants, pacemakers and Motion artefacts (i.e., 208 cases). The cohort was then categorized into two groups: “abnormal”, which consisted of cases with acute infarct, and “normal”, which consisted of cases with no diseases. The cohort of acute brain infarcts was divided into three parts, which were a train set, a validation set, and a test set, with proportions of 80%, 10%, and 10%, correspondingly. As the medical dataset collected is limited in number and considering the overfitting issue when ingested into the proposed SMC-CNN, we have applied a data augmentation and synthesis strategy to overcome the problem of inaccurate model prediction. Henceforth, we have increased the cohort size by using geometrical translation and artificial image generation from the selected MRI sequences.

## 8.2.2 Standard Augmentation Techniques

To expand the size of the medical cohort obtained, the MRI sequences are passed through a series of geometric translations (refer Figure 5.9 in Chapter 5). This transformation will aid ML and DL models to enhance performance by minimizing overfitting problems. The DL algorithms need to be trained with a massive sample



of data to provide a reliable diagnostic outcome. In the medical domain, obtaining a vast amount of clinician-annotated data is tedious and time-intensive. The medical datasets are usually small in size or restricted to private medical institutes. Henceforth, we apply data augmentation strategies to randomly transform the original MRI sequences to increase the data cohort size. Section 5.6 in Chapter 5, provides a comprehensive explanation of the process for creating augmented MRI scans using standard augmentation techniques.

### 8.2.3 Synthetic Data Generated using DCGAN

Synthetic data is artificially annotated information that is simulated by the algorithms to obtain the alternate mimicked data from the original data cohort. Synthetic data in health informatics assists in the data scarcity problem and allows us to expand the cohort size with less time than the manual collection of expert-annotated data. For our study, we have incorporated the Deep Convolutional Generative Adversarial Network (DCGAN) to produce synthetic data from the MRI cohort collected from KMC Hospital. The DCGAN module comprises two major modules, namely the Generator and the Discriminator modules. The generator module creates the synthetic image by ingesting additional noise from the original image, and the discriminator module then classifies it as an actual or artificial image. Section 5.6 in Chapter 5, offers a thorough explanation of the procedure for generating synthetic MRI scans using DCGAN.

## 8.3 Methodology

In this section, we give an in-depth overview of the proposed methodology for predicting acute brain infarct from MRI sequences, including DWI, T2-Flair, ADC, and SWI. To begin with, we present a contour-based brain segmentation technique to segment the brain contours from the MRI sequences. Further, we propose two stacked multi-channel convolutional neural networks (i.e., SMC-CNN-M and SMC-CNN-I) for predicting disease from multiple MRI sequences and individual MRI sequences. Lastly, we integrate Grad-CAM to facilitate the visualization of diseases in the MRI sequences and to evaluate the model's potential to predict acute infarct.

### 8.3.1 Contour-based Brain Segmentation for MRI Sequences

The high-resolution MRI sequences like DWI, T2-flair, ADC, and SWI comprise non-brain tissue, which needs to be discarded before ingesting the MRI sequence into the SMC-CNN as the principal region of interest is the brain. So, we propose contour-based brain segmentation to extract the brain contours from the MRI sequences. The Algorithm 4 provides the step-by-step procedure followed for contour-based brain segmentation from MRI sequences.

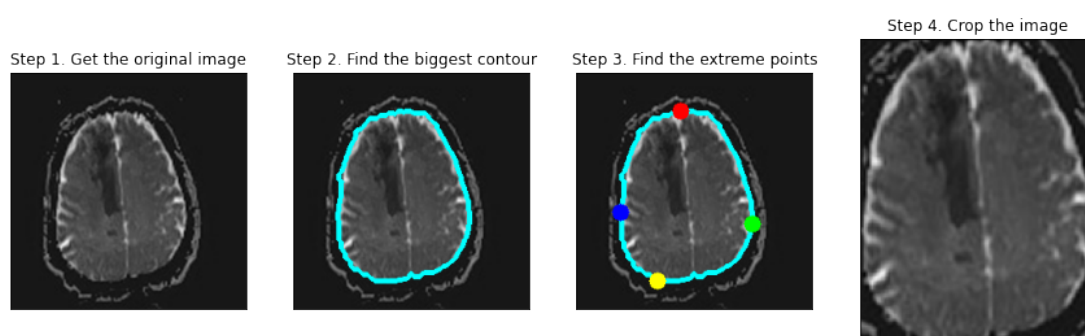


Figure 8.1: Contour-based brain segmentation of MRI sequence. From left to right: Step 1 indicates the original MRI Image, step 2 represents the biggest contour obtained, step 3 indicates the extreme points on the left, right, top and bottom of the contour and step 4 represents the cropped image.

To begin with, we convert the RGB MR Image sequences into grayscale and apply Gaussian blur (Young and van Vliet, 1995) to mute any noise in the image by blurring it slightly. Further, we use the same threshold value for every pixel in the grayscale image to obtain a clear partition between the foreground and background in an image. The pixel value is set to zero if it is less than the given threshold; otherwise, it will be assigned the maximum value (i.e., Max-value = 255). Morphological erosion and dilation are applied to remove the insignificant pixels on the contour boundaries and enhance the white pixels in the contour area. We capture the most prominent contour out of all the contours obtained in the MRI sequence. To segment and crop the region of interest from the MRI sequences, we locate the extreme points of the largest contour: the left, right, top, and bottom. These points serve as a reference for the segmentation and cropping processes. Figure 8.1 shows the contour-based brain segmentation of the MRI sequences.

**Algorithm 4:** Contour-based brain segmentation for MRI Sequences**Input:** MRI Imaging Sequences like DWI, T2-Flair, ADC and SWI**Output:** Segmented brain contour from the MRI

```

1 initialization
2 Function Crop-Brain-Contour(I):
   /* Convert the image I to grayscale and blur it slightly using
   Gaussian blur to mute any noise. */
3 Convert RGB → Grayscale
4 G ← GaussianBlur (Grayscale)
   /* For every pixel in grayscale image, same threshold value is applied.
   The pixel value is set to zero if it is less than the given threshold;
   otherwise, it will be assigned with the maximum value. */
5 T ← Threshold (G, Min-value=45,
6 Max-value=255)
   /* Remove the pixels on the contour boundaries using morphological
   erosion and increase the size of the white pixel to enhance the
   contour area by dilation technique. */
7 T ← Erode (T, kernel, iteration=2)
8 T ← Dilate (T, kernel, iteration=2)
   /* Find the contours in the MRI sequences and capture the largest
   one. */
9 C ← Find-Contours (T)
10 C ← Grab-Contours (C)
11 C ← Max (C)
12 Find the extreme point (Left, right, top and bottom) of the largest
   contour C.
13 Draw Contours on the MRI Image.
14 Crop the brain contour from the MRI Image.
15 return Cropped Image;
16 End Function

```

### 8.3.2 Stacked Multi-Channel Convolution Neural Network (SMC-CNN)

We propose two stacked multi-channel convolutional Neural Networks (i.e., SMC-CNN-M and SMC-CNN-I) to predict the acute infarct from multiple and individual MRI sequences. The segmented MRI sequences, including DWI, T2-Flair,

ADC, and SWI, are passed through the four identical convolutional Neural Network channels to extract features from each MRI sequence. The multi-channel imaging features are fused using the concatenation layer. The combined features are subsequently fed into a DNN that is fully connected to facilitate the prediction of acute infarct. The overall architecture of the SMC-CNN-M: Stacked Multi-Channel Convolution Neural Network for Predicting Acute Infarct from Multiple MRI Sequences is shown in Figure 8.2. The SMC-CNN-M contains four parallel channels for DWI, T2-flair, ADC, and SWI MRI sequences to retrieve multi-channel features. The individual MRI sequences with a size of  $240 \times 240 \times 3$  are passed through each of the four channels to obtain the imaging features of four different MRI sequences. We use five layers of convolution and five layers of pooling in every channel to extract imaging features from the MRI sequences. The feature size of 12800 is obtained from each channel and is concatenated to form a feature size of 51200. The fully connected DNN receives these multi-channel visual features as input and produces predicted diagnostic outcomes.

We also propose the SMC-CNN-I: Stacked Multi-Channel Convolution Neural Network for predicting acute infarct from individual MRI sequences. The individual segmented MRI sequence is ingested through four parallel multi-channel and multi-scale CNNs to extract multi-level features with varied receptive fields. The features retrieved from the four levels are fused using the concatenation technique, and the resulting feature set is then fed into a fully connected DNN for prediction of infarcts in the brain MRI. The overall architecture of SMC-CNN-I is presented in Figure 8.3. The SMC-CNN-I contains four parallel layers with varied filter sizes (i.e., 32, 64, 32, and 64) to retrieve multi-scale features from the MRI sequences. The individual MRI sequences with a size of  $240 \times 240 \times 3$  are passed through each of the four layers to obtain the imaging features. Every channel contains five convolution layers and five pooling layers with alternate filters of 32 and 64. Likewise, we have employed four convolution neural networks in parallel to extract features from the MRI sequences. The four stacked CNNs produce the output imaging features of sizes 800, 1600, 800, and 1600. We concatenate the flattened features from all four stacked CNNs to obtain fused multi-level features of size 4800. The fused feature is then fed into a DNN that consists of two hidden layers and one final layer, which predicts acute infarct from the MRI sequences.

The convolution and pooling layers are the main modules used in our proposed models. The MRI image sequence can be represented as a high-dimensional matrix comprising feature vectors. The imaging input undergoes a linear operation known as “convolution”, whereby the input MRI imaging data is multiplied with an array

of weights called a filter or kernel. The element-wise multiplication is employed to the kernel and MR image to obtain the activation map. During the convolution operation, the filter is moved across the input MRI sequence from left to right and top to bottom. This movement is referred to as strides, and in our study, we have used the strides of 1. The convolution operation is defined as follows:

$$Z_{l,m,n}^i = (W_n^i)^T \cdot Y_{l,m}^i + b_n^i \quad (8.1)$$

Where  $Z_{l,m,n}^i$  represents the  $i^{th}$  layer feature value at the location  $(l, m)$  of the  $n^{th}$  feature map. Here,  $W_n^i$  indicates the kernel weights of the  $i^{th}$  layer,  $b_n^i$  depicts the bias terms of the  $i^{th}$  layer and  $Y_{l,m}^i$  is the imaging input at the position  $(l, m)$  of the  $i^{th}$  layer. In the SMC-CNN-M model, we have used a filter size of 32 in every channel. In SMC-CNN-I, we have varied the filter size between 32 and 64 between the channels to extract the multi-scale features from the individual MRI sequences. We have selected the ReLU activation function in the convolution layer operation to prevent model overfitting (Hara *et al.*, 2015). The ReLU function improves the sparsity of the network by setting some of the neuron output to zero. Additionally, it decreases the correlation between parameters and reduces overfitting. In the proposed frameworks, ReLU allows each neuron to have a stronger filtering effect, resulting in more efficient computation. The ReLU operation can be defined as follows:

$$f(y) = \begin{cases} y & \text{if } y > 0 \\ 0 & \text{if } y < 0 \end{cases} \quad (8.2)$$

Here,  $f(y)$  is the output of the ReLU activation function, and  $y$  is the input to the activation function. The maximum pooling method is utilized to extract the highest value from the feature tiles generated by the convolutional layer and use it as the output in the pooling layer. This approach helps reduce the feature map's spatial dimensions and retain the essential information by eliminating redundant features. In this study, a pool size of  $2 \times 2$  and stride of 1 are used for the downsampling operation. The mapping from input  $Y_i^k$  in the  $k$ th layer to output  $Z_i^k$  is performed through a neuron, defined as:

$$Z_i^k = \max\{Y_i^k\} \quad (8.3)$$

The final features from each of the parallel convolution and pooling layers are combined through a concatenation operation, which combines the feature weights.

The concatenation operation combines the output of multiple pooling layers into a single tensor, where each pooling layer's output is represented as a slice along a new axis. This allows the network to preserve the spatial information from different pooling layers and combine it for use in the next layer in the network. The features obtained from the concatenation operation are then processed by a DNN for acute infarct prediction. The flattened clinical features represented by  $C = c_1, c_2, c_3, \dots, c_n$  are fed into DNN, and  $Y_i$  indicates  $i^{th}$  output produced from every layer, defined as,

$$Y_i = \phi(W_1 \cdot c_1 + W_2 \cdot c_2 + \dots + W_n \cdot c_n) \quad (8.4)$$

Here, the weight parameters are denoted by  $W_1, W_2, \dots, W_n$  and they are used along with the non-linear activation function  $\phi$ . The Rectified Linear Unit (ReLU) activation function is applied on two hidden layers (i.e., 256 and 128), and the sigmoid activation function is utilized for acute infarct prediction.

The following is the rationale behind selecting different layers, channels, and hyperparameters: The SMC-CNN-M and SMC-CNN-I architectures utilize five layers of convolution and pooling in each of the four parallel channels to extract features from the input MRI sequences. This design choice was made to enable multi-level feature extraction with increasing receptive field sizes. By employing multiple levels of convolution and pooling layers with varying filter sizes and strides, the models are able to capture both low-level and high-level features from the input MRI sequences, leading to improved accuracy in predicting acute infarct. To determine the optimal number of layers and parameters, we employed grid search approaches (Bergstra and Bengio, 2012). Through empirical evaluation, we found that the current proposed model outperformed other configurations. Overall, the use of multi-level convolution and pooling layers with varying filter sizes and strides enables our SMC-CNN-M and SMC-CNN-I architectures to effectively capture features from multiple MRI sequences and achieve improved predictive performance for acute infarct.

### 8.3.3 Acute Brain Infarct Visualization using Grad-CAM

The SMC-CNN will extract multi-channel features from the MRI sequences. The features that were obtained are fed into a DNN-based fully connected network for obtaining probability scores for two classes (i.e., no disease and acute infarct) at the softmax layer. The class that has the maximum probability score is utilized to determine the final outcome of the disease. We incorporate Grad-CAM (Selvaraju



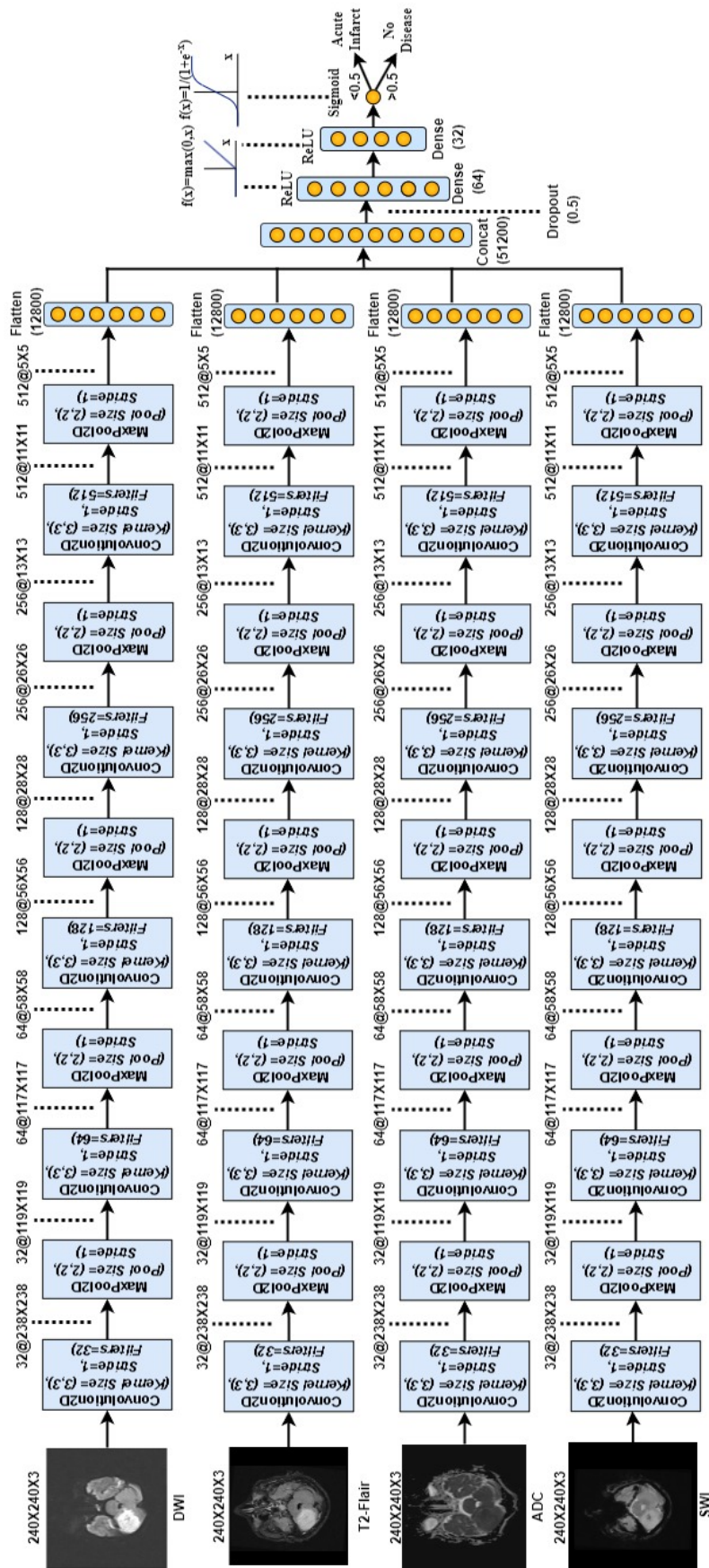


Figure 8.2: SMC-CNN-M: Stacked Multi-Channel Convolution Neural Network to predict acute infarct from the four MRI Sequences, including DWI, T2-Flair, ADC, and SWI





*et al.*, 2016) for visualizing the acute infarct from the MRI sequences (i.e., DWI, ADC, T2-flair and ADC). The grad-CAM generates a heatmap and localizes the disease region in particular MRI sequences, and allows us to achieve an explainable model by providing information about the model's capability to achieve the desired diagnostic outcome. The gradients from the final convolution layer of the SMC-CNN (i.e., either SMC-CNN-M or SMC-CNN-I) are extracted and used to create heatmaps at the regions in an MR image with acute infarcts. The regions with maximum weights (or gradients) significantly impact the prediction outcome. The backpropagation operation is performed with acute infarct =1 and no disease =0. The gradient weights are updated by measuring the Global Average Pooling (GAP) of the gradient across all the features in each channel, as shown below:

$$Z_k = \frac{1}{f_{Height} \times f_{Width}} \sum_{m=1}^{f_{Height}} \sum_{n=1}^{f_{Width}} w_{j(m,n)} \quad (8.5)$$

In this context,  $Z_k$  stands for the one-dimensional feature corresponding to the  $k^{th}$  dimension extracted from the GAP method. Additionally,  $f_{Height}$  and  $f_{Width}$  represent the height and width of the two-dimensional activation map, while  $w_j$  denotes the  $j^{th}$  activation map produced by the SMC-CNN at a specific location  $(m, n)$ . These updated gradient weights are then added to the feature map through multiplication. Finally, the output score is calculated for both the acute infarct and healthy control variables, as depicted below:

$$S_T = \frac{1}{f_{Height} \times f_{Width}} \sum_k W_k^T F_k \quad (8.6)$$

In equation 8.6, the  $S_T$  represent the score of the proposed SMC-CNN network in target variable  $T$ ;  $W_k^T$  indicates the gradient weight  $k^{th}$  feature map in target variable  $T$  and  $F_k$  is the  $k^{th}$  feature map. To create the class discrimination positioning map, we calculate the partial derivative of the target class score  $S_T$  with respect to the feature map of a specific layer in the network  $F_k$  using the following formula:

$$\nabla_k^T = \frac{\partial S_T}{\partial F_k} \quad (8.7)$$

The map produced emphasizes the areas of the MR image input that hold greater significance for the network's classification prediction. In equation 8.7,  $\nabla_k^T$  depicts the gradient of the  $k^{th}$  feature map. Further, the total sum obtained is further ingested into the ReLU activation function to produce the MR image with a

heatmap showcasing the disease outcome.

$$L^T = \text{ReLU}\left(\sum_j \nabla_k^T F_k\right) \quad (8.8)$$

Here,  $L^T$  indicates the heatmap for the target class  $T$ . The visual depiction of the disease visualization from the MRI sequence is shown in the Figure. 8.4.

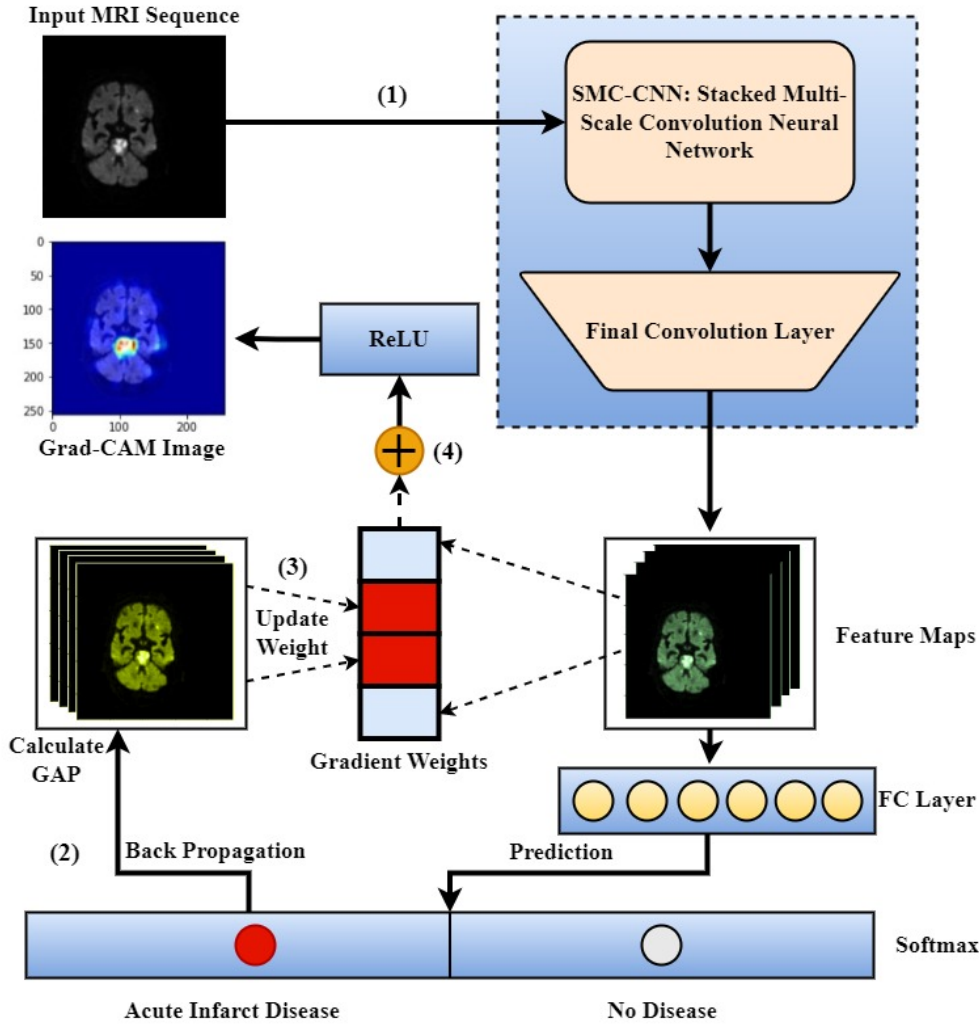


Figure 8.4: Disease Visualization using Gradient-weighted Class Activation Mapping (Grad-CAM): (1) MRI Sequences (like DWI, T2-Flair, ADC and SWI) are given as input to the proposed SMC-CNN (i.e., either SMC-CNN-M or SMC-CNN-I) framework to obtain the acute infarct prediction, (2) The backpropagation operation is performed with acute infarct = 1, and no disease = 0, (3) The global Average Pooling (GAP) of the gradient is calculated for each channel, and the weights of the SMC-CNN are updated, (4) The gradient weights and the feature maps are added and multiplied and given as input to the rectified linear activation function.

