Synthesizing Global Context and Local Precision for Pathological Image Analysis of Diverse Imaging Modalities

Ph.D. Thesis

By

Chandravardhan Singh Raghaw



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

May 2026

Synthesizing Global Context and Local Precision for Pathological Image Analysis of Diverse Imaging Modalities

A THESIS

submitted to the

INDIAN INSTITUTE OF TECHNOLOGY INDORE

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

DOCTOR OF PHILOSOPHY

by

Chandravardhan Singh Raghaw



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

May 2026

INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled **Synthesizing Global Context and Local Precision for Pathological Image Analysis of Diverse Imaging Modalities** in the partial fulfillment of the requirements for the award of the degree of **Doctor of Philosophy** and submitted in the **Department of Computer Science and Engineering, Indian Institute of Technology Indore,** is an authentic record of my own work carried out during the time period from July 2022 to May 2026 under the supervision of Dr. Nagendra Kumar, Assistant Professor, Indian Institute of Technology Indore, Indore, India.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

25/Aug/26

Signature of the Student with Date

(Chandravardhan Singh Raghaw)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Nagendra Kumar 25/08/2026

Signature of Thesis Supervisor with Date

(Dr. Nagendra Kumar)

ACKNOWLEDGEMENT

I would like to take this opportunity to express my heartfelt gratitude to a number of persons who in one or the other way contributed by making this time as learnable, enjoyable, and bearable. At first, I would like to thank my supervisor **Dr. Nagendra Kumar** who was a constant source of inspiration during my work. Without his constant guidance and research directions, this research work could not be completed. His continuous support and encouragement has motivated me to remain streamlined in my research work.

I am thankful to **Dr. Ranveer Singh** and **Dr. Mohd. Arshad**, my research committee members for taking out some valuable time to evaluate my progress all these years. Their good comments and suggestions helped me to improve my work at various stages. I am also grateful to **Dr. Ranveer Singh**, HOD of Computer Science and Engineering for his help and support.

My sincere acknowledgement and respect to **Prof. Suhas Joshi**, Director, Indian Institute of Technology Indore for providing me the opportunity to explore my research capabilities at Indian Institute of Technology Indore.

I would like to appreciate the fine company of my dearest colleagues, labmates, friends, undergraduate students who have also supported me in my research work. I am also grateful to the institute staff for their unfailing support and assistance.

I would like to express my heartfelt respect to my parents for their love, care and support they have provided to me throughout my life.

Finally, I am thankful to all who directly or indirectly contributed, helped, and supported me.

Chandravardhan Singh Raghaw

Dedicated to My Family

ABSTRACT

Modern clinical practice relies on non-invasive imaging to support accurate diagnosis and treatment monitoring. Techniques such as chest X-rays (CXR), magnetic resonance imaging (MRI), computed tomography (CT), whole slide imaging (WSI), and colonoscopy images allow clinicians to examine internal body structures without the usual risks associated with surgical procedures. The wider use of digital imaging has also increased the volume of data that must be reviewed each day. Interpreting these images requires time, attention, and domain expertise. In many clinical settings, high workload can affect interpretation quality, and this issue is worsened by the shortage of trained radiologists in many regions today. Hospitals worldwide remain understaffed.

Deep learning has improved automated systems' ability to analyze medical images for tasks such as classification and segmentation. These methods learn relevant patterns directly from data and have shown good performance in controlled settings. However, the reliability in routine clinical settings remains limited. Medical images often differ in quality and appearance across patient variability and imaging devices. Variations in contrast, noise, and acquisition conditions reduce the model performance. In many cases, diseased and healthy regions look similar, which makes accurate prediction difficult. Another challenge is the lack of large, well-annotated datasets. Expert labeling is time-consuming, introduces inconsistencies, and is further complicated by class imbalance, which affects model training. In addition, many models focus on fine details and the overall structure separately, rather than combining both.

This thesis addresses these limitations by developing methods that improve feature learning, data efficiency, and localization accuracy across different imaging modalities. First, a global and local feature aggregation approach is proposed using transformer and convolutional architectures. This design captures both structural context and detailed patterns to improve classification results in tasks such as leukemia detection and pneumonia analysis. Second, data-efficient learning frameworks are introduced to reduce dependency on large labeled datasets. By leveraging semi-supervised and self-supervised learning strategies, these methods make better use of available data and improve robustness in limited-annotation settings. Third, boundary-aware mechanisms are developed to improve segmentation performance where lesion boundaries are unclear or indistinguishable. These methods refine predictions to capture the extent of pathological regions in detail. Fourth, multi-scale feature

fusion techniques are designed to handle large variations in lesion size and appearance by integrating information across different resolutions.

In addition, the proposed methods include interpretability components that highlight relevant regions in the image and help clinicians understand model predictions. The approaches are evaluated on multiple clinical tasks, including leukemia classification, pneumonia detection, knee osteoarthritis severity grading, liver and tumor segmentation, and polyp segmentation. Results show the improved performance and consistent behavior across diverse datasets.

Overall, this thesis presents a structured approach to the pathological image analysis that focuses on improving diagnosis accuracy, handling limited data, and capturing both local and global information. The proposed methods are designed to support practical computer-aided diagnosis systems that can be applied in real clinical settings.

Keywords: Medical Image Analysis, Computer-Aided Diagnosis, Deep Learning, Multi-Modal Imaging, Global–Local Feature Learning, Medical Image Segmentation, Data-Efficient Learning

List of Publications

In Refereed Journals

Published / Accepted:

1. **C.S. Raghaw**, A. Sharma, S. Bansal, M.Z.U. Rehman, and N. Kumar. *CoTCoNet: An optimized coupled transformer-convolutional network with an adaptive graph reconstruction for leukemia detection*, Computers in Biology and Medicine, vol. 179, p. 108821, 2024 (Elsevier), DOI = "https://doi.org/10.1016/j.compbiomed.2024.108821".

2. **C.S. Raghaw**, P.S. Bhore, M.Z.U. Rehman, and N. Kumar. *An Explainable Contrastive-based Dilated Convolutional Network with Transformer for Pediatric Pneumonia Detection*, Applied Soft Computing, Vol. 167, p. 112258, 2024 (Elsevier), DOI = "https://doi.org/10.1016/j.asoc.2024.112258".

3. **C.S. Raghaw**, A. Yadav, J.S. Sanjotra, S. Dangi, and N. Kumar. *MNet-SAt: A Multi-scale Network with Spatial-enhanced Attention for segmentation of polyps in colonoscopy*, Biomedical Signal Processing and Control, Vol. 102, p. 107363, 2025 (Elsevier), DOI = "https://doi.org/10.1016/j.bspc.2024.107363".

4. **C.S. Raghaw**, J.S. Sanjotra, M.Z.U. Rehman, S. Bansal, S.S. Dar, and N. Kumar. *T-MPEDNet: Unveiling the Synergy of Transformer-aware Multiscale Progressive Encoder-Decoder Network with Feature Recalibration for Tumor and Liver Segmentation*, Biomedical Signal Processing and Control, Vol. 110, p. 108225, 2025 (Elsevier), DOI = "https://doi.org/10.1016/j.bspc.2025.108225".

5. **C.S. Raghaw**, A. Parwal, S.S. Dar, P. Darade, and N. Kumar. *H-SemiS: Hierarchical Fusion of Semi and Self-Supervised Learning for Knee Osteoarthritis Severity Grading*, Expert Systems with Applications, p. 132279, 2026 (Elsevier), DOI = "https://doi.org/10.1016/j.eswa.2026.132279"

6. **C.S. Raghaw**, T. Sahu, P. Darade, S.S. Dar, and N. Kumar. *Quantum-Guided Textural–Spatial Fusion Network for Thyroid Nodule Classification in Ultrasound Imaging*, International Journal of Imaging Systems and Technology, p. e70428, 2026 (Wiley), DOI = "https://doi.org/10.1002/ima.70428"

In Refereed Conferences

1. **C.S. Raghaw**, S.S. Dar, A. N. Sharma, S. Modi, N. Kumar. *CliQ-RRG: Clinical-Knowledge Guided Disease-aware Visual–Textual Alignment for QA-Style Radiology Report Generation*, Findings of the Association for Computational Linguistics (EMNLP), 2026.

Contents

Abstract i

List of Publications iv

Table of Contents vi

List of Figures xiv

List of Tables xviii

List of Algorithms xix

List of Abbreviations and Acronyms xx

1 Introduction 1

1.1 Research Gaps and Challenges 2

1.2 Research Objective 5

1.3 Thesis Contributions 6

1.4 Thesis Organization 8

2 Literature Survey 12

2.1 Medical Image Classification 13

2.1.1 Background 13

2.1.2 Challenges 13

2.1.3 Existing Literature 14

2.2 Medical Image Segmentation 16

2.2.1 Background 16

2.2.2 Challenges . . . 17
2.2.3 Existing Literature . . . 18
2.3 Explainability in Medical Imaging . . . 20
2.4 Summary . . . 20

3 Coupled Transformer-Convolutional Network with an Adaptive Graph Reconstruction for Leukemia Detection 23
3.1 Introduction . . . 24
3.2 Methodology . . . 25
3.2.1 Framework Overview . . . 25
3.2.2 Preprocessing . . . 25
3.2.3 Proposed Framework . . . 27
3.3 Experimental Settings . . . 37
3.3.1 Benchmark Datasets . . . 37
3.3.2 Evaluation Metrics . . . 39
3.4 Experimental Results . . . 40
3.4.1 Quantitative Analysis . . . 40
3.4.2 Qualitative Analysis . . . 43
3.4.3 Ablation Analysis . . . 44
3.4.4 Cross-Dataset Analysis . . . 48
3.4.5 Model Complexity Analysis . . . 49
3.5 Conclusion . . . 50

4 Contrastive-based Dilated Convolutional Network with Transformer for Pediatric Pneumonia Detection 53
4.1 Introduction . . . 54
4.2 Methodology . . . 55
4.2.1 Framework Overview . . . 55
4.2.2 Preprocessing . . . 55
4.2.3 Proposed Framework . . . 62
4.3 Experimental Settings . . . 71
4.3.1 Benchmark Datasets . . . 72
4.3.2 Evaluation Metrics . . . 73

4.4 Experimental Results . . . 74
4.4.1 Quantitative Analysis . . . 74
4.4.2 Qualitative Analysis . . . 78
4.4.3 Ablation Analysis . . . 79
4.4.4 Cross-Dataset Analysis . . . 83
4.4.5 Model Complexity Analysis . . . 85
4.5 Conclusion . . . 86

5 Hierarchical Fusion of Semi and Self-Supervised Learning for Knee Osteoarthritis Severity Grading 88
5.1 Introduction . . . 89
5.2 Methodology . . . 90
5.2.1 Framework Overview . . . 90
5.2.2 Preprocessing . . . 91
5.2.3 Proposed Framework . . . 91
5.3 Experimental Settings . . . 106
5.3.1 Benchmark Datasets . . . 106
5.3.2 Evaluation Metrics . . . 108
5.4 Experimental Results . . . 108
5.4.1 Quantitative Analysis . . . 109
5.4.2 Qualitative Analysis . . . 111
5.4.3 Ablation Analysis . . . 112
5.4.4 Cross-Dataset Analysis . . . 116
5.4.5 Model Complexity Analysis . . . 119
5.5 Conclusion . . . 120

6 Transformer-aware Multiscale Progressive Encoder-Decoder Network for Tumor and Liver Segmentation 122
6.1 Introduction . . . 123
6.2 Methodology . . . 124
6.2.1 Framework Overview . . . 124
6.2.2 Preprocessing . . . 125
6.2.3 Proposed Framework . . . 125

6.3 Experimental Settings . . . 138
6.3.1 Benchmark Dataset . . . 138
6.3.2 Evaluation Metrics . . . 140
6.4 Experimental Results . . . 141
6.4.1 Quantitative Analysis . . . 141
6.4.2 Qualitative Analysis . . . 146
6.4.3 Ablation Analysis . . . 147
6.4.4 Cross-Dataset Analysis . . . 149
6.4.5 Model Complexity Analysis . . . 152
6.5 Conclusion . . . 152

7 A Multiscale Network with Spatial-enhanced Attention for Segmentation of Polyps in Colonoscopy 155
7.1 Introduction . . . 156
7.2 Methodology . . . 157
7.2.1 Framework Overview . . . 157
7.2.2 Preprocessing . . . 157
7.2.3 Proposed Framework . . . 159
7.3 Experimental Settings . . . 167
7.3.1 Benchmark Dataset . . . 168
7.3.2 Evaluation Metrics . . . 168
7.4 Experimental Results . . . 168
7.4.1 Quantitative Analysis . . . 169
7.4.2 Qualitative Analysis . . . 171
7.4.3 Ablation Analysis . . . 172
7.4.4 Cross-Dataset Analysis . . . 175
7.4.5 Model Complexity Analysis . . . 177
7.5 Conclusion . . . 178

8 Conclusion and Future Directions 180
8.1 Summary of the Thesis . . . 181
8.1.1 Coupled Transformer–Convolutional Network with an Adaptive Graph Reconstruction for Leukemia Detection . . . 181

8.1.2 Contrastive–based Dilated Convolutional Network with Transformer for Pediatric Pneumonia Detection 181
8.1.3 Hierarchical Fusion of Semi and Self-Supervised Learning for Knee Osteoarthritis Severity Grading 182
8.1.4 Transformer-aware Multiscale Progressive Encoder-Decoder Network for Tumor and Liver Segmentation 183
8.1.5 A Multiscale Network with Spatial-enhanced Attention for Segmentation of Polyps in Colonoscopy 183
8.2 Future Work 186

Bibliography **189**

List of Figures

1.1 Challenges in diverse medical imaging modalities. (a-b) morphological indistinguishability in normal versus cancerous cells, (c-d) fuzzy and blurred boundaries, (e) variations in polyp size, (f) under-exposed X-rays, (g) overlapping nuclei, (h) small tumor size, (i) multiple polyps, and (j) data skewness. 3

3.1 Proposed architecture of Convolutional Network (CoTCoNet) for leukemia detection in whole slide imaging. 26

3.2 Proposed architecture of Global Feature Module (GMod). 28

3.3 Proposed architecture of Spatial Feature Module (SMod). 30

3.4 Flowchart of the Sine Cosine Algorithm (SCA). 36

3.5 Sample dataset images for leukemia classification. 38

3.6 Comparative analysis of CoTCoNet against existing baselines across four datasets using ROC curves, a radar chart, and PR curves. 44

3.7 Qualitative analysis of CoTCoNet framework. 44

4.1 The workflow of our proposed XCCNet framework for detecting pediatric pneumonia in chest X-rays. 56

4.2 Proposed architecture of Explainable Contrastive-based Dilated Convolutional Network with Transformer (XCCNet) for pediatric pneumonia detection in chest X-rays. 57

4.3 Proposed architecture of Adversarial-based Data Augmentation (ADA) unit. 58

4.4 Proposed architecture of ResUNet++ for lung segmentation. 61

4.5 Proposed architecture of ResNet-based rib suppression model. 62

4.6 Comparison between (a) standard 3×3 convolutions and (b) dilated convolutions (dilation rate 2). 64
4.7 Sample dataset images for pediatric pneumonia classification. 72
4.8 Qualitative analysis of XCCNet framework. 79
4.9 Impact of varying CLAHE parameters on chest X-rays: (a)–(e) Normal case and (f)–(j) Pneumonic case 81
4.10 Comparative analysis of model parameters, inference time, and accuracy for XCCNet framework against existing baselines on benchmark datasets. . . . 85

5.1 The workflow of our proposed H-SemiS framework for knee osteoarthritis severity grading in knee X-rays. 90
5.2 Proposed architecture of Masked Image Reconstruction (MI-Rec). 92
5.3 Proposed architecture of Similarity-aware Reconstructed Image Labeler (SiRL). 96
5.4 Proposed architecture of Quantum-infused Teacher-Student (Q-TeSt) framework. 97
5.5 Proposed architecture of Quantum Convolutional Network (QCN). 100
5.6 Proposed architecture of Hierarchical Multi-classifier (HiM) module. 105
5.7 Sample dataset images for knee osteoarthritis classification. 107
5.8 Quantitative comparison of the proposed H-SemiS framework against existing baselines on (a) OAI dataset and (b) DKXI dataset. 110
5.9 Confusion matrices for classification on the OAI and DKXI datasets. 111
5.10 Qualitative analysis of H-SemiS framework. 112
5.11 KXR reconstruction examples across different KL grades. 113
5.12 Mask ratio ablation study on H-SemiS performance. 114
5.13 Impact of labeled data percentage on Teacher-Student framework accuracy. 115
5.14 Performance analysis of H-SemiS under different decomposition techniques. 117
5.15 Performance comparison: ROC and PR curves for OP and KO datasets. . . 118
5.16 Comparative analysis of model parameters, inference time, and accuracy for the H-SemiS framework against existing baselines on benchmark datasets. . 118

6.1 The workflow of our proposed T-MPEDNet framework for segmenting liver and tumors in computed tomography and magnetic resonance imaging. . . . 124
6.2 Visualization of the processing pipeline within the preprocessing block. . . 125
6.3 Proposed architecture of the Progressive Feature Encoder-Decoder (PED) block. 126
6.4 Proposed architecture of Compressive Channel Recalibration (CCR) module. 129
6.5 Proposed architecture of Multi-Head Dynamic Contextual Attention (MHDCA) module. 131
6.6 Proposed architecture of Multi-Scale Atrous Spatial (MSAS) module. . . . 134
6.7 Proposed architecture of Morphological Boundary Refinement (MBR) module. 135
6.8 Sample dataset images for liver and tumor segmentation. 138
6.9 Comparative analysis of T-MPEDNet against existing baselines across three datasets using DSC, IoU, Accuracy, and HD95. 145
6.10 Qualitative analysis of T-MPEDNet framework. 147
6.11 Visual results showing the contribution of key components in T-MPEDNet. 148
6.12 Impact analysis curves of T-MPEDNet components for liver and tumor segmentation on the LiTS dataset using Dice similarity coefficient. 149
6.13 Comparative analysis of model parameters, inference time, and accuracy for the T-MPEDNet framework against existing baselines on benchmark datasets. 151

7.1 The workflow of our proposed MNet-SAt framework for polyp segmentation in colonoscopy images. 158
7.2 Proposed architecture of Multiscale Network with Spatial-enhanced Attention (MNet-SAt) for polyp segmentation in colonoscopy images. 160
7.3 Edge maps of the Sobel operator in x-y directions 161
7.4 Proposed architecture of Multi-Head Spatial-Enhanced Attention (MHSEAt). 164
7.5 Proposed architecture of Channel-Enhanced Atrous Spatial Pyramid Pooling (CE-ASPP). 166
7.6 Comparative analysis of MNet-SAt against existing baselines across two datasets using ROC curves and PR curves. 169
7.7 Qualitative analysis of MNet-SAt framework. 171

7.8 Comparative analysis of model parameters, inference time, and accuracy for the MNet-SAt framework against existing baselines on benchmark datasets. 177

List of Tables

3.1 An overview of benchmark datasets for leukemia classification. 38

3.2 Evaluation metrics of the CoTCoNet framework for leukemia detection in whole slide imaging. 39

3.3 Quantitative analysis of CoTCoNet framework with existing baselines on the C-NMC 2019 dataset. **Bold**: best and underlined: second-best scores. . . . 41

3.4 Quantitative analysis of CoTCoNet framework with existing baselines on the ALL-IDB1 dataset. **Bold**: best and underlined: second-best scores. 41

3.5 Quantitative analysis of CoTCoNet framework with existing baselines on the ALL-IDB2 dataset. **Bold**: best and underlined: second-best scores. 42

3.6 Quantitative analysis of CoTCoNet framework with existing baselines on the SN-AM dataset. **Bold**: best and underlined: second-best scores. 43

3.7 Ablation analysis of CoTCoNet framework with different image enhancement techniques on the C-NMC 2019 dataset. 45

3.8 Ablation analysis of CoTCoNet framework with different image segmentation techniques on the C-NMC 2019 dataset. 46

3.9 Ablation analysis of CoTCoNet framework with different image augmentation techniques on the C-NMC 2019 dataset. 47

3.10 Ablation analysis of CoTCoNet framework with different feature extraction techniques on the C-NMC 2019 dataset. 47

3.11 Ablation analysis of CoTCoNet framework with different graph-based feature reconstruction and feature selection techniques on the C-NMC 2019 dataset. 48

3.12 Cross-Dataset analysis of CoTCoNet framework with different training and testing datasets. 49

3.13 Model Complexity Analysis of the CoTCoNet framework comparing parameters and inference times with existing baselines. 50

4.1 An overview of benchmark datasets for pediatric pneumonia detection. . . . 72

4.2 Evaluation metrics of the XCCNet framework for pediatric pneumonia detection in chest X-rays. 74

4.3 Quantitative analysis of XCCNet framework with existing baselines on the Kermany dataset. **Bold**: best and underlined: second-best scores. 75

4.4 Quantitative analysis of XCCNet framework with existing baselines on the VinDR-PCXR dataset. **Bold**: best and underlined: second-best scores. . . . 76

4.5 Quantitative analysis of XCCNet framework with existing baselines on the NIH-Pediatric dataset. **Bold**: best and underlined: second-best scores. . . . 77

4.6 Quantitative analysis of XCCNet framework with existing baselines on the Trivedi dataset. **Bold**: best and underlined: second-best scores. 78

4.7 Ablation analysis of XCCNet framework with different adversarial-based data augmentation techniques on the Kermany dataset. 80

4.8 Ablation analysis of XCCNet framework with different CXR enhancer techniques on the Kermany dataset. 80

4.9 Ablation analysis of XCCNet framework with different CLAHE parameters on the Kermany dataset. 81

4.10 Ablation analysis of XCCNet framework with different segmentators on the Kermany dataset. 82

4.11 Ablation analysis of XCCNet framework with different rib suppressor techniques on the Kermany dataset. 82

4.12 Ablation analysis of XCCNet framework with different feature extraction techniques on the Kermany dataset. 83

4.13 Ablation analysis of XCCNet framework with different feature fusion techniques on the Kermany dataset. 84

4.14 Cross-Dataset analysis of XCCNet framework with different training and testing datasets. 84

4.15 Model Complexity Analysis of the XCCNet framework comparing parameters and inference times with existing baselines. 86

5.1 An overview of benchmark datasets for knee osteoarthritis classification. . . 106

5.2 Evaluation metrics of H-SemiS framework osteoarthritis severity grading. . 108

5.3 Quantitative analysis of H-SemiS framework with existing baselines on the OAI and DKXI datasets. **Bold**: best and underlined: second-best scores. . . 109

5.4 Ablation analysis of H-SemiS framework with different masking ratios and reconstruction techniques on the OAI dataset. 113

5.5 Ablation analysis of H-SemiS framework with different thresholds for labeling reconstructed samples on the OAI dataset. 114

5.6 Ablation analysis of H-SemiS framework with different percentages of labeled samples and architectural components on the OAI dataset. 116

5.7 Evaluation of H-SemiS generalizability across varied training and testing samples. 117

5.8 Model Complexity Analysis of the H-SemiS framework, comparing parameters and inference times with existing baselines. 119

6.1 Progressive Encoder-Decoder Network architecture of T-MPEDNet with feature dimensions and layer configurations. 137

6.2 An overview of benchmark datasets for liver and tumor segmentation. . . . 140

6.3 Quantitative analysis of T-MPEDNet framework with existing baselines on the LiTS dataset. **Bold**: best and underlined: second-best scores. 142

6.4 Quantitative analysis of T-MPEDNet framework with existing baselines on the 3DIRCADb dataset. **Bold**: best and underlined: second-best scores. . . 143

6.5 Quantitative analysis of T-MPEDNet framework with existing baselines on the ATLAS dataset. **Bold**: best and underlined: second-best scores. 144

6.6 Ablation analysis of the T-MPEDNet framework with the contribution of each T-MPEDNet component on the LiTS dataset. 147

6.7 Cross-Dataset analysis of T-MPEDNet framework with different training and testing datasets. 150
6.8 Model Complexity Analysis of the T-MPEDNet framework comparing parameters, inference times, and memory with existing baselines. 151

7.1 Feature map dimensions across different stages of the MNet-SAt architecture. 161
7.2 Quantitative analysis of MNet-SAt framework with existing baselines on the Kvasir-SEG dataset. **Bold**: best and underlined: second-best scores. 170
7.3 Quantitative analysis of MNet-SAt framework with existing baselines on the CVC-ClinicDB dataset. **Bold**: best and underlined: second-best scores. . . 170
7.4 Ablation analysis of MNet-SAt framework with different encoder-decoder backbones on Kvasir-SEG dataset. 172
7.5 Ablation analysis of MNet-SAt framework with different multi-scale feature aggregator components on Kvasir-SEG dataset. 173
7.6 Ablation analysis of MNet-SAt framework with different spatial-enhanced attention components on Kvasir-SEG dataset. 174
7.7 Ablation analysis of MNet-SAt framework with different channel-enhanced atrous spatial pyramid pooling components on Kvasir-SEG dataset. 174
7.8 Ablation analysis of MNet-SAt framework with different optimizers on Kvasir-SEG dataset. 175
7.9 Ablation analysis of MNet-SAt framework with different components on the Kvasir-SEG dataset. 175
7.10 Cross-Dataset analysis of MNet-SAt framework with training on Kvasir-SEG and testing on CVC-ClinicDB. 176
7.11 Cross-Dataset analysis of MNet-SAt framework with training on CVC-ClinicDB and testing on Kvasir-SEG. 177
7.12 Model Complexity Analysis of the MNet-SAt framework comparing parameters, FLOPS, MACs, training time, and inference times with existing baselines. 178

List of Algorithms

3.1 COUPLED TRANSFORMER CONVOLUTIONAL NETWORK 34
4.1 CXP MODULE 63
4.2 FEATURE EXTRACTOR AND CLASSIFICATION MODULE 69

List of Abbreviations and Acronyms

ADA	Adversarial-based Data Augmentation
B-ALL	B-lineage Acute Lymphoblastic Leukemia
BCE	Binary Cross-Entropy
CAD	Computer-Aided Diagnosis
CLAHE	Contrast Limited Adaptive Histogram Equalization
CLIP	Contrastive Language-Image Pretraining
CNN	Convolutional Neural Network
CoTCoNet	Coupled Transformer Convolutional Network
CoTFx	Contrastive-based Transformer Feature Extractor
CT	Computed Tomography
CXR	Chest X-Ray
DHFFC	Dual Hybrid Feature Fusion and Classification
DL	Deep Learning
GMod	Global Feature Module
GraFR	Graph-based Feature Reconstruction
Grad-CAM	Gradient-weighted Class Activation Mapping
H-SemiS	Hierarchical Semi-Supervised framework
HDLC	Hierarchical Deep Learning Classifier
HiM	Hierarchical Multi-classifier
HQ-TeSt	Hierarchical Quantum-infused Teacher–Student

JSN	Joint Space Narrowing
KL	Kellgren-Lawrence
KOA	Knee Osteoarthritis
KXR	Knee X-Ray
LIME	Local Interpretable Model-agnostic Explanations
MI-Rec	Masked Image Reconstruction
MLP	Multi-Layer Perceptron
MNet-SAt	Multi-scale Network with Spatial-enhanced Attention
MRI	Magnetic Resonance Imaging
MSA	Multi-head Self-Attention
QCN	Quantum Convolutional Network
Q-TeSt	Quantum-infused Teacher-Student
ROI	Region of Interest
SAM	Segment Anything Model
SCA	Sine-Cosine Algorithm
SE	Squeeze-and-Excitation
SFx	Spatial Feature Extraction
SMod	Spatial Feature Module
T-MPEDNet	Transformer-based Multiscale Progressive Encoder-Decoder Network
ViT	Vision Transformer
WSI	Whole-Slide Imaging
XAI	Explainable Artificial Intelligence
XCCNet	Explainable Contrastive-based Dilated Convolutional Network with Transformer

Chapter 1

Introduction

CONTENTS

1.1 Research Gaps and Challenges 2
1.2 Research Objective 5
1.3 Thesis Contributions 6
1.4 Thesis Organization 8

In modern medical practice, the ability to assess the internal state of the human body without invasive procedures is essential for accurate diagnosis and early monitoring of treatment. Reliable clinical judgment of a patient's condition depends on the ability to observe internal anatomical structures and pathological changes during routine evaluation. The response to ongoing treatment procedures also depends on careful assessment of these clinical findings. Although traditional surgical procedures provide direct access for visual inspection and tissue sampling, such approaches are invasive and affect the patient's overall health in some cases. As a result, pathological image analysis has shifted slightly toward digital analysis of different imaging modalities, including Chest X-Ray (CXR), Knee X-Ray (KXR), Whole-Slide Imaging (WSI) in microscopy, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and colonoscopy images. This phase transition supports a broad view of disease diagnosis using improved internal visualization of human anatomy. This avoids the risks linked with the complex invasive procedures [1].

The use of medical imaging in clinical practice has grown steadily, generating a massive volume of images that require precise analysis. Interpreting these scans is not simple be-

cause it demands time, careful observation, and expert knowledge. Research indicates that when clinicians spend less time reviewing individual images, the likelihood of diagnostic errors increases [2]. In high-pressure medical environments, this reality results in a heavier workload and less attention per case. This situation is further complicated by the shortage of skilled radiologists, which limits the ability to review cases promptly and accurately [3]. As a result, delays in diagnosis are common, and many healthcare facilities struggle to maintain consistent and reliable radiology services, which affects the quality of patient care and clinical outcomes.

To address these challenges, Computer-Aided Diagnosis (CAD) has emerged as an important tool to support clinical decision-making and help to reduce the shortage of expert clinicians. This progress is driven by Deep Learning (DL), in which neural networks learn to identify complex pathological patterns directly from raw imaging data [4]. Medical artificial intelligence has seen rapid development in areas such as radiology and pathology that rely on image interpretation, supported by improvements in model design and the availability of large annotated datasets [5]. Architectures such as Convolutional Neural Network (CNN) and transformer-based methods have shown superior performance across different imaging modalities. Incorporating these systems into daily clinical routines enables healthcare providers to improve diagnostic consistency, reduce the burden on radiologists, and support timely, accurate patient care.

1.1 Research Gaps and Challenges

Deep Learning (DL) has led to clear improvements in pathological image analysis [6,7], and many CAD report strong performance on specific classification [8–10] and segmentation tasks [11–13]. However, several research gaps limit the reliable use in clinical practice. Current supervised learning methods depend on large labeled datasets, which are difficult and expensive to obtain in medical domains such as Knee Osteoarthritis (KOA) and pediatric pneumonia. Where each scan consists of complex anatomy. Furthermore, many existing architectures struggle with data distribution challenges and fail to generalize across different patient populations or scanner variabilities. Another major gap involves the lack of interpretability in DL decisions, as many high-performing methods remain black boxes. The

lack of interpretability makes it difficult for clinicians to trust the outputs in critical medical cases. Additionally, standard CNNs fail to capture long-range contextual relationships and thus cannot map complex, irregular structures, such as liver tumors and tiny polyps. Finally, the common issues of class imbalance and noisy annotations in medical datasets degrade the performance of classification frameworks, underscoring the need for semi-supervised approaches that require fewer data samples. Despite impressive progress in CAD systems, significant obstacles and technical gaps remain and require focused attention to address real-world clinical challenges.

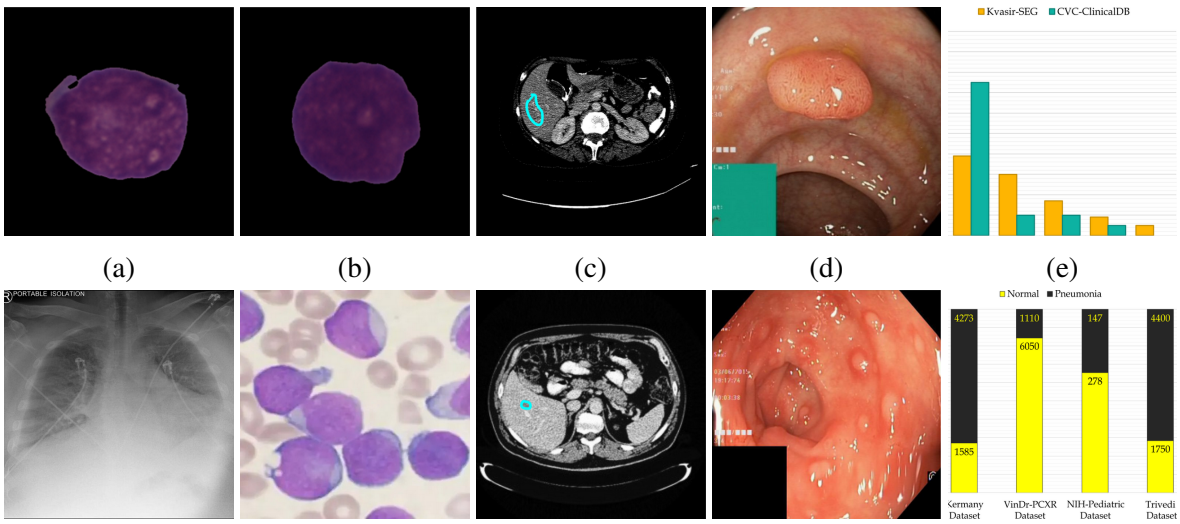


Figure 1.1: Challenges in diverse medical imaging modalities. (a-b) morphological indistinguishability in normal versus cancerous cells, (c-d) fuzzy and blurred boundaries, (e) variations in polyp size, (f) under-exposed X-rays, (g) overlapping nuclei, (h) small tumor size, (i) multiple polyps, and (j) data skewness.

- **Ambiguous Morphological Patterns and Contextual Inconsistency**

  One of the major challenges in pathological image analysis is the presence of ambiguous morphological patterns. Pathological regions such as tumors, nodules, and abnormal cells that appear similar to the surrounding healthy tissues. The issue becomes more challenging in high-resolution multimodal imaging data. Variations in staining, acquisition settings, and anatomical structures can further reduce visual consistency. As illustrated in Figure 1.1, morphological similarity between normal and diseased regions can lead to incorrect predictions. In addition, many existing methods

focus either on local pixel-level details or on global structural information, without effectively combining the two. However, in clinical practice, detailed diagnosis analyzes the finer details within the broader anatomical context. The lack of such joint learning remains a key limitation in current approaches.

- **Limited Labeled Data, Class Imbalance, and Annotation Noise**

  Another important challenge is the limited availability of high-quality annotated medical data. Creating reliable annotations requires expert knowledge, which is both time-consuming and expensive. Manual data curation and labeling are labor-intensive processes that restrict the size of labeled datasets. In many cases, certain disease categories are underrepresented, resulting in class imbalance and leading to normal samples dominating the dataset and increasing the risk of overfitting. This imbalance biases the model toward the majority classes and reduces its sensitivity to minority classes. Furthermore, differences in expert interpretation introduce annotation noise, thereby affecting label consistency. As shown in Figure 1.1, data distribution skew is a common issue across datasets. These factors make it difficult to train DL methods and highlight the need for approaches that can learn under limited supervision with fewer samples.