## 8.4 Experimental Setup

This section offers an overview of the parameter configuration for the various models used and outlines the assessment measures used to evaluate the proposed model. Radiologists typically rely on all four MRI sequences (DWI, SWI, ADC, and T2-flair) to determine the presence of acute infarcts in an image. However, we have found no previous research that has utilized all four sequences for predicting acute infarcts. Therefore, our study is the first of its kind, and we have conducted a benchmark comparison of our proposed model with state-of-the-art deep learning models.

### 8.4.1 Parameter Configuration of Proposed SMC-CNN and State-of-the-art Deep Learning Models

The experimental analysis for the research study was conducted using a server with the technical configuration of an NVIDIA Tesla M40 accelerator, 128GB RAM, a 24GB graphics processing unit (GPU), a 3TB hard drive, and a Linux operating system. We employed Python 3.6 programming language and leveraged the open-source Keras software package, and the Tensorflow library (Abadi and *et al.*, 2015). In the field of medical image analysis, there are many ML and DL models available that have been widely used for various medical imaging tasks. However, we did not find any specific specialized algorithm or model that has been exclusively developed for acute brain infarct prediction. Most of the reviewed methods implemented existing ML and DL techniques on acute Infarct or stroke on a private dataset. Therefore, to assess the effectiveness of the SMC-CNN models in predicting acute Infarct from MRI sequences, we have employed eight pre-trained deep learning models like VGG19 (Simonyan and Zisserman, 2015b), VGG16 (Simonyan and Zisserman, 2015b), ResNet50 (He *et al.*, 2016b), MobileNet (Howard *et al.*, 2017b), InceptionV3 (Szegedy *et al.*, 2016b), EfficientNetB2 (Tan and Le, 2021), DenseNet121 (Huang *et al.*, 2017b), and Xception (Chollet, 2017b) as a reference point for comparison. We have modified the hyperparameters of the pre-trained models to make them suitable for predicting acute Infarct from MRI sequences. These pre-trained models are considered state-of-the-art in the domain of DL and were initialized with weights that were previously trained on the ImageNet dataset (Deng *et al.*, 2009).

The use of ImageNet pre-trained weights helps to address the issue of the large cohort required for training in deep learning. To fine-tune the models for

the acute infarct prediction task, we froze the later (or top) layers and retrained the earlier (or bottom) layers, including the input layer, using MRI sequences collected from KMC hospital. Fine-tuning a pre-trained model for a specific task involves modifying the weights of the model's later layers. This is done while keeping the initial layers fixed, containing general features useful for a wide range of tasks. By using this process, the amount of training data necessary to attain high performance on a particular task is minimized, as the initial layers have already acquired significant features from the ImageNet dataset. We have optimized the hyperparameters of all eight standard DL algorithms to obtain the optimal diagnostic outcome for predicting acute brain infarct. By comparing the performance of our proposed SMC-CNN model with these baseline models, we can evaluate the effectiveness of our approach for acute infarct prediction from MRI sequences. To find the best possible combination of hyperparameters for our proposed model as well as for the standard DL algorithms, we have used a technique called grid search (Bergstra and Bengio, 2012). This technique involves systematically testing different values of each hyperparameter within a defined range and evaluating the model's performance for each variety of hyperparameters. The goal is to identify the optimal combination of hyperparameters that leads to the best performance for a given disease prediction task. The fine-tuned hyperparameter configuration of the proposed SMC-CNN and state-of-the-art Deep learning models is presented in the Table 8.2. In both our proposed models and the baseline DL techniques, we utilized early stopping (Bai *et al.*, 2021), and trained them for 50 epochs. To predict the binary outcome (i.e., Acute Infarct and no disease), we utilized the binary cross-entropy loss function for all the DL algorithms.

### 8.4.2 Evaluation Metrics

We have employed six performance metrics to measure its effectiveness in our investigation of how well the SMC-CNN performs on MRI sequences obtained from a private hospital. These assessment measures include Accuracy (Acc), Precision (Pre), Recall (Rec), F1-Score (F1), Cohen's kappa ( $\kappa$ ), and AUROC. We will establish these metrics by utilizing fundamental terminologies such as True Positive, True Negative, False Positive, and False Negative. Our research is specifically focused on binary classification, which involves categorizing MRI sequences into two classes: "No disease" or "healthy" and "Acute Infarct". We have provided clear definitions for the terms mentioned above to understand them better.

- **True Positive** ( $True_{+ve}$ ) refers to an MRI sample that is accurately clas-

Table 8.2: The detailed parameter configuration of proposed SMC-CNN and state-of-the-art Deep learning models

Models	Input Size	Number of Layers	Activation Function	Stride	Dropout rate	Optimizer	Learning rate	Batch Size	Epochs
VGG19	$240 \times 240 \times 3$	19	ReLU	1	0.5	RMSProp	0.0001	32	50
VGG16	$240 \times 240 \times 3$	16	ReLU	1	0.5	RMSProp	0.0001	32	50
ResNet50	$240 \times 240 \times 3$	50	ReLU	2	0.5	Adam	0.0001	32	50
MobileNet	$240 \times 240 \times 3$	28	ReLU6	1	0.5	Adam	0.0001	32	50
InceptionV3	$240 \times 240 \times 3$	159	ReLU	2	0.5	Adam	0.0001	32	50
EfficientNetB2	$240 \times 240 \times 3$	23	Swish	1	0.5	RMSProp	0.0001	32	50
DenseNet121	$240 \times 240 \times 3$	121	ReLU	1	0.5	RMSprop	0.0001	32	50
Xception	$240 \times 240 \times 3$	126	ReLU	1	0.5	RMSProp	0.0001	32	50
SMC-CNN-I	$240 \times 240 \times 3$	10	ReLU	1	0.5	Adam	0.001	32	50
SMC-CNN-M	$240 \times 240 \times 3$	10	ReLU	1	0.5	Adam	0.001	32	50

sified as “Acute Infarct”, indicating that it belongs to the correct category.

- **True Negative** ( $True_{-ve}$ ), on the other hand, pertains to an MRI sample that is accurately classified as “Normal” or “Healthy”, indicating that it also belongs to the correct category.
- **False Positive** ( $False_{+ve}$ ) denotes an MRI sample belonging to the “Normal” or “Healthy” class but is wrongly categorized as “Acute Infarct”.
- **False Negative** ( $False_{-ve}$ ), meanwhile, depicts an MRI case that is incorrectly classified as “Acute Infarct” when it actually belongs to the “Normal” or “Healthy” class.

To summarize, True Positive and True Negative refer to accurate classifications, whereas False Positive and False Negative indicate inaccurate categorizations. True Positive and True Negative are crucial in assessing the classification model’s accuracy, while False Positive and False Negative play a significant role in revealing the model’s shortcomings.

$$Acc = \frac{True_{+ve} + True_{-ve}}{True_{+ve} + True_{-ve} + False_{+ve} + False_{-ve}} \quad (8.9)$$

$$Pre = \frac{True_{+ve}}{True_{+ve} + False_{+ve}} \quad (8.10)$$

$$Rec = \frac{True_{+ve}}{True_{+ve} + False_{-ve}} \quad (8.11)$$

$$F1 = 2 \times \frac{Pre \times Rec}{Pre + Rec} \quad (8.12)$$

$$X = (True_{+ve} \times True_{-ve} - False_{-ve} \times False_{+ve}) \quad (8.13)$$

$$Y = (True_{+ve} + False_{+ve}) \times (False_{+ve} + True_{-ve}) \quad (8.14)$$

$$Z = (True_{+ve} + False_{-ve}) \times (False_{-ve} + True_{-ve}) \quad (8.15)$$

$$\kappa = \frac{2 \times X}{Y + Z} \quad (8.16)$$

. Eq. 8.9 pertains to the model’s accuracy, which quantifies the overall number of accurate forecasts produced by the proposed SMC-CNN model. It is worth emphasizing that a model’s high accuracy rate does not always ensure that it can accurately classify labels in cases where there is an uneven distribution or class

imbalance among the cohort. Therefore, when evaluating the performance of a model, it is essential to consider other metrics that take into account class imbalance. These metrics can provide a more comprehensive evaluation of a model's ability to classify instances accurately, particularly in datasets with class imbalance. The assessment of a model's performance heavily relies on precision and recall, which provide crucial information. Precision evaluates the model's ability to predict the abnormal class accurately, and can be represented by equation 8.10. This equation shows that precision is determined by the proportion of accurate predictions for acute infarct samples, relative to the total number of predictions generated by the proposed model. On the other hand, as demonstrated in equation 8.11, recall represents the ratio of correctly predicted acute infarct cases to the total number of acute infarct cases. Both precision and recall evaluate the performance of the proposed SMC-CNN models in reducing  $False_{+ve}$  and  $False_{-ve}$  predictions. To assess a model's performance in scenarios with class imbalance, the F1-score considers  $False_{+ve}$  and  $False_{-ve}$ , and achieves a balance between precision and recall by using equation 8.12 to calculate the harmonic mean. This metric is valuable in evaluating the model's effectiveness. Cohen's kappa is a commonly used metric for measuring agreement between two raters, and it can also be employed to assess the effectiveness of a classification model. The kappa coefficient can be calculated using the confusion matrix, which includes data on the number of  $True_{+ve}$ ,  $False_{+ve}$ ,  $True_{-ve}$ , and  $False_{-ve}$ , as demonstrated in equation 8.16. Cohen's kappa then measures the effectiveness of a model by comparing the level of agreement between its predictions and the true values to what would be expected by chance. A higher kappa coefficient signifies a more substantial agreement between the model's predictions and the actual values, while a lower coefficient indicates weaker agreement. The AUROC metric evaluates the ability of binary classification to distinguish between Acute brain infarct sample and healthy sample by plotting True Positive Rate (TPR) against False Positive Rate (FPR) across multiple thresholds. A higher AUROC value, nearing 1, indicates precise classification of MRI sequences as either healthy or having an acute brain infarct, while a lower value, nearing 0, suggests inadequate distinction between the two classes.

## 8.5 Results and Discussion

This part presents the findings of our experimental analysis of the proposed SMC-CNN models (i.e., SMC-CNN-I and SMC-CNN-M) on both standard augmented

and synthetic MRI sequences. We conducted an ablation study to assess the efficacy of the SMC-CNN-M model by varying the input MRI sequences. Our study also includes a qualitative analysis of the SMC-CNN, whereby we visualize and localize acute infarcts in the brain regions. This helps us gain a comprehensive perception of the model's strengths and weaknesses and its potential for future developments in the medical field.

### 8.5.1 Quantitative Analysis

The proposed model was subjected to a detailed benchmarking experiment by applying it to the collected MRI sequences. The evaluation of the SMC-CNN-I model's effectiveness with the standard deep learning model for DWI, T2-Flair, ADC and SWI MRI sequences is shown in Table. 8.3, Table. 8.4, Table. 8.5 and Table. 8.6, respectively. The results of our proposed SMC-CNN-I model for acute infarct prediction from MRI sequences show promising outcomes. Based on the empirical analysis, it can be seen that the SMC-CNN-I model that was proposed performs better than the current leading baseline models such as VGG-16, VGG-19, ResNet-50, MobileNet, Inception V3, EfficientNetB2, DenseNet121, and Xception. This result signifies the potential of our proposed model in accurately predicting acute infarcts from various MRI sequences, including DWI, T2-Flair, ADC, and SWI. The accuracy scores achieved for DWI, T2-Flair, ADC, and SWI were 0.9824, 0.9427, 0.9583, and 0.94229, respectively. These scores serve as evidence that the proposed model is effective when used with standard augmented data. The accuracy score depicts the proportion of cases predicted correctly in relation to the overall cases evaluated. Hence, a higher accuracy score suggests that our proposed model can accurately identify acute infarcts and distinguish them from no disease cases when MRI sequences are given as input. We obtained an accuracy of 0.9830, 0.9834, 0.9883, 0.9710 for DWI, T2-Flair, ADC, and SWI, respectively, showcasing that the proposed SMC-CNN-I model can generalize on a broader range of inputs. It also suggests that the proposed SMC-CNN-I model has learned to identify important patterns or relationships from the MRI sequences that are significant to acute infarct disease prediction. Improved synthetic data accuracy proves that the model can be applied to real-world data.

The precision values for DWI, T2-Flair, ADC and SWI are reported as 0.9826, 0.9452, 0.9604, and 0.94312, respectively. The precision metric measures the proportion of accurately categorized acute infarcts samples to the overall instances categorized as acute infarct. Precision with higher values indicate that the pre-



Table 8.3: Benchmarked performance evaluation of proposed SMC-CNN-I model and the baseline DL techniques for predicting acute brain infarct from DWI MRI sequences obtained from the KMC hospital.

Augmentation Techniques	Models	Accuracy	Precision	Recall	F1-Score	Cohen's Kappa	AUROC
Standard Augmentation	VGG19	0.9684	0.9686	0.9684	0.9684	0.9367	0.9932
	VGG16	0.9631	0.9632	0.9631	0.9631	0.9261	0.9960
	ResNet50	0.9315	0.9355	0.9315	0.9314	0.8632	0.9687
	MobileNet	0.9578	0.9578	0.9578	0.9578	0.9156	0.9965
	InceptionV3	0.9368	0.9368	0.9368	0.9368	0.8731	0.9807
	EfficientNetB2	0.5736	0.3291	0.5736	0.4182	0.0000	0.8109
	DenseNet121	0.9578	0.9597	0.9578	0.9578	0.9157	0.9972
	Xception	0.9685	0.9695	0.9685	0.9686	0.9369	0.9982
	<b>Proposed SMC-CNN-I</b>	<b>0.9824</b>	<b>0.9826</b>	<b>0.9824</b>	<b>0.9824</b>	<b>0.9642</b>	<b>0.9996</b>
	Synthetic Data Generated using DCGAN	VGG19	0.9738	0.9741	0.9738	0.9737	0.9463
VGG16		0.9742	0.9743	0.9742	0.9843	0.9683	0.9965
ResNet50		0.8219	0.8505	0.8219	0.8167	0.6381	0.9281
MobileNet		0.9738	0.9742	0.9738	0.9737	0.9470	0.9980
InceptionV3		0.9738	0.9743	0.9738	0.9738	0.9475	0.9990
EfficientNetB2		0.4659	0.2171	0.4659	0.2962	0.0000	0.7086
DenseNet121		0.9581	0.9583	0.9581	0.9581	0.9159	0.9976
Xception		0.9685	0.9695	0.9685	0.9686	0.9369	0.9982
<b>Proposed SMC-CNN-I</b>		<b>0.9830</b>	<b>0.9831</b>	<b>0.9830</b>	<b>0.9830</b>	<b>0.9661</b>	<b>0.9990</b>

Table 8.4: Benchmarked performance evaluation of proposed SMC-CNN-I model and the baseline DL techniques for predicting acute brain infarct from T2-Flair MRI sequences obtained from the KMC hospital.

Augmentation Techniques	Models	Accuracy	Precision	Recall	F1-Score	Cohen's Kappa	AUROC
Standard Augmentation	VGG19	0.9242	0.9242	0.9242	0.9242	0.8478	0.9606
	VGG16	0.8888	0.8888	0.8888	0.8887	0.7750	0.9510
	ResNet50	0.6767	0.7122	0.6767	0.6496	0.3191	0.8654
	MobileNet	0.9344	0.9344	0.9344	0.9344	0.8776	0.9700
	InceptionV3	0.9141	0.9141	0.9141	0.9141	0.8282	0.9726
	EfficientNetB2	0.4797	0.2302	0.4797	0.3111	0.0000	0.4241
	DenseNet121	0.9000	0.9104	0.9090	0.9089	0.8176	0.9780
	Xception	0.9343	0.9345	0.9343	0.9342	0.8665	0.9777
	<b>Proposed SMC-CNN-I</b>	<b>0.9427</b>	<b>0.9452</b>	<b>0.9427</b>	<b>0.9426</b>	<b>0.8855</b>	<b>0.9810</b>
	Synthetic Data Generated using DCGAN	VGG19	0.9627	0.9645	0.9627	0.9629	0.9242
VGG16		0.9767	0.9768	0.9767	0.9767	0.9518	0.9956
ResNet50		0.9162	0.9162	0.9162	0.9162	0.8285	0.9785
MobileNet		0.9720	0.9723	0.9720	0.9721	0.9436	0.9887
InceptionV3		0.9720	0.9729	0.9720	0.9721	0.9421	0.9827
EfficientNetB2		0.5488	0.3012	0.5488	0.3889	0.0000	0.2821
DenseNet121		0.9581	0.9583	0.9581	0.9581	0.9159	0.9976
Xception		0.9581	0.9597	0.9581	0.9578	0.9133	0.9843
<b>Proposed SMC-CNN-I</b>		<b>0.9834</b>	<b>0.9835</b>	<b>0.9834</b>	<b>0.9834</b>	<b>0.9663</b>	<b>0.9983</b>

Table 8.5: Benchmarked performance evaluation of proposed SMC-CNN-I model and the baseline DL techniques for predicting acute brain infarct from ADC MRI sequences obtained from the KMC hospital.

Augmentation Techniques	Models	Accuracy	Precision	Recall	F1-Score	Cohen's Kappa	AUROC
Standard Augmentation	VGG19	0.9191	0.9194	0.9191	0.9192	0.8381	0.9689
	VGG16	0.8737	0.8738	0.8737	0.8735	0.7460	0.9639
	ResNet50	0.7626	0.7770	0.7626	0.7540	0.5041	0.8500
	MobileNet	0.8787	0.8799	0.8787	0.8789	0.7560	0.9468
	InceptionV3	0.9292	0.9292	0.9292	0.9292	0.8582	0.9746
	EfficientNetB2	0.4545	0.2066	0.4545	0.2840	0.0000	0.4952
	DenseNet121	0.9343	0.9343	0.9343	0.9341	0.8685	0.9718
	Xception	0.9191	0.9208	0.9191	0.9191	0.8385	0.9736
	<b>Proposed SMC-CNN-I</b>	<b>0.9583</b>	<b>0.9604</b>	<b>0.9583</b>	<b>0.9580</b>	<b>0.8750</b>	<b>0.9833</b>
	VGG19	0.9702	0.9713	0.9702	0.9704	0.9321	0.9970
Synthetic Data Generated using DCGAN	VGG16	0.9642	0.9664	0.9642	0.9647	0.9125	0.9928
	ResNet50	0.8333	0.8850	0.8333	0.8423	0.6334	0.9763
	MobileNet	0.9821	0.9821	0.9821	0.9821	0.9575	0.9920
	InceptionV3	0.9107	0.9205	0.9107	0.9128	0.7928	0.9876
	EfficientNetB2	0.5991	0.3589	0.5991	0.4489	0.0000	0.6071
	DenseNet121	0.9880	0.9882	0.9880	0.9880	0.9704	0.9935
	Xception	0.8869	0.9204	0.8869	0.8920	0.7457	0.9985
	<b>Proposed SMC-CNN-I</b>	<b>0.9883</b>	<b>0.9804</b>	<b>0.9883</b>	<b>0.9880</b>	<b>0.9875</b>	<b>0.9999</b>

Table 8.6: Benchmarked performance evaluation of proposed SMC-CNN-I model and the baseline DL techniques for predicting acute brain infarct from SWI MRI sequences obtained from the KMC hospital.

Augmentation Techniques	Models	Accuracy	Precision	Recall	F1-Score	Cohen's Kappa	AUROC
Standard Augmentation	VGG19	0.8838	0.8889	0.8838	0.8823	0.7603	0.9592
	VGG16	0.8989	0.9092	0.8989	0.8980	0.7967	0.9715
	ResNet50	0.5656	0.3199	0.5656	0.4087	0.0000	0.6422
	MobileNet	0.9444	0.9452	0.9444	0.9442	0.8865	0.9767
	InceptionV3	0.8838	0.8839	0.8838	0.8838	0.7676	0.9453
	EfficientNetB2	0.5454	0.2975	0.5454	0.3850	0.0000	0.4373
	DenseNet121	0.9191	0.9196	0.9191	0.9190	0.8371	0.9721
	Xception	0.9393	0.9394	0.9393	0.9393	0.8775	0.9798
	<b>Proposed SMC-CNN-I</b>	<b>0.9429</b>	<b>0.9431</b>	<b>0.9429</b>	<b>0.9428</b>	<b>0.8841</b>	<b>0.9833</b>
	Synthetic Data Generated using DCGAN	VGG19	0.9339	0.9350	0.9339	0.9340	0.8668
VGG16		0.9600	0.9647	0.9647	0.9647	0.9268	0.9937
ResNet50		0.6079	0.7399	0.6079	0.5790	0.2739	0.8799
MobileNet		0.9635	0.9635	0.9635	0.9635	0.9365	0.9875
InceptionV3		0.9647	0.9658	0.9647	0.9645	0.9269	0.9941
EfficientNetB2		0.7559	0.5714	0.7559	0.6508	0.0000	0.1778
DenseNet121		0.9383	0.9383	0.9383	0.9382	0.8740	0.9814
Xception		0.9559	0.9577	0.9559	0.9561	0.9084	0.9971
<b>Proposed SMC-CNN-I</b>		<b>0.9710</b>	<b>0.9714</b>	<b>0.9710</b>	<b>0.9710</b>	<b>0.9420</b>	<b>0.9966</b>

sented model accurately identifies acute infarct cases without generating many false positives. It is also observed that the precision of 0.9831, 0.9835, 0.9883 and 0.9714 for DWI, T2-Flair, ADC, and SWI sequences, respectively, on synthetic data, indicating the model's superiority in identifying the critical features that distinguish acute infarct lesions from normal tissue with a minimal number of false positives. This can potentially reduce the likelihood of false-positive diagnoses that could lead to unnecessary interventions and procedures. The recall values for DWI, T2-Flair, ADC and SWI are reported as 0.9824, 0.9427, 0.9583, and 0.9429, respectively. The recall metric measures the proportion of actual acute brain infarct instances that are correctly identified by a SMC-CNN-I model out of all the acute brain infarct instances present in the dataset. High recall values indicate that the model correctly identifies most acute infarct cases without missing too many positive cases. This is a critical factor in medical imaging applications where missing a positive case could severely affect patient outcomes. The higher recall of 0.983, 0.9834, 0.9883 and 0.9710 for DWI, T2-Flair, ADC, and SWI, respectively, is produced for synthetic data. This demonstrates its generalizability, allowing it to be applied to real-world MRI data with acute infarct. The F1-score values for DWI, T2-Flair, ADC and SWI are reported as 0.9824, 0.9426, 0.958, and 0.9428, respectively. The F1-score is a metric that balances precision and recall by calculating their harmonic mean. F1-score with higher values indicate that the SMC-CNN-I model is delivering superior outcome in both precision and recall without overemphasizing one metric at the expense of the other. It is also seen that increased F1-score of 0.983, 0.9834, 0.988 and 0.9710 is obtained for DWI, T2-Flair, ADC, and SWI, respectively on synthetic data. The F1-score is a metric that balances precision and recall by taking their harmonic mean. This allows the proposed model to better identify patterns and features in the data that are indicative of acute infarcts, resulting in improved F1-score. Additionally, the synthetic data can be produced with known ground truth, allowing for more accurate evaluation of the model's performance compared to real-world data where ground truth may be more difficult to determine.