- **Low-Contrast Pathological Margins and Boundary Indistinctness**

  Unclear, low-contrast boundaries further complicate the accurate identification of pathological regions. In modalities such as CT, MRI, and CXR, the transition from healthy to diseased tissue is often gradual, complicating the precise localization of lesions. Additional factors, such as imaging noise, external artifacts, and overlapping anatomical structures, degrade the boundary visibility. As depicted in Figure 1.1, fuzzy and blurred boundaries are commonly observed, especially in tumor and organ segmentation tasks. These issues directly affect both detection and segmentation performance, as existing methods fail to identify the true pathological area. Improving boundary identification is an essential requirement for clinical applications such as surgical planning and radiotherapy.

- **Morphological Heterogeneity and Multiscale Variability**

  Pathological structures exhibit significant variation in size, shape, and appearance

across different patients and imaging modalities. Small lesions such as early-stage tumors are often difficult to detect due to their limited visibility. In comparison to the larger abnormalities that contain multiple regions and require broader contextual understanding. As shown in Figure 1.1, variations in object size, presence of multiple lesions, and heterogeneous structures are common in medical images. In addition, factors such as varying cell density, organ structure, and imaging resolution further increase the overall complexity. Methods that rely on a single scale representation cannot capture the broader context of the anatomical region. Therefore, handling multiscale variability remains an important challenge for developing robust and generalizable medical image analysis systems.

These challenges highlight gaps in current CAD systems. Existing methods address these gaps but fail to provide a unified solution that works efficiently in real clinical settings. Therefore, there is a need for approaches that effectively learn global and local features from limited, imbalanced data to improve boundary localization and capture multiscale variation. The research presented in this thesis is designed to address these limitations and to develop reliable, generalizable methods for pathological image analysis.

1.2 Research Objective

The primary goal of this research is to bridge the gap between complex medical data and reliable clinical decision-making. To achieve this, the thesis is structured around the following four key objectives:

(i) **Objective 1: Global–Local Feature Synthesis**

The first objective of this thesis is to develop methods that learn unified representations by combining the global anatomical context with fine-grained pathological details. In many clinical scenarios, understanding the broader structural layout is an important task that identifies the fine-grained local abnormalities. This objective focuses on designing a feature synthesis strategy to bridge the gap that produces a reliable clinical diagnosis. The approach is evaluated on Acute Lymphoblastic Leukemia using WSI and on the Pediatric Pneumonia using CXR scans, where both contextual and localized feature representations are important for reliable diagnosis.

(ii) **Objective 2: Data-Efficient Learning**

The second objective addresses the challenge of limited and noisy data annotation in medical imaging. Instead of relying on large, labeled datasets, this work aims to develop a learning framework that produces reliable results with minimal supervision and remains robust to label noise and class imbalance. The focus is on KOA severity grading from KXR, where annotations are often subjective and costly to obtain. This objective explores various strategies to reduce label dependency while maintaining the strong predictive performance.

(iii) **Objective 3: Precise Boundary Refinement**

The third objective focuses on improving lesion localization accuracy in cases where boundaries are unclear or poorly defined. Medical images, particularly CT and MRI scans, exhibit low contrast between regions of interest and surrounding tissues, which makes segmentation difficult. This objective aims to design progressive-decoding mechanisms that refine predictions stepwise, enabling precise identification of liver and tumor regions. The goal is to identify the boundary while accurately preserving structural consistency.

(iv) **Objective 4: Multi-Scale Feature Fusion**

The fourth objective aims to address the variability in lesion shape and size, including appearance across different imaging conditions. Pathological findings range from very small anomalies to large, diffuse regions, requiring a method that can operate effectively across multiple scales. This objective focuses on developing scale-aware feature fusion methods that can capture both coarse and fine details in a unified framework. The approach is applied to colorectal polyp segmentation in colonoscopy images, where detecting lesions of varying sizes is essential for reliable clinical assessment.

1.3 Thesis Contributions

This thesis presents a set of methods to address challenges in pathological image analysis identified in the motivation and objectives. The research focuses on improving the synergy

between global contextual understanding and local morphological precision across various medical imaging modalities, including blood smears, CXR, KXR, abdominal CT/MRI scans, and colonoscopic frames. A brief overview of our research contributions is provided below, and more details are available in the later chapters.

First, we present a global–local feature synthesis by designing a transformer-convolutional architecture. Recognizing that CNN excel at capturing local textures but often miss long-range dependencies, we designed frameworks such as Coupled Transformer Convolutional Network (CoTCoNet) [14] and Explainable Contrastive-based Dilated Convolutional Network with Transformer (XCCNet) [15]. These architectures capture global anatomical context and fine pathological details such as cellular changes in leukemia and localized opacities in pediatric pneumonia. This dual-path design improves the framework's ability to learn complex visual patterns by combining the overall structural understanding of an image with long-range precision to local features.

Second, to address the challenge of limited annotated medical data samples, we introduce data-efficient learning frameworks that leverage semi-supervised and self-supervised strategies. High-quality labeled medical data are expensive and time-consuming to obtain. As a result, datasets are often small or contain noisy labels. Using the Hierarchical Semi-Supervised framework (H-SemiS) framework [16], we efficiently leverage unlabeled data via hierarchical classification and staged pseudo-labeling. These approaches enable effective learning from scarce and imperfect labels while reducing the impact of class imbalance and annotation variability. This leads to stable and generalizable performance.

Third, we present boundary-aware mechanisms for precise lesion localization, specifically for detecting small tumors in low-contrast settings. In frameworks such as Transformer-based Multiscale Progressive Encoder-Decoder Network (T-MPEDNet) [17], we incorporate progressive refinement techniques and feature recalibration modules that address the indistinct transitions between healthy and diseased tissues. By leveraging morphologically inspired operations and multiscale encoder-decoder architectures, we enhance the framework's ability to focus on edge-specific features. This ensures that the spatial features and precise boundaries of small, heterogeneous lesions, such as early-stage hepatic tumors, are captured accurately. This is important for effective surgical planning.

Fourth, to handle the inherent variability in the size and shape of diverse pathological

structures, we propose multiscale feature fusion techniques. Pathological findings can range from small anomalies to large-scale organ identification. Using the Multi-scale Network with Spatial-enhanced Attention (MNet-SAt) [18] framework, which is designed for the specific tasks such as automated polyp segmentation in colonoscopy. We align and combine features across different resolutions using spatially enhanced attention. This allows the method to maintain high sensitivity for small and indistinguishable abnormalities. It also preserves the structural integrity of larger lesions across diverse imaging modalities and patient populations.

Fifth, the thesis incorporates clinical interpretability into the proposed frameworks by integrating explainable analysis techniques. In medical applications, it is important to achieve superior performance to provide a clear reason behind method predictions. The proposed methods include mechanisms to highlight relevant regions and support visual interpretation, enabling clinicians to understand better and trust the model's outputs. This ensures that predictions are accurate, transparent, and clinically meaningful. This supports reliable, trustworthy decision-making in real-world diagnostic settings.

Finally, this thesis demonstrates the generalizability of the proposed methods across multiple imaging modalities, including WSI, CXR, CT, MRI, and endoscopic imaging. Despite differences in imaging characteristics and the disease presentation, the core design principles remain the same. This shows that the proposed approaches can be applied broadly to different clinical tasks in CAD.

1.4 Thesis Organization

The remainder of this thesis is structured to provide a comprehensive overview of our research, moving from the foundational literature to core methodological contributions and concluding with a summary of our main findings. The organization is as follows:

Chapter 1 establishes the foundation of this research by introducing the core motivation and primary objectives. It outlines the significant challenges currently facing pathological image analysis. This provides a detailed summary of the thesis's key technical contributions.

Chapter 2 provides a comprehensive review of existing work in medical image classification and segmentation. It analyzes current approaches, highlights their limitations, and

further identifies the research gaps that motivate the proposed methods.

Chapter 3 presents a global–local feature-synthesis framework for leukemia classification using WSI. In this chapter, the proposed method CoTCoNet, learns transformer-based global features with convolutional local feature extraction and a graph-based feature reconstruction module to capture complex hematological patterns. It also incorporates leukocyte segmentation and synthetic sample generation to address data limitations. Extensive experiments demonstrate strong performance across multiple datasets, achieving high accuracy and improved generalization compared to existing methods. This chapter is based on our work published in *Computers in Biology and Medicine (2024)* [14].

Chapter 4 extends the idea of global–local feature learning to pediatric pneumonia detection from CXR. In this chapter, the proposed XCCNet framework learns feature representations using dilated convolutions and transformer-based representations, supported by a contrastive learning framework that improves feature discrimination. Additional components, such as contrast enhancement with data augmentation and an explainable module, further improve the diagnostic performance. Experimental results across multiple datasets show consistent improvement in accuracy and reliability. This chapter is based on our work published in *Applied Soft Computing (2024)* [15].

Chapter 5 focuses on data-efficient learning for KOA severity grading using KXR. In this chapter, the proposed H-SemiS framework builds on the semi-supervised and self-supervised learning within a hierarchical structure to handle limited labeled data and multi-class imbalance. It also includes reconstruction-based learning and noise-aware label refinement to improve the robustness. The method demonstrates strong performance across multi-class and binary datasets, even with a small fraction of annotated data. This chapter is based on our work published in *Expert Systems with Applications (2026)* [16].

Chapter 6 addresses precise boundary refinement in liver and tumor segmentation using CT and MRI scans. In this chapter, the proposed T-MPEDNet framework employs a progressive encoder–decoder design with transformer-based contextual learning and multiscale feature extraction. A dedicated boundary refinement module improves segmentation accuracy in challenging scenarios with low contrast and indistinct tumor boundaries. Experimental evaluations show clear improvements over existing segmentation approaches. This chapter is based on our work published in *Biomedical Signal Processing and Control (2025)* [17].

Chapter 7 presents a multiscale feature fusion framework for polyp segmentation in colonoscopy images. In this chapter, the proposed MNet-SAt method utilizes multiscale feature aggregation with spatially enhanced attention to capture diverse polyp characteristics. It also includes edge-aware mechanisms that preserve boundary details and improve the localization. The framework achieves strong performance across benchmark datasets, demonstrating its effectiveness in handling complex polyp structures. This chapter is based on our work published in *Biomedical Signal Processing and Control (2025)* [18].

Chapter 8 serves as the conclusion of this thesis. It summarizes the key research findings and discusses the main contribution of each chapter within the limit of CAD-based approaches. Finally, it outlines potential future directions in the research domain, including multimodal clinical data integration with 3D volumetric analysis and interpretability enhancements for real-time, lightweight deployment.

Chapter 2

Literature Survey

CONTENTS

2.1 Medical Image Classification 13
2.1.1 Background 13
2.1.2 Challenges 13
2.1.3 Existing Literature 14
2.2 Medical Image Segmentation 16
2.2.1 Background 16
2.2.2 Challenges 17
2.2.3 Existing Literature 18
2.3 Explainability in Medical Imaging 20
2.4 Summary 20

This chapter reviews the existing literature for pathological image analysis with an emphasis on Deep Learning (DL) methods for clinical imaging data. It outlines the key challenges in this field and analyzes existing research in light of these challenges. The focus of this thesis is on two primary tasks: medical image classification and medical image segmentation. The following sections provide a detailed discussion of these two tasks.

2.1 Medical Image Classification

2.1.1 Background

Medical image classification is an important step in pathological image analysis, aiming to categorize images according to specific diseases and imaging modality. Automating this process can assist the radiologists and physicians. It also provides a second opinion, which helps them interpret complex cases in detail. In recent years, DL has emerged as a suitable approach for interpretation and has replaced traditional machine learning methods due to its ability to learn features directly from data. However, medical image classification remains a more challenging task than general image classification. Medical images often exhibit high intra-class variation and strong inter-class similarity, especially in high-resolution multimodal data. In addition, manual annotation is a time-consuming process and requires expert knowledge. This limits the availability of large labeled datasets for model training. Earlier methods relied on handcrafted, domain-specific features that do not generalize across diverse datasets. Additionally, such features fail to capture the clinical meaning of the images, creating a gap between low-level visual patterns and high-level semantic understanding. Therefore, a modern approach is needed to learn meaningful representations that bridge the gap. This improves the classification performance in real-world clinical settings.

2.1.2 Challenges

Building a reliable classification system for medical images is more challenging than for natural images due to the complexity and variability of clinical data. Although DL has improved performance, several issues still affect the model's reliability. Conventional handcrafted features are insufficient to represent medical images, as pathological patterns are complex and domain-specific. Even with DL methods that capture the clinically meaningful features, it remains difficult due to the gap between low-level visual patterns and high-level clinical understanding. Many methods focus either on local details or global context, but not both. In addition, medical images often exhibit high intra-class variation and strong inter-class similarity, where diseased and healthy tissues may appear visually similar, making classification difficult.

Another challenge lies in the data and generalization aspects. Medical images often suffer from low contrast, noise, and overlapping structures, as well as variations in acquisition settings that affect feature extraction and model performance. The availability of labeled data is limited by the need for expert annotation, leading to small datasets with class imbalance and annotation inconsistencies. These factors increase the risk of overfitting and degrade the model's robustness. Furthermore, models trained on specific datasets may not generalize across different imaging conditions due to variations in data distribution across imaging protocols and patient populations.

2.1.3 Existing Literature

Research in medical image classification has progressed from traditional feature engineering methods to modern DL approaches. Early techniques relied on manually designed features, which struggled to generalize across the different datasets. In contrast, DL models learn meaningful representations directly from data samples. This section reviews developments in this area, with a focus on recent methods for addressing challenges. The challenges include structural ambiguity, variability in medical images, and limited availability of labeled data samples.

2.1.3.1 CNN-based Methods

Convolutional Neural Network (CNN) have long been the standard for identifying pathological patterns due to their ability to capture local spatial hierarchies. These models use learnable filters to identify key features such as cell edges, textures, and organ boundaries. Several studies have explored CNN-based architectures for disease classification. Lightweight models based on ShuffleNet and MobileNet have been designed to improve efficiency while maintaining accuracy [19, 20]. Other works have evaluated standard architectures, including AlexNet, ResNet, DenseNet, and VGG, for leukemia detection and classification. These methods, combined with data augmentation, help reduce overfitting [21]. Capsule-based models and ensemble frameworks have also been proposed to enhance feature representation and improve classification performance [22, 23]. In pneumonia detection CNN-based approaches using depth-wise separable convolutions and customized activation functions have

demonstrated promising results in Chest X-Ray (CXR) analysis [24, 25]. Federated learning frameworks have been further explored to enable training across distributed datasets [26]. Despite these advances, CNN-based methods mainly focus on local features and struggle to capture long-range dependencies, which are important for learning complex pathological patterns in data.

2.1.3.2 Transformer-based Methods

Transformer-based models have attracted attention in medical imaging for their ability to learn global relationships within an image using self-attention mechanisms. These models are effective in capturing long-range dependencies that are missed by convolutional-based methods. Recent studies have applied vision transformers to medical image classification tasks such as pneumonia detection and musculoskeletal analysis [27, 28]. Pre-trained transformer models have been used to extract global representations from CXR, and other approaches combine transformer features with convolutional components to retain spatial information [29]. Models such as MedViT introduce hierarchical designs that integrate transformer blocks with feature processing for improved performance across different datasets [30]. These approaches highlight the potential of transformers to learn the global context in medical imaging data. However, transformer-based methods struggle to capture the fine-grained local details that are important for identifying fine-grained pathological changes.

2.1.3.3 Hybrid CNN–Transformer Methods

To address the limitations of individual models, recent work has focused on combining CNNs with transformers to leverage both local and global features. Such hybrid approaches aim to improve the representation quality by combining the complementary strengths of both architectures [31]. Several frameworks use parallel and sequential designs to combine convolutional and transformer features. For example, in methods where CNN branches extract local features and transformer modules capture global context, the resulting features are fused for the final prediction [32]. Another approach employs hierarchical architectures that integrate convolutional blocks with transformer layers to improve classification performance [33]. Hybrid models are also applied to cell classification and pneumonia detection

by combining multi-scale local features with global dependencies, which leads to improved results [34, 35]. By bridging the gap between local and global features, these hybrid architectures better mimic the workflow of human experts. However, it can also introduce redundancy by making it vital to develop efficient fusion modules that maintain the method's performance and interpretability.

2.1.3.4 Learning under Limited Annotations

A major limitation in medical image classification is the availability of labeled data. To address this, several studies have explored learning strategies effective with limited annotations. Semi-supervised learning methods leverage both labeled and unlabeled data to improve the performance of the model. Techniques such as consistency regularization, pseudo-labeling, and teacher-student frameworks are widely adopted [36, 37]. Methods like Mean Teacher and FixMatch improve the prediction stability by applying consistency under data augmentation [38]. Other recent approaches focus on handling imbalanced data distributions [39,40]. In medical imaging methods for tasks such as Knee Osteoarthritis (KOA) grading and Magnetic Resonance Imaging (MRI)-based analysis [41, 42] are complex in nature. In addition, class imbalance remains a key issue, especially in diseases like leukemia, where abnormal samples are very few. Techniques such as synthetic sample generation and distribution-aware learning have been proposed to reduce this problem [43–45]. Despite these efforts, effectively utilizing unlabeled data and handling imbalance in real clinical settings remain open challenges. This highlights the need for more robust learning frameworks that can generalize well under limited and imperfect supervision.

2.2 Medical Image Segmentation

2.2.1 Background

Computer-Aided Diagnosis (CAD) plays an important role in improving the efficiency and consistency of medical image analysis. Among its applications, medical image segmentation is essential for identifying regions of interest such as organs, tissues, and lesions. It is widely used to measure tissue volumes, study anatomical structures to support diagnosis, and assist

in surgical planning. Similar to general image segmentation, the goal is to divide the image into meaningful and non-overlapping regions based on specific criteria. However, tumor and lesion segmentation is more challenging than standard organ segmentation. Tumors often share visual characteristics with surrounding healthy tissue, which makes accurate localization difficult in high-resolution images. Manual annotation is time-consuming and relies on expert judgment, and leads to variability. Traditional methods and earlier approaches often struggle to handle such complexity. Recent DL models have improved segmentation performance, but challenges remain in identifying clear boundaries and handling variability across patients and imaging conditions. Accurate segmentation is important for estimating tumor size, tracking disease progression, and planning treatments such as radiotherapy. Therefore, developing reliable and precise segmentation methods remains an important requirement for effective Computer-Aided Diagnosis (CAD) systems.