The Cohen's Kappa values for DWI, T2-Flair, ADC, and SWI are reported as 0.9642, 0.8855, 0.875, and 0.8841, respectively. Cohen's Kappa metric quantifies the agreement between the model's predictions and the actual values while taking into account the agreement that would be expected by chance. High Cohen's kappa values indicate that the proposed classifier is performing significantly better than random guessing. The AUROC values for DWI, T2-Flair, ADC, and SWI are reported as 0.9996, 0.9810, 0.9833, and 0.9833, respectively. The AUROC

metric gauges the ability of the SMC-CNN-I model to distinguish between acute infarct and healthy samples across a range of thresholds. High AUROC values showcase that the proposed classifier can accurately categorize acute infarct and no disease instances across a range of possible threshold values. We have also achieved higher cohens kappa value and AUROC on synthetic data indicating the SMC-CNN-I model's ability to learn the underlying patterns and features of acute infarcts from the synthetic data and can generalize well to real-world data. This demonstrates the effectiveness and robustness of the proposed deep learning model for acute infarct prediction from MRI sequences.

Based on these quantitative results, it is determined that the SMC-CNN-I model performs better for DWI MRI sequences when compared to T2-Flair, ADC, and SWI images. This means that the model is more accurate, has higher precision and recall, better F1-score, higher Cohen's kappa, and better AUROC for DWI MRI sequences than for the other types of MR sequences. One possible reasoning could be that DWI images provide higher contrast between the infarcted and healthy tissues, making it easier for the model to distinguish between the two. Another reason could be that the proposed SMC-CNN-I model is optimized for extracting features from DWI sequences, making it more effective for predicting acute infarcts in these types of MR images. Figure 8.5 depicts the learning curve of the SMC-CNN-I model. The increase in accuracy and decrease in loss suggests that the proposed model successfully learns and identifies critical features to distinguish acute infarct lesions from normal tissue in MRI sequences.

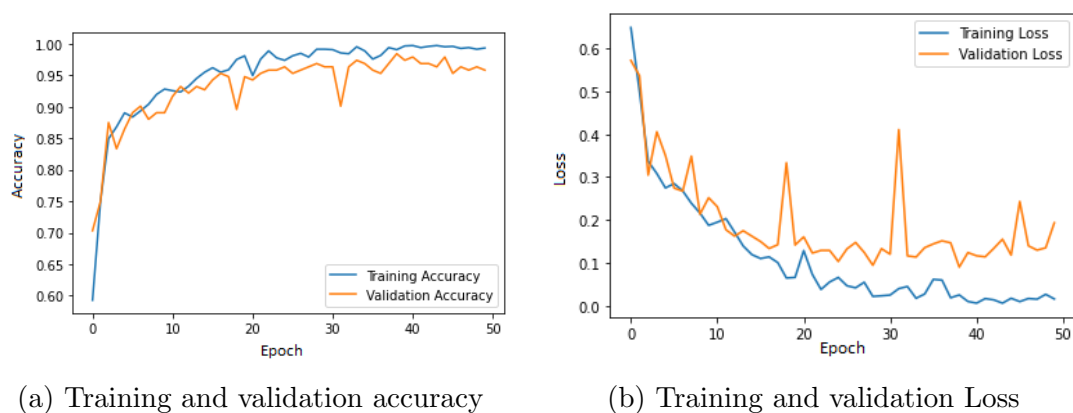


Figure 8.5: The learning curve of proposed SMC-CNN-I model

### 8.5.2 Ablation Study

To assess the impact of varying input MRI sequences on the efficacy of the SMC-CNN-M model, we performed an ablation study. The outcomes of the ablation study are presented in Table 8.7, which indicates that fusing the DWI+T2-flair+ADC+SWI imaging feature provides better performance compared to other combinations of MRI sequence fusion for both standard augmentation and synthetic data. The empirical evaluation further revealed that multi-image fusion with DWI+t2-flair+ADC+SWI gives better performance than individual MRI sequence analysis. However, it is worth noting that individual DWI analysis performs competitively compared to multi-image fusion, indicating that significant features are available in the DWI sequence for acute infarct prediction.

Our experiment suggests that multi-fusion of MRI sequences can yield better diagnostic performance compared to individual MRI analysis. By combining various features, the model can accurately predict acute infarcts from other types of brain strokes. This approach increases the model’s accuracy and can reduce the likelihood of misdiagnosis. Therefore, our research highlights the importance of considering multiple MRI sequences in medical image analysis. Our findings suggest that combining various MRI sequences can lead to better diagnostic performance and further highlight the significance of the DWI sequence in acute infarct prediction. This knowledge can help improve medical diagnosis and patient outcomes, particularly in cases of acute stroke, where timely diagnosis is crucial. The confusion matrix in Figure 8.6 illustrates the  $True_{+ve}$ ,  $False_{+ve}$ ,  $True_{-ve}$ , and  $False_{-ve}$  values of the SMC-CNN-M model when utilized on both the standard and synthetic datasets of MRI sequences, encompassing DWI, ADC, T2-flair, and SWI.

Our study evaluated the performance of the SMC-CNN-M and SMC-CNN-I models in predicting acute infarct from multiple and individual MRI sequences, respectively. Both models were trained and tested on the same dataset, but with different input MRI sequences. The SMC-CNN-M model used DWI, T2-Flair, ADC, and SWI sequences, while the SMC-CNN-I model used individual segmented MRI sequences. After evaluating the models using qualitative metrics, we found that the SMC-CNN-M model produced higher accuracy than the SMC-CNN-I model. One possible explanation for this result is that the SMC-CNN-M model was able to capture and leverage the complementary information from multiple MRI sequences, leading to more accurate predictions. On the other hand, the SMC-CNN-I model may have struggled to capture the same level of information

from individual MRI sequences alone.

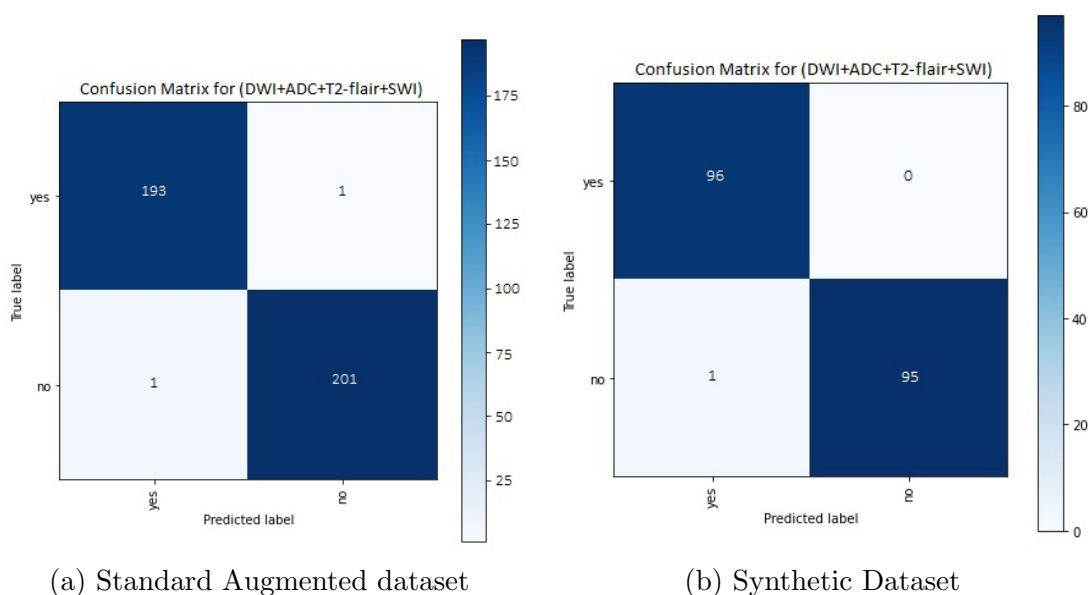


Figure 8.6: Confusion Matrix of proposed SMC-CNN-M model on DWI, T2-flair, ADC and SWI MRI sequences

### 8.5.3 Qualitative Analysis

To assess the efficacy of our presented SMC-CNN (i.e., SMC-CNN-I & SMC-CNN-M) models in predicting acute infarcts, we conducted disease visualization from DWI, T2-Flair, ADC, and SWI MRI images, enabling us to explain the model's predictions. We integrated a Grad-CAM-based deep-learning model, where the output from the final convolution layers of our SMC-CNN models was fed as input to Grad-CAM. The Grad-CAM generated heatmaps, which were superimposed onto the original images based on the gradients generated. The disease visualization results from DWI, T2-Flair, ADC, and SWI MRI sequences are illustrated in Figure 8.7. In the case of DWI and T2-Flair images, the colour red indicates higher disease localization, while blue indicates negative disease localization, and vice versa for ADC and SWI images. The Grad-CAM images generated from DWI and T2-Flair sequences offer superior qualitative results due to their hyper-intensity nature, making them ideal for accurately representing the exact location of acute infarct lesions.

This approach gave us a comprehensive understanding of our model's ability to predict acute infarcts, making it more explainable and easier to interpret. The outcomes of our investigation showcase that the SMC-CNN model we developed



Table 8.7: Ablation study of the proposed SMC-CNN-M by varying the MRI input sequence

Augmentation Techniques	MRI Sequences	Accuracy	Precision	Recall	F1-Score	Cohen's Kappa	AUROC
Stanadard Augmentation	DWI+T2-Flair+ADC+SWI	0.9949	0.9948	0.9948	0.9948	0.9888	0.9996
	DWI+T2-Flair+ADC	0.9771	0.9768	0.9773	0.9771	0.9692	0.9883
	DWI+T2-Flair	0.9642	0.9640	0.9643	0.9640	0.9527	0.9844
Synthetic data generated using DCGAN	DWI+T2-Flair+ADC+SWI	0.9953	1.0	0.9897	0.9948	0.9758	0.9942
	DWI+T2-Flair+ADC	0.9880	0.9873	0.9877	0.9881	0.9671	0.9951
	DWI+T2-Flair	0.9738	0.9741	0.9738	0.9742	0.9621	0.9899

can achieve performance levels comparable to those of expert radiologists. Our findings demonstrate that the SMC-CNN models are a promising tool for medical professionals, mainly when there is a resource shortage, and the model can help increase efficiency in radiology workflows. By being explainable, it can provide insights into how it arrived at its predictions, making it easier for radiologists to understand and interpret the results. This can lead to more accurate diagnoses and better patient outcomes. Overall, our research highlights the potential of the SMC-CNN models to serve as an essential tool in medical imaging and radiology.

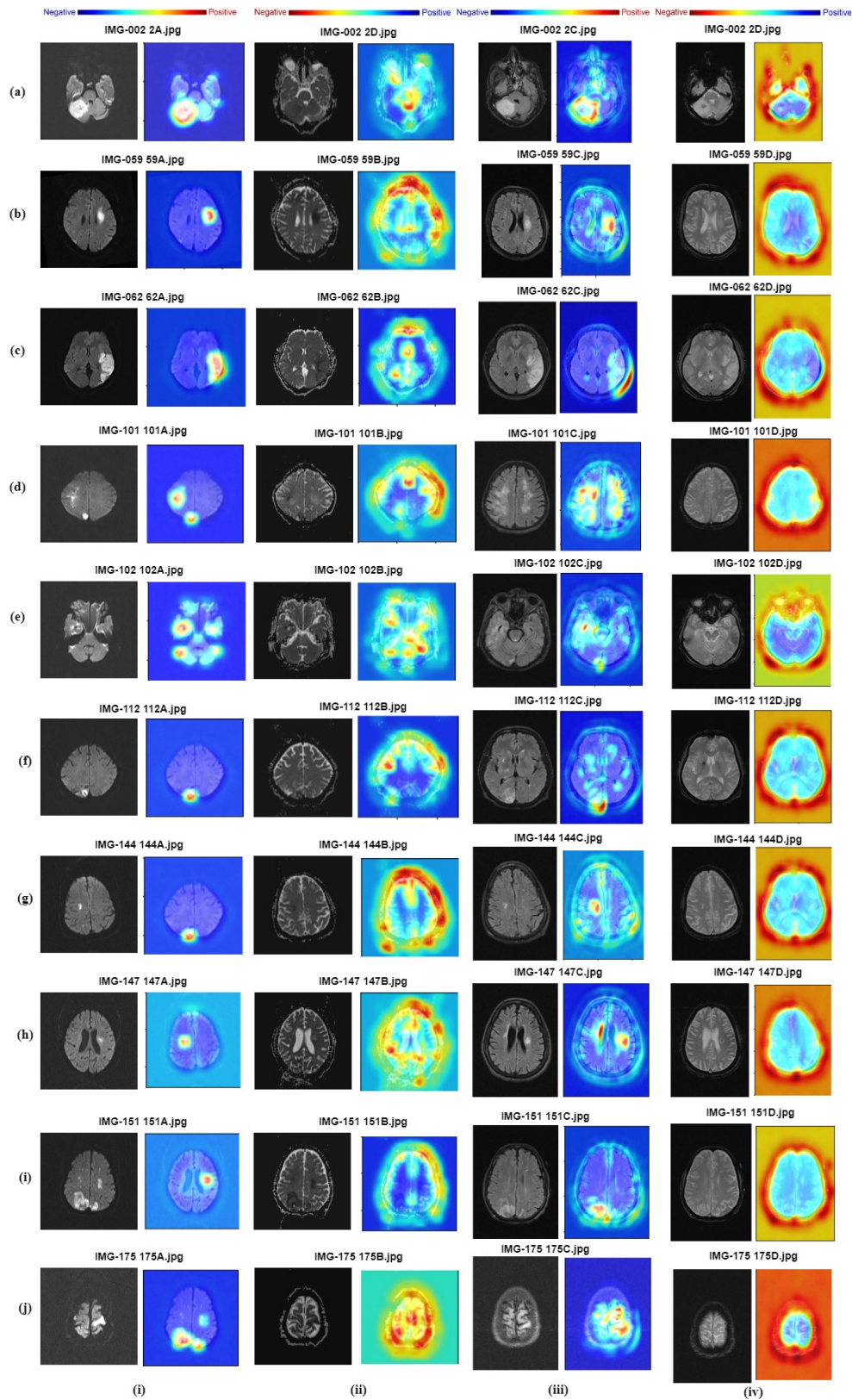


Figure 8.7: The disease visualization of acute infarct in (i) DWI, (ii) ADC (iii) T2-Flair (iv) SWI MRI sequences using Grad-CAM technique. Every First Image in the sequences indicates the original image; second image indicates the heatmap generated Grad-CAM Image. For DWI and T2-Flair MRI sequences the red indicates the higher disease localization, and the blue indicates the negative disease localization and vice versa in ADC and SWI MRI sequences. (a)-(j) represents the 10 different cases with acute infarct.

## 8.6 Discussion

The study conducted has yielded several significant findings that shed light on the subject under investigation:

- The proposed SMC-CNN (i.e., SMC-CNN-I and SMC-CNN-M) models outperformed the current baseline deep learning techniques, such as VGG-16, VGG-19, ResNet-50, MobileNet, Inception V3, EfficientNetB2, DenseNet121, and Xception, in accurately predicting acute infarcts from various MRI sequences, including DWI, T2-Flair, ADC, and SWI. Therefore, the suggested model has the capability to enhance the diagnosis of acute infarcts in medical imaging.
- The SMC-CNN-I model was found to be more effective in identifying acute infarcts in DWI MRI sequences when compared to T2-Flair, ADC, and SWI images. This could be because DWI images provide higher contrast between infarcted and healthy tissues, making it easier for the model to distinguish between the two. Additionally, the proposed SMC-CNN-I model may be optimized for extracting features from DWI sequences, making it more effective for predicting acute infarcts in these types of MR images.
- The outcomes of the empirical analysis showcase that the DWI MRI sequence is a superior imaging technique compared to other MRI sequences for identifying early acute brain infarcts. The study results present that DWI is more sensitive in identifying initial alterations in brain tissue caused by stroke. This implies that DWI can provide earlier and more accurate stroke diagnoses, leading to prompt and effective treatment and ultimately improving patient outcomes.
- The quantitative analysis of the proposed SMC-CNN model demonstrated high accuracy scores, precision values, recall values, F1-score values, Cohen's Kappa values, and AUROC values. These results show that the model can accurately identify acute infarcts and distinguish them from no-disease cases when MRI sequences are given as input.
- The ablation study conducted on the SMC-CNN-M model revealed that fusing the DWI+T2-flair+ADC+SWI imaging feature provides better performance compared to other combinations of MRI sequences. This finding indicates that a combination of different MRI sequences can enhance the precision of the model in identifying acute infarcts.

- The qualitative analysis of the SMC-CNN models provides a comprehensive perception of the model's strengths and weaknesses and its potential for future developments in the medical field. By visualizing and localizing acute infarcts in the brain regions, the model can assist medical practitioners in better understanding the location and extent of infarcts in the brain, thus facilitating better treatment decisions.
- The performance of the proposed SMC-CNN-I and SMC-CNN-M models was evaluated on synthetic MRI data created using DCGAN. The results demonstrated a considerable improvement in the ability of the model to learn and identify important patterns, leading to an accurate prediction of infarcts. Furthermore, the successful application of the model to synthetic data suggests its potential for use in real-world data, highlighting its generalizability.

## 8.7 Summary

In this chapter, two novel deep learning techniques were proposed: stacked multi-channel CNNs to predict acute infarct from individual and multiple MRI sequences, including DWI, ADC, T2-flair, and SWI. The proposed SMC-CNN-I and SMC-CNN-M model have been evaluated through a benchmarking experiment by comparing them with the baseline DL models. The MRI sequences were collected from KMC Hospital (India) and annotated by expert radiologists. Both the proposed models have outperformed the state-of-the-art baseline models such as VGG-16, VGG-19, ResNet-50, MobileNet, Inception V3, EfficientNetB2, DenseNet121, and Xception for DWI, T2-Flair, ADC, and SWI MRI sequences, demonstrating the potential of the model to predict acute infarcts from various MRI sequences accurately. The obtained accuracy, precision, recall, F1-score, and Cohen's kappa values demonstrate the effectiveness and reliability of the proposed model in identifying critical features that distinguish acute infarct lesions from normal tissue, reducing the likelihood of false-positive diagnoses. The improved performance of the proposed models on synthetic data shows that the model can be applied to real-world data, allowing for a more accurate evaluation of the model's performance compared to real-world data, where ground truth may be more difficult to determine. We conducted an ablation study on the SMC-CNN-M model, varying the input MRI sequences. Our findings show that combining DWI, T2-flair, ADC, and SWI imaging features resulted in better performance

compared to other combinations of MRI sequence fusion for both standard augmentation and synthetic data. Through empirical evaluation, we observed that multi-image fusion with DWI, T2-flair, ADC, and SWI outperformed individual MRI sequence analysis. However, individual DWI analysis still showed competitive performance, indicating that significant features are present in the DWI sequence for acute infarct prediction. Our experiment supports the idea that fusing multiple MRI sequences will lead to better diagnostic performance compared to analyzing individual MRI sequences. We also performed qualitative analysis by applying grad-CAM for disease visualization showcasing the models ability to predict the location of the infarct lesion. The multi-fusion deep learning framework we proposed has demonstrated superior performance in predicting acute infarct from MRI sequences.

## Publications

*(based on study proposed in this chapter)*

1. Shashank Shetty, Ananthanarayana V. S., Ajit Mahale, and Suba Arul Devi, Stacked Multi-Channel Convolution Neural Network for predicting Acute Brain Infarct from Magnetic Resonance Imaging Sequences, *IEEE Access* [Indexed: SCIE & Scopus, IF: 3.476] (*Status: Revisions Submitted*)

## Chapter 9

# Conclusion & Future Work

### 9.1 Conclusion

The use of medical data is crucial in developing CRS, which can potentially transform the delivery and management of personalized medicine. Computer-assisted CRS is pivotal in helping clinicians with the prognosis and treatment process and delivers essential benefits that are fundamental to the healthcare industry. When it comes to clinical decision-making, a CRS can act as a consultant for less experienced healthcare providers or as an additional viewpoint for seasoned clinicians, providing specialist insights that can be valuable to both. CRSs offer accurate prognostic recommendations and suggest economical and efficient treatments that can be advantageous to the intended patient group. However, building a CRS with a critical emphasis on the system's performance has posed considerable obstacles during the design and development phases. After a comprehensive analysis of the literature, it has been discovered that there is considerable potential for creating an AI-powered clinical recommendation system that employs a variety of healthcare data sources. Various research gaps have been recognized, particularly in areas such as predicting diseases from unstructured radiology text, interpreting unstructured radiology images, and integrating multimodal radiology data.

An in-depth examination of the literature has revealed a significant opportunity to enhance the performance of disease prediction systems, particularly with regards to addressing the challenges associated with unstructured radiology free-text reports, which are addressed in our first research objective. Towards this, in Chapter 4, we have proposed a UM-TES framework comprising clinical Clinical Knowledge-based Text modelling techniques with a deep learning framework to predict pulmonary diseases in radiology free-text reports. To model the text in the diagnostic reports, the GloVe Embedding model was used in conjunction

with a knowledge base. The textual features were then processed using the DDR-CNN model to reduce their dimensionality. The final step was to apply a DNN to predict any abnormalities in the reports. Through our experimentation, we observed that the proposed UM-TES word embedding technique yielded superior performance when compared to state-of-the-art NLP models. Additionally, we evaluated the performance of the DNN classifier against that of a standard machine learning-based classifier and determined that the former achieved better results. Our observation revealed that the improved performance of UM-TES is attributed to the integration of a radiology knowledge base, which enhances prediction accuracy even when the training cohort is small in size. Consequently, the proposed model can be implemented in scenarios where data are scarce, which is often the case in the medical domain, where cohorts are institution-specific or restricted to specific domains. The framework consists of several key factors that contribute to its superior performance compared to state-of-the-art NLP models:

- *Integration of Knowledge Base:* One of the main factors contributing to the improved performance of UM-TES is the incorporation of a radiology knowledge base. This integration enhances the prediction accuracy, particularly when dealing with small training cohorts.
- *Semantic Embeddings from GloVe:* To represent the diagnostic reports effectively, we employed the GloVe word embedding model. GloVe captures semantic relationships between words in a corpus and encodes their meanings and contexts in continuous vector spaces. Applying this technique to radiology reports enables the model to understand intricate connections between medical terms and phrases associated with pulmonary diseases, thereby improving its comprehension of the data.
- *Better Word Representations:* UM-TES benefits from using dense and meaningful word embeddings provided by GloVe. Unlike traditional sparse representations like bag-of-words, these embeddings offer more informative input to the deep neural network. As a result, the model can learn more relevant patterns and dependencies present in the radiology reports.
- *Capturing Contextual Information:* DNNs excel at capturing complex patterns and contextual information within data. In the context of radiology reports, DNNs can recognize language patterns indicative of various pulmonary diseases. By learning from the semantic embeddings and hierarchical structures within the text data, DNNs become adept at understanding



the subtleties in medical language.