2.2.2 Challenges

Developing accurate medical image segmentation methods remains difficult due to complex internal anatomy, especially for tumor and lesion analysis. One major limitation is the availability of labeled data, as expert annotation is time-consuming and often yields small datasets with class imbalance, in which abnormal regions occupy only a small portion of an image. In addition, medical images frequently suffer from low contrast-noise and variations in acquisition protocols. This also affects model performance while making reliable predictions. For example, in Computed Tomography (CT) and MRI scans, there are faint intensity differences between organs and surrounding tissues, which make segmentation more challenging. Variations in scanners and imaging settings also introduce inconsistencies across datasets. Tumor segmentation is further complicated by large variability in size, shape, location, and the number of lesions within and across diverse patients.

Another challenge lies in the models' ability to capture both local and global information effectively. Patch-based and convolutional approaches often focus on local regions, which leads to high computational cost and limited contextual understanding. However, recent DL models can extract rich features and select the information to maintain spatial consistency across the segmentation map. Boundary identification is also challenging due to blurred or indistinct organ boundaries, especially in colonoscopy images of liver tumors and polyps,

where surrounding tissues appear similar. Additional factors such as uneven illumination, noise artifacts, and occlusions further degrade the segmentation's overall quality. These challenges highlight the need for methods that handle the data variability with improved boundary accuracy and capture multi-scale information for accurate segmentation.

2.2.3 Existing Literature

Medical image segmentation has progressed from traditional geometric and rule-based methods to advanced DL approaches. Earlier techniques depended on handcrafted features and struggled with complex anatomical variations. Recent methods focus on accurately identifying organs, tissues, and lesions across different imaging modalities by learning features directly from data. This section reviews key developments in segmentation methods, with an emphasis on architectural designs that improve performance and handle clinical variability.

2.2.3.1 Encoder-Decoder based Methods

Encoder-decoder architectures form the foundation of many medical image segmentation methods because they can learn hierarchical feature representations. Such models utilize encoders to capture contextual information and decoders to remap spatial details, with skip connections that retain fine-grained features. This structure allows efficient learning of both global context and local spatial information simultaneously. The U-Net architecture and its variants are widely adopted in medical imaging tasks [46–49]. Recent work has extended this framework by introducing multi-scale decoding and feature-refinement strategies. For example, LMIS and X-Net improve segmentation by capturing variations in structure and texture [50, 51]. Dual-channel and quasi-recurrent encoder-decoder designs have also been proposed to model both spatial and contextual dependencies [52, 53]. Other approaches incorporate edge-aware mechanisms and multi-plane feature integration to improve segmentation quality [54, 55]. In addition, the slice-based fusion and 3D U-Net extensions have been explored to maintain anatomical consistency across volumetric data [56, 57]. Despite these improvements, the encoder-decoder model struggles with boundary preservation and is limited in capturing complex variations in tumor regions, especially in low-contrast and noisy settings.

2.2.3.2 Attention based Methods

Attention mechanisms enhance feature representation by focusing on relevant regions and by suppressing irrelevant information, thereby improving the model's ability to capture spatial and contextual relationships in medical images. These mechanisms help networks learn the diagnostically important structures from surrounding background tissues and imaging noise. Several studies incorporated attention blocks into segmentation networks to improve focus on the relevant region. For example, non-local attention and hybrid attention frameworks capture the long-range dependencies and improve the identification of tumor regions [58–60]. Models such as UgCmA-Net and SBCNet integrate attention with multi-task learning and boundary enhancement to improve segmentation performance [61, 62]. Other approaches include TD-Net and GCHA-Net that combine global and local attention modules to refine feature learning [63, 64]. Channel-based and spatial-based attention mechanisms, such as Squeeze-and-Excitation (SE) blocks and hybrid attention networks, are used to improve feature discrimination [65–67]. In polyp segmentation, attention-based models such as PraNet and FocusNet improve boundary localization by focusing on key regions [68, 69]. While attention-based methods improve the segmentation performance, they still overlook the structural and geometric relationships, which are important for identifying complex pathological regions and preserving anatomical consistency.

2.2.3.3 Multi-scale Methods

Multi-scale feature learning is important in medical image segmentation, as it captures information at different resolutions. This helps to handle variations in lesion size, shape, and appearance across imaging modalities. Recent approaches integrate multi-scale feature extraction within the encoder-decoder frameworks. For example, MSML-AttUNet and RMAU-Net combine attention mechanisms with multi-scale representations to improve segmentation accuracy [70, 71]. Other multi-scale methods utilize atrous convolution, pyramid pooling, and residual connections to extract features across multiple receptive fields [72, 73]. Methods such as FRA-Unet and KIU-Net further enhance segmentation by combining multi-stage processing and edge-aware feature learning [74, 75]. In the polyp segmentation, the multi-scale fusion approach improves the detection of small and irregular structures by integrating

features from different layers [76–78]. However, existing multi-scale methods struggle to integrate features across scales and maintain consistency between local and global contexts, thereby limiting segmentation performance in complex clinical settings.

2.3 Explainability in Medical Imaging

Explainability methods help understand and interpret DL decisions in medical imaging tasks. Several studies have explored the use of explainability techniques in medical image classification, including pneumonia detection from CXRs. For example, hybrid-based approaches combine convolutional networks and transformers to generate the attention maps, which support the interpretation of the prediction [79]. Other approaches remove irrelevant background regions and apply visualization methods such as Gradient-weighted Class Activation Mapping (Grad-CAM) to highlight important lung regions [80, 81]. In addition, multiple visualization techniques, including Ablation-CAM, Eigen-CAM, and Grad-CAM++, are used to analyze model behavior and improve prediction outcomes [82–85]. These methods help determine whether the model learns meaningful clinical features or irrelevant patterns that affect the final decision.

Despite these advances, many existing approaches rely on a single explanation technique and are limited to specific architectures, thereby failing to provide a complete understanding of final decisions. Visual explanations can also vary across methods, making interpretation difficult in practice. To address this, recent work explores combining multiple techniques, including Grad-CAM, Grad-CAM++, Score-CAM, and Local Interpretable Model-agnostic Explanations (LIME), for improved interpretability [81, 85–87]. These approaches aim to improve the interpretability and trust in automated systems. Overall explainability remains an active area in Explainable Artificial Intelligence (XAI), with ongoing efforts to provide consistent, clinically meaningful insights into the final model's predictions.

2.4 Summary

This chapter reviewed recent work in medical image classification and segmentation, focusing on advances in DL approaches. Although these methods have improved performance,

several challenges remain, including limited annotated data, gaps in feature representation, boundary precision, and variability across datasets. XAI techniques are also discussed to support precise clinical understanding. These findings highlight existing limitations and motivate the methods developed in the subsequent chapters.

Chapter 3

Coupled Transformer-Convolutional Network with an Adaptive Graph Reconstruction for Leukemia Detection

CONTENTS

3.1 Introduction . . . 24
3.2 Methodology . . . 25
3.2.1 Framework Overview . . . 25
3.2.2 Preprocessing . . . 25
3.2.3 Proposed Framework . . . 27
3.3 Experimental Settings . . . 37
3.3.1 Benchmark Datasets . . . 37
3.3.2 Evaluation Metrics . . . 39
3.4 Experimental Results . . . 40
3.4.1 Quantitative Analysis . . . 40
3.4.2 Qualitative Analysis . . . 43
3.4.3 Ablation Analysis . . . 44
3.4.4 Cross-Dataset Analysis . . . 48
3.4.5 Model Complexity Analysis . . . 49
3.5 Conclusion . . . 50

3.1 Introduction

This chapter focuses on leukemia classification to address the first objective of this thesis: developing global–local feature synthesis methods for medical image analysis. The goal is to design a framework that captures both global structural information and fine-grained cellular details for accurate classification. Leukemia detection from blood smear samples is challenging because leukemic and normal cells are visually similar. This includes staining variations and overlapping nuclei. In addition, differences in cell density and image quality further complicate feature extraction. Traditional approaches rely on local features and global context individually, but fail to learn them jointly. This limits the ability to capture complex hematological patterns within the cells. This highlights the need for precise methods that learn joint feature representations to improve the classification performance.

To address these challenges, this chapter presents the Coupled Transformer Convolutional Network (CoTCoNet) for leukemia classification. The framework begins with leukocyte preprocessing using the LeuSAM module to isolate relevant cell regions and enhance feature visibility. To mitigate data scarcity and class imbalance, the LeuGAN module generates high-quality synthetic leukocyte samples while preserving essential characteristics. The core framework leverages a transformer-based Global Feature Module with a convolution-based Spatial Feature Module to learn complementary representations. A Graph-based Feature Reconstruction module further refines these features by reconstructing hidden patterns, enabling richer feature representation. In addition, a population-based optimization strategy is employed to select the most discriminative features for classification.

The organization of this chapter is structured to provide a detailed technical analysis of the proposed research. It begins by detailing the methodology in Section 3.2, which covers the CoTCoNet pipeline, the LeuSAM segmentation module, and the architectural design of the coupled Global and Spatial Feature Modules. Section 3.3 defines the experimental settings, including a summary of the four benchmark datasets and the evaluation metrics employed. Section 3.4 presents the experimental results, providing a quantitative comparison against state-of-the-art methods and qualitative analysis. Finally, Section 3.5 summarizes the study's key findings and clinical implications.

3.2 Methodology

This section presents the proposed framework for leukemia classification. It describes the overall pipeline, including preprocessing and framework design. The framework's key components learn meaningful representations, as detailed in the following subsections.

3.2.1 Framework Overview

This section includes a comprehensive breakdown of the proposed **Co**upled **T**ransformer **Co**nvolution **Net**work (CoTCoNet) framework, as illustrated in Figure 3.1, which comprises five distinct sub-modules: (A) leukocyte pre-processing and segmentation undertakes the initial processing and segmentation of cell images; (B) generative adversarial network-driven cell synthesis responsible for generating synthetic cells; (C) global and spatial deep feature extraction extracts joint global and spatial features, coupled with the identification of hidden features; (D) population-based optimization for feature selection selects and refines the most optimized set of features; (E) hierarchical deep learning classifier categorizes cells as either normal or leukemic.

3.2.2 Preprocessing

In this section, we explore the preprocessing of leukocytes, including enhancement and segmentation, as illustrated in Figure 3.1. For the enhancement of leukocytes, we first employ Contrast Limited Adaptive Histogram Equalization Contrast Limited Adaptive Histogram Equalization (CLAHE) for image contrast amplification and visibility enhancement. After that, we sharpen filters to emphasize the edges and features of the microscopic images. Secondly, we identify and capture Region-of-Interest Region of Interests (ROIs) that are more likely to exhibit variations indicative of leukemia while excluding any superfluous sections such as plasma. Subsequently, we leverage the fine-tuned Leukemia Segment Anything Model (LeuSAM) to perform segmentation and extraction of leukocyte cells from the ROI images, resulting in segmented leukocytes $\mathbf{x}_s$.

Fine-tuned LeuSAM: Our proposed LeuSAM model adapts the pre-trained Segment Anything Model Segment Anything Model (SAM) [88] for the task of leukocyte cell segmen-

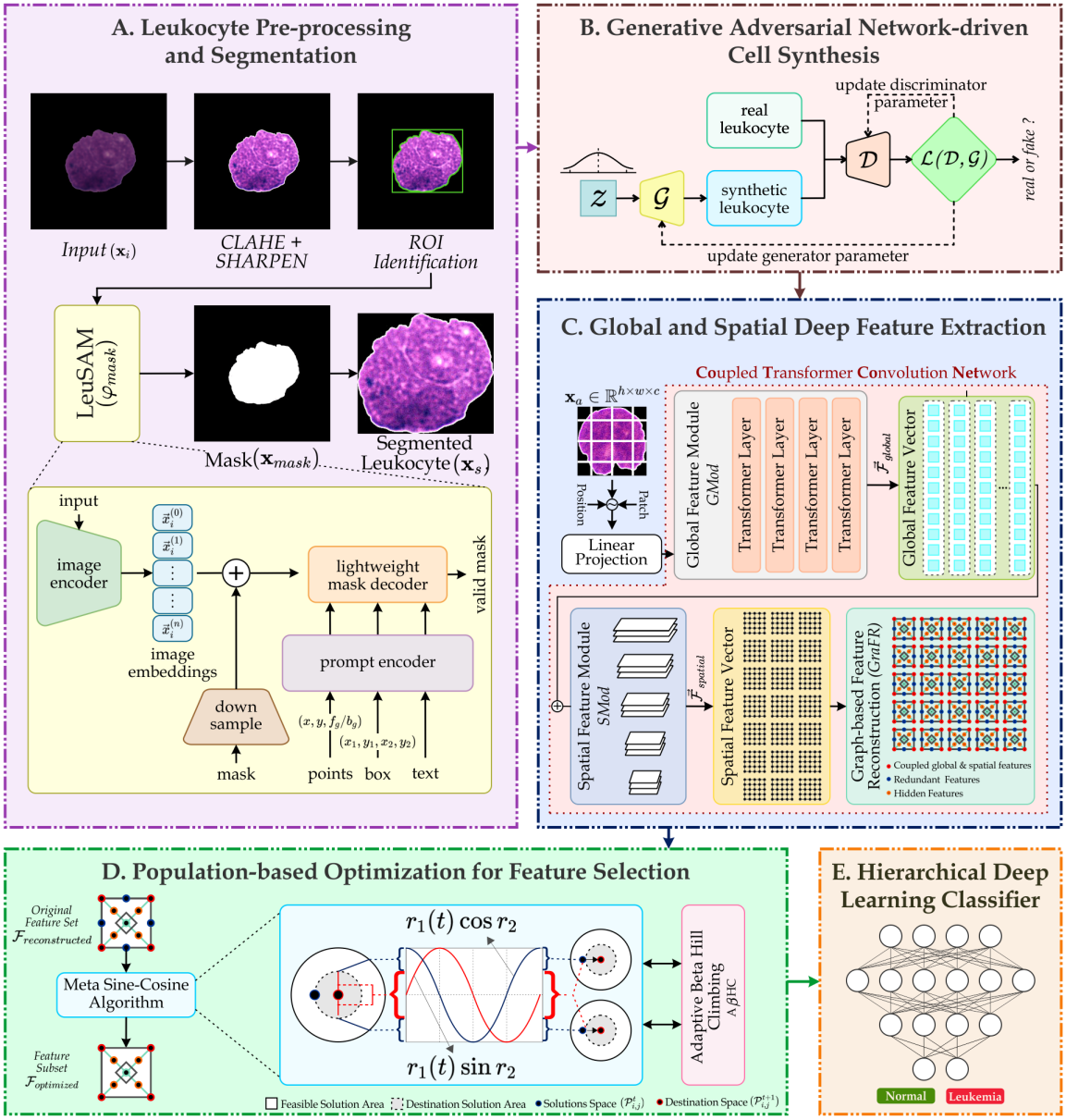


Figure 3.1: Proposed architecture of Convolutional Network (CoTCoNet) for leukemia detection in whole slide imaging.

tation. SAM incorporates an image encoder, a prompt encoder, and a mask decoder. It was originally trained on the high-resolution image and segmentation mask dataset, SA-1B [88]. While training the entire SAM offers a powerful approach, we prioritize efficiency by fine-tuning only the mask decoder. This leverages the pre-trained image encoder's feature extraction capabilities while freezing its weights to prevent overfitting. The fine-tuned decoder translates the extracted features into accurate pixel-wise segmentation masks for blood cells within blood smear samples. This approach enables LeuSAM to effectively transfer its

learned feature extraction abilities from the general domain to the specific task of blood cell segmentation while simultaneously refining the mask generation process.

3.2.3 Proposed Framework

This section unveils the arranged configuration of the Coupled Transformer Convolutional Network (CoTCoNet) framework, as illustrated in Figure 3.1. The CoTCoNet comprises three modules: Global Feature Module Global Feature Module (GMod), Spatial Feature Module Spatial Feature Module (SMod), and Graph-based Feature Reconstruction Graph-based Feature Reconstruction (GraFR) module. This framework involves leveraging a well-designed transformer-based GMod to capture the overarching global features effectively. At the same time, the SMod module capitalizes on the extraction capabilities of the Convolutional Neural Network (CNN) by inheriting the spatial features. Subsequently, the combined features undergo a reconstruction process utilizing a Graph-based Feature Reconstruction (GraFR) module to extract the pertinent features of each data point by evaluating the similarity score of feature vectors. These three modules form a cohesive network that collaboratively leverages the extracted features.

3.2.3.1 Global Feature Module

We design a transformer-based network called Global Feature Module (GMod) that focuses on extracting global information from the input images to classify leukemic cells accurately. The GMod module is schematically illustrated in Figure 3.2.

(a) Patch Decomposition: The standard transformer accepts a 1-D sequence of token embeddings as its primary input; however, the manipulation of 2-D synthetic leukocyte images involves a reshaping operation of input $\mathbf{x}_a \in \mathbb{R}^{h \times w \times c}$ to $\mathbf{x}_p \in \mathbb{R}^{n \times (p^2 \cdot c)}$, where (h, w, c) corresponds to height, width, channels and (p, p) indicates the patch size. The overall count of these patches, referred to as n, is computed as $(h \times w)/p^2$, furnished as input to the stacked transformer layers.