Intending to design and develop AI-powered CRS for disease prediction from an unstructured medical image, in Chapter 5 as a second research objective, we have proposed a lightweight and explainable deep learning network named UMVES, a Multi-Scale Chest X-ray Network that consists of MSDL and DS-CNN layers to predict the pulmonary diseases from the CXR obtained from the publicly available Open-I dataset and the CXR data collected from the private medical hospital. The MSDL layer captures the multi-scale features with the help of a broader receptive field, and the DS-CNN layer learns the imaging features by adjusting lesser parameters. The quantitative and qualitative analyses of the proposed UM-VES model are performed on both CXR datasets. The experimental validation was observed through evaluation metrics like accuracy, precision, recall, F1-score, MCC, and AUROC. The experimental results show that the proposed model outperformed baseline deep learning techniques and existing state-of-the-art approaches. The MSDL layer in the proposed model has significantly impacted the prediction outcome by capturing the Multi-scale features from the CXR. The grad-CAM method is employed to visualize the pulmonary abnormalities from the CXR and to check the model's ability to arrive at a decision. The obtained grad-CAM CXR samples are compared with the CXRs labelled by expert radiologists. It is observed that the UM-VES can reach a performance level similar to that of the radiologists. This study also presents RAD-DCGAN for generating synthetic images from radiology X-ray and MRI cohorts collected from a private medical hospital. We have conducted a comprehensive qualitative analysis of the proposed RAD-DCGAN compared with conventional data augmentation techniques like rotation, zooming, brightness, and shearing. The eight state-of-the-art deep learning classifiers are used to check the efficacy of the data generated from the proposed RAD-DCGAN and the traditional data augmentation techniques. The detailed investigation shows that the synthetic data generated through the proposed RAD-DCGAN has achieved a significantly higher classification accuracy of 3-4% compared to the data generated through basic data augmentation strategies. This superior performance is due to the higher-resolution synthetic images generated with additional information, which aids the classifier's performance. The framework consists of several other key factors that contribute to its superior performance:

- *Capturing Multi-Scale Information:* The MSDL layer is designed to capture multi-scale features from the input data effectively. It does this by applying

dilated convolutions with different dilation rates in parallel. Dilated convolutions increase the receptive field without introducing additional parameters, allowing the model to capture context over larger spatial scales. By using multiple dilated convolutions in parallel, the MSDL layer can gather information at various scales, which is particularly advantageous in medical imaging tasks where diseases may present at different scales within the images.

- *Enhanced Feature Representation:* The MSDL layer allows the UMVES model to encode rich and diverse information from the input data. This helps in representing complex patterns and structures in medical images accurately. The capability of capturing multi-scale information enables the model to detect subtle pulmonary abnormalities, leading to improved diagnostic performance.
- *Flexibility in Receptive Field Size:* By adjusting the dilation rates, the MSDL layer can control the size of the receptive field for each parallel convolution. This adaptability enables the UMVES model to focus on relevant features and adapt to different CXR image characteristics, making it more versatile and robust in handling diverse datasets.
- *Efficient Hierarchical Processing:* The parallel structure of the MSDL layer facilitates hierarchical feature extraction. It allows the model to learn features at multiple levels of abstraction, starting from fine details to broader context. This hierarchical processing helps the UMVES model comprehend the complex anatomical and pathological structures present in CXR images.
- *Reduced Parameters with features retained:* The DS-CNN layer used in UMVES is a lightweight and efficient variant of the traditional convolutional neural network. It reduces the number of parameters while still learning important imaging features effectively. This makes the UMVES model computationally efficient and allows it to process CXR images more quickly. Additionally, the DS-CNN's ability to learn meaningful representations from medical images contributes to the model's superior performance.

Upon conducting an extensive review, it has been discovered that the analysis of multimodal medical data can significantly improve the performance of disease prediction systems. To address the challenges related to integrating this data into a single space, our third objective in Chapter 6 was focused on finding solutions to these shortcomings. After performing a comprehensive investigation on two multimodal clinical datasets, it was found that multimodal learning provides a benefit

over unimodal learning when performing the classification of radiology chest X-rays with associated clinical free-text notes. With regards to the two proposed multimodal fusion strategies, CBP-MMFN performs better than the DHP-MMFN model across publicly available Indiana University cohorts and data collected from the KMC hospital. The superior results in CBP-MMFN are obtained because of the intermodal dynamics between the textual and imaging modalities. The Bilinear interaction map generated from the outer product of visual and textual features in CBP-MMFN generates a far more expressive multimodal feature representation, encoding more tensor correlation than the simple concatenation operation and element-wise product. Hence, the discriminative features extracted from the CBP-MMFN model provide a significant performance gain over the uni-modal models and the DHP-MMFN model. The unimodal text-only model (UM-TES) has given more promising results than the proposed unimodal image-only (UM-VES) model. The two major reasons for it are as follows:

- Incorporating a clinical knowledge base helps to jointly learn word vectors from the cohort and knowledge base, which increases the vocabulary size and allows learning infrequent clinical words.
- It has been found that radiology reports have more discriminative features than chest X-rays. This is because the annotators have focused on the text being assigned to the labels of the radiology reports.

We also observed that the existing state-of-the-art multimodal fusion techniques applied to radiology images and their associated reports are either straightforward concatenation or late fusion techniques like averaging, which ignore intermodal interaction among the two modalities. The proposed multimodal medical tensor fusion techniques outperform the existing state-of-the-art techniques. The proposed models focus on inter-modal dynamics, which find the tensor correlation between the textual and imaging modalities. The experimental results prove that the multimodal representation obtained from the proposed model has more expressive features than the traditional concatenation strategy. In conclusion, we presented the unimodal text-only model (i.e., UM-TES) and the unimodal image-only model (UM-VES) to predict abnormalities from radiology reports and Chest X-rays. We proposed two Multimodal Medical Tensor Fusion Networks (i.e., CBP-MMFN and DHP-MMFN) for predicting abnormalities from a radiology chest X-ray and its associated reports. We evaluated both proposed multimodal and unimodal models on two multimodal radiology cohorts: a) publicly available Indiana University dataset, b) Real-time data collected from KMC private hospitals.

After a thorough investigation, we found that multimodal models have better performance than models using a single modality. We also compared the proposed multimodal fusion models with the state-of-the-art fusion models for predicting abnormalities in the radiology cohort. Our proposed model has achieved superior performance. We conclude that multimodal learning leads to competitive performance in predicting abnormalities from radiology chest x-rays with associated reports. The superior performance of the overall proposed fusion models can be attributed to the following key factors:

- *Intermodal Interaction:* In traditional concatenation-based multimodal fusion, the features from different modalities are simply combined in a linear manner, which might not effectively leverage the intermodal dependencies present in the data. On the other hand, the DHP-MMFN and CBP-MMFN are advanced techniques that provide intermodal interaction by capturing non-linear dependencies between the multimodal text and visual features.
- *Bilinear interaction map:* The Bilinear interaction map generated from the outer product of visual and textual features in CBP-MMFN generates a far more expressive multimodal feature representation, encoding more tensor correlation than the simple concatenation operation and element-wise product. Hence, the discriminative features extracted from the CBP-MMFN model provide a significant performance gain over the uni-modal models and the DHP-MMFN model.
- *Use of Specialized Modules:* The framework gains a significant edge by utilizing specialized modules, namely UM-VES and UM-TES, to extract discriminative visual and textual features from CXR and Radiology reports.

In Chapter 7, as part of our third research objective, we aimed to develop a deep learning-based model that can accurately and automatically generate diagnostic reports from CXR images. To achieve this, we employed a cross-modal retrieval technique that retrieves radiology reports from the image. Our approach, which utilized the beam search method, outperformed existing models in generating robust diagnostic reports. This can be attributed to the encoder of our proposed network, which extracted multi-channel visual features and discriminative text features based on knowledge. Compared to existing models, our approach showed superior results in terms of BLEU4 scores, which is a standard metric used to compare the accuracy of generated text to the ground truth. In addition, we created a dynamic web portal that allows for the easy uploading of frontal and

lateral CXR images and provides the corresponding diagnostic reports as output. This feature greatly simplifies the report writing process for radiologists, as it automates the process and saves time. The key factors that influence superior performance are as follows:

- *Sequence-to-Sequence Mapping:* The encoder-decoder architecture is specifically designed for sequence-to-sequence mapping, which is a natural fit for report generation tasks. The encoder processes the input (e.g., chest X-ray scans) and encodes the information into a fixed-length vector or context representation. The decoder then uses this context representation to generate a variable-length sequence (e.g., diagnostic reports) based on the encoded information.
- *Handling Variable-Length Outputs:* In report generation, the length of the output text (reports) can vary significantly based on the complexity of the input. The encoder-decoder architecture can handle this variability, as the decoder is capable of generating sequences of different lengths, making it well-suited for tasks where the length of the output is not predetermined.
- *Semantic Understanding:* The encoder module learns to extract meaningful representations from the input data. In the case of chest X-ray scans, the encoder can understand the relevant features and patterns in the images, capturing important diagnostic information that is useful for generating accurate and contextually relevant reports.
- *End-to-End Learning:* The entire encoder-decoder module can be trained end-to-end using backpropagation. This means that the model learns to optimize both the encoding and decoding processes simultaneously, leading to more effective representations and better report generation.
- *Handling Rare or Unseen Cases:* The encoder-decoder architecture can handle cases that were not explicitly seen during training, as the model learns to generalize from the patterns and structures in the data. This adaptability is crucial in medical report generation, where rare conditions or unique cases may arise.
- *Use of Specialized Modules:* The framework employs specialized modules like UM-VES, UM-TES, and LSTM. Each module likely plays a crucial role in the overall performance.

In Chapter 8, we address the challenge of fusing imaging features from multimodal medical images and representing them in a common space for disease prediction. To tackle this problem, we propose two novel deep learning techniques that utilize SMC-CNNs. These techniques are designed to predict acute infarct from individual and multiple MRI sequences, such as DWI, ADC, T2-flair, and SWI. The proposed SMC-CNN-I and SMC-CNN-M models have been evaluated through a benchmarking experiment by comparing them with the baseline DL models. The MRI sequences were collected from KMC Hospital (India) and annotated by expert radiologists. Both the proposed models have outperformed the state-of-the-art baseline models such as VGG-16, VGG-19, ResNet-50, MobileNet, Inception V3, EfficientNetB2, DenseNet121, and Xception for DWI, T2-Flair, ADC, and SWI MRI sequences, demonstrating the potential of the model to predict acute infarcts from various MRI sequences accurately. The obtained accuracy, precision, recall, F1-score, and Cohen's kappa values demonstrate the effectiveness and reliability of the proposed model in identifying critical features that distinguish acute infarct lesions from normal tissue, reducing the likelihood of false-positive diagnoses. The improved performance of the proposed models on synthetic data shows that the model can be applied to real-world data, allowing for a more accurate evaluation of the model's performance compared to real-world data, where ground truth may be more difficult to determine. We conducted an ablation study on the SMC-CNN-M model, varying the input MRI sequences. Our findings show that combining DWI, T2-flair, ADC, and SWI imaging features resulted in better performance compared to other combinations of MRI sequence fusion for both standard augmentation and synthetic data. Through empirical evaluation, we observed that multi-image fusion with DWI, T2-flair, ADC, and SWI outperformed individual MRI sequence analysis. However, individual DWI analyses still showed competitive performance, indicating that significant features are present in the DWI sequence for acute infarct prediction. Our experiment supports the idea that fusing multiple MRI sequences will lead to better diagnostic performance compared to analyzing individual MRI sequences. We also performed qualitative analysis by applying grad-CAM for disease visualization showcasing the model's ability to predict the location of the infarct lesion. The multi-fusion deep learning framework we proposed has demonstrated superior performance in predicting acute infarct from MRI sequences. The key factors that influence superior performance are as follows:

- *Multi-Channel Fusion:* The SMC-CNN models are designed to fuse infor-

mation from multiple MRI sequences, including DWI, T2-Flair, ADC, and SWI. The fusion of information from these different sequences might enhance the model's ability to capture complementary features and patterns, leading to better predictions.

- *Deep Convolutional Neural Networks (CNNs)*: The core building blocks of the SMC-CNN models are deep CNNs. These networks are well-known for their ability to automatically learn hierarchical features from data, which is crucial for medical image analysis tasks like infarct detection. The effective use of CNNs might be a significant factor in achieving superior performance.

## 9.2 Future Work

The thesis proposes an intelligent framework for predicting diseases using multimodal medical data, which includes radiology text and images. The proposed framework utilizes advanced ML and DL techniques to extract meaningful semantic features from the medical data, which are then used for disease prediction. However, there is still scope for improvement and future research in this area. Other medical modalities, such as structured clinical data or genomic data, can be incorporated into the framework to obtain even better semantic features. A potential opportunity to enhance prognostic decision-making exists by leveraging additional sources of unstructured radiology text, such as medical prescriptions and discharge summaries. In addition to leveraging unstructured text, our aim is to enhance the performance of the image-only model. By doing so, we can potentially improve the accuracy and reliability of radiology image analysis. Furthermore, we plan to expand the scope of our model by applying it to other types of diagnostic images such as MRI, ultrasound, and CT scans. This will allow us to further evaluate the effectiveness and robustness of our proposed model across different modalities and ultimately improve its overall utility in clinical settings. We have leveraged the Grad-CAM technique to enhance the interpretability of our proposed models, allowing doctors to visualize the areas of the input image that were most significant in making a given prediction, including true and false positives. By providing these interpretable results, doctors can better comprehend the decision-making process of our models and identify potential areas for improvement or biases. This, in turn, fosters greater trust and acceptance of automated systems in clinical settings. Moving forward, we plan to explore additional visualization techniques to further understand the ability of our models to

make decisions. Our extensive literature review has revealed a significant shortage of high-quality, diverse multimodal medical data available for AI research. This scarcity emphasizes the urgent need for concerted efforts to collect, curate, and expertly annotate such data, in order to conduct comprehensive studies that can yield actionable insights and improved health outcomes. The potential use of models inspired by social networks, text, and linkage exploitation, along with attention mechanisms, holds promise for enhancing multimodal clinical data fusion in the context of pulmonary disease prediction from CXR and radiology text reports. As we move forward with our research, we are keen to explore and incorporate these innovative approaches into our future work.



# Appendix A

## Publications based on Research Work

### A.1 Journal Publications

1. Shashank Shetty, Ananthanarayana V S., and Ajit Mahale (2022). MS-CheXNet: An Explainable and Lightweight Multi-Scale Dilated Network with Depthwise Separable Convolution for Prediction of Pulmonary Abnormalities in Chest Radiographs, *Multidisciplinary Digital Publishing Institute (MDPI) Mathematics*, 10, no. 19: 3646. <https://doi.org/10.3390/math10193646> [Indexed: SCIE & Scopus, IF: 2.592] (*Status: Published Online*)
2. Shashank Shetty, Ananthanarayana V. S., and Ajit Mahale (2022), Comprehensive Review of Multimodal Medical Data Analysis: Open Issues and Future Research Directions, *Acta Informatica Pragensia (AIP)*, 11(3), 423-457, <https://doi.org/10.18267/j.aip.202> [Indexed: Scopus, IF: 1.15] (*Status: Published Online*)
3. Shashank Shetty, Ananthanarayana V. S., and Ajit Mahale (2023), Multimodal medical tensor fusion network-based DL framework for abnormality prediction from the radiology CXRs and clinical text reports. *Multimedia Tools and Applications*, Springer Publisher, <https://doi.org/10.1007/s11042-023-14940-x> [Indexed: SCIE & Scopus, IF: 2.577] (*Status: Published Online*)
4. Shashank Shetty, Ananthanarayana V S., and Ajit Mahale. (2023). Diagnostic Performance Evaluation of Deep Learning-Based Medical Text Modelling to Predict the Pulmonary Diseases from the Unstructured Radiology Free-Text Reports. *Acta Informatica Pragensia*. Volume 12, Issue 2, <https://doi.org/10.18267/j.aip.214>, <https://aip.vse.cz/corproof.php?tartkey=>

[aip-000000-0483](#) [Indexed: Scopus, IF: 1.15] (*Status: Published Online*)

5. Shashank Shetty, Ananthanarayana V. S., and Ajit Mahale, Cross-Modal Deep Learning-based Clinical Recommendation System for Radiology Report Generation from Chest X-rays, *International Journal of Engineering*, [Indexed: ESCI & Scopus, IF: 1.64] (*Status: Published Online*)
6. Shashank Shetty, Ananthanarayana V. S., Ajit Mahale, and Suba Arul Devi, Stacked Multi-Channel Convolution Neural Network for predicting Acute Brain Infarct from Magnetic Resonance Imaging Sequences, *IEEE Access* [Indexed: SCIE & Scopus, IF: 3.476] (*Status: Revisions Submitted*)

## A.2 Conference Publications

1. Shashank Shetty, Ananthanarayana V S., and Ajit Mahale. (2020) Medical Knowledge-Based Deep Learning Framework for Disease Prediction on Unstructured Radiology Free-Text Reports Under Low Data Condition, *21st EANN (Engineering Applications of Neural Networks) 2020 Conference. EANN 2020., vol 2. Springer, Cham, Halkidiki, Greece.* [https://doi.org/10.1007/978-3-030-48791-1\\_27](https://doi.org/10.1007/978-3-030-48791-1_27) [Core Ranked Conference - Springer Proceedings] (*Status: Published Online*)
2. Shashank Shetty, Ananthanarayana V S., and Ajit Mahale. Data Augmentation vs. Synthetic Data Generation: An Empirical Evaluation for Enhancing Radiology Image Classification, *IEEE 17th International Conference on Industrial and Information Systems (ICIIS'23)*. [Core Ranked Conference] (*Status: Accepted For Presentation*)

## References

- Abadi, M. and et. al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. URL <https://www.tensorflow.org/>. Software available from tensorflow.org.
- Abdellatif, A., J. Bouaud, C. Lafuente-Lafuente, J. Belmin, and B. Séroussi (2021). Computerized decision support systems for nursing homes: A scoping review. *Journal of the American Medical Directors Association*, 22(5), 984–994. URL <https://doi.org/10.1016/j.jamda.2021.01.080>.
- Abiyev, R. and M. Ma'aitah (2018). Deep convolutional neural networks for chest diseases detection. *Journal of Healthcare Engineering*, 2018, 1–11.
- Acosta, J. N., G. J. Falcone, P. Rajpurkar, and E. J. Topol (2022). Multimodal biomedical AI. *Nature Medicine*, 28(9), 1773–1784. URL <https://doi.org/10.1038/s41591-022-01981-2>.
- Adali, T., Y. Levin-Schwartz, and V. D. Calhoun (2015). Multimodal data fusion using source separation: Application to medical imaging. *Proceedings of the IEEE*, 103(9), 1494–1506. URL <https://doi.org/10.1109/jproc.2015.2461601>.
- Agarap, A. F. (2018). Deep Learning using Rectified Linear Units (ReLU). *arXiv e-prints*, arXiv:1803.08375.
- Albawi, S., T. A. Mohammed, and S. Al-Zawi, Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*. 2017.
- Alfarghaly, O., R. Khaled, A. Elkorany, M. Helal, and A. Fahmy (2021a). Automated radiology report generation using conditioned transformers. 24, 100557. URL <https://doi.org/10.1016/j.imu.2021.100557>.

- Alfarghaly, O., R. Khaled, A. Elkorany, M. Helal, and A. Fahmy (2021b). Automated radiology report generation using conditioned transformers. *Informatomics in Medicine Unlocked*, 24, 100557. ISSN 2352-9148. URL <https://www.sciencedirect.com/science/article/pii/S2352914821000472>.
- Algaze, C. A., M. Wood, N. M. Pageler, P. J. Sharek, C. A. Longhurst, and A. Y. Shin (2016). Use of a checklist and clinical decision support tool reduces laboratory use and improves cost. *Pediatrics*, 137(1). URL <https://doi.org/10.1542/peds.2014-3019>.
- Alqahtani, A., H. U. Khan, S. Alsubai, M. Sha, A. Almadhor, T. Iqbal, and S. Abbas (2022). An efficient approach for textual data classification using deep learning. *Frontiers in Computational Neuroscience*, 16. URL <https://doi.org/10.3389/fncom.2022.992296>.
- Amal, S., L. Safarnejad, J. A. Omiye, I. Ghanzouri, J. H. Cabot, and E. G. Ross (2022). Use of multi-modal data and machine learning to improve cardiovascular disease care. *Frontiers in Cardiovascular Medicine*, 9. ISSN 2297-055X. URL <https://www.frontiersin.org/articles/10.3389/fcvm.2022.840262>.
- An, J., X. Lu, and H. Duan, Integrated visualization of multi-modal electronic health record data. In *2008 2nd International Conference on Bioinformatics and Biomedical Engineering*. 2008a.
- An, J., X. Lu, and H. Duan, Integrated visualization of multi-modal electronic health record data. In *2008 2nd International Conference on Bioinformatics and Biomedical Engineering*. 2008b.
- Anesi, G. L. and M. P. Kerlin (2021). The impact of resource limitations on care delivery and outcomes: routine variation, the coronavirus disease 2019 pandemic, and persistent shortage. *Current Opinion in Critical Care*, 27(5), 513–519. URL <https://doi.org/10.1097/mcc.0000000000000859>.
- Araujo, A., W. D. Norris, and J. Sim (2019). Computing receptive fields of convolutional neural networks. *Distill*.
- Arenillas, J. F., A. Rovira, C. A. Molina, E. Grive, J. Montaner, and J. Alvarez-Sabin (2002). Prediction of early neurological deterioration using diffusion- and perfusion-weighted imaging in hyperacute middle cerebral artery ischemic stroke. *Stroke*, 33(9), 2197–2205. URL <https://doi.org/10.1161/01.str.0000027861.75884.df>.

- Aydin, F., M. Zhang, M. Ananda-Rajah, and G. Haffari (2019a). Medical multimodal classifiers under scarce data condition. *CoRR*, abs/1902.08888. URL <http://arxiv.org/abs/1902.08888>.
- Aydin, F., M. Zhang, M. Ananda-Rajah, and G. Haffari (2019b). Medical multimodal classifiers under scarce data condition. *CoRR*, abs/1902.08888. URL <http://arxiv.org/abs/1902.08888>.
- Bai, Y.-L., E. Yang, B. Han, Y. Yang, J. Li, Y. Mao, G. Niu, and T. Liu, Understanding and improving early stopping for learning with noisy labels. In *Neural Information Processing Systems*. 2021.
- Banerjee, I., M. Sofela, J. Yang, J. Chen, N. Shah, R. Ball, A. Mushlin, M. Desai, J. Bledsoe, T. Amrhein, D. Rubin, R. Zamanian, and M. Lungren (2019). Development and performance of the pulmonary embolism result forecast model (perform) for computed tomography clinical decision support. *JAMA Network Open*, 2, e198719.
- Barik, D. and A. Thorat (2015). Issues of unequal access to public health in india. *Frontiers in Public Health*, 3. URL <https://doi.org/10.3389/fpubh.2015.00245>.
- Barillot, C., D. Lemoine, L. L. Briquer, F. Lachmann, and B. Gibaud (1993). Data fusion in medical imaging: merging multimodal and multipatient images, identification of structures and 3d display aspects. *European Journal of Radiology*, 17(1), 22–27. URL [https://doi.org/10.1016/0720-048x\(93\)90024-h](https://doi.org/10.1016/0720-048x(93)90024-h).
- Bayrak, S., E. Yucel, and H. Takci (2022). Epilepsy radiology reports classification using deep learning networks. *Computers, Materials & Continua*, 70(2), 3589–3607. URL <https://doi.org/10.32604/cmc.2022.018742>.
- Bekiesińska-Figatowska, M. (2015). Artifacts in magnetic resonance imaging. *Polish Journal of Radiology*, 80, 93–106. URL <https://doi.org/10.12659/pjr.892628>.
- Bengio, Y. (2009). Learning deep architectures for ai. *Foundations*, 2, 1–55.
- Bergstra, J. and Y. Bengio (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb), 281–305. ISSN ISSN 1533-7928. URL <http://www.jmlr.org/papers/v13/bergstra12a.html>.