(b) Linear Transformation and Positional Embedding: We apply the trainable linear projection to map the vectorized token patches $\mathbf{x}_p$ into a latent embedding space with d_L dimensionality. The token sequence denoted as $\mathbf{x}_p = [\mathbf{x}_p^{(1)}, \mathbf{x}_p^{(2)} \cdots \mathbf{x}_p^{(N)}]$ is initially structured as an

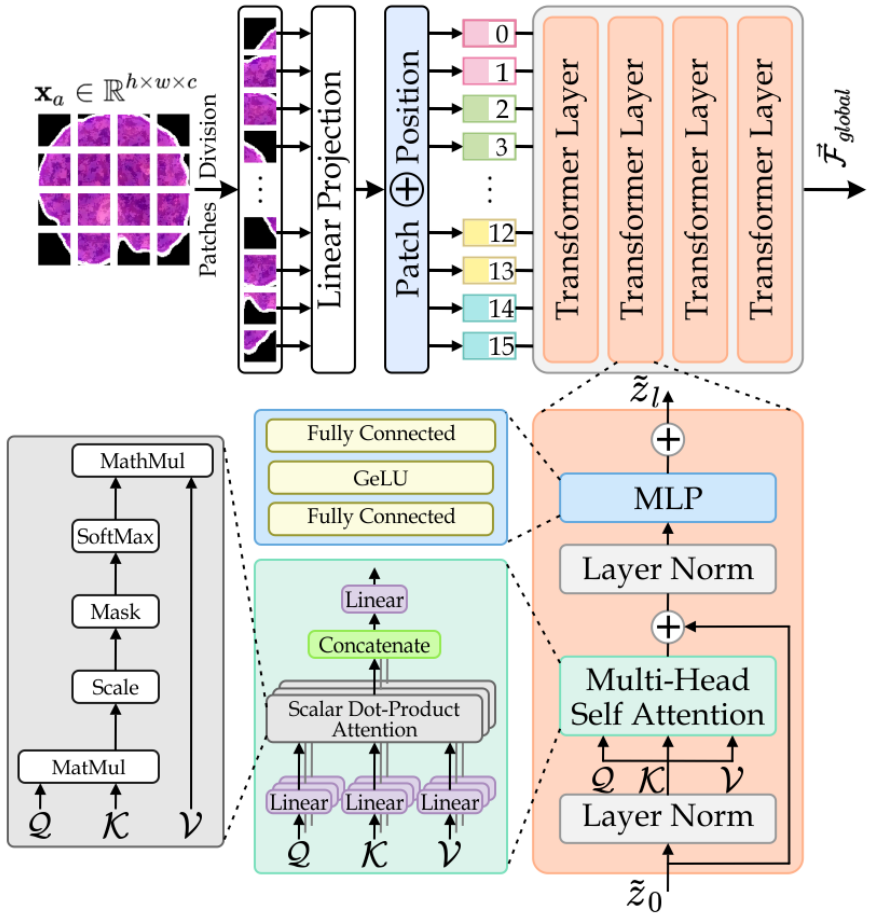


Figure 3.2: Proposed architecture of Global Feature Module (GMod).

array after incorporating the learnable embedding and the position embedding the resultant $\tilde{z}_0$, [refer Equation (3.1)] is directed towards the transformer, whose output is given by $\tilde{z}_L$, where L signifies the number of transformer blocks. Subsequently, we integrate position embeddings into the token embedding to inherit the positional information as follows:

$$\tilde{z}_0 = [\mathbf{x}_{class}\mathbf{E};\ \mathbf{x}_p^{(1)}\mathbf{E};\ \mathbf{x}_p^{(2)}\mathbf{E};\ \cdots;\ \mathbf{x}_p^{(N)}\mathbf{E}] + \mathbf{E}_{pos} \tag{3.1}$$

where, $\mathbf{E} \in \mathbb{R}^{(p^2 \cdot c) \times d_L}$ denotes the patch embedding projection, and $\mathbf{E}_{pos} \in \mathbb{R}^{(n+1) \times d_L}$ represents the positional embedding.

(c) Self-Attention Mechanism: The resulting token embeddings $\tilde{z}_0$ are introduced as input to a transformer that consists of a Multihead Self-Attention Multi-head Self-Attention (MSA) blocks and Multi-Layer Perceptrons (MLP) blocks [refer Equation (3.2) and Equation (3.3)]. The MLP block includes a dual fully connected dense layer with a Gaussian

error Linear Unit (GeLU) residing between them. The Layer Normalization (LN), as defined in Equation (3.4), is employed prior to the MSA and MLP blocks, while residual concatenation is utilized after each block. The procedure for the forward calculation is outlined as follows.

$$\tilde{z}_l = MSA\big(LN(z_{l-1})\big) + z_{l-1}, \quad l = 1, \ldots, L \tag{3.2}$$

$$z_l = MLP\big(LN(\tilde{z}_l)\big) + \tilde{z}_l, \quad l = 1, \ldots, L \tag{3.3}$$

$$y = LN(z_L^0) \tag{3.4}$$

The core mechanism of the transformer is MSA, which aims to capture correlated global contextual information effectively. To utilize diverse implications, three adaptable weight matrices ($\mathbf{W}^{\mathcal{Q}}$, $\mathbf{W}^{\mathcal{K}}$, $\mathbf{W}^{\mathcal{V}}$) are established. Simultaneously, $\mathbf{W}^{\mathcal{Q}} \in \mathbb{R}^{h\times w\times c}$, $\mathbf{W}^{\mathcal{K}} \in \mathbb{R}^{h\times w\times c}$, $\mathbf{W}^{\mathcal{V}} \in \mathbb{R}^{h\times w\times c}$ are initialized randomly. Leveraging these trainable weight matrices facilitates a linear transformation of tokens into a 3-D invariant matrix comprising Queries ($\mathcal{Q} \in \mathbb{R}^{h\times w\times c}$), Keys ($\mathcal{K} \in \mathbb{R}^{h\times w\times c}$), and Values ($\mathcal{V} \in \mathbb{R}^{h\times w\times c}$). Subsequently, the computation of attention scores involves the entire information encompassed within $\mathcal{Q}$, $\mathcal{K}$, $\mathcal{V}$, and the weights are calculated using the softmax function. Concisely, the SA mechanism can be encapsulated in Equation (3.5):

$$SA = Attention(\mathcal{Q}, \mathcal{K}, \mathcal{V}) = softmax\left(\frac{\mathcal{Q}\mathcal{K}^T}{\sqrt{d_{\mathcal{K}}}}\right)\mathcal{V} \tag{3.5}$$

where, $d_{\mathcal{K}}$ is the dimension of $\mathcal{K}$. The MSA module is equipped with multiple weight matrices to facilitate the mapping of $\mathcal{Q}$, $\mathcal{K}$, and $\mathcal{V}$ components; this plays a crucial role in computing MSA values through uniform operational procedures. Then, the outcomes from individual attention heads are combined through concatenation. The linear transformation represented in Equation (3.6) is the final step in deriving the MSA outcomes:

$$MSA(\mathcal{Q}, \mathcal{K}, \mathcal{V}) = Concat(SA_1, SA_2, \ldots, SA_H)\mathbf{W} \tag{3.6}$$

where, H represent total number of heads, $\mathbf{W} \in \mathbb{R}^{h\times d_{\mathcal{K}}\times d_{\mathcal{W}}}$ denotes the parameter matrix of he linear layers and $d_{\mathcal{W}} = N + 1$ corresponds to token number.

The pre-trained MSA module generates the feature matrix and is directed into the MLP

module. Also, the MLP is followed by Layer Normalization (LN) to amplify gradient magnitudes, mitigating the gradient vanishing concerns and accelerating the training process. In the transformer module, the dimensions of the input z_{l-1} and the output $\tilde{z}_{\updownarrow}$ from the l^{th} transformer module are identical. The trainable embedding z_L^0, comprising a set of global leukocyte features $\vec{\mathcal{F}}_{global}$ is fed into the Spatial Module consisting of several convolutions for capturing spatial features.

3.2.3.2 Spatial Feature Module

Transformer-based Global Feature Module (GMod) excels at capturing global features, but it has limitations when learning intricate spatial patterns as effectively as Convolutional Neural Networks (CNNs). To address this, CoTCoNet incorporates a convolution-based Spatial Feature Module (SMod) designed to efficiently extract spatial features at various scales. As illustrated in Figure 3.3, the SMod architecture comprises five cascaded Spatial Feature Learning (SFL) blocks denoted by $\varphi_{SFL(i)}$, where $i \in [1, 2, 3, 4, 5]$. This design allows for concurrent extraction and processing of spatial information alongside the global features obtained from GMod.

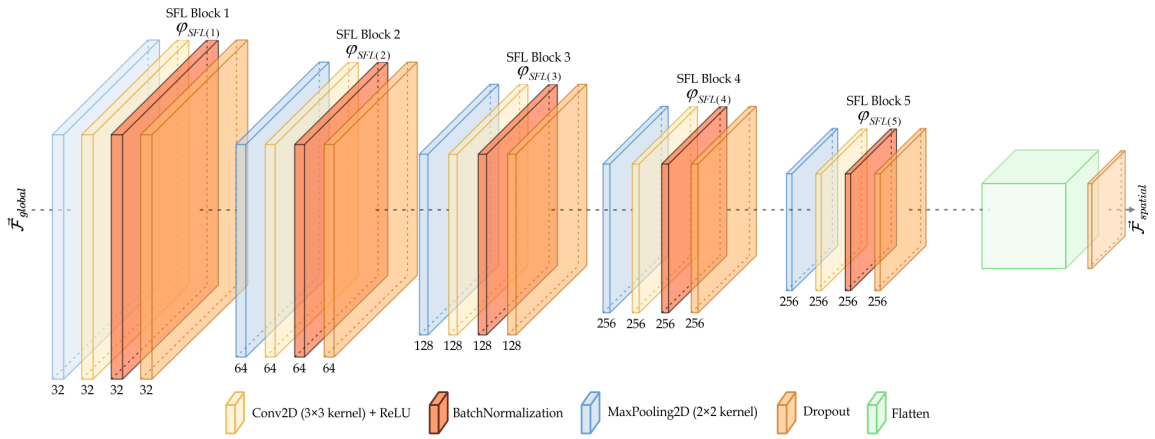


Figure 3.3: Proposed architecture of Spatial Feature Module (SMod).

The SFL block's functionality is mathematically formalized in Equation (3.7), represented by $\varphi^{f}_{SFL(i)}(\ddot{\mathcal{C}})$. This denotes the output of an SFL block with f filters. Within this equation, $\varkappa_s^{\tilde{k},\tilde{k}}$ represents a max-pooling operation with a kernel size of $\tilde{k} \times \tilde{k}$ and stride s, while $\Psi^{k,k,f}_{SFL(i)}(\ddot{\mathcal{C}})$ designates a convolution operation applied to the input feature map $(\ddot{\mathcal{C}})$ with f filters and a kernel size of $k \times k$. Additionally, β represents the batch normalization

layer, and ρ denotes the ReLU activation function.

$$\varphi^{f}_{SFL(i)}(\ddot{\mathcal{C}}) \;=\; \varkappa_2^{2,2}\left(\rho\left(\beta\left(\Psi^{3,3,f}(\ddot{\mathcal{C}})\right)\right)\right) \tag{3.7}$$

The SFL block extracts spatial features through a series of operations. First, a max-pooling layer (kernel size of 2×2, stride value 2) downsamples the input, reducing its spatial dimensions. Next, a convolutional layer with a 3×3 kernel and a variable number of filters (f_i where $i \in [32, 64, 128, 256, 256]$) extracts fine-grained spatial features like textures, corners, and edges from the downsampled input. Batch normalization then stabilizes training and improves generalization by normalizing activations. ReLU activation introduces non-linearity, enhancing feature representation. Finally, a dropout layer (rate of 0.2) mitigates overfitting by randomly deactivating neurons, promoting robust feature learning. The concatenated output of the SFL blocks yields a rich spatial feature representation ($\vec{\mathcal{F}}_{spatial}$) for the Graph-based Feature Reconstruction module (GraFR). GraFR leverages aggregated information from SFL layers to improve feature quality, ultimately enhancing the framework's leukocyte cell classification accuracy.

3.2.3.3 Graph-based Feature Reconstruction

The morphological similarity of leukocytes creates challenges for accurate classification. To address this, a Graph-based Feature Reconstruction module (GraFR) is employed. GraFR upscales lower-dimensional feature maps, revealing hidden or unobserved features. It enhances the representation of each node's features by leveraging the features of neighboring nodes. GraFR bridges local and global feature information by connecting lower-level and higher-level feature maps. The GraFR module follows four key steps: (1) Generating nodes from feature maps, (2) Calculating edge weights, (3) Selecting hidden features, and (4) Reconstructing the feature map.

(a) Node Generation for Feature Maps: Let $\mathbb{G}(V_g, E_g)$ be the graph, where V_g is the set of nodes, and E_g is the set of edges. Each node v in the graph is associated with a one-dimensional (1D) flattened array x_v that contains coupled global and spatial features. This can be represented as $\vec{x_v} = [x_{v,1}, x_{v,2}, \cdots, x_{v,d}]$ where d represents the dimensionality of the feature vectors.

(b) Edge Weight Computation: The edge weights in the graph are calculated using the Euclidean distance between the feature vectors of nodes. This means that for any two nodes u and v in V_g, the Euclidean distance $\ddot{d}_{uv}$ between their feature vectors x_u and x_v is given in Equation (3.8):

$$\ddot{d}_{uv}(u,v) = \sqrt{\sum_{i=1}^{\bar{N}} \|\vec{x}_{u_i} - \vec{x}_{v_i}\|^2} \tag{3.8}$$

where, $\vec{x}_{u_i}$ and $\vec{x}_{u_i}$ are the i^{th} elements of x_u and x_v, and $\bar{N}$ represents the number of of nodes, respectively.

(c) Hidden Feature Selection Criteria: The criteria for selecting the hidden features $\hbar$ aim to identify nodes with the highest similarity scores. The node incorporates global and spatial features while considering all features from lower-dimensional representations. The similarity score $\nabla_{u \leftrightarrow v}$ between nodes u and v is the inverse of the edge weight, as shown in Equation (3.9).

$$\nabla_{u \leftrightarrow v} = \left(\frac{1}{\ddot{d}_{uv}}\right) \tag{3.9}$$

A higher similarity score indicates a stronger similarity between core and hidden features. For inclusion of the feature node, each node i in V_g is considered if it possesses the highest similarity score across all other nodes in V_g, as shown in the following Equation (3.10).

$$\hbar = \{i \in V_g \mid \forall v \in V_g,\ (\nabla_{i \leftrightarrow v} \geq \nabla_{v \leftrightarrow i})\} \tag{3.10}$$

The set of hidden features $\hbar$ includes node i if this condition is satisfied.

(d) Feature Map Reconstruction: The output of the Feature Map Reconstruction is a vector $\mathcal{F}_{reconstructed}$ that summarizes all features and highlights hidden features of the nodes in the graph. This can be obtained by weighing the contribution of each node's feature vector based on the similarity score. Equation (3.11) is used to compute the reconstructed features $\mathcal{F}_{reconstructed}$.

$$\mathcal{F}_{reconstructed} = \frac{\sum_{v=V_g} \nabla_{u \leftrightarrow v} \cdot x_v}{\sum_{v=V_g} \nabla_{u \leftrightarrow v}} \tag{3.11}$$

This reconstructed feature $\mathcal{F}_{reconstructed}$ incorporates information from all nodes in the graph, with nodes that have higher similarity scores contributing more significantly to the final representation of reconstructed features. The algorithmic overview of the CoTCoNet is outlined

in Algorithm 3.1. We first extract global features from the input leukocyte samples using lines 1-14. In this stage, L and H represent the number of transformer layers and the number of heads employed within the transformer architecture, respectively. Subsequently, lines 15-24 focus on extracting spatial features by utilizing the Spatial Feature Learning (SFL) blocks. Finally, lines 25-31 perform the reconstruction of the integrated global and spatial features. The calling sequence for the GMod, SMod, and GraFR functions is then presented in lines 32-34.

3.2.3.4 Population-based Optimization for Feature Selection

This section explains how a population-based meta-heuristic algorithm is applied for feature selection and optimization. Since population-based optimization techniques randomly search for the optima of optimization problems, there is no certainty of discovering a solution in a single execution. However, as the optimization iterations and the number of random solutions increase, the likelihood of identifying the global optimum rises. Furthermore, a local search technique called Adaptive β-Hill Climbing (AβHC) is employed to refine the obtained feature subset. The optimized feature set learns to map between the feature set and output classes. This mapping facilitates the Hierarchical Deep Learning Classifier Hierarchical Deep Learning Classifier (HDLC) to perform the final distinction between normal cells and leukocyte cells, leveraging the optimized feature set as its input.

In this work, a population-based meta-heuristic algorithm, namely Sine-Cosine Algorithm Sine-Cosine Algorithm (SCA) [89], is applied for feature selection and optimization. SCA follows iterative stochastic and population-based methods that imitate the harmonious behavior of sine and cosine functions to explore the solution space and find the optimal destination space. SCA typically consists of two primary stages: exploration and exploitation. The exploration phase combines random solutions with a high degree of randomness to identify promising regions within the search space. In contrast, the exploitation phase alters the solutions with gradual modifications to reduce randomness in the variations.

To initiate the optimization procedure using SCA, the search element alters its self-

Algorithm 3.1 COUPLED TRANSFORMER CONVOLUTIONAL NETWORK

Input: Leukocyte samples $\mathbf{x}_a \in \mathbb{R}^{h \times w \times c}$.
Output: Reconstructed features $\vec{\mathcal{F}}_{reconstructed}$

function GMOD ($\mathbf{x}_a \in \mathbb{R}^{h \times w \times c}$)
  $\mathbf{x}_p^{n \times (p^2 \cdot c)} \Leftarrow \mathbf{x}_a^{h \times w \times c}$
  $\tilde{z}_0 \Leftarrow [\mathbf{x}_{class}\mathbf{E};\ \mathbf{x}_p^{(1)}\mathbf{E};\ \mathbf{x}_p^{(2)}\mathbf{E};\ \cdots;\ \mathbf{x}_p^{(N)}\mathbf{E}] + \mathbf{E}_{pos}$
  for $l = 1\ to\ L$ **do**
    for $h = 1\ to\ H$ **do**
      $\mathcal{Q}_h^l \Leftarrow \mathbf{W}_l^{\mathcal{Q}} \cdot z_{l-1}$; $\mathcal{K}_h^l \Leftarrow \mathbf{W}_l^{\mathcal{K}} \cdot z_{l-1}$;
      $\mathcal{V}_h^l \Leftarrow \mathbf{W}_l^{\mathcal{V}} \cdot z_{l-1}$
      $SA_h^l \Leftarrow \textit{softmax}\left(\frac{(\mathcal{Q}_h^l)\cdot(\mathcal{K}_h^l)^T}{\sqrt{d_{\mathcal{K}}}}\right) \cdot \mathcal{V}_h^l$
    end for
    $\tilde{z}_l \Leftarrow \textit{Concat}\left(SA_1^l, SA_2^l \ldots SA_H^l\right)$
    $z_l = \textit{MLP}(LN(\tilde{z}_l)) + \tilde{z}_l$
  end for
  $\vec{\mathcal{F}}_{global} \Leftarrow LN\left(\textit{Concat}(z_1^l, z_2^l \ldots z_H^l\right)$
  return $\vec{\mathcal{F}}_{global}$
end function
function SMOD ($\vec{\mathcal{F}}_{global}$)
  for Each $\varphi_{SFL(i)},\ i \in \{1, 2 \ldots 5\}$ **do**
    $\textit{MXP} \Leftarrow \varkappa_2^{2\times 2}\ [\vec{\mathcal{F}}_{global}]$
    $\textit{CONV} \Leftarrow \Psi^{3,3,f_i}(\ddot{\mathcal{C}})$ [*MXP*],
    where $f_i \in \{32, 64, 128, 256, 256\}$
    $\textit{BN} \Leftarrow \beta$ [*CONV*]
    $\textit{DRP} \Leftarrow \textit{BN}$
  end for
  $\vec{\mathcal{F}}_{spatial} \Leftarrow Flatten$ [*DRP*]
  return $\vec{\mathcal{F}}_{spatial}$
end function
function GRAFR ($\vec{\mathcal{F}}_{spatial}$)
  $\vec{x_v} \Leftarrow [x_{v,1}, x_{v,2}, \cdots, x_{v,d}]$
  $\ddot{d}_{uv}(u, v) \Leftarrow \sqrt{\sum_{i=1}^{\bar{N}} \|\vec{x}_{u_i} - \vec{x}_{v_i}\|^2}$
  $\hbar \Leftarrow \{i \in V_g \mid \forall v \in V_g,\ (\nabla_{i\leftrightarrow v} \geq \nabla_{v\leftrightarrow i})\}$,
  where $\nabla_{u\leftrightarrow v} \Leftarrow \left(\frac{1}{\ddot{d}_{uv}}\right)$
  $\mathcal{F}_{reconstructed} \Leftarrow \frac{\sum_{v=V_g} \nabla_{u\leftrightarrow v} \cdot x_v}{\sum_{v=V_g} \nabla_{u\leftrightarrow v}}$
  return $\mathcal{F}_{reconstructed}$
end function
$\vec{\mathcal{F}}_{global} \leftarrow$ GMOD($\mathbf{x}_a \in \mathbb{R}^{h \times w \times c}$)
$\vec{\mathcal{F}}_{spatial} \leftarrow$ SMOD($\vec{\mathcal{F}}_{global}$)
$\mathcal{F}_{reconstructed} \leftarrow$ GRAFR($\vec{\mathcal{F}}_{spatial}$)

position based on the sine and cosine functions as given in Equation (3.12).

$$\mathcal{P}_{i,j}^{t+1} = \begin{cases} \mathcal{P}_{i,j}^{t} + r_{1,j}^{t} \times \sin\left(r_{2,j}^{t}\right) \times |r_{3,j}^{t} D_{j}^{t} - \mathcal{P}_{i,j}^{t}|, r_{4,j}^{t} < 0.5 \\ \mathcal{P}_{i,j}^{t} + r_{1,j}^{t} \times \cos\left(r_{2,j}^{t}\right) \times |r_{3,j}^{t} D_{j}^{t} - \mathcal{P}_{i,j}^{t}|, r_{4,j}^{t} \geq 0.5 \end{cases} \tag{3.12}$$

Here, $\mathcal{P}_{i,j}^{t}$ is the position of current solution in j^{th} dimension of i^{th} search element at t^{th} iteration. $r_{2,j}^{t}$, $r_{3,j}^{t}$ and $r_{4,j}^{t}$ are uniformly distributed random numbers, D_{j}^{t} represents the position of j^{th} dimension of destination point (best solution) at t^{th} iteration and $||$ signify the absolute value. A random number $r_{1,j}^{t}$ facilitates the transition from exploration to exploitation of the search space, which is determined by Equation (3.13).

$$r_{1,j}^{t} = \alpha - t\frac{\alpha}{T} \tag{3.13}$$

Here, α, t and T represents the constant number, the t^{th} iteration and total number of iterations, respectively.

The value of $r_{1,j}^{t}$ decides if the search area is for exploitation (destination solution region) $(r_{1,j}^{t} \in [-1, 1])$ or exploration (feasible solution region) $(r_{1,j}^{t} \in [-1, -2]$ or $r_{1,j}^{t} \in [1, 2])$. Meanwhile, the stochastic variable $(r_{2,j}^{t}$ defines the search agent's movement towards or away from the destination point, bounded within $[0, 2\pi]$, in sync with a complete cycle of sine and cosine functions. $(r_{3,j}^{t}$ balances the exploration and exploitation rates by bringing a random weight between $(0, 2)$. Furthermore, $r_{3,j}^{t}$ introduces a stochastic step size for the destination point, either emphasizing $(r_{3,j}^{t}$>1) or not emphasizing $(r_{3,j}^{t}$<1) its impact. Finally, the parameter $r_{4,j}^{t}$ evenly transitions between the sine and cosine components, as given in Equation (3.12). The flowchart of feature optimization utilizing the SCA algorithm is presented in Figure 3.4.

After the identification of the most efficient features through the meta-heuristic algorithm SCA, further enhancement of exploitation ability can be achieved by integrating the local search technique named Adaptive β-Hill Climbing (AβHC). AβHC is a feature optimization algorithm utilizing local search-based techniques. These search techniques are guided by a pair of control parameters $\mathcal{N}_{\text{HC}}$ and β_{HC}, respectively. By adjusting these parameters, the search technique finds the optimal trade-off between exploitation and exploration.

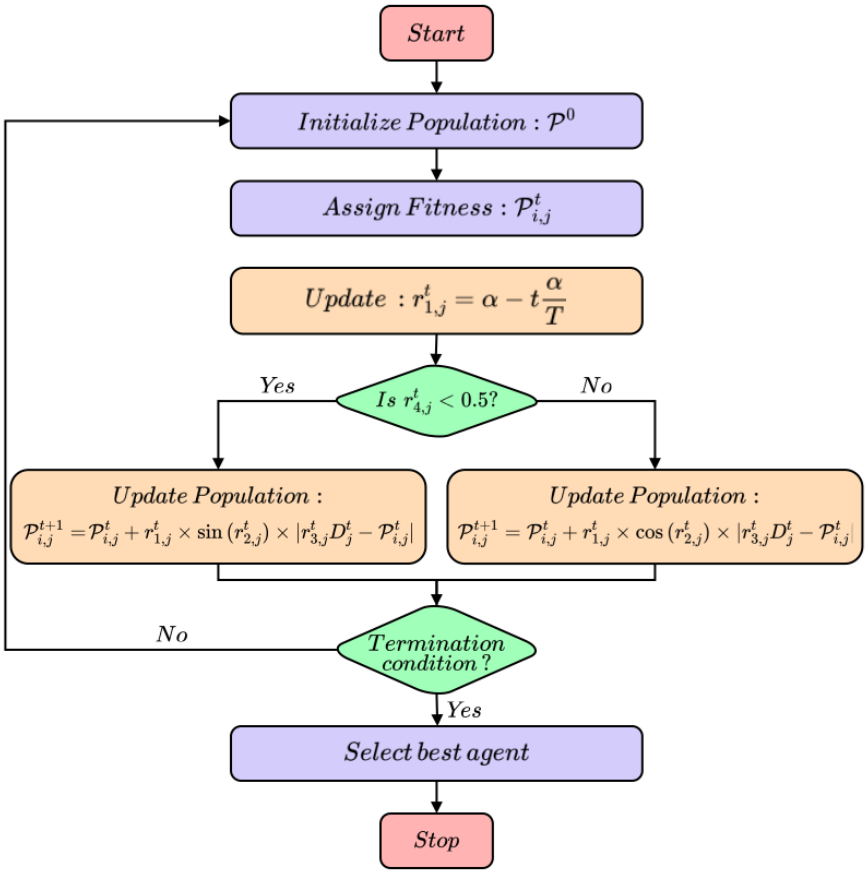


Figure 3.4: Flowchart of the Sine Cosine Algorithm (SCA).

Fine-tuning these parameters plays a significant role in the optimization because it helps to enhance the convergence rate. The parameter $\mathcal{N}_{\text{HC}}$ is initially set to a value close to 1, but it is gradually decreased as the search process iterates. This allows the algorithm to dynamically adjust $\mathcal{N}_{\text{HC}}$ to improve the search performance, as given in Equation (3.14).

$$\mathcal{N}^t_{\text{HC}} = 1 - \frac{t^{\frac{1}{P}}}{T^{\frac{1}{P}}_{max}} \tag{3.14}$$

Here, $\mathcal{N}^t_{\text{HC}}$ represents the value of $\mathcal{N}_{\text{HC}}$ at time t, P is constant used to linearly decrease the value of $\mathcal{N}_{\text{HC}}$ to a value close to 0 and T_{max} represents the upper limit of iterations for AβHC algorithm.

Moreover, the β parameter undergoes deterministic adaptation within a defined range $\in [\beta^{min}_{\text{HC}}, \beta^{max}_{\text{HC}}]$, mathematically expressed in Equation (3.15).

$$\beta^t_{\text{HC}} = \beta^{min}_{\text{HC}} + t \times \frac{\beta^{max}_{\text{HC}} - \beta^{min}_{\text{HC}}}{T_{max}} \tag{3.15}$$

Here, β_{HC}^{t} denotes the rate of β_{HC} at time t, β_{HC}^{min} and β_{HC}^{max} represents the minimum and maximum value of β_{HC} respectively, T_{max} is the total number of iterations and t signifies the current time.

3.2.3.5 Hierarchical Deep Learning Classifier

The Hierarchical Deep Learning Classifier (HDLC) is a multi-stage DL classifier that utilizes the potential of the optimized reconstructed feature set $\mathcal{F}_{reconstructed}$. HDLC focuses on the hematological features of cells to effectively classify both normal and leukocyte cells. HDLC is structured into two phases: an initial convolutional phase and a subsequent fully-connected phase. During the initial phase, a 1D convolutional layer employs 128 filters with dimensions of 3×3. It computes the dot product between these filters and the input feature set denoted by $\mathcal{F}_{reconstructed}$. The convolution layer internally employs a non-linear *ReLU* activation function to preserve the essential information. The output of a 1D convolution undergoes a linear transformation to produce a one-dimensional vector and is subsequently input into a sequence of six fully connected dense layers in the second stage. The neurons in the fully connected layer learn non-linear associations between input features and the desired output. The final output can be interpreted as the likelihood of the input being a leukocyte cell, which is the final prediction of the classifier.

3.3 Experimental Settings

This section describes the datasets used to evaluate the proposed CoTCoNet framework for leukemic cell classification. It provides details of the benchmark dataset and outlines the evaluation metrics used to assess model performance.

3.3.1 Benchmark Datasets

To evaluate the generalizability and performance of the proposed CoTCoNet framework, we employed four publicly available datasets, namely, C-NMC 2019 [90], ALL-IDB1 [91], ALL-IDB2 [91], and SN-AM datasets [92]. Figure 3.5 provides visual examples of random leukocyte samples used in the CoTCoNet evaluation. For a comprehensive assessment, we randomly partitioned the leukocyte samples within each dataset into an 80% for training,

10% for validation, and 10% independent testing split, as detailed in Table 3.1. We now provide a brief summary of each dataset.

Table 3.1: An overview of benchmark datasets for leukemia classification.

Dataset	Original Leukocytes		Synthetic Leukocytes		Total Leukocytes	
	Cancer Class	Normal Class	Cancer Class	Normal Class	Cancer Class	Normal Class
C-NMC 2019	8491	4037	NIL	4454	8491	8491
ALL-IDB1	49	59	NIL	NIL	49	59
ALL-IDB2	130	130	NIL	NIL	130	130
SN-AM	30	30	NIL	NIL	30	30

'NIL' represents the absence of synthetic leukocyte generation.

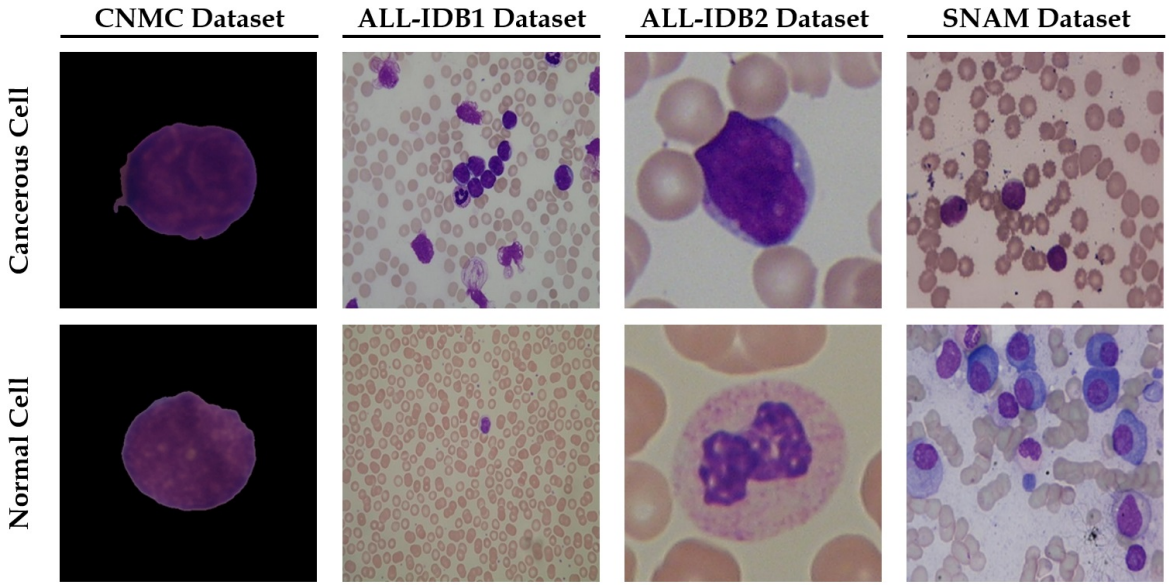


Figure 3.5: Sample dataset images for leukemia classification.

(a) C-NMC 2019: It is also known as the ALL Challenge dataset of ISBI 2019, a collection of single-cell microscopic images of bone marrow from healthy participants and participants with Acute Lymphoblastic Leukemia (ALL). C-NMC 2019 was created to broaden the scope of research on differentiating leukemic blasts from normal cells, as both appear similar under a microscope. The leukemic cells in the dataset are labeled as "all", and normal cells are labeled as "hem".

(b) ALL-IDB1: The acute Lymphoblastic Leukemia Image Database (ALL-IDB1) for the Image Processing initiative resulted in datasets for segmentation and image classification.

It consists of multi-cell blood microscopic images and is divided into two main categories: ALL-positive (blood samples with leukemic cells), labeled as "ImXXX_1.jpg" and ALL-negative (blood samples from non-ALL patients), labeled as "ImXXX_0.jpg". Expert oncologists have labeled the lymphocytes in these blood samples.

(c) ALL-IDB2: The images in this dataset are cropped areas of interest of normal and leukemic cells from ALL-IDB1. It is a single-cell blood microscopic image for evaluating leukemic cell classification systems.

(d) SN-AM: The microscopic images in this dataset are collected from patients using a bone marrow aspiration procedure. The multi-cell within bone marrow slides of patients diagnosed with B-lineage Acute Lymphoblastic Leukemia (B-lineage Acute Lymphoblastic Leukemia (B-ALL) and Multiple Myeloma (MM).

3.3.2 Evaluation Metrics

To assess the performance of our framework, we adopt widely used evaluation metrics for leukemic cell classification, such as accuracy, precision, recall, and $F1$-Score. We focus on accuracy and $F1$-Score because our dataset is not skewed towards the majority or minority class. We also consider precision and recall to visualize the model's ability to make accurate positive predictions and identify all relevant positive instances. Table 3.2 shows the mathematical representations of the metrics.

Table 3.2: Evaluation metrics of the CoTCoNet framework for leukemia detection in whole slide imaging.