- Berner, E. S., *Clinical Decision Support Systems: Theory and Practice*. Springer Publishing Company, Incorporated, 2010, 2nd edition. ISBN 1441922237, 9781441922236.
- Bharati, S., P. Podder, and M. R. H. Mondal (2020). Hybrid deep learning for detecting lung diseases from x-ray images. *Informatics in Medicine Unlocked*, 20, 100391. URL <https://doi.org/10.1016/j.imu.2020.100391>.
- Blackmore, C. C., R. S. Mecklenburg, and G. S. Kaplan (2011). Effectiveness of clinical decision support in controlling inappropriate imaging. *Journal of the American College of Radiology*, 8(1), 19–25. URL <https://doi.org/10.1016/j.jacr.2010.07.009>.
- Bleyer, W. (1997). The u.s. pediatric cancer clinical trials programmes: International implications and the way forward. *European Journal of Cancer*, 33(9), 1439–1447. URL [https://doi.org/10.1016/s0959-8049\(97\)00249-9](https://doi.org/10.1016/s0959-8049(97)00249-9).
- Bloice, M. D., C. Stocker, and A. Holzinger (2017a). Augmentor: An image augmentation library for machine learning. *CoRR*, abs/1708.04680. URL <http://arxiv.org/abs/1708.04680>.
- Bloice, M. D., C. Stocker, and A. Holzinger (2017b). Augmentor: An image augmentation library for machine learning. *Journal of Open Source Software*, 2(19), 432. URL <https://doi.org/10.21105/joss.00432>.
- Boikanyo, K., A. M. Zungeru, B. Sigweni, A. Yahya, and C. Lebekwe (2023). Remote patient monitoring systems: Applications, architecture, and challenges. *Scientific African*, 20, e01638. ISSN 2468-2276. URL <https://www.sciencedirect.com/science/article/pii/S2468227623000959>.
- Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.
- Boonn, W. W. and C. P. Langlotz (2009). Radiologist use of and perceived need for patient data access. *Journal of digital imaging*, 22(4), 357—362. ISSN 0897-1889. URL <https://europepmc.org/articles/PMC3043710>.
- Brady, A., R. Ó. Laoide, P. McCarthy, and R. McDermott (2012). Discrepancy and error in radiology: concepts, causes and consequences. *Ulster Med. J.*, 81(1), 3–9.

- Brady, A. P. (2016). Error and discrepancy in radiology: inevitable or avoidable? *Insights into Imaging*, 8(1), 171–182. URL <https://doi.org/10.1007/s13244-016-0534-1>.
- Brant-Zawadzki, M., D. Atkinson, M. Detrick, W. G. Bradley, and G. Scidmore (1996). Fluid-attenuated inversion recovery (FLAIR) for assessment of cerebral infarction. *Stroke*, 27(7), 1187–1191. URL <https://doi.org/10.1161/01.str.27.7.1187>.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
- Bridge, C. P., B. C. Bizzo, J. M. Hillis, J. K. Chin, D. S. Comeau, R. Gauriau, F. Macruz, J. Pawar, F. T. C. Noro, E. Sharaf, M. S. Takahashi, B. Wright, J. F. Kalafut, K. P. Andriole, S. R. Pomerantz, S. Pedemonte, and R. G. González (2022). Development and clinical application of a deep learning model to identify acute infarct on magnetic resonance imaging. *Scientific Reports*, 12(1). URL <https://doi.org/10.1038/s41598-022-06021-0>.
- Cabana, M. D., C. S. Rand, N. R. Powe, A. W. Wu, M. H. Wilson, P.-A. C. Abboud, and H. R. Rubin (1999). Why don't physicians follow clinical practice guidelines? *JAMA*, 282(15), 1458. URL <https://doi.org/10.1001/jama.282.15.1458>.
- Calloway, S., H. A. Akilo, and K. Bierman (2013). Impact of a clinical decision support system on pharmacy clinical interventions, documentation efforts, and costs. *Hospital Pharmacy*, 48(9), 744–752. URL <https://doi.org/10.1310/hpj4809-744>.
- Candemir, S., S. Rajaraman, G. Thoma, and S. Antani, Deep learning for grading cardiomegaly severity in chest x-rays: An investigation. In *2018 IEEE Life Sciences Conference (LSC)*. IEEE, 2018. URL <https://doi.org/10.1109/lsc.2018.8572113>.
- Carvalho, R., J. Pedrosa, and T. Nedelcu, Multimodal multi-tasking for skin lesion classification using deep neural networks. In G. Bebis, V. Athitsos, T. Yan, M. Lau, F. Li, C. Shi, X. Yuan, C. Mousas, and G. Bruder (eds.), *Advances in Visual Computing*. Springer International Publishing, Cham, 2021. ISBN 978-3-030-90439-5.

- Castro, S., E. Tseytlin, O. Medvedeva, K. Mitchell, S. Visweswaran, T. Bekhuis, and R. Jacobson (2017). Automated annotation and classification of bi-rads assessment from radiology reports. *Journal of Biomedical Informatics*, 69.
- Chandra, B. S., C. S. Sastry, and S. Jana (2019). Robust heartbeat detection from multimodal data via CNN-based generalizable information fusion. *IEEE Transactions on Biomedical Engineering*, 66(3), 710–717. URL <https://doi.org/10.1109/tbme.2018.2854899>.
- Chanumolu, R., L. Alla, P. Chirala, N. C. Chennampalli, and B. P. Kolla, Multi-modal medical imaging using modern deep learning approaches. In *2022 IEEE VLSI Device Circuit and System (VLSI DCS)*. 2022.
- Chapman, B., S. Lee, H. Kang, and W. Chapman (2011). Document-level classification of ct pulmonary angiography reports based on an extension of the context algorithm. *Journal of biomedical informatics*, 44, 728–37.
- Chaudhary, A., A. Hazra, and P. Chaudhary, Diagnosis of chest diseases in x-ray images using deep convolutional neural network. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. 2019.
- Chaudhury, S., L. Dey, I. Verma, and E. Hassan, Mining multimodal data. In *Pattern Recognition and Big Data*. WORLD SCIENTIFIC, 2016, 581–604. URL [https://doi.org/10.1142/9789813144552\\_0017](https://doi.org/10.1142/9789813144552_0017).
- Chen, L.-C., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille (2018). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4), 834–848.
- Chen, R., J. C. Ho, and J.-M. S. Lin (2020). Extracting medication information from unstructured public health data: a demonstration on data from population-based and tertiary-based samples. *BMC Medical Research Methodology*, 20(1). URL <https://doi.org/10.1186/s12874-020-01131-7>.
- Chen, X. and X. Lin (2014). Big data deep learning: Challenges and perspectives. *IEEE Access*, 2, 514–525.
- Chen, Z., Y. Shen, Y. Song, and X. Wan, Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Asso-*



- ciation for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 2021. URL <https://aclanthology.org/2021.acl-long.459>.
- Chien, S.-C., Y.-L. Chen, C.-H. Chien, Y.-P. Chin, C. H. Yoon, C.-Y. Chen, H.-C. Yang, and Y.-C. J. Li (2022). Alerts in clinical decision support systems (CDSS): A bibliometric review and content analysis. *Healthcare*, 10(4), 601. URL <https://doi.org/10.3390/healthcare10040601>.
- Cho, I., J.-H. Lee, J. Choi, H. Hwang, and D. W. Bates (2016). National rules for drug–drug interactions: Are they appropriate for tertiary hospitals? *Journal of Korean Medical Science*, 31(12), 1887. URL <https://doi.org/10.3346/jkms.2016.31.12.1887>.
- Chollet, F. (2016). Xception: Deep learning with depthwise separable convolutions. *CoRR*, abs/1610.02357. URL <http://arxiv.org/abs/1610.02357>.
- Chollet, F., Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017a.
- Chollet, F., Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 2017b. ISSN 1063-6919. URL <https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.195>.
- Cohen, J. P., M. Hashir, R. Brooks, and H. Bertrand (2020). On the limits of cross-domain generalization in automated x-ray prediction. URL <https://arxiv.org/abs/2002.02497>.
- Cohen, M. (2007). Accuracy of information on imaging requisitions: Does it matter? *Journal of the American College of Radiology : JACR*, 4, 617–21.
- Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. P. Kuksa (2011). Natural language processing (almost) from scratch. *CoRR*, abs/1103.0398. URL <http://arxiv.org/abs/1103.0398>.
- Comfere, N., O. Sokumbi, V. Montori, A. LeBlanc, L. Prokop, M. Murad, and J. Tilburt (2013). Provider-to-provider communication in dermatology and implications of missing clinical information in skin biopsy requisition forms: A systematic review. *International journal of dermatology*, 53.

- Comito, C., D. Falcone, and A. Forestiero (2022). Ai-driven clinical decision support: Enhancing disease diagnosis exploiting patients similarity. *IEEE Access*, 10, 6878–6888.
- Cornu, P., S. Phansalkar, D. L. Seger, I. Cho, S. Pontefract, A. Robertson, D. W. Bates, and S. P. Slight (2018). High-priority and low-priority drug–drug interactions in different international electronic health record systems: A comparative study. *International Journal of Medical Informatics*, 111, 165–171. URL <https://doi.org/10.1016/j.ijmedinf.2017.12.027>.
- Cristea, M., G. G. Noja, P. Stefea, and A. L. Sala (2020). The impact of population aging and public health support on EU labor markets. *International Journal of Environmental Research and Public Health*, 17(4), 1439. URL <https://doi.org/10.3390/ijerph17041439>.
- Croon, R. D., L. V. Houdt, N. N. Htun, G. Štiglic, V. V. Abeele, and K. Verbert (2021). Health recommender systems: Systematic review. *Journal of Medical Internet Research*, 23(6), e18035. URL <https://doi.org/10.2196/18035>.
- Dahl, F. A., T. Rama, P. Hurlen, P. H. Brekke, H. Husby, T. Gundersen, Ø. Nytrø, and L. Øvreid (2021). Neural classification of norwegian radiology reports: using NLP to detect findings in CT-scans of children. *BMC Medical Informatics and Decision Making*, 21(1). URL <https://doi.org/10.1186/s12911-021-01451-8>.
- Dash, S., S. K. Shakyawar, M. Sharma, and S. Kaushik (2019). Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*, 6(1). URL <https://doi.org/10.1186/s40537-019-0217-0>.
- Davis, M. F., S. Sriram, W. S. Bush, J. C. Denny, and J. L. Haines (2013). Automated extraction of clinical traits of multiple sclerosis in electronic medical records. *Journal of the American Medical Informatics Association*, 20(e2), e334–e340. URL <https://doi.org/10.1136/amiajnl-2013-001999>.
- Dean, E. R. and M. L. Scoggins (2012). Essential elements of patient positioning: A review for the radiology nurse. *Journal of Radiology Nursing*, 31(2), 42–52. URL <https://doi.org/10.1016/j.jradnu.2011.08.002>.
- Dean, N., B. Jones, J. Jones, J. Ferraro, H. Post, D. Aronsky, C. Vines, T. Allen, and P. Haug (2015). Impact of an electronic clinical decision support tool

- for emergency department patients with pneumonia. *Annals of emergency medicine*, 66.
- Demner-Fushman, D., M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald (2015). Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2), 304–310. URL <https://doi.org/10.1093/jamia/ocv080>.
- Demner-Fushman, D., M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. K. Antani, G. R. Thoma, and C. J. McDonald (2016). Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association : JAMIA*, 23 2, 304–10.
- den Broeck, J. V., S. A. Cunningham, R. Eeckels, and K. Herbst (2005). Data cleaning: Detecting, diagnosing, and editing data abnormalities. *PLoS Medicine*, 2(10), e267. URL <https://doi.org/10.1371/journal.pmed.0020267>.
- Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009.
- Devarakonda, M. and C.-H. Tsou, Automated problem list generation from electronic medical records in ibm watson. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15. AAAI Press, 2015. ISBN 0-262-51129-0. URL <http://dl.acm.org/citation.cfm?id=2888116.2888262>.
- Diaz, O., K. Kushibar, R. Osuala, A. Linardos, L. Garrucho, L. Igual, P. Radeva, F. Prior, P. Gkontra, and K. Lekadir (2021). Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools. *Physica Medica*, 83, 25–37. URL <https://doi.org/10.1016/j.ejmp.2021.02.007>.
- Diogo, V. S., H. A. Ferreira, and D. P. and (2022). Early diagnosis of alzheimer's disease using machine learning: a multi-diagnostic, generalizable approach. *Alzheimer's Research & Therapy*, 14(1). URL <https://doi.org/10.1186/s13195-022-01047-y>.
- Dramburg, S., M. M. Fernández, E. Potapova, and P. M. Matricardi (2020). The potential of clinical decision support systems for prevention, diagnosis,

- and monitoring of allergic diseases. *Frontiers in Immunology*, 11. URL <https://doi.org/10.3389/fimmu.2020.02116>.
- Dunnmon, J., D. Yi, C. Langlotz, C. Ré, D. Rubin, and M. Lungren (2018). Assessment of convolutional neural networks for automated classification of chest radiographs. *Radiology*, 290, 181422.
- Dutta, S., W. J. Long, D. F. Brown, and A. T. Reisner (2013). Automated detection using natural language processing of radiologists recommendations for additional imaging of incidental findings. *Annals of Emergency Medicine*, 62(2), 162–169. ISSN 0196-0644. URL <http://www.sciencedirect.com/science/article/pii/S0196064413001054>.
- Dvornik, N., J. Mairal, and C. Schmid (2019). On the importance of visual context for data augmentation in scene understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1.
- Embi, P. J., A. Jain, J. Clark, and C. M. Harris (2005). Development of an electronic health record-based clinical trial alert system to enhance recruitment at the point of care. *AMIA Annu. Symp. Proc.*, 231–235.
- Eslami, S., N. F. de Keizer, D. A. Dongelmans, E. de Jonge, M. J. Schultz, and A. Abu-Hanna (2012). Effects of two different levels of computerized decision support on blood glucose regulation in critically ill patients. *International Journal of Medical Informatics*, 81(1), 53–60. URL <https://doi.org/10.1016/j.ijmedinf.2011.10.004>.
- Evans, R. S. (2016). Electronic health records: Then, now, and in the future. *Yearbook of Medical Informatics*, 25(S 01), S48–S61. URL <https://doi.org/10.15265/iys-2016-s006>.
- Fang, G., Z. Huang, and Z. Wang (2022). Predicting ischemic stroke outcome using deep learning approaches. *Frontiers in Genetics*, 12. URL <https://doi.org/10.3389/fgene.2021.827522>.
- Faris, H., M. Habib, M. Faris, H. Elayan, and A. Alomari (2021). An intelligent multimodal medical diagnosis system based on patients’ medical questions and structured symptoms for telemedicine. *Informatics in Medicine Unlocked*, 23, 100513. URL <https://doi.org/10.1016/j.imu.2021.100513>.

- Fazli, S., S. Dahne, W. Samek, F. Bieszmann, and K.-R. Muller (2015). Learning from more than one data source: Data fusion techniques for sensorimotor rhythm-based brain–computer interfaces. *Proceedings of the IEEE*, 103(6), 891–906. URL <https://doi.org/10.1109/jproc.2015.2413993>.
- Felfernig, A. and B. Gula, An empirical study on consumer behavior in the interaction with knowledge-based recommender applications. In *The 8th IEEE International Conference on E-Commerce Technology and The 3rd IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services (CEC/EEE'06)*. 2006.
- Fiszman, M., W. W. Chapman, D. Aronsky, R. S. Evans, and P. J. Haug (2000). Automatic Detection of Acute Bacterial Pneumonia from Chest X-ray Reports. *Journal of the American Medical Informatics Association*, 7(6), 593–604. ISSN 1067-5027. URL <https://doi.org/10.1136/jamia.2000.0070593>.
- French, B., R. Boddepalli, and R. Govindarajan (2016). Acute ischemic stroke: Current status and future directions. *Missouri medicine*, 113, 480–486.
- Freund, Y. and R. E. Schapire, Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning, ICML'96*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1996. ISBN 1558604197.
- Frid-Adar, M., I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321, 321–331. URL <https://doi.org/10.1016/j.neucom.2018.09.013>.
- Fukui, A., D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. *CoRR*, abs/1606.01847. URL <http://arxiv.org/abs/1606.01847>.
- Gajbhiye, G., A. Nandedkar, and I. Faye, *Automatic Report Generation for Chest X-Ray Images: A Multilevel Multi-attention Approach*. 2020. ISBN 978-981-15-4014-1, 174–182.
- García, S., S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera (2016). Big data preprocessing: methods and prospects. *Big Data Analytics*, 1(1). URL <https://doi.org/10.1186/s41044-016-0014-0>.

- Gehrmann, S., F. Dernoncourt, Y. Li, E. T. Carlson, J. T. Wu, J. Welt, J. Foote, E. T. Moseley, D. W. Grant, P. D. Tyler, and L. A. Celi, Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *In PloS one*. 2018.
- Giełczyk, A., A. Marciniak, M. Tarczewska, and Z. Lutowski (2022). Pre-processing methods in chest x-ray image classification. *PLOS ONE*, 17(4), 1–11. URL <https://doi.org/10.1371/journal.pone.0265949>.
- Gold, R., C. Sheppler, D. Hessler, A. Bunce, E. Cottrell, N. Yosuf, M. Pisciotta, R. Gunn, M. Leo, and L. Gottlieb (2021). Using electronic health record-based clinical decision support to provide social risk-informed care in community health centers: Protocol for the design and assessment of a clinical decision support tool. *JMIR Research Protocols*, 10(10), e31733. URL <https://doi.org/10.2196/31733>.
- Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial networks. 2014a.
- Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014b). Generative adversarial networks. URL <https://arxiv.org/abs/1406.2661>.
- Gouda, W., M. Almurafeh, M. Humayun, and N. Z. Jhanjhi (2022). Detection of COVID-19 based on chest x-rays using deep learning. *Healthcare*, 10(2), 343. URL <https://doi.org/10.3390/healthcare10020343>.
- Goyal, M., J. M. Ospel, M. Kappelhof, and A. Ganesh (2021). Challenges of outcome prediction for acute stroke treatment decisions. *Stroke*, 52(5), 1921–1928. URL <https://doi.org/10.1161/strokeaha.120.033785>.
- Greenspan, H., B. van Ginneken, and R. M. Summers (2016). Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, 35(5), 1153–1159.
- Griner, D., R. Zhang, X. Tie, C. Zhang, J. W. Garrett, K. Li, and G.-H. Chen, COVID-19 pneumonia diagnosis using chest x-ray radiograph and deep learning. In M. A. Mazurowski and K. Drukker (eds.), *Medical Imaging 2021: Computer-Aided Diagnosis* volume11597. International Society for Optics and Photonics, SPIE, 2021. URL <https://doi.org/10.1117/12.2581972>.

- Gu, J., Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, 77, 354–377. ISSN 0031-3203. URL <https://www.sciencedirect.com/science/article/pii/S0031320317304120>.
- Güler, M. G. and E. Geçici (2020). A decision support system for scheduling the shifts of physicians during COVID-19 pandemic. *Computers & Industrial Engineering*, 150, 106874. URL <https://doi.org/10.1016/j.cie.2020.106874>.
- Gulshan, V., L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22), 2402. URL <https://doi.org/10.1001/jama.2016.17216>.
- Gundogdu, B., U. Pamuksuz, J. H. Chung, J. M. Telleria, P. Liu, F. Khan, and P. J. Chang (2021). Customized impression prediction from radiology reports using bert and lstms. *IEEE Transactions on Artificial Intelligence*, 1–1.
- Haberman, S., J. Feldman, Z. O. Merhi, G. Markenson, W. Cohen, and H. Minkoff (2009). Effect of clinical-decision support on documentation compliance in an electronic medical record. *Obstetrics & Gynecology*, 114(2), 311–317. URL <https://doi.org/10.1097/aog.0b013e3181af2cb0>.
- Hak, F., T. Guimarães, and M. Santos (2022). Towards effective clinical decision support systems: A systematic review. *PLOS ONE*, 17(8), e0272846. URL <https://doi.org/10.1371/journal.pone.0272846>.
- Hamidinekoo, A., T. Pieciak, M. Afzali, O. Akanyeti, and Y. Yuan (2021). Glioma classification using multimodal radiology and histology data, 508–518. URL [https://doi.org/10.1007/978-3-030-72087-2\\_45](https://doi.org/10.1007/978-3-030-72087-2_45).
- Hanna, T. N., M. E. Zygmunt, R. Peterson, D. Theriot, H. Shekhani, J.-O. Johnson, and E. A. Krupinski (2018). The effects of fatigue from overnight shifts on radiology search patterns and diagnostic performance. *Journal of the American College of Radiology*, 15(12), 1709–1716. URL <https://doi.org/10.1016/j.jacr.2017.12.019>.