Evaluation Metrics	**Mathematical Formulation**
Accuracy	$\frac{TP_{leuk} + TN_{nor}}{TP_{leuk} + FP_{leuk} + TN_{nor} + FN_{nor}}$
Precision	$\frac{TP_{leuk}}{TP_{leuk} + FP_{leuk}}$
Recall	$\frac{TP_{leuk}}{TP_{leuk} + FN_{nor}}$
F1-Score	$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

Here, TP_{leuk} is True Positive, a leukemia sample correctly predicted as leukemia. TN_{nor} is True Negative, a non-leukemia sample correctly predicted as non-leukemia. FP_{leuk} is False

Positive, a non-leukemia sample incorrectly predicted as leukemia. FN_{nor} is False Negative, a leukemia sample incorrectly predicted as non-leukemia.

3.4 Experimental Results

This section presents the experimental results of the proposed method. It includes quantitative evaluation against existing approaches, followed by qualitative analysis to visualize model performance. Ablation analysis is conducted to assess the contribution of individual components. Cross-dataset evaluation is also performed to examine generalization, along with model complexity analysis.

3.4.1 Quantitative Analysis

We evaluate the performance of the proposed CoTCoNet framework by comparing it with existing methods across various datasets. To maintain a fair comparison, we reproduce the results of existing methods under identical data-splitting criteria (see Section 3.3.1). This approach ensures methodological consistency throughout the comparative analysis.

(a) C-NMC 2019 Dataset: Table 3.3 highlights the performance of our proposed method compared to state-of-the-art leukemic cell classification techniques. Among segmentation-based methods, Khandekar *et al.* [93] achieve competitive results, exhibiting only a minor accuracy decline of 1.84% compared to CoTCoNet. In addressing the class imbalance, Khan *et al.* [43] achieve the highest accuracy within their category, but with a larger accuracy drop of 15.24%. Deep learning-based classification methods demonstrate performance closer to CoTCoNet, with Chen *et al.* [21] achieving the second-best accuracy at a 0.73% difference. This distinction arises from CoTCoNet's innovative integration of transformer-based and convolutional feature extractors, empowering it to simultaneously learn both long-range dependencies and fine-grained details within the leukocyte sample. This leads to a more comprehensive and informative feature representation.

(b) ALL-IDB1 Dataset: Table 3.4 demonstrates CoTCoNet's performance on the ALL-IDB1 [91] dataset compared to existing methods. Among segmentation-based methods, Gokulkannan *et al.* [94] and Khandekar *et al.* [93] exhibit similar performance, with ac-

Table 3.3: Quantitative analysis of CoTCoNet framework with existing baselines on the C-NMC 2019 dataset. **Bold**: best and underlined: second-best scores.

	Methods	Acc	Pre	Rec	F1
I	Gokulkannan *et al.* [94]	0.8304	0.8762	0.8750	0.8756
	Khandekar *et al.* [93]	0.9710	0.9720	0.9701	0.9710
II	Khan *et al.* [43]	0.8367	0.8724	0.8986	0.8853
	Bashir *et al.* [95]	0.7782	0.8231	0.8538	0.8382
	Depto *et al.* [96]	0.7729	0.7803	0.7729	0.7766
III	Chen *et al.* [21]	0.9821	**0.9842**	0.9896	0.9869
	Dhalla *et al.* [22]	0.9523	0.9404	0.9504	0.9454
	Mohammed *et al.* [23]	0.9629	0.9793	0.9458	0.9623
	Das *et al.* [97]	0.9529	0.9282	0.9765	0.9517
	Jawahar *et al.* [98]	0.9100	0.9300	0.9800	0.9600
	Gehlot *et al.* [99]	0.9630	0.9554	0.9219	0.9480
Ours	CoTCoNet	**0.9894**	0.9800	**0.9988**	**0.9893**

I: Segmentation-based Classification, II: Imbalanced Classification, III: Deep Learning-based Classification

curacy declines within 1.67%-2.16% relative to CoTCoNet. In addressing the class imbalance, Khan *et al.* [43] and Bashir *et al.* [95] achieve equal accuracies but with a 7.15% drop compared to CoTCoNet. Among deep learning-based classification methods, Das *et al.* [97] demonstrate exceptional performance, showing only a 0.61% accuracy difference. Methods like Gehlot *et al.* [99] and CoTCoNet achieve superior results due to their emphasis on deep feature learning. However, CoTCoNet's unique advantage stems from integrating complex feature learning with graph-based feature reconstruction. This empowers CoTCoNet to precisely classify leukocyte samples, offering a significant advantage over other techniques.

Table 3.4: Quantitative analysis of CoTCoNet framework with existing baselines on the ALL-IDB1 dataset. **Bold**: best and underlined: second-best scores.

	Methods	Acc	Pre	Rec	F1
I	Gokulkannan *et al.* [94]	0.9833	0.9873	0.9984	0.9928
	Khandekar *et al.* [93]	0.9787	0.9844	0.9844	0.9844
II	Khan *et al.* [43]	0.9285	0.8571	0.9839	0.9161
	Bashir *et al.* [95]	0.9285	0.8571	0.9587	0.9051
	Depto *et al.* [96]	0.9118	0.9444	0.8947	0.9189
III	Chen *et al.* [21]	0.9815	0.9727	0.9981	0.9852
	Dhalla *et al.* [22]	0.9914	0.9934	0.9914	0.9924
	Mohammed *et al.* [23]	0.9412	0.8889	**1.0000**	0.9412
	Das *et al.* [97]	0.9939	**1.0000**	0.9865	0.9932
	Jawahar *et al.* [98]	0.9412	0.8750	**1.0000**	0.9333
	Gehlot *et al.* [99]	**1.0000**	**1.0000**	**1.0000**	**1.0000**
Ours	CoTCoNet	**1.0000**	**1.0000**	**1.0000**	**1.0000**

I: Segmentation-based Classification, II: Imbalanced Classification, III: Deep Learning-based Classification

(c) ALL-IDB2 Dataset: Our analysis of Table 3.5 demonstrates CoTCoNet's superiority on the ALL-IDB2 [91] dataset. Segmentation-based classification methods show an accuracy decline of 2.13%-3.9% relative to CoTCoNet. In addressing the class imbalance, the decline ranges from 3.13%-18.33%. This large interval primarily reflects the comparable performance of Bashir *et al.* [95], who utilize transfer learning (AlexNet) and randomized data augmentation to mitigate overfitting. Among deep learning-based methods, Chen *et al.* [21] achieve strong results, though slightly below CoTCoNet. CoTCoNet's superior performance stems from its strategic integration of feature selection and optimization techniques. This allows CoTCoNet to effectively identify the most discriminative features within the leukocyte image data, leading to a more robust and accurate classification model.

Table 3.5: Quantitative analysis of CoTCoNet framework with existing baselines on the ALL-IDB2 dataset. **Bold**: best and underlined: second-best scores.

	Methods	Acc	Pre	Rec	F1
I	Gokulkannan *et al.* [94]	0.9610	0.8914	0.9613	0.9250
	Khandekar *et al.* [93]	0.9787	0.9844	0.9844	0.9844
II	Khan *et al.* [43]	0.8492	0.8870	0.8029	0.8429
	Bashir *et al.* [95]	0.9687	**1.0000**	0.9333	0.9655
	Depto *et al.* [96]	0.8167	0.8659	0.8417	0.8536
III	Chen *et al.* [21]	0.9928	0.9826	0.9857	0.9841
	Dhalla *et al.* [22]	0.9880	0.9819	0.9810	0.9814
	Mohammed *et al.* [23]	0.9750	**1.0000**	0.9524	0.9756
	Das *et al.* [97]	0.9821	0.9845	0.9795	0.9820
	Jawahar *et al.* [98]	0.8250	0.9500	0.7600	0.8444
	Gehlot *et al.* [99]	0.9250	0.9500	0.9048	0.9268
Ours	CoTCoNet	**1.0000**	**1.0000**	**1.0000**	**1.0000**

I: Segmentation-based Classification, II: Imbalanced Classification, III: Deep Learning-based Classification

(d) SN-AM Dataset: Table 3.6 demonstrates CoTCoNet's performance on the SN-AM [92] dataset compared to state-of-the-art leukemic cell classification techniques. Among segmentation-based and deep learning classification methods, Khandekar *et al.* [93] and Gehlot *et al.* [99] exhibit similar performance, with a 7.69% accuracy decline compared to CoTCoNet. Gokulkannan *et al.* [94] follows closely within the segmentation-based category. Notably, CoTCoNet and Chen *et al.* [21]. However, CoTCoNet effectively tackles class imbalance by integrating a GAN-driven approach. It further employs segmentation to isolate leukocyte cells, facilitating robust feature-to-class mapping. This mapping allows the feature-learning mechanism to pinpoint discriminative features, resulting in superior separa-

tion of leukemic and non-leukemic classes. Consequently, our proposed framework excels on the SN-AM [92] dataset, outperforming existing techniques. Figure 3.6 shows the Receiver Operating Characteristic (ROC) curve, performance comparison of different methods, and the Precision-Recall (PR) curves across various datasets.

Table 3.6: Quantitative analysis of CoTCoNet framework with existing baselines on the SN-AM dataset. **Bold**: best and underlined: second-best scores.

	Methods	Acc	Pre	Rec	F1
I	Gokulkannan *et al.* [94]	0.9091	**1.0000**	0.8462	0.9167
	Khandekar *et al.* [93]	0.9231	**1.0000**	0.8750	0.9333
II	Khan *et al.* [43]	0.8182	0.8273	0.8889	0.8570
	Bashir *et al.* [95]	0.7973	0.8364	0.8778	0.8566
	Depto *et al.* [96]	0.8769	0.8142	0.9521	0.8778
III	Chen *et al.* [21]	**1.0000**	**1.0000**	**1.0000**	**1.0000**
	Dhalla *et al.* [22]	0.8636	0.8182	0.9000	0.8572
	Mohammed *et al.* [23]	0.8462	**1.0000**	0.7778	0.8750
	Das *et al.* [97]	0.7692	0.7143	0.8333	0.7692
	Jawahar *et al.* [98]	0.8462	0.8571	0.8571	0.8571
	Gehlot *et al.* [99]	0.9231	0.8571	**1.0000**	0.9231
Ours	CoTCoNet	**1.0000**	**1.0000**	**1.0000**	**1.0000**

I: Segmentation-based Classification, II: Imbalanced Classification, III: Deep Learning-based Classification

3.4.2 Qualitative Analysis

For the visual explanation of the decision reasoning of our CoTCoNet framework, we adopt Local Interpretable Model-agnostic Explanations LIME [87] to gain insights into the specific regions that significantly contribute to the classification process. Figure 3.7 illustrates 20 microscopic blood smear images sourced from the C-NMC 2019 [90] dataset; each image is accompanied by highlighted pertinent features generated by the LIME technique. The first ten images in the first row have been diagnosed as positive for leukemia by our CoTCoNet framework, while the images in the third row were diagnosed as negative. However, if no highly suspicious blood cells are detected, the model classifies the image as negative. This visualization underscores the interpretability of our CoTCoNet framework, providing valuable insights into its decision-making process.

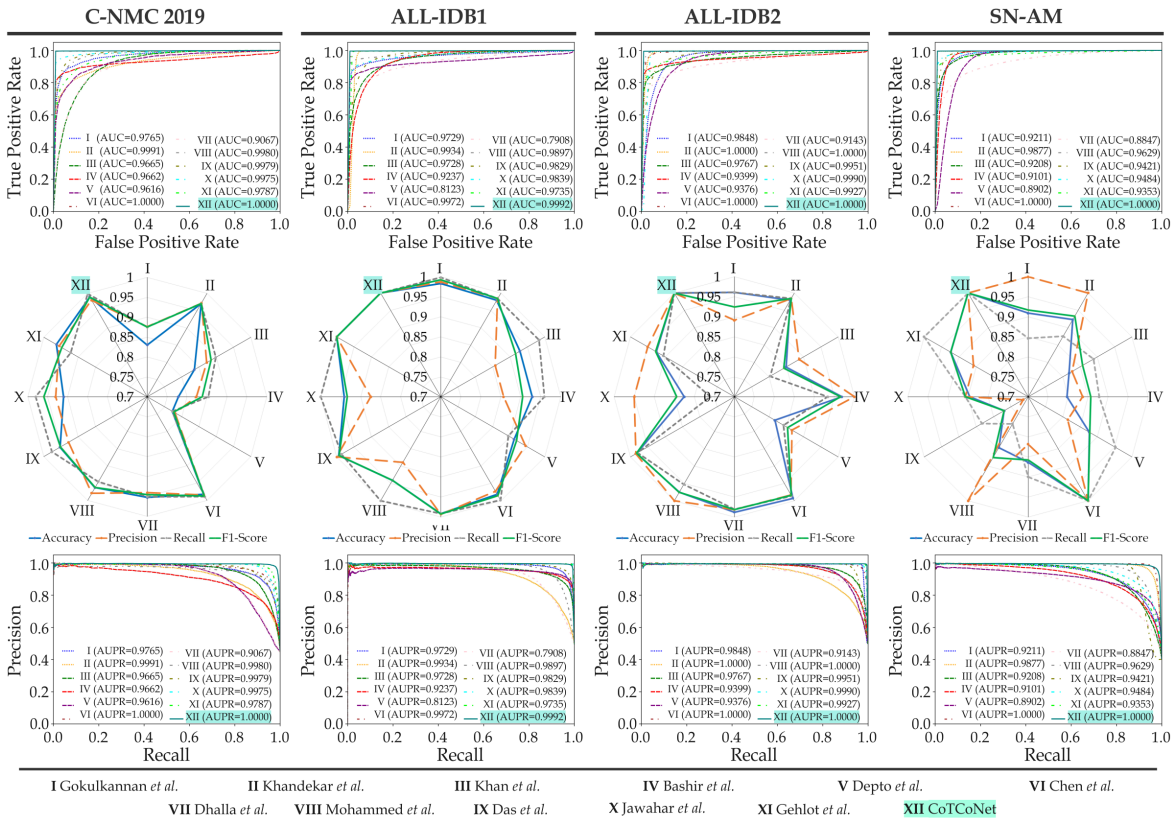


Figure 3.6: Comparative analysis of CoTCoNet against existing baselines across four datasets using ROC curves, a radar chart, and PR curves.

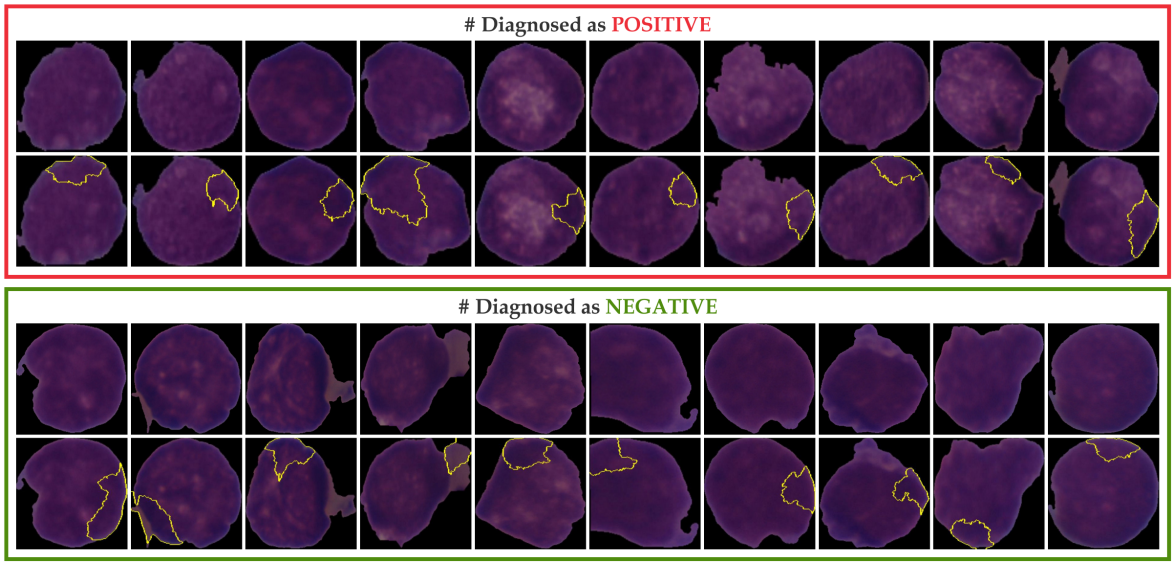


Figure 3.7: Qualitative analysis of CoTCoNet framework.

3.4.3 Ablation Analysis

To assess the contributions of the key modules within the CoTCoNet framework, we systematically removed each module one at a time on the C-NMC 2019 Dataset [90]. This ablation

study showed that all proposed components and strategies actively enhance the overall performance of our framework. Furthermore, the additional technique or module under examination is indicated within parentheses (e.g., CoTCoNet [technique/module]). Specifically, we explored the effects of different components, such as Image Enhancement, Segmentation, Data Augmentation, Feature Extraction, Graph-based Feature Reconstruction, and Feature Selection. The pipeline for our proposed model is comprehensively described in Section 3.2.2.

(a) Effect of Image Enhancement: The ablation study presented in Table 3.7 demonstrates that without applying any image enhancement technique, the accuracy of CoTCoNet decreases by 8.83%. While Fuzzy [100], MirNet [101], and Histogram Equalization (HE) [102] offer improvement, they still result in accuracy losses ranging from 3% to 6%. Furthermore, employing only Contrast Limited Adaptive Histogram Equalization (CLAHE) or Sharpen techniques individually yields accuracy reductions of 2.83% and 1.06%, respectively, compared to the proposed approach. These findings highlight the limitations of traditional methods in fully addressing noise and contrast variations inherent in leukocyte images. Our proposed framework, CoTCoNet (CLAHE + Sharpen), effectively mitigates these challenges. CLAHE provides targeted contrast enhancement, followed by a sharpening step for improved feature visibility. This sequential strategy outperforms other methods, facilitating the framework's extraction of highly discriminative features from leukocyte samples. Consequently, image enhancement directly contributes to the superior classification accuracy achieved by the CoTCoNet framework.