- Hara, K., D. Saito, and H. Shouno, Analysis of function of rectified linear unit used in deep learning. *In 2015 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2015. URL <https://doi.org/10.1109/ijcnn.2015.7280578>.
- Hartung, M. P., I. C. Bickle, F. Gaillard, and J. P. Kanne (2020). How to create a great radiology report. *RadioGraphics*, 40(6), 1658–1670. URL <https://doi.org/10.1148/rg.2020200020>.
- Hashmi, M. F., S. Katiyar, A. G. Keskar, N. D. Bokde, and Z. W. Geem (2020). Efficient pneumonia detection in chest xray images using deep transfer learning. *Diagnostics*, 10(6), 417. URL <https://doi.org/10.3390/diagnostics10060417>.
- Hassanpour, S., G. Bay, and C. Langlotz (2017). Characterization of change and significance for clinical findings in radiology reports through natural language processing. *Journal of Digital Imaging*, 30.
- Hassanzadeh, T., D. Essam, and R. Sarker (2020). An evolutionary denseres deep convolutional neural network for medical image segmentation. *IEEE Access*, 8, 212298–212314.
- He, K., X. Zhang, S. Ren, and J. Sun (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385. URL <http://arxiv.org/abs/1512.03385>.
- He, K., X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition. 2016a.
- He, K., X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition. *In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016b.
- He, Z., W. Chen, Z. Li, M. Zhang, W. Zhang, and M. Zhang, See: Syntax-aware entity embedding for neural relation extraction. *In AAAI*. 2018.
- Hearst, M. A. (1998). Support vector machines. *IEEE Intelligent Systems*, 13(4), 18–28. ISSN 1541-1672. URL <https://doi.org/10.1109/5254.708428>.
- Heidarvinchek, F., R. McConville, C. Morgan, R. McNaney, A. Masullo, M. Mirmehdi, A. L. Whone, and I. Craddock (2021). Multimodal classification



- of parkinson's disease in home environments with resiliency to missing modalities. *Sensors*, 21(12), 4133. URL <https://doi.org/10.3390/s21124133>.
- Helal Uddin, M., M. N. Hossain, M. S. Islam, M. A. A. Zubaer, and S.-H. Yang (2022). Detecting covid-19 status using chest x-ray images and symptoms analysis by own developed mathematical model: A model development and analysis approach. *COVID*, 2(2), 117–137. ISSN 2673-8112. URL <https://www.mdpi.com/2673-8112/2/2/9>.
- Helmons, P. J., B. O. Suijkerbuijk, P. V. N. Panday, and J. G. Kosterink (2015). Drug-drug interaction checking assisted by clinical decision support: a return on investment analysis. *Journal of the American Medical Informatics Association*, 22(4), 764–772. URL <https://doi.org/10.1093/jamia/ocu010>.
- Henriksson, A., M. Kvist, H. Dalianis, and M. Duneld (2015). Identifying adverse drug event information in clinical notes with distributional semantic representations of context. *Journal of Biomedical Informatics*, 57, 333–349. URL <https://doi.org/10.1016/j.jbi.2015.08.013>.
- Hermessi, H., O. Mourali, and E. Zagrouba (2021). Multimodal medical image fusion review: Theoretical background and recent advances. *Signal Processing*, 183, 108036. URL <https://doi.org/10.1016/j.sigpro.2021.108036>.
- Hermier, M. and N. Nighoghossian (2004). Contribution of susceptibility-weighted imaging to acute stroke assessment. *Stroke*, 35(8), 1989–1994. URL <https://doi.org/10.1161/01.str.0000133341.74387.96>.
- Higgins, T. L., A. Deshpande, M. D. Zilberberg, P. K. Lindenauer, P. B. Imrey, P.-C. Yu, S. D. Haessler, S. S. Richter, and M. B. Rothberg (2020). Assessment of the accuracy of using ICD-9/i diagnosis codes to identify pneumonia etiology in patients hospitalized with pneumonia. *JAMA Network Open*, 3(7), e207750. URL <https://doi.org/10.1001/jamanetworkopen.2020.7750>.
- Hilmizen, N., A. Bustamam, and D. Sarwinda, The multimodal deep learning for diagnosing covid-19 pneumonia from chest ct-scan and x-ray images. In *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*. 2020a.
- Hilmizen, N., A. Bustamam, and D. Sarwinda, The multimodal deep learning for diagnosing COVID-19 pneumonia from chest CT-scan and x-ray images. In *2020 3rd International Seminar on Research of Information Technology and*

- Intelligent Systems (ISRITI)*. IEEE, 2020b. URL <https://doi.org/10.1109/isriti51436.2020.9315478>.
- Hinton, G. (2018). Deep Learning—A Technology With the Potential to Transform Health Care. *JAMA*, 320(11), 1101–1102. ISSN 0098-7484. URL <https://doi.org/10.1001/jama.2018.11100>.
- Hinton, G. E., N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2012). Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580. URL <http://arxiv.org/abs/1207.0580>.
- Holschneider, M., R. Kronland-Martinet, J. Morlet, and P. Tchamitchian (1989). A real-time algorithm for signal analysis with the help of the wavelet transform. *Wavelets, Time-Frequency Methods and Phase Space*, -1, 286.
- Hopper, K. D., C. J. Kasales, M. A. V. Slyke, T. A. Schwartz, T. R. TenHave, and J. A. Jozefiak (1996). Analysis of interobserver and intraobserver variability in CT tumor measurements. *American Journal of Roentgenology*, 167(4), 851–854. URL <https://doi.org/10.2214/ajr.167.4.8819370>.
- Hosmer, D. W. and S. Lemeshow, *Applied logistic regression*. John Wiley and Sons, 2000. ISBN 0471356328, 9780471356325.
- Hosny, A., C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. W. L. Aerts (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8), 500–510. URL <https://doi.org/10.1038/s41568-018-0016-5>.
- Howard, A. G., M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam (2017a). Mobilenets: Efficient convolutional neural networks for mobile vision applications. URL <https://arxiv.org/abs/1704.04861>.
- Howard, A. G., M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam (2017b). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861. URL <http://arxiv.org/abs/1704.04861>.
- Huang, G., Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017a.

- Huang, G., Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, Densely connected convolutional networks. *In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017b.
- Huang, G., Z. Liu, and K. Q. Weinberger (2016). Densely connected convolutional networks. *CoRR*, abs/1608.06993. URL <http://arxiv.org/abs/1608.06993>.
- Huang, S.-C., A. Pareek, R. Zamanian, I. Banerjee, and M. P. Lungren (2020). Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. *Scientific Reports*, 10(1). URL <https://doi.org/10.1038/s41598-020-78888-w>.
- Hussain, S., I. Mubeen, N. Ullah, S. S. U. D. Shah, B. A. Khan, M. Zahoor, R. Ullah, F. A. Khan, and M. A. Sultan (2022). Modern diagnostic imaging technique applications and risk factors in the medical field: A review. *BioMed Research International*, 2022, 1–19. URL <https://doi.org/10.1155/2022/5164970>.
- Hwang, E. J., S. Park, K.-N. Jin, J. I. Kim, S. Y. Choi, J. H. Lee, J. M. Goo, J. Aum, J.-J. Yim, C. M. Park, D. L.-B. A. D. A. Development, and E. Group (2018). Development and Validation of a Deep Learning-based Automatic Detection Algorithm for Active Pulmonary Tuberculosis on Chest Radiographs. *Clinical Infectious Diseases*, 69(5), 739–747. ISSN 1058-4838. URL <https://doi.org/10.1093/cid/ciy967>.
- Iakovidis, D. and C. Smailis (2012). A semantic model for multimodal data mining in healthcare information systems. *Studies in health technology and informatics*, 180, 574–8.
- Ioffe, S. and C. Szegedy (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167. URL <http://arxiv.org/abs/1502.03167>.
- Iqbal, S., A. N. Qureshi, J. Li, and T. Mahmood (2023). On the analyses of medical images using traditional machine learning techniques and convolutional neural networks. *Archives of Computational Methods in Engineering*, 30(5), 3173–3233. URL <https://doi.org/10.1007/s11831-023-09899-9>.
- Irvin, J., P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S.

- Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng (2019). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 590–597. URL <https://ojs.aaai.org/index.php/AAAI/article/view/3834>.
- Jaeger, S., S. Candemir, S. Antani, Y.-X. J. Wang, P.-X. Lu, and G. R. Thoma (2014). Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4 6, 475–7.
- Jia, T., C. Wang, Z. Tian, B. Wang, and F. Tian (2022). Design of digital and intelligent financial decision support system based on artificial intelligence. *Computational Intelligence and Neuroscience*, 2022, 1–7. URL <https://doi.org/10.1155/2022/1962937>.
- Jiang, B., G. Zhu, Y. Xie, J. Heit, H. Chen, Y. Li, V. Ding, A. Eskandari, P. Michel, G. Zaharchuk, and M. Wintermark (2021). Prediction of clinical outcome in patients with large-vessel acute ischemic stroke: Performance of machine learning versus SPAN-100. *American Journal of Neuroradiology*, 42(2), 240–246. URL <https://doi.org/10.3174/ajnr.a6918>.
- Jimmy, B. and J. Jose (2011). Patient medication adherence: Measures in daily practice. *Oman Medical Journal*, 26(3), 155–159. URL <https://doi.org/10.5001/omj.2011.38>.
- Jindal, R. and S. Taneja (2015). A lexical approach for text categorization of medical documents. *Procedia Computer Science*, 46, 314–320. ISSN 1877-0509. URL <http://www.sciencedirect.com/science/article/pii/S1877050915000903>. Proceedings of the International Conference on Information and Communication Technologies, ICICT 2014, 3-5 December 2014 at Bolgatty Palace Island Resort, Kochi, India.
- Jing, B., P. Xie, and E. Xing, On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018. URL <https://doi.org/10.18653/v1/p18-1240>.
- Jing, B., P. Xie, and E. P. Xing, On the automatic generation of medical imaging reports. In *ACL*. 2017.

- Johnson, A., T. Pollard, S. Berkowitz, N. Greenbaum, M. Lungren, C.-y. Deng, R. Mark, and S. Horng (2019a). MIMIC-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6, 317.
- Johnson, A. E. W., T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. Deng, R. G. Mark, and S. Horng (2019b). MIMIC-CXR: A large publicly available database of labeled chest radiographs. *CoRR*, abs/1901.07042. URL <http://arxiv.org/abs/1901.07042>.
- Johnson, E., W. C. Baughman, and G. Ozsoyoglu, Mixing domain rules with machine learning for radiology text classification. 2014.
- Jonas, J. B., T. Aung, R. R. Bourne, A. M. Bron, R. Ritch, and S. Panda-Jonas (2017). Glaucoma. *The Lancet*, 390(10108), 2183–2193. ISSN 0140-6736. URL [https://doi.org/10.1016/S0140-6736\(17\)31469-1](https://doi.org/10.1016/S0140-6736(17)31469-1).
- Joseph, N., S. Bernadin, D. Hodges, and P. Sekhar, A case study on using unstructured data analysis methods to identify local covid-19 hotspots. In *SoutheastCon 2021*. IEEE, 2021. URL <https://doi.org/10.1109/southeastcon45413.2021.9401921>.
- Joshi, A., A. Papanastassiou, K. P. Vives, D. D. Spencer, L. H. Staib, and X. Papademetris, Light-sensitive visualization of multimodal data for neurosurgical applications. In *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, 2010. URL <https://doi.org/10.1109/isbi.2010.5490128>.
- Jung, E., M. Luna, and S. H. Park, Conditional gan with an attention-based generator and a 3d discriminator for 3d medical image generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI*. Springer-Verlag, Berlin, Heidelberg, 2021. ISBN 978-3-030-87230-4.
- Kabir, Y., M. Dojat, B. Scherrer, F. Forbes, and C. Garbay, Multimodal mri segmentation of ischemic stroke lesions. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2007. ISSN 1094-687X.
- Kang, D.-W., S.-I. Sohn, K.-S. Hong, K.-H. Yu, Y.-H. Hwang, M.-K. Han, J. Lee, J.-M. Park, A.-H. Cho, H.-J. Kim, D.-E. Kim, Y.-J. Cho, J. Koo, S.-C. Yun,

- S. U. Kwon, H.-J. Bae, and J. S. Kim (2012). Reperfusion therapy in unclear-onset stroke based on MRI evaluation (RESTORE). *Stroke*, 43(12), 3278–3283. URL <https://doi.org/10.1161/strokeaha.112.675926>.
- Kelly, B. S., C. Judge, S. M. Bollard, S. M. Clifford, G. M. Healy, A. Aziz, P. Mathur, S. Islam, K. W. Yeom, A. Lawlor, and R. P. Killeen (2022). Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). *European Radiology*, 32(11), 7998–8007. URL <https://doi.org/10.1007/s00330-022-08784-6>.
- Kermary, D. S., K. Zhang, and M. H. Goldbaum, Labeled optical coherence tomography (oct) and chest x-ray images for classification. 2018.
- Keselman, A. and C. A. Smith (2012). A classification of errors in lay comprehension of medical documents. *Journal of Biomedical Informatics*, 45(6), 1151–1163. URL <https://doi.org/10.1016/j.jbi.2012.07.012>.
- Khan, M. I. and A. Banerji (2014). Health care management in india: Some issues and challenges. *Journal of Health Management*, 16(1), 133–147. URL <https://doi.org/10.1177/0972063413518690>.
- Kharazmi, P., S. Kalia, H. Lui, Z. J. Wang, and T. K. Lee (2017). A feature fusion system for basal cell carcinoma detection through data-driven feature learning and patient profile. *Skin Research and Technology*, 24(2), 256–264. URL <https://doi.org/10.1111/srt.12422>.
- Kieu, S. T. H., A. Bade, M. H. A. Hijazi, and H. Kolivand (2020). A survey of deep learning for lung disease detection on medical images: State-of-the-art, taxonomy, issues and future directions. *Journal of Imaging*, 6(12), 131. URL <https://doi.org/10.3390/jimaging6120131>.
- Kim, D.-Y., K.-H. Choi, J.-H. Kim, J. Hong, S.-M. Choi, M.-S. Park, and K.-H. Cho (2023). Deep learning-based personalised outcome prediction after acute ischaemic stroke. *Journal of Neurology, Neurosurgery & Psychiatry*, jnnp-2022-330230. URL <https://doi.org/10.1136/jnnp-2022-330230>.
- Kingma, D. P. and J. Ba (2014). Adam: A method for stochastic optimization. URL <https://arxiv.org/abs/1412.6980>.
- Kingma, D. P. and J. Ba, Adam: A method for stochastic optimization. *In ICLR (Poster)*. 2015.

- Kline, A., H. Wang, Y. Li, S. Dennis, M. Hutch, Z. Xu, F. Wang, F. Cheng, and Y. Luo (2022). Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(1). URL <https://doi.org/10.1038/s41746-022-00712-8>.
- Koutkias, V. and J. B. and (2018). Contributions from the 2017 literature on clinical decision support. *Yearbook of Medical Informatics*, 27(01), 122–128. URL <https://doi.org/10.1055/s-0038-1641222>.
- Kreimeyer, K., M. Foster, A. Pandey, N. Arya, G. Halford, S. F. Jones, R. Forshee, M. Walderhaug, and T. Botsis (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *Journal of Biomedical Informatics*, 73, 14–29. URL <https://doi.org/10.1016/j.jbi.2017.07.012>.
- Krizhevsky, A., I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems* volume 25. Curran Associates, Inc., 2012a. URL <https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>.
- Krizhevsky, A., I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems* volume 25. Curran Associates, Inc., 2012b. URL <https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>.
- Krupinski, E. A. (2010). Current perspectives in medical image perception. *Attention, Perception & Psychophysics*, 72(5), 1205–1217. URL <https://doi.org/10.3758/app.72.5.1205>.
- Krupinski, E. A., E. T. Hall, S. Jaw, B. Reiner, and E. Siegel (2011). Influence of radiology report format on reading time and comprehension. *Journal of Digital Imaging*, 25(1), 63–69. URL <https://doi.org/10.1007/s10278-011-9424-8>.
- Kumar, E. and P. Jayadev, *Deep Learning for Clinical Decision Support Systems: A Review from the Panorama of Smart Healthcare*. 2020. ISBN 978-3-030-33965-4, 79–99.



- Kunhimangalam, R., S. Ovallath, and P. K. Joseph (2014). A clinical decision support system with an integrated EMR for diagnosis of peripheral neuropathy. *Journal of Medical Systems*, 38(4). URL <https://doi.org/10.1007/s10916-014-0038-9>.
- Kurniawan, H. and M. Pechenizkiy, Towards the stress analytics framework: Managing, mining, and visualizing multi-modal data for stress awareness. *In 2014 IEEE 27th International Symposium on Computer-Based Medical Systems*. IEEE, 2014. URL <https://doi.org/10.1109/cbms.2014.129>.
- Kusakunniran, W., S. Karnjanapreechakorn, T. Siriapisith, P. Borwarnginn, K. Sutassananon, T. Tongdee, and P. Saiviroonporn (2021). COVID-19 detection and heatmap generation in chest x-ray images. *Journal of Medical Imaging*, 8(S1). URL <https://doi.org/10.1117/1.JMI.8.S1.014001>.
- Kwok, R., M. Dinh, D. Dinh, and M. Chu (2009). Improving adherence to asthma clinical guidelines and discharge documentation from emergency departments: Implementation of a dynamic and integrated electronic decision support system. *Emergency Medicine Australasia*, 21(1), 31–37. URL <https://doi.org/10.1111/j.1742-6723.2008.01149.x>.
- Lahat, D., T. Adali, and C. Jutten (2015). Multimodal data fusion: An overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9), 1449–1477. URL <https://doi.org/10.1109/jproc.2015.2460697>.
- Lao, Q., T. Fevens, and B. Wang, Leveraging disease progression learning for medical image recognition. *In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2018.
- Lee, H., E.-J. Lee, S. Ham, H.-B. Lee, J. S. Lee, S. U. Kwon, J. S. Kim, N. Kim, and D.-W. Kang (2020). Machine learning approach to identify stroke within 4.5 hours. *Stroke*, 51(3), 860–866. URL <https://doi.org/10.1161/strokeaha.119.027611>.
- Lee, M. S., Y. S. Kim, M. Kim, M. Usman, S. S. Byon, S. H. Kim, B. I. Lee, and B.-D. Lee (2021). Evaluation of the feasibility of explainable computer-aided detection of cardiomegaly on chest radiographs using deep learning. *Scientific Reports*, 11(1). URL <https://doi.org/10.1038/s41598-021-96433-1>.



- Lependu, P., S. V. Iyer, A. Bauer-Mehren, R. Harpaz, Y. T. Ghebremariam, J. P. Cooke, and N. H. Shah (2013). Pharmacovigilance using clinical text. *AMIA Summits Transl. Sci. Proc.*, 2013, 109.
- Leslie, A., A. Jones, and P. Goddard (2000). The influence of clinical information on the reporting of ct by radiologists. *The British journal of radiology*, 73, 1052–5.
- Levin, D., U. Aladl, G. Germano, and P. Slomka (2005). Techniques for efficient, real-time, 3d visualization of multi-modality cardiac data using consumer graphics hardware. *Computerized Medical Imaging and Graphics*, 29(6), 463–475. URL <https://doi.org/10.1016/j.compmedimag.2005.02.007>.
- Li, D., Z. Liu, L. Luo, S. Tian, and J. Zhao (2022). Prediction of pulmonary fibrosis based on x-rays by deep neural network. *Journal of Healthcare Engineering*, 2022, 1–13. URL <https://doi.org/10.1155/2022/3845008>.
- Li, H. and Y. Fan, Early prediction of alzheimer’s disease dementia based on baseline hippocampal MRI and 1-year follow-up cognitive measures using deep recurrent neural networks. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019. URL <https://doi.org/10.1109/isbi.2019.8759397>.
- Li, Y., J. Zhao, Z. Lv, and Z. Pan (2021). Multimodal medical supervised image fusion method by CNN. *Frontiers in Neuroscience*, 15. URL <https://doi.org/10.3389/fnins.2021.638976>.
- Liang, S., D. Beaton, S. R. Arnott, T. Gee, M. Zamyadi, R. Bartha, S. Symons, G. M. MacQueen, S. Hassel, J. P. Lerch, E. Anagnostou, R. W. Lam, B. N. Frey, R. Milev, D. J. Müller, S. H. Kennedy, and C. J. M. S. and (2021). Magnetic resonance imaging sequence identification using a metadata learning approach. *Frontiers in Neuroinformatics*, 15. URL <https://doi.org/10.3389/fninf.2021.622951>.
- Lin, M., Q. Chen, and S. Yan (2014). Network in network. *CoRR*, abs/1312.4400.
- Lin, S., Z. Han, D. Li, J. Zeng, X. Yang, X. Liu, and F. Liu (2020). Integrating model- and data-driven methods for synchronous adaptive multi-band image fusion. *Information Fusion*, 54, 145–160. URL <https://doi.org/10.1016/j.inffus.2019.07.009>.

- Lipton, J. A., R. J. Barendse, A. F. Schinkel, K. M. Akkerhuis, M. L. Simoons, and E. J. Sijbrands (2011). Impact of an alerting clinical decision support system for glucose control on protocol compliance and glycemic control in the intensive cardiac care unit. *Diabetes Technology & Therapeutics*, 13(3), 343–349. URL <https://doi.org/10.1089/dia.2010.0100>.
- Litjens, G., T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88. ISSN 1361-8415. URL <https://www.sciencedirect.com/science/article/pii/S1361841517301135>.
- Liu, F., C. You, X. Wu, S. Ge, S. Wang, and X. Sun (2021). Auto-encoding knowledge graph for unsupervised medical report generation. *CoRR*, abs/2111.04318. URL <https://arxiv.org/abs/2111.04318>.
- Liu, G., T. H. Hsu, M. B. A. McDermott, W. Boag, W. Weng, P. Szolovits, and M. Ghassemi (2019a). Clinically accurate chest x-ray report generation. *CoRR*, abs/1904.02633. URL <http://arxiv.org/abs/1904.02633>.
- Liu, G., T. M. Hsu, M. McDermott, W. Boag, W.-H. Weng, P. Szolovits, and M. Ghassemi (2019b). Clinically accurate chest x-ray report generation.
- Liu, S. and W. Deng, Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. 2015.
- Lopez, K., S. J. Fodeh, A. Allam, C. A. Brandt, and M. Krauthammer (2020). Reducing annotation burden through multimodal learning. *Frontiers in Big Data*, 3, 19. ISSN 2624-909X. URL <https://www.frontiersin.org/article/10.3389/fdata.2020.00019>.
- Mackintosh, N., M. Terblanche, R. Maharaj, A. Xyrichis, K. Franklin, J. Keddie, E. Larkins, A. Maslen, J. Skinner, S. Newman, J. H. D. S. Magalhaes, and J. Sandall (2016). Telemedicine with clinical decision support for critical care: a systematic review. *Systematic Reviews*, 5(1). URL <https://doi.org/10.1186/s13643-016-0357-7>.
- Madai, V. I., I. Galinovic, U. Grittner, O. Zaro-Weber, A. Schneider, S. Z. Martin, F. C. v. Samson-Himmelstjerna, K. L. Stengl, M. A. Mutke, W. Moeller-Hartmann, M. Ebinger, J. B. Fiebach, and J. Sobesky (2014). DWI intensity

- values predict FLAIR lesions in acute ischemic stroke. *PLoS ONE*, 9(3), e92295. URL <https://doi.org/10.1371/journal.pone.0092295>.
- Mahbub, M., S. Srinivasan, I. Danciu, A. Peluso, E. Begoli, S. Tamang, and G. D. Peterson (2022). Unstructured clinical notes within the 24 hours since admission predict short, mid & long-term mortality in adult ICU patients. *PLOS ONE*, 17(1), e0262182. URL <https://doi.org/10.1371/journal.pone.0262182>.
- Majkowska, A., S. Mittal, D. F. Steiner, J. J. Reicher, S. M. McKinney, G. E. Duggan, K. Eswaran, P.-H. Cameron Chen, Y. Liu, S. R. Kalidindi, A. Ding, G. S. Corrado, D. Tse, and S. Shetty (2020). Chest radiograph interpretation with deep learning models: Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*, 294(2), 421–431. ISSN 0033-8419. URL <https://doi.org/10.1148/radiol.2019191293>.
- Manssour, I., S. Furuie, L. Nedel, and C. Freitas, A multimodal visualization framework for medical data. In *Proceedings 13th Brazilian Symposium on Computer Graphics and Image Processing (Cat. No.PR00878)*. IEEE Comput. Soc, 2000. URL <https://doi.org/10.1109/sibgra.2000.895844>.
- McCague, C., S. Ramlee, M. Reinius, I. Selby, D. Hulse, P. Piyatissa, V. Bura, M. Crispin-Ortuzar, E. Sala, and R. Woitek (2023). Introduction to radiomics for a clinical audience. *Clinical Radiology*, 78(2), 83–98. ISSN 0009-9260. URL <https://www.sciencedirect.com/science/article/pii/S000992602200705X>. Special Issue Section: Artificial Intelligence and Machine Learning.
- McDonald, R. J., K. M. Schwartz, L. J. Eckel, F. E. Diehn, C. H. Hunt, B. J. Bartholmai, B. J. Erickson, and D. F. Kallmes (2015). The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Academic Radiology*, 22(9), 1191–1198. ISSN 1076-6332. URL <https://www.sciencedirect.com/science/article/pii/S1076633215002457>.
- McEvoy, D., T. K. Gandhi, A. Turchin, and A. Wright (2018). Enhancing problem list documentation in electronic health records using two methods: the example of prior splenectomy. *BMJ Quality & Safety*, 27(1), 40–47. ISSN 2044-5415. URL <https://qualitysafety.bmj.com/content/27/1/40>.

- McEvoy, D. S., D. F. Sittig, T.-T. Hickman, S. Aaron, A. Ai, M. Amato, D. W. Bauer, G. M. Fraser, J. Harper, A. Kennemer, M. A. Krall, C. U. Lehmann, S. Malhotra, D. R. Murphy, B. O’Kelley, L. Samal, R. Schreiber, H. Singh, E. J. Thomas, C. V. Vartian, J. Westmorland, A. B. McCoy, and A. Wright (2016). Variation in high-priority drug-drug interaction alerts across institutions and electronic health records. *Journal of the American Medical Informatics Association*, 24(2), 331–338. URL <https://doi.org/10.1093/jamia/ocw114>.
- McPhail, S. (2016). Multimorbidity in chronic disease: impact on health care resources and costs. *Risk Management and Healthcare Policy*, Volume 9, 143–156. URL <https://doi.org/10.2147/rmhp.s97248>.
- Meystre, S. M., P. M. Heider, Y. Kim, D. B. Aruch, and C. D. Britten (2019). Automatic trial eligibility surveillance based on unstructured clinical data. *International Journal of Medical Informatics*, 129, 13–19. URL <https://doi.org/10.1016/j.ijmedinf.2019.05.018>.
- Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013a). Efficient estimation of word representations in vector space. URL <https://arxiv.org/abs/1301.3781>.
- Mikolov, T., I. Sutskever, K. Chen, G. Corrado, and J. Dean (2013b). Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546. URL <http://arxiv.org/abs/1310.4546>.
- Miles, K. (2011). Can imaging help improve the survival of cancer patients? *Cancer Imaging*, 11(1A), S86–S92. URL <https://doi.org/10.1102/1470-7330.2011.9022>.
- Moghadam, S. T., F. Sadoughi, F. Velayati, S. J. Ehsanzadeh, and S. Poursharif (2021). The effects of clinical decision support system for prescribing medication on patient outcomes and physician practice performance: a systematic review and meta-analysis. *BMC Medical Informatics and Decision Making*, 21(1). URL <https://doi.org/10.1186/s12911-020-01376-8>.
- Mohd, T. K., J. Carvalho, and A. Y. Javaid, Multi-modal data fusion of voice and EMG data for robotic control. In *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*. IEEE, 2017. URL <https://doi.org/10.1109/uemcon.2017.8249063>.

- Mondal, A. K., J. Dolz, and C. Desrosiers (2018). Few-shot 3d multi-modal medical image segmentation using generative adversarial learning. URL <https://arxiv.org/abs/1810.12241>.
- Mucherino, A., P. J. Papajorgji, and P. M. Pardalos, *k-Nearest Neighbor Classification*. Springer New York, New York, NY, 2009. ISBN 978-0-387-88615-2, 83–106. URL [https://doi.org/10.1007/978-0-387-88615-2\\_4](https://doi.org/10.1007/978-0-387-88615-2_4).
- Mujtaba, G., L. Shuib, N. Idris, W. L. Hoo, R. G. Raj, K. Khowaja, K. Shaikh, and H. F. Nweke (2019). Clinical text classification research trends: Systematic literature review and open issues. *Expert Systems with Applications*, 116, 494–520. ISSN 0957-4174. URL <http://www.sciencedirect.com/science/article/pii/S0957417418306110>.
- Nakamura, Y., S. Hanaoka, Y. Nomura, T. Nakao, S. Miki, T. Watadani, T. Yoshikawa, N. Hayashi, and O. Abe (2021). Automatic detection of actionable radiology reports using bidirectional encoder representations from transformers. *BMC Medical Informatics and Decision Making*, 21(1). URL <https://doi.org/10.1186/s12911-021-01623-6>.
- Narayan, S. (1997). The generalized sigmoid activation function: Competitive supervised learning. *Information Sciences*, 99(1), 69–82. ISSN 0020-0255. URL <https://www.sciencedirect.com/science/article/pii/S0020025596002009>.
- Nasir, N., A. Kansal, F. Barneih, O. Al-Shaltone, T. Bonny, M. Al-Shabi, and A. A. Shammaa (2023). Multi-modal image classification of COVID-19 cases using computed tomography and x-rays scans. *Intelligent Systems with Applications*, 17, 200160. URL <https://doi.org/10.1016/j.iswa.2022.200160>.
- Nazari-Farsani, S., Y. Yu, R. Duarte Armindo, M. Lansberg, D. S. Liebeskind, G. Albers, S. Christensen, C. S. Levin, and G. Zaharchuk (2023). Predicting final ischemic stroke lesions from initial diffusion-weighted images using a deep neural network. *NeuroImage: Clinical*, 37, 103278. ISSN 2213-1582. URL <https://www.sciencedirect.com/science/article/pii/S2213158222003436>.
- Nedjah, N., I. Santos, and L. Mourelle (2019). Sentiment analysis using convolutional neural network via word embeddings. *Evolutionary Intelligence*.

- Nguyen, H. T. N., D. Nie, T. Badamdorj, Y. Liu, Y. Zhu, J. Truong, and L. Cheng (2021). Automated generation of accurate & fluent medical x-ray reports. *CoRR*, abs/2108.12126. URL <https://arxiv.org/abs/2108.12126>.
- Nicolson, A., J. Dowling, and B. Koopman (2022). Improving chest x-ray report generation by leveraging warm-starting. *CoRR*, abs/2201.09405. URL <https://arxiv.org/abs/2201.09405>.
- Nie, D., H. Zhang, E. Adeli, L. Liu, and D. Shen, 3d deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients. In S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*. Springer International Publishing, Cham, 2016. ISBN 978-3-319-46723-8.
- Nischitha and N. B. Padmavathi, Fusion of multimodal abdominal cancerous images and classification using support vector machine. In *2017 International Conference on Intelligent Sustainable Systems (ICISS)*. IEEE, 2017. URL <https://doi.org/10.1109/iss1.2017.8389411>.
- Nunes, N., Deep learning for automatic classification of multi-modal information corresponding to chest radiology reports. 2019.
- Nunes, N., B. Martins, N. A. da Silva, F. Leite, and M. J. Silva, A multi-modal deep learning method for classifying chest radiology exams. In *Progress in Artificial Intelligence*. Springer International Publishing, 2019, 323–335. URL [https://doi.org/10.1007/978-3-030-30241-2\\_28](https://doi.org/10.1007/978-3-030-30241-2_28).
- Obuchowicz, R., M. Oszust, and A. Piorkowski (2020). Interobserver variability in quality assessment of magnetic resonance images. *BMC Medical Imaging*, 20(1). URL <https://doi.org/10.1186/s12880-020-00505-z>.
- Ogbole, G., C. Okorie, M. Owolabi, O. Ogun, A. Adeyinka, and A. Ogunniyi (2015). Role of diffusion-weighted imaging in acute stroke management using low-field magnetic resonance imaging in resource-limited settings. *West African Journal of Radiology*, 22(2), 61. URL <https://doi.org/10.4103/1115-3474.162168>.
- ONC (2022). Office-based physician electronic health record adoption. URL <https://dashboard.healthit.gov/quickstats/pages/physician-ehr-adoption-trends.php>.

- Ouahab, A. (2021). Multimodal convolutional neural networks for detection of covid-19 using chest x-ray and CT images. *Optical Memory and Neural Networks*, 30(4), 276–283. URL <https://doi.org/10.3103/s1060992x21040044>.
- Ovbiagele, B. and M. N. Nguyen-Huynh (2011). Stroke epidemiology: Advancing our understanding of disease mechanism and therapy. *Neurotherapeutics*, 8(3), 319–329. URL <https://doi.org/10.1007/s13311-011-0053-1>.
- Ozkara, B. B., M. Karabacak, O. Hamam, R. Wang, A. Kotha, N. Khalili, M. Hoesinyazdi, M. M. Chen, M. Wintermark, and V. S. Yedavalli (2023). Prediction of functional outcome in stroke patients with proximal middle cerebral artery occlusions using machine learning models. *Journal of Clinical Medicine*, 12(3). ISSN 2077-0383. URL <https://www.mdpi.com/2077-0383/12/3/839>.
- Pandey, M., Z. Xu, E. Sholle, G. Maliakal, G. Singh, Z. Fatima, D. Larine, B. C. Lee, J. Wang, A. R. van Rosendael, L. Baskaran, L. J. Shaw, J. K. Min, and S. J. Al’Aref (2020). Extraction of radiographic findings from unstructured thoracoabdominal computed tomography reports using convolutional neural network based natural language processing. *PLOS ONE*, 15(7), e0236827. URL <https://doi.org/10.1371/journal.pone.0236827>.
- Pandeya, Y. R. and J. Lee (2021). Deep learning-based late fusion of multimodal information for emotion classification of music video. *Multimedia Tools and Applications*, 80(2), 2887–2905. ISSN 1573-7721. URL <https://doi.org/10.1007/s11042-020-08836-3>.
- Pandya, M. D., P. D. Shah, and S. Jardosh, Medical image diagnosis for disease detection: A deep learning approach. In *U-Healthcare Monitoring Systems*. Elsevier, 2019, 37–60. URL <https://doi.org/10.1016/b978-0-12-815370-3.00003-7>.
- Parmar, C., J. D. Barry, A. Hosny, J. Quackenbush, and H. J. Aerts (2018). Data analysis strategies in medical imaging. *Clinical Cancer Research*, 24(15), 3492–3499. URL <https://doi.org/10.1158/1078-0432.ccr-18-0385>.
- Pasa, F., V. Golkov, F. Pfeiffer, D. Cremers, and D. Pfeiffer (2019). Efficient deep network architectures for fast chest x-ray tuberculosis screening and visualization. *Scientific Reports*, 9(1). URL <https://doi.org/10.1038/s41598-019-42557-4>.



- Pennington, J., R. Socher, and C. Manning, Glove: Global vectors for word representation. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 2014. URL <https://www.aclweb.org/anthology/D14-1162>.
- Person, M., M. Jensen, A. Smith, and H. Gutierrez (2019). Multimodal fusion object detection system for autonomous vehicles. *Journal of Dynamic Systems, Measurement, and Control*, 141.
- Phansalkar, S., A. A. Desai, D. Bell, E. Yoshida, J. Doole, M. Czochanski, B. Middleton, and D. W. Bates (2012). High-priority drug–drug interactions for use in electronic health records. *Journal of the American Medical Informatics Association*, 19(5), 735–743. URL <https://doi.org/10.1136/amiajnl-2011-000612>.
- Pichardo-Lowden, A. R., P. Haidet, G. E. Umpierrez, E. B. Lehman, F. T. Quigley, L. Wang, C. M. Rafferty, C. J. DeFlicht, and V. M. Chinchilli (2022). Clinical decision support for glycemic management reduces hospital length of stay. *Diabetes Care*, 45(11), 2526–2534. URL <https://doi.org/10.2337/dc21-0829>.
- Pincay, J., L. Terán, and E. Portmann, Health recommender systems: A state-of-the-art review. *In 2019 Sixth International Conference on eDemocracy eGovernment (ICEDEG)*. 2019.
- Pinho, E. and C. Costa, Extensible architecture for multimodal information retrieval in medical imaging archives. *In 2016 12th International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*. 2016.
- Polikar, R., C. Tilley, B. Hillis, and C. M. Clark, Multimodal EEG, MRI and PET data fusion for alzheimer's disease diagnosis. *In 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 2010. URL <https://doi.org/10.1109/iembs.2010.5627621>.
- Polnaszek, B., A. Gilmore-Bykovskiy, M. Hovanes, R. Roiland, P. Ferguson, R. Brown, and A. J. Kind (2016). Overcoming the challenges of unstructured data in multisite, electronic medical record-based abstraction. *Medical Care*, 54(10), e65–e72. URL <https://doi.org/10.1097/mlr.000000000000108>.
- Pons, E., L. M. M. Braun, M. G. M. Hunink, and J. A. Kors (2016). Natural language processing in radiology: A systematic review. *Radiology*, 279(2), 329–343. URL <https://doi.org/10.1148/radiol.16142770>.



- Pool, F. and S. Goergen (2010). Quality of the written radiology report: A review of the literature. *Journal of the American College of Radiology*, 7(8), 634–643. URL <https://doi.org/10.1016/j.jacr.2010.03.016>.
- Powers, W. J., A. A. Rabinstein, T. Ackerson, O. M. Adeoye, N. C. Bambakidis, K. Becker, J. Biller, M. Brown, B. M. Demaerschalk, B. Hoh, E. C. Jauch, C. S. Kidwell, T. M. Leslie-Mazwi, B. Ovbiagele, P. A. Scott, K. N. Sheth, A. M. Southerland, D. V. Summers, and D. L. Tirschwell (2018). 2018 guidelines for the early management of patients with acute ischemic stroke: A guideline for healthcare professionals from the american heart association/american stroke association. *Stroke*, 49(3). URL <https://doi.org/10.1161/str.000000000000158>.
- Pruszydlo, M. G., S. U. Walk-Fritz, T. Hoppe-Tichy, J. Kaltschmidt, and W. E. Haefeli (2012). Development and evaluation of a computerised clinical decision support system for switching drugs at the interface between primary and tertiary care. *BMC Medical Informatics and Decision Making*, 12(1). URL <https://doi.org/10.1186/1472-6947-12-137>.
- Purushotham, S., C. Meng, Z. Che, and Y. Liu (2018). Benchmarking deep learning models on large healthcare datasets. *Journal of Biomedical Informatics*, 83, 112–134. URL <https://doi.org/10.1016/j.jbi.2018.04.007>.
- Purwar, S., R. K. Tripathi, R. Ranjan, and R. Saxena (2019). Detection of microcytic hypochromia using cbc and blood film features extracted from convolution neural network by different classifiers. *Multimedia Tools and Applications*, 79(7-8), 4573–4595. URL <https://doi.org/10.1007/s11042-019-07927-0>.
- Purwar, S., R. K. Tripathi, R. Ranjan, and R. Saxena (2020). Detection of microcytic hypochromia using cbc and blood film features extracted from convolution neural network by different classifiers. *Multimedia Tools and Applications*, 79(7), 4573–4595. ISSN 1573-7721. URL <https://doi.org/10.1007/s11042-019-07927-0>.
- Puttagunta, M. and S. Ravi (2021). Medical image analysis based on deep learning approach. *Multimedia Tools and Applications*. URL <https://doi.org/10.1007/s11042-021-10707-4>.
- Qiu, S., G. H. Chang, M. Panagia, D. M. Gopal, R. Au, and V. B. Kolachalama (2018). Fusion of deep learning models of MRI scans, mini-mental state exam-

- ination, and logical memory test enhances diagnosis of mild cognitive impairment. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10(1), 737–749. URL <https://doi.org/10.1016/j.dadm.2018.08.013>.
- Qiu, W., H. Kuang, J. M. Ospel, M. D. Hill, A. M. Demchuk, M. Goyal, and B. K. Menon (2021). Automated prediction of ischemic brain tissue fate from multiphase computed tomographic angiography in patients with acute ischemic stroke using machine learning. *Journal of Stroke*, 23(2), 234–243. URL <https://doi.org/10.5853/jos.2020.05064>.
- Radford, A., L. Metz, and S. Chintala (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. URL <https://arxiv.org/abs/1511.06434>.
- Radford, A., L. Metz, and S. Chintala (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434.
- Rahaman, S., M. M. Islam, and M. S. Hossain, A belief rule based clinical decision support system framework. In *2014 17th International Conference on Computer and Information Technology (ICCIT)*. 2014.
- Rajkomar, A., S. Lingam, A. G. Taylor, M. Blum, and J. Mongan (2016). High-throughput classification of radiographs using deep convolutional neural networks. *Journal of Digital Imaging*, 30(1), 95–101. URL <https://doi.org/10.1007/s10278-016-9914-9>.
- Rajpurkar, P., J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz, B. N. Patel, K. W. Yeom, K. Shpanskaya, F. G. Blankenberg, J. Seekins, T. J. Amrhein, D. A. Mong, S. S. Halabi, E. J. Zucker, A. Y. Ng, and M. P. Lungren (2018). Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLOS Medicine*, 15(11), 1–17. URL <https://doi.org/10.1371/journal.pmed.1002686>.
- Rajpurkar, P., J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Y. Ding, A. Bagul, C. P. Langlotz, K. S. Shpanskaya, M. P. Lungren, and A. Y. Ng (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *CoRR*, abs/1711.05225. URL <http://arxiv.org/abs/1711.05225>.

- Ramachandram, D. and G. W. Taylor (2017). Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34, 96–108.
- Ramesh, T. and V. Santhi (2020). Exploring big data analytics in health care. *International Journal of Intelligent Networks*, 1, 135–140. ISSN 2666-6030. URL <https://www.sciencedirect.com/science/article/pii/S2666603020300154>.
- Ramirez-Alonso, G., O. Prieto-Ordaz, R. López-Santillan, and M. Montes-Y-Gómez (2022). Medical report generation through radiology images: An overview. *IEEE Latin America Transactions*, 20(6), 986–999.
- Rana, M. and M. Bhushan (2022). Machine learning and deep learning approach for medical image analysis: diagnosis to detection. *Multimedia Tools and Applications*, 82(17), 26731–26769. URL <https://doi.org/10.1007/s11042-022-14305-w>.
- Reda, I., A. Khalil, M. Elmogy, A. Aboelfetouh, A. Shalaby, M. Abou-El-Ghar, A. Elmaghraby, M. Ghazal, and A. El-Baz (2018). Deep learning role in early diagnosis of prostate cancer. *Technology in Cancer Research Treatment*, 17, 153303461877553.
- Rehman, A., S. Naz, and I. Razzak (2021). Leveraging big data analytics in health-care enhancement: trends, challenges and opportunities. *Multimedia Systems*, 28(4), 1339–1371. URL <https://doi.org/10.1007/s00530-020-00736-8>.
- Ren, S., K. He, R. Girshick, and J. Sun (2017). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), 1137–1149.
- Ribeiro, R. T., R. T. Marinho, and J. M. Sanches (2013). Classification and staging of chronic liver disease from multimodal data. *IEEE Transactions on Biomedical Engineering*, 60(5), 1336–1344. URL <https://doi.org/10.1109/tbme.2012.2235438>.
- Rouse, M. (2018). Hitech (health information technology for economic and clinical health) act of 2009'. URL <https://searchhealthit.techtarget.com/definition/HITECH-Act>.
- Roweis, S. T. and L. K. Saul (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500), 2323–2326. ISSN 0036-8075. URL <https://science.sciencemag.org/content/290/5500/2323>.

- Ruder, S. (2016). An overview of gradient descent optimization algorithms. *CoRR*, abs/1609.04747. URL <http://arxiv.org/abs/1609.04747>.
- Ryoo, S. and H. J. Kim (2014). Activities of the korean institute of tuberculosis. *Osong Public Health and Research Perspectives*, 5, S43–S49. URL <https://doi.org/10.1016/j.phrp.2014.10.007>.
- Saeedi, S., S. Rezayi, H. Keshavarz, and S. R. N. Kalhori (2023). MRI-based brain tumor detection using convolutional deep learning methods and chosen machine learning techniques. *BMC Medical Informatics and Decision Making*, 23(1). URL <https://doi.org/10.1186/s12911-023-02114-6>.
- Sagheer, S. V. M. and S. N. George (2020). A review on medical image denoising algorithms. *Biomedical Signal Processing and Control*, 61, 102036. URL <https://doi.org/10.1016/j.bspc.2020.102036>.
- Sahoo, A. K., C. Pradhan, R. K. Barik, and H. Dubey (2019). DeepReco: Deep learning based health recommender system using collaborative filtering. *Computation*, 7(2), 25. URL <https://doi.org/10.3390/computation7020025>.
- Sai, R., S. B. K, and B K Tripathy (2021). Automated medical report generation on chest x-ray images using co-attention mechanism. URL <http://rgdoi.net/10.13140/RG.2.2.26061.15843>.
- Salem, H. A., G. Caddeo, J. McFarlane, K. Patel, L. Cochrane, D. Soria, M. Henley, and J. Lund (2018). A multicentre integration of a computer-led follow-up of prostate cancer is valid and safe. *BJU International*, 122(3), 418–426. URL <https://doi.org/10.1111/bju.14157>.
- Sammut, C. and G. I. Webb (eds.), *TF-IDF*. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8, 986–987.
- Sandfort, V., K. Yan, P. J. Pickhardt, and R. M. Summers (2019). Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Scientific Reports*, 9(1). URL <https://doi.org/10.1038/s41598-019-52737-x>.
- Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3). URL <https://doi.org/10.1007/s42979-021-00592-x>.

- Saxena, A., S. Singh Tomar, G. Jain, and R. Gupta, Deep learning based diagnosis of diseases using image classification. In *2021 11th International Conference on Cloud Computing, Data Science Engineering (Confluence)*. 2021.
- Schaut, M., M. Schaefer, U. Trost, and A. Sander (2022). Integrated antibiotic clinical decision support system (CDSS) for appropriate choice and dosage: an analysis of retrospective data. *Germs*, 12(2), 203–213. URL <https://doi.org/10.18683/germs.2022.1323>.
- Schmidhuber, J. (2014). Deep learning in neural networks: An overview. *CoRR*, abs/1404.7828. URL <http://arxiv.org/abs/1404.7828>.
- Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017.
- Selvaraju, R. R., A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra (2016). Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391. URL <http://arxiv.org/abs/1610.02391>.
- Shachor, Y., H. Greenspan, and J. Goldberger (2020). A mixture of views network with applications to multi-view medical imaging. *Neurocomputing*, 374, 1–9. URL <https://doi.org/10.1016/j.neucom.2019.09.027>.
- Sharma, A., R. Kumar, and V. Jaiswal, Classification of heart disease from mri images using convolutional neural network. In *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*. 2021.
- Shetty, S., A. V. S, and A. Mahale (2022). Comprehensive review of multimodal medical data analysis: Open issues and future research directions. *Acta Informatica Pragensia*, 11(3), 423–457. URL <https://doi.org/10.18267/j.aip.202>.
- Shi, T., H. Jiang, and B. Zheng (2022). C2ma-net: Cross-modal cross-attention network for acute ischemic stroke lesion segmentation based on ct perfusion scans. *IEEE Transactions on Biomedical Engineering*, 69(1), 108–118.
- Shin, B., F. H. Chokshi, T. Lee, and J. D. Choi (2017). Classification of radiology reports using neural attention models. *CoRR*, abs/1708.06828. URL <http://arxiv.org/abs/1708.06828>.

- Shortliffe, E. H. and B. G. Buchanan (1975). A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23(3), 351–379. ISSN 0025-5564. URL <https://www.sciencedirect.com/science/article/pii/0025556475900474>.
- Siddique, M. A. B., S. Sakib, and M. A. Rahman, Performance analysis of deep autoencoder and nca dimensionality reduction techniques with knn, enn and svm classifiers. In *2019 2nd International Conference on Innovation in Engineering and Technology (ICIET)*. 2019.
- Silva, S. T., F. Hak, and J. Machado (2022). Rule-based clinical decision support system using the OpenEHR standard. *Procedia Computer Science*, 201, 726–731. URL <https://doi.org/10.1016/j.procs.2022.03.098>.
- Simonyan, K. and A. Zisserman, Very deep convolutional networks for large-scale image recognition. In Y. Bengio and Y. LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. 2015a. URL <http://arxiv.org/abs/1409.1556>.
- Simonyan, K. and A. Zisserman, Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*. 2015b.
- Sinaga, A. S. R. M. and R. E. Putra, Predictive analytic healthcare sector using classification machine learning algorithm. In *2022 International Symposium on Information Technology and Digital Innovation (ISITDI)*. IEEE, 2022. URL <https://doi.org/10.1109/isitdi55734.2022.9944492>.
- Singh, A. and R. Parida (2022). Decision-making models for healthcare supply chain disruptions: Review and insights for post-pandemic era. *International Journal of Global Business and Competitiveness*, 17(2), 130–141. URL <https://doi.org/10.1007/s42943-021-00045-5>.
- Singh, H., G. D. Schiff, M. L. Graber, I. Onakpoya, and M. J. Thompson (2016). The global burden of diagnostic errors in primary care. *BMJ Quality & Safety*, 26(6), 484–494. URL <https://doi.org/10.1136/bmjqs-2016-005401>.
- Sippo, D., G. Warden, K. Andriole, R. Lacson, I. Ikuta, R. Birdwell, and R. Khorasani (2013). Automated extraction of bi-rads final assessment categories from

- radiology reports with natural language processing. *Journal of digital imaging*, 26.
- Sirshar, M., M. F. K. Paracha, M. U. Akram, N. S. Alghamdi, S. Z. Y. Zaidi, and T. Fatima (2022). Attention based automated radiology report generation using CNN and LSTM. *PLOS ONE*, 17(1), e0262209. URL <https://doi.org/10.1371/journal.pone.0262209>.
- Sittig, D. F., A. Wright, and B. Middleton (2016). Clinical decision support: a 25 year retrospective and a 25 year vision. *Yearbook of Medical Informatics*, 25(S 01), S103–S116. URL <https://doi.org/10.15265/iys-2016-s034>.
- Sivic, J. and A. Zisserman (2009). Efficient visual search of videos cast as text retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4), 591–606. URL <https://doi.org/10.1109/tpami.2008.111>.
- Soenksen, L. R., Y. Ma, C. Zeng, L. Boussioux, K. V. Carballo, L. Na, H. M. Wiberg, M. L. Li, I. Fuentes, and D. Bertsimas (2022). Integrated multi-modal artificial intelligence framework for healthcare applications. *npj Digital Medicine*, 5(1). URL <https://doi.org/10.1038/s41746-022-00689-4>.
- Song, J., J. Zheng, P. Li, X. Lu, G. Zhu, and P. Shen (2021). An effective multimodal image fusion method using MRI and PET for alzheimer's disease diagnosis. *Frontiers in Digital Health*, 3. URL <https://doi.org/10.3389/fdgth.2021.637386>.
- Spasic, I. and G. Nenadic (2020). Clinical text data in machine learning: Systematic review. *JMIR Medical Informatics*, 8(3), e17984. URL <https://doi.org/10.2196/17984>.
- Spasić, I., Özlem Uzuner, and L. Zhou (2020). Emerging clinical applications of text analytics. *International Journal of Medical Informatics*, 134, 103974. URL <https://doi.org/10.1016/j.ijmedinf.2019.103974>.
- Spasov, S. E., L. Passamonti, A. Duggento, P. Lio, and N. Toschi, A multi-modal convolutional neural network framework for the prediction of alzheimer's disease. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018. URL <https://doi.org/10.1109/embc.2018.8512468>.



- Sreejith, S., H. Khanna Nehemiah, and A. Kannan (2022). A clinical decision support system for polycystic ovarian syndrome using red deer algorithm and random forest classifier. *Healthcare Analytics*, 2, 100102. ISSN 2772-4425. URL <https://www.sciencedirect.com/science/article/pii/S2772442522000442>.
- Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56), 1929–1958. URL <http://jmlr.org/papers/v15/srivastava14a.html>.
- Stivaros, S. M., A. Gledson, G. Nenadic, X.-J. Zeng, J. Keane, and A. Jackson (2010). Decision support systems for clinical radiological practice — towards the next generation. *The British Journal of Radiology*, 83(995), 904–914. URL <https://doi.org/10.1259/bjr/33620087>.
- Sturmberg, J. P. and J. Bircher (2019). Better and fulfilling healthcare at lower costs: The need to manage health systems as complex adaptive systems. *F1000Research*, 8, 789. URL <https://doi.org/10.12688/f1000research.19414.1>.
- Sun, L., J. Chen, Y. Xu, M. Gong, K. Yu, and K. Batmanghelich (2022). Hierarchical amortized GAN for 3d high resolution medical image synthesis. *IEEE Journal of Biomedical and Health Informatics*, 26(8), 3966–3975. URL <https://doi.org/10.1109/jbhi.2022.3172976>.
- Sundararaman, A., S. V. Ramanathan, and R. Thati (2018). Novel approach to predict hospital readmissions using feature selection from unstructured data with class imbalance. *Big Data Research*, 13, 65–75. URL <https://doi.org/10.1016/j.bdr.2018.05.004>.
- Sutton, R. T., D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine*, 3(1). URL <https://doi.org/10.1038/s41746-020-0221-y>.
- Suzuki, K. and Y. Chen (eds.), *Artificial Intelligence in Decision Support Systems for Diagnosis in Medical Imaging*. Springer International Publishing, 2018. URL <https://doi.org/10.1007/978-3-319-68843-5>.



- Syeda-Mahmood, T., F. Wang, D. Beymer, A. Amir, M. Richmond, and S. Hashmi, AALIM: Multimodal mining for cardiac decision support. *In 2007 Computers in Cardiology*. IEEE, 2007. URL <https://doi.org/10.1109/cic.2007.4745458>.
- Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich (2014). Going deeper with convolutions. *CoRR*, abs/1409.4842. URL <http://arxiv.org/abs/1409.4842>.
- Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, Rethinking the inception architecture for computer vision. *In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016a.
- Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, Rethinking the inception architecture for computer vision. *In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016b.
- Tajbakhsh, N., J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, 35(5), 1299–1312.
- Tan, M. and Q. V. Le (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. URL <https://arxiv.org/abs/1905.11946>.
- Tan, M. and Q. V. Le, Efficientnetv2: Smaller models and faster training. *In ICMLvolume139 of Proceedings of Machine Learning Research*. PMLR, 2021.
- Tan, W., P. Tiwari, H. M. Pandey, C. Moreira, and A. K. Jaiswal (2020). Multi-modal medical image fusion algorithm in the era of big data. *Neural Computing and Applications*. URL <https://doi.org/10.1007/s00521-020-05173-2>.
- Tang, Y.-X., Y.-B. Tang, Y. Peng, K. Yan, M. Bagheri, B. Redd, C. Brandon, Z. lu, M. Han, J. Xiao, and R. Summers (2020). Automated abnormality classification of chest radiographs using deep convolutional neural networks. *npj Digital Medicine*, 3.
- Tayefi, M., P. Ngo, T. Chomutare, H. Dalianis, E. Salvi, A. Budrionis, and F. Godtlielsen (2021). Challenges and opportunities beyond structured data in analysis of electronic health records. *WIREs Computational Statistics*, 13(6). URL <https://doi.org/10.1002/wics.1549>.

- Thinaharan, N., V. Thiagarasu, and and (2019). A rule based clinical decision support system for healthcare industry. *Indian Journal of Science and Technology*, 12(12), 1–10. URL <https://doi.org/10.17485/ijst/2019/v12i12/142291>.
- Torres-Velázquez, M., W.-J. Chen, X. Li, and A. B. McMillan (2021). Application and construction of deep learning networks in medical imaging. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 5(2), 137–159.
- Tran, T. N. T., M. Atas, A. Felfernig, and M. Stettinger (2017). An overview of recommender systems in the healthy food domain. *Journal of Intelligent Information Systems*, 50(3), 501–526. URL <https://doi.org/10.1007/s10844-017-0469-0>.
- Tran, T. N. T., A. Felfernig, C. Trattner, and A. Holzinger (2020). Recommender systems in the healthcare domain: state-of-the-art and research issues. *Journal of Intelligent Information Systems*, 57(1), 171–201. URL <https://doi.org/10.1007/s10844-020-00633-6>.
- Trivedi, H., J. Mesterhazy, B. Laguna, T. Vu, and J. Sohn (2017). Automatic determination of the need for intravenous contrast in musculoskeletal mri examinations using ibm watson’s natural language processing algorithm. *Journal of Digital Imaging*, 31.
- Trzcinski, T., Multimodal social media video classification with deep neural networks. In R. S. Romaniuk and M. Linczuk (eds.), *Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2018* volume10808. International Society for Optics and Photonics, SPIE, 2018. URL <https://doi.org/10.1117/12.2501679>.
- van der Putten, N., R. Vinke, M. Citroen, R. Cornet, E. van Mulligen, and A. den Boer, Integrating medical images, biosignals, and alphanumeric data in a cardiological department. In *Computers in Cardiology 1995*. 1995.
- van Timmeren, J. E., D. Cester, S. Tanadini-Lang, H. Alkadhi, and B. Baessler (2020). Radiomics in medical imaging—“how-to” guide and critical reflection. *Insights into Imaging*, 11(1). URL <https://doi.org/10.1186/s13244-020-00887-2>.
- Varoquaux, G. and V. Cheplygina (2022). Machine learning for medical imaging: methodological failures and recommendations for the future. *npj Digital Medicine*, 5(1). URL <https://doi.org/10.1038/s41746-022-00592-y>.

- Vesdapunt, N. and N. Covavisaruch (2018). Automatic stroke lesions segmentation in diffusion-weighted MRI. *CoRR*, abs/1803.10385. URL <http://arxiv.org/abs/1803.10385>.
- Vinod, S., M. Naveen, A. K. Patra, and A. A. R. John, Accelerating towards larger deep learning models and datasets – a system platform view point. *In 2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. 2020.
- Vonbach, P., A. Dubied, S. Krähenbühl, and J. H. Beer (2008). Prevalence of drug–drug interactions at hospital entry and during hospital stay of patients in internal medicine. *European Journal of Internal Medicine*, 19(6), 413–420. URL <https://doi.org/10.1016/j.ejim.2007.12.002>.
- Voorhees, E. M. (2013). The trec medical records track, 239–246. URL <https://doi.org/10.1145/2506583.2506624>.
- Wang, L., X. Chen, L. Zhang, L. Li, Y. Huang, Y. Sun, and X. Yuan (2023). Artificial intelligence in clinical decision support systems for oncology. *International Journal of Medical Sciences*, 20(1), 79–86. URL <https://doi.org/10.7150/ijms.77205>.
- Wang, P., P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. W. Cottrell (2017a). Understanding convolution for semantic segmentation. *CoRR*, abs/1702.08502. URL <http://arxiv.org/abs/1702.08502>.
- Wang, Q., J. Xu, H. Chen, and B. He, Two improved continuous bag-of-word models. *In 2017 International Joint Conference on Neural Networks (IJCNN)*. 2017b.
- Wang, X., Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017c. URL <https://doi.org/10.1109/cvpr.2017.369>.
- Wang, X., Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, *ChestX-ray: Hospital-Scale Chest X-ray Database and Benchmarks on Weakly Supervised Classification and Localization of Common Thorax Diseases*. Springer International Publishing, Cham, 2019. ISBN 978-3-030-13969-8, 369–392. URL [https://doi.org/10.1007/978-3-030-13969-8\\_18](https://doi.org/10.1007/978-3-030-13969-8_18).

- Wang, X., Y. Peng, L. Lu, Z. Lu, and R. M. Summers (2018a). Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays.
- Wang, X., Y. Wang, C. Gao, K. Lin, and Y. Li (2018b). Automatic diagnosis with efficient medical case searching based on evolving graphs. *IEEE Access*, 6, 53307–53318. URL <https://doi.org/10.1109/access.2018.2871769>.
- Weibel, N., S. Ashfaq, A. Calvitti, J. D. Hollan, and Z. Agha, Multimodal data analysis and visualization to study the usage of electronic health records. *In 2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*. 2013.
- Werdiger, F., A. Bivard, and M. Parsons (2022). Artificial intelligence in acute ischemic stroke, 1503–1518. URL [https://doi.org/10.1007/978-3-030-64573-1\\_287](https://doi.org/10.1007/978-3-030-64573-1_287).
- Willeminck, M. J., W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin, and M. P. Lungren (2020). Preparing medical imaging data for machine learning. *Radiology*, 295(1), 4–15. URL <https://doi.org/10.1148/radiol.2020192224>.
- Williams, S., H. L. Horsfall, J. P. Funnell, J. G. Hanrahan, D. Z. Khan, W. Muirhead, D. Stoyanov, and H. J. Marcus (2021). Artificial intelligence in brain tumour surgery—an emerging paradigm. *Cancers*, 13(19), 5010. URL <https://doi.org/10.3390/cancers13195010>.
- Winder, A. J., M. Wilms, K. Amador, F. Flottmann, J. Fiehler, and N. D. Forkert (2022). Predicting the tissue outcome of acute ischemic stroke from acute 4d computed tomography perfusion imaging using temporal features and deep learning. *Frontiers in Neuroscience*, 16. URL <https://doi.org/10.3389/fnins.2022.1009654>.
- Wissler, L., M. Almashraee, D. Monett, and A. Paschke (2014). The gold standard in corpus annotation. *IEEE GSC*. URL <http://rgdoi.net/10.13140/2.1.4316.3523>.
- Wong, C., M. Peters, J. Tilburt, and N. Comfere (2015). Dermatopathologists’ Opinions About the Quality of Clinical Information in the Skin Biopsy Requisition Form and the Skin Biopsy Care Process: A Semiquantitative Assessment.

- American Journal of Clinical Pathology*, 143(4), 593–597. ISSN 0002-9173. URL <https://doi.org/10.1309/AJCPHPG6DQFBKKUR>.
- Wood, C., M. Cross, and P. La, Multimodal information retrieval, extraction and generation for use in the health domain. *In 1998 Second International Conference. Knowledge-Based Intelligent Electronic Systems. Proceedings KES'98 (Cat. No.98EX111)* volume3. 1998.
- Woolen, S. A., C. J. Kim, A. M. Hernandez, A. Becker, A. J. Martin, E. Kuoy, W. C. Pevec, and S. Tutton (2023). Radiology environmental impact: What is known and how can we improve? *Academic Radiology*, 30(4), 625–630. URL <https://doi.org/10.1016/j.acra.2022.10.021>.
- Wu, M., X. Zhong, Q. Peng, M. Xu, S. Huang, J. Yuan, J. Ma, and T. Tan (2019). Prediction of molecular subtypes of breast cancer using BI-RADS features based on a “white box” machine learning approach in a multi-modal imaging setting. *European Journal of Radiology*, 114, 175–184. URL <https://doi.org/10.1016/j.ejrad.2019.03.015>.
- Xi, X., H. Xu, H. Shi, C. Zhang, H. Y. Ding, G. Zhang, Y. Tang, and Y. Yin (2017). Robust texture analysis of multi-modal images using local structure preserving ranklet and multi-task learning for breast tumor diagnosis. *Neurocomputing*, 259, 210–218. URL <https://doi.org/10.1016/j.neucom.2016.06.082>.
- Xu, B., R. Huang, and M. Li (2016). Revise saturated activation functions. *CoRR*, abs/1602.05980. URL <http://arxiv.org/abs/1602.05980>.
- Xu, B., N. Wang, T. Chen, and M. Li (2015). Empirical evaluation of rectified activations in convolutional network. *CoRR*, abs/1505.00853. URL <http://arxiv.org/abs/1505.00853>.
- Xue, Y., T. Xu, L. Rodney Long, Z. Xue, S. Antani, G. R. Thoma, and X. Huang, Multimodal recurrent model with attention for automated radiology report generation. *In A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger (eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Springer International Publishing, Cham, 2018. ISBN 978-3-030-00928-1.
- Yala, A., C. Lehman, T. Schuster, T. Portnoi, and R. Barzilay (2019). A deep learning mammography-based model for improved breast cancer risk predic-

- tion. *Radiology*, 292(1), 60–66. URL <https://doi.org/10.1148/radiol.2019182716>.
- Yamashita, R., M. Nishio, R. K. G. Do, and K. Togashi (2018). Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 9(4), 611–629. ISSN 1869-4101. URL <https://doi.org/10.1007/s13244-018-0639-9>.
- Yang, P., G. Bi, J. Qi, X. Wang, Y. Yang, and L. Xu (2021). Multimodal wearable intelligence for dementia care in healthcare 4.0: a survey. *Information Systems Frontiers*. URL <https://doi.org/10.1007/s10796-021-10163-3>.
- Yang, S., X. Wu, S. Ge, S. K. Zhou, and L. Xiao (2022). Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical Image Analysis*, 80, 102510. URL <https://doi.org/10.1016/j.media.2022.102510>.
- Yao, Y., L. Rosasco, and A. Caponnetto (2007). On early stopping in gradient descent learning. *Constructive Approximation*, 26(2), 289–315. ISSN 1432-0940. URL <https://doi.org/10.1007/s00365-006-0663-2>.
- Yasaka, K. and O. Abe (2018). Deep learning and artificial intelligence in radiology: Current applications and future directions. *PLOS Medicine*, 15(11), 1–4. URL <https://doi.org/10.1371/journal.pmed.1002707>.
- Yi, S., G. Zhang, C. Qian, Y. Lu, H. Zhong, and J. He (2022). A multimodal classification architecture for the severity diagnosis of glaucoma based on deep learning. *Frontiers in Neuroscience*, 16. URL <https://doi.org/10.3389/fnins.2022.939472>.
- Yoo, Y., L. Y. W. Tang, D. K. B. Li, L. Metz, S. Kolind, A. L. Traboulsee, and R. C. Tam (2017). Deep learning of brain lesion patterns and user-defined clinical and MRI features for predicting conversion to multiple sclerosis from clinically isolated syndrome. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 7(3), 250–259. URL <https://doi.org/10.1080/21681163.2017.1356750>.
- Yoon, H. J., Y. J. Jeong, H. Kang, J. E. Jeong, and D.-Y. Kang (2019). Medical image analysis using artificial intelligence. *Progress in Medical Physics*, 30(2), 49. URL <https://doi.org/10.14316/pmp.2019.30.2.49>.

- Young, I. T. and L. J. van Vliet (1995). Recursive implementation of the gaussian filter. *Signal Process.*, 44(2), 139–151. URL <http://dblp.uni-trier.de/db/journals/sigpro/sigpro44.html#YoungV95>.
- Yu, F. and V. Koltun (2016). Multi-scale context aggregation by dilated convolutions.
- Yu, Y., Y. Xie, T. Thamm, E. Gong, J. Ouyang, S. Christensen, M. Marks, M. Lansberg, G. Albers, and G. Zaharchuk (2021). Tissue at risk and ischemic core estimation using deep learning in acute stroke. *American Journal of Neuroradiology*, 42(6), 1030–1037. URL <https://doi.org/10.3174/ajnr.a7081>.
- Yuan, J., H. Liao, R. Luo, and J. Luo (2019). Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. URL <https://arxiv.org/abs/1907.09085>.
- Zech, J. R., M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11), e1002683. URL <https://doi.org/10.1371/journal.pmed.1002683>.
- Zeng, Y., C. Long, W. Zhao, and J. Liu (2022). Predicting the severity of neurological impairment caused by ischemic stroke using deep learning based on diffusion-weighted images. *Journal of Clinical Medicine*, 11(14). ISSN 2077-0383. URL <https://www.mdpi.com/2077-0383/11/14/4008>.
- Zhang, D., F. Ren, Y. Li, L. Na, and Y. Ma (2021). Pneumonia detection from chest x-ray images based on convolutional neural network. *Electronics*, 10(13). ISSN 2079-9292. URL <https://www.mdpi.com/2079-9292/10/13/1512>.
- Zhang, D. and D. Shen, Semi-supervised multimodal classification of alzheimer's disease. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, 2011. URL <https://doi.org/10.1109/isbi.2011.5872715>.
- Zhang, L., M. Wang, M. Liu, and D. Zhang (2020). A survey on deep learning for neuroimaging-based brain disorder analysis. *Frontiers in Neuroscience*, 14. URL <https://doi.org/10.3389/fnins.2020.00779>.
- Zhang, L., Y. Wang, Z. Peng, Y. Weng, Z. Fang, F. Xiao, C. Zhang, Z. Fan, K. Huang, Y. Zhu, W. Jiang, J. Shen, and R. Zhan (2022). The progress



- of multimodal imaging combination and subregion based radiomics research of cancers. *International Journal of Biological Sciences*, 18(8), 3458–3469. URL <https://doi.org/10.7150/ijbs.71046>.
- Zhang, R., L. Zhao, W. Lou, J. M. Abrigo, V. C. T. Mok, W. C. W. Chu, D. Wang, and L. Shi (2018a). Automatic segmentation of acute ischemic stroke from dwi using 3-d fully convolutional densenets. *IEEE Transactions on Medical Imaging*, 37(9), 2149–2160.
- Zhang, Y., D. Y. Ding, T. Qian, C. D. Manning, and C. P. Langlotz (2018b). Learning to summarize radiology findings. *CoRR*, abs/1809.04698. URL <http://arxiv.org/abs/1809.04698>.
- Zhao, B., Z. Liu, G. Liu, C. Cao, S. Jin, H. Wu, and S. Ding (2021). Deep learning-based acute ischemic stroke lesion segmentation method on multimodal MR images using a few fully labeled subjects. *Computational and Mathematical Methods in Medicine*, 2021, 1–13. URL <https://doi.org/10.1155/2021/3628179>.
- Zhou, X., Y. Li, and W. Liang (2021). Cnn-rnn based intelligent recommendation for online medical pre-diagnosis support. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(3), 912–921.
- Zhu, H., L. Jiang, H. Zhang, L. Luo, Y. Chen, and Y. Chen (2021). An automatic machine learning approach for ischemic stroke onset time identification based on dwi and flair imaging. *NeuroImage: Clinical*, 31, 102744.
- Zou, X.-L., Y. Ren, D.-Y. Feng, X.-Q. He, Y.-F. Guo, H.-L. Yang, X. Li, J. Fang, Q. Li, J.-J. Ye, L.-Q. Han, and T.-T. Zhang (2020). A promising approach for screening pulmonary hypertension based on frontal chest radiographs using deep learning: A retrospective study. *PLOS ONE*, 15(7), 1–13. URL <https://doi.org/10.1371/journal.pone.0236378>.



## Bio-data

**Name:** Shashank

**Current Address:** Research Scholar,  
Department of Information Technology,  
NITK Surathkal  
Mangaluru, Karnataka  
India - 575025.

**Permanent Address:** S/O Chandrashekar Shetty  
3-17J(3A), "Chandrama"  
Kaneerthota, Koteka Post  
Mangalore, Karnataka  
India - 575022.

**Email:** shashankshetty06@gmail.com

**Mobile No:** +91-8197903771

**Qualification:** Ph.D. in Information Technology  
Department of Information Technology  
National Institute of Technology Karnataka, Surathkal  
Mangaluru, Karnataka  
India - 575025.

M.Tech in Computer Science & Engineering  
NMAM Institute of Technology, Nitte, Karkala,  
Udupi, Karnataka  
India - 574110.

B.Tech in Computer Science & Engineering  
Srinivas Institute of Technology, Valachil  
Mangalore, Karnataka  
India - 574143.

**Research Area:** Machine Learning, Deep Learning, Healthcare Analytics,  
Multimodal Medical Data Analytics, Cloud Computing, Web Technology