Table 3.7: Ablation analysis of CoTCoNet framework with different image enhancement techniques on the C-NMC 2019 dataset.

Methods	Acc	Pre	Rec	F1
CoTCoNet (No Enhancement)	0.9011	0.8681	0.9294	0.8977
CoTCoNet (Fuzzy)	0.9211	0.9741	0.8807	0.9251
CoTCoNet (MirNET)	0.9499	0.8999	1.0000	0.9473
CoTCoNet (HE)	0.9541	0.9753	0.9356	0.9550
CoTCoNet (Sharpen)	0.9611	0.9435	0.9780	0.9604
CoTCoNet (CLAHE)	0.9788	0.9647	0.9927	0.9785
CoTCoNet (CLAHE + Sharpen)	**0.9894**	**0.9800**	**0.9988**	**0.9893**

(b) Effect of Segmentation: Our proposed framework, CoTCoNet (fine-tuned LeuSAM), achieves superior performance as detailed in Table 3.8. Replacing the fine-tuned LeuSAM

with CoTCoNet (UNet [46]) and CoTCoNet (UNet++ [103]) significantly reduces accuracy by 12.57% and 9.27%, respectively. This highlights the importance of fine-tuning LeuSAM within the CoTCoNet framework. Furthermore, fine-tuning SAM within CoTCoNet leads to a remarkable 7.85% accuracy improvement compared to using SAM without fine-tuning. These results demonstrate the effectiveness of incorporating the fine-tuned LeuSAM for precise leukocyte cell segmentation. By eliminating extraneous information, LeuSAM allows the framework to focus on the essential features of the leukocytes (e.g., size, shape, texture). By concentrating on segmented leukocyte regions, the model can extract more precise and discriminative features crucial for accurately classifying different types of leukocytes, which is vital for medical diagnosis.

Table 3.8: Ablation analysis of CoTCoNet framework with different image segmentation techniques on the C-NMC 2019 dataset.

Methods	Acc	Pre	Rec	F1
CoTCoNet (UNet)	0.8657	0.8092	0.9124	0.8577
CoTCoNet (UNet++)	0.8920	0.6568	0.9839	0.7878
CoTCoNet (w\o fine-tuning LeuSAM)	0.9136	0.9047	0.9385	0.9213
CoTCoNet (fine-tuned LeuSAM)	**0.9894**	**0.9800**	**0.9988**	**0.9893**

(c) Effect of Augmentation: Data augmentation analysis in Table 3.9 reveals that training with imbalanced leukemia samples significantly reduces accuracy by 16.88%, indicating overfitting. In contrast, employing traditional augmentation techniques like scaling, cropping, translation, flipping, and rotation improves performance by 5.61% compared to models without augmentation. Notably, incorporating LeuGAN into our framework, denoted as CoTCoNet (LeuGAN), yields the most substantial performance gains of 16.88% and 11.27% improvements over models trained without augmentation and with traditional augmentation, respectively. These results underscore the power of GAN-based augmentation to enhance our framework's performance. LeuGAN, by synthesizing high-resolution leukocyte samples, significantly expands and diversifies the training dataset. This diversification improves the framework's ability to generalize to unseen examples. Furthermore, the interconnected feedback loop crucially ensures that the generated synthetic samples preserve essential hematological features. Consequently, the framework effectively learns discriminative characteristics for accurate leukocyte classification.

Table 3.9: Ablation analysis of CoTCoNet framework with different image augmentation techniques on the C-NMC 2019 dataset.

Methods	Acc	Pre	Rec	F1
CoTCoNet (No Augmentation)	0.8206	0.5457	0.8024	0.6496
CoTCoNet (Traditional Augmentation)	0.8767	0.8499	0.8919	0.8704
CoTCoNet (LeuGAN)	**0.9894**	**0.9800**	**0.9988**	**0.9893**

(d) Effect of Feature Extraction: To evaluate the significance of feature extractors, we conducted an ablation study focusing on transformer-based and convolutional-based extractors (without Feature Reconstruction and Selection). Results in Table 3.10 reveal a 13-15% performance decrease when using transformer-based extractors like ViT-B/32, ViT-L/32, and ViT-H/14. Notably, GMod outperforms these standard transformer-based extractors by 3-7%. Similarly, SMod demonstrates a 1.59% accuracy improvement compared to VGG-16. Moreover, applying SMod earlier results in higher accuracy than most feature extractors, except for the proposed CoTCoNet (GMod + SMod) framework. This is because SMod examines tiny details within a small image region. When applied first, SMod filters out long-range dependencies essential for accurate leukocyte classification. In contrast, GMod excels at capturing global relationships within the image. If low-level details dominate the initial input from SMod, GMod struggles to learn meaningful global features. However, applying GMod first provides a robust foundation of global features, and SMod refines these features by adding localized details, enriching the leukocyte representation. This complementary approach leads to superior classification performance.

Table 3.10: Ablation analysis of CoTCoNet framework with different feature extraction techniques on the C-NMC 2019 dataset.

Methods	Acc	Pre	Rec	F1
CoTCoNet (ViT-B/32)	0.8581	0.8269	0.8819	0.8535
CoTCoNet (ViT-L/32)	0.8422	0.8996	0.7703	0.8299
CoTCoNet (ViT-H/14)	0.8351	0.6914	0.9702	0.8075
CoTCoNet (VGG-16)	0.8528	0.8492	0.8303	0.8396
CoTCoNet (GMod)	0.8651	0.8422	0.8827	0.8620
CoTCoNet (SMod)	0.8687	0.9184	0.8092	0.8604
CoTCoNet (SMod + GMod)	0.9629	0.9793	0.9458	0.9623
CoTCoNet (GMod + SMod)	**0.9894**	**0.9800**	**0.9988**	**0.9893**

(e) Effect of Graph-based Feature Reconstruction and Feature Selection: Table 3.11 highlights the superiority of the CoTCoNet framework, which integrates a graph-based fea-

ture reconstruction module (GraFR) and a population-based algorithm for feature selection and optimization. When CoTCoNet is limited to the Global Feature Module (GMod) and Spatial Feature Module (SMod) only, i.e., CoTCoNet (w\o graph + w\o selection), it fails to achieve optimal performance in classifying leukemic cells. Results show that incorporating the GraFR module alone offers a marginal performance gain of 0.75%. Crucially, adding feature selection boosts performance by 4.3%, while the combined approach, CoTCoNet (with graph + with selection), yields the greatest improvement of 4.59%. This demonstrates the synergistic effect of GraFR and feature selection in enhancing CoTCoNet's performance and feature representation. The GraFR module effectively uncovers hidden biological features by upscaling unobserved sample features to a higher-dimensional feature map. Similarity scores between neighbor nodes allow GraFR to emphasize these discriminative features, ultimately improving framework performance.

Table 3.11: Ablation analysis of CoTCoNet framework with different graph-based feature reconstruction and feature selection techniques on the C-NMC 2019 dataset.

Methods	Acc	Pre	Rec	F1
CoTCoNet (w/o graph + w/o selection)	0.9435	0.9764	0.9160	0.9453
CoTCoNet (with graph + w/o selection)	0.9510	0.9494	0.9988	0.9734
CoTCoNet (w/o graph + with selection)	0.9865	0.9741	0.9988	0.9863
CoTCoNet (with graph + with selection)	**0.9894**	**0.9800**	**0.9988**	**0.9893**

3.4.4 Cross-Dataset Analysis

To assess CoTCoNet's generalizability, we performed a cross-dataset evaluation on both unicellular and multicellular leukocyte samples. This involved training on one dataset and testing on another, including a concatenated dataset. Table 3.12 demonstrates CoTCoNet's performance under various training and testing distributions. Our analysis reveals that CoTCoNet achieves high accuracy when trained and tested on the same large, diverse dataset. However, its performance declines when trained on smaller datasets and tested on large ones. For example, when trained on C-NMC 2019 [90], CoTCoNet achieves an optimal performance of 0.9894 but suffers an 11.39% drop in accuracy to 0.8755 when tested on the concatenated dataset of C-NMC 2019 [90] and ALL-IDB2 [91]. Notably, CoTCoNet's performance improves by 3.71% when trained and tested on the concatenated dataset. Similarly,

training on the smaller ALL-IDB2 [91] dataset and testing on the larger combined dataset of C-NMC 2019 [90] and ALL-IDB2 [91] leads to a significant performance decline. The results indicate there is potential to refine the transferability of knowledge between these two datasets. However, training CoTCoNet on the combined ALL-IDB2 [91] and SN-AM [92] dataset yields the second-best performance, suggesting benefits from combining data from different domains. Overall, these findings demonstrate that CoTCoNet exhibits comparable accuracy on concatenated datasets compared to single datasets. This implies that combining data from diverse sources can enhance CoTCoNet's robustness and ability to generalize unseen data by providing a broader and richer learning experience.

Table 3.12: Cross-Dataset analysis of CoTCoNet framework with different training and testing datasets.

Training	**Testing**											
	Category-1: Uni-Cellular											
	U1				**U2**				**U1 + U2**			
	Acc	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**
U1	0.9894	0.9800	0.9980	0.9893	0.9236	0.9416	0.9357	0.9386	0.8755	0.8923	0.9014	0.8968
U2	0.6132	0.6218	0.5985	0.6099	1.0000	1.0000	1.0000	1.0000	0.5892	0.5744	0.5976	0.5858
U1 + U2	0.9603	0.9617	0.9453	0.9534	0.9347	0.9239	0.9406	0.9322	0.9523	0.9455	0.9378	0.9416
	Category-2: Multi-Cellular											
	M1				**M2**				**M1 + M2**			
	Acc	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**
M1	1.0000	1.0000	1.0000	1.0000	0.8750	0.8038	1.0000	0.8912	0.9214	0.9004	0.9281	0.9140
M2	0.7363	0.6984	0.7206	0.7093	1.0000	1.0000	1.0000	1.0000	0.6963	0.6812	0.7016	0.6912
M1 + M2	0.8957	0.8615	0.9074	0.8839	0.7703	0.7326	0.7914	0.7609	0.9754	0.9806	0.9753	0.9779

U1: C-NMC 2019, U2: ALL-IDB2, M1: ALL-IDB1, M2: SN-AM.

3.4.5 Model Complexity Analysis

We analyze CoTCoNet's model complexity from two perspectives. Table 3.13 compares its computational requirements (number of network parameters and inference time) against existing methods. Compared to these methods, CoTCoNet exhibits a parameter size of 97.84M, slightly below Chen *et al.* [21] but larger than others. Methods by Gehlot *et al.* [99] and Jawahar *et al.* [98] demonstrate the lowest resource requirements and inference times. CoTCoNet's complex architecture, with numerous layers and filters, drives its computational

demands. This complexity is designed to capture the intricate details within leukocyte samples, leading to high framework accuracy. Despite its computational complexity, CoTCoNet achieves the best classification performance across Table 3.3, Table 3.4, Table 3.5, and Table 3.6. This is evidenced by its top scores in accuracy, precision, recall, and F1-score. An inference time of 41.60ms further suggests suitability for near real-time applications. Overall, these results demonstrate a favorable balance between CoTCoNet's classification performance and its computational costs. Future research could explore optimization techniques to enhance efficiency while maintaining accuracy.

Table 3.13: Model Complexity Analysis of the CoTCoNet framework comparing parameters and inference times with existing baselines.

	Methods	**Parameters (M)**	**Inference Time (ms)**
I	Gokulkannan *et al.* [94]	54.64	45.53
	Khandekar *et al.* [93]	64.47	25.30
II	Khan *et al.* [43]	10.83	32.17
	Bashir *et al.* [95]	62.30	23.09
	Depto *et al.* [96]	5.43	8.70
III	Chen *et al.* [21]	99.42	63.57
	Dhalla *et al.* [22]	7.81	20.40
	Mohammed *et al.* [23]	20.80	8.81
	Das *et al.* [97]	24.50	28.61
	Jawahar *et al.* [98]	0.73	2.31
	Gehlot *et al.* [99]	0.10	1.08
Ours	CoTCoNet	97.84	41.60

I: Segmentation-based Classification, II: Imbalanced Classification, III: Deep Learning-based Classification
M: million, ms: milliseconds

3.5 Conclusion

In this study, we propose a novel framework called **Co**upled **T**ransformer **Co**nvolutional **Net**work (CoTCoNet), which hierarchically integrates a deep convolution network with a powerful transformer for the classification of leukemia within microscopic blood smear slides. CoTCoNet effectively leverages the capabilities of transformers and deep convolutional networks to efficiently extract both global and spatial features from blood cell images. Furthermore, we incorporate graph-based feature reconstruction to enhance feature knowledge, leveraging the advantages of complex, hard-to-see biological features of leukocytes. The aforementioned technical framework comprises several advantages: (a) the proposed

method acquires features from indistinguishable cell images across different classes, significantly enhancing model flexibility in clinical practice. Moreover, the preservation of the structural integrity of cells in blood smear images yields comprehensive and discriminative features; (b) the proposed framework utilizes a Population-based Meta-Heuristic Algorithm to identify optimal features, effectively reducing optimization errors; (c) the proposed method is capable of generating interpretable attention maps to reinforce the prediction results, thereby facilitating hematopathologists in their decision-making process when utilizing AI algorithms; (d) the proposed framework demonstrates superior classification performance across four publicly available and diverse datasets. Extensive benchmarking across four public datasets confirms that the CoTCoNet framework outperforms existing methods in leukocyte classification. Its potential clinical utility lies in its ability to accurately discriminate between leukemic and normal cells.

Chapter 4

Contrastive-based Dilated Convolutional Network with Transformer for Pediatric Pneumonia Detection

CONTENTS

4.1 Introduction . . . 54
4.2 Methodology . . . 55
4.2.1 Framework Overview . . . 55
4.2.2 Preprocessing . . . 55
4.2.3 Proposed Framework . . . 62
4.3 Experimental Settings . . . 71
4.3.1 Benchmark Datasets . . . 72
4.3.2 Evaluation Metrics . . . 73
4.4 Experimental Results . . . 74
4.4.1 Quantitative Analysis . . . 74
4.4.2 Qualitative Analysis . . . 78
4.4.3 Ablation Analysis . . . 79
4.4.4 Cross-Dataset Analysis . . . 83
4.4.5 Model Complexity Analysis . . . 85
4.5 Conclusion . . . 86

4.1 Introduction

This chapter addresses pediatric pneumonia detection by introducing methods that capture global and local features in chest radiographs. The aim is to learn both the global context and the fine-grained spatial patterns from chest X-rays (CXR). This joint learning is important for reliable diagnosis. Pneumonia is the major cause of mortality among children, and therefore, early detection is essential for timely treatment. CXRs are widely used for diagnosis but require expert knowledge, which is time-consuming. Pediatric CXRs also pose challenges, including low contrast and overlapping anatomical structures, and variability in lung appearance due to improper growth and acquisition conditions. Class imbalance and limited labeled data also affect the model performance. Existing approaches rely either on convolutional networks for spatial feature extraction or on transformer-based methods for global context. However, often ignores the combined effect of both. These limitations reduce the ability to capture complex disease patterns. Such challenges motivate the need for a framework that learn global and local representations while improving feature quality with interpretability.

To address these limitations, this chapter presents an Explainable Contrastive-based Dilated Convolutional Network with Transformer (XCCNet) for pediatric pneumonia detection. The framework leverages contrastive learning, transformer-based global modeling, and convolutional spatial feature extraction to learn complementary patterns from CXRs. Contrastive learning aligns global and local features by including intra-class similarity and inter-class separation. A preprocessing module enhances image quality by improving contrast and extracting the lung region. To address data imbalance, an adversarial data-generation unit produces synthetic CXRs. In addition, a rib suppression mechanism reduces interference from bony structures. An explainability module generates attention maps that highlight important regions and support the interpretation of model predictions.

The remainder of this chapter is organized as follows. Section 4.2 details the methodology, including the XCCNet pipeline, adversarial data augmentation, and feature extraction modules. Section 4.3 describes the datasets and evaluation metrics. Section 4.4 presents quantitative comparisons with state-of-the-art methods and qualitative analysis. Finally, Section 4.5 summarizes the findings and discusses the relevance of the proposed framework.

4.2 Methodology

This section presents the proposed framework for pediatric pneumonia classification. It describes the overall pipeline, including preprocessing and framework design. The framework identified the key components of joint representations for learning meaningful representations, as detailed in the following subsections.

4.2.1 Framework Overview

Figure 4.1 details the architectural overview of the XCCNet framework, designed for detecting pediatric pneumonia from chest X-rays (CXRs). XCCNet comprises four integral modules: (a) a chest X-ray preprocessing module, which refines raw CXRs to improve feature extraction, aiding both spatial and contrastive-based transformer feature extraction. It reduces data skewness, amplifies low-radiation CXRs, extracts lung information, and selectively reduces rib visibility while preserving internal soft tissue details; (b) a spatial feature extraction module utilizes dilated convolutions to capture fine-grained spatial features from CXRs, such as shape, color, texture, and hierarchical information. Dilated convolutions enlarge the receptive field while maintaining feature map resolution, allowing for the extraction of richer spatial details; (c) a contrastive-based transformer feature extractor module combines a transformer architecture to explore global contextual dependencies with contrastive learning. The contrastive learning loss function promotes feature learning by focusing on intra-class similarity and inter-class dissimilarity, leading to improved feature discrimination. Additionally, a cross-entropy loss is integrated to guide the network toward accurate classification based on the learned features. Finally, (d) a dual hybrid feature fusion and classification module fuses spatial and global features with the same dimensions along the channel dimension. The fused features are fed into a deep classifier to categorize the input CXR as normal or pneumonia.

4.2.2 Preprocessing

This section details various image processing techniques employed within the Chest X-ray Preprocessing (CXP) module to prepare input Chest X-rays (CXR). The CXP module acts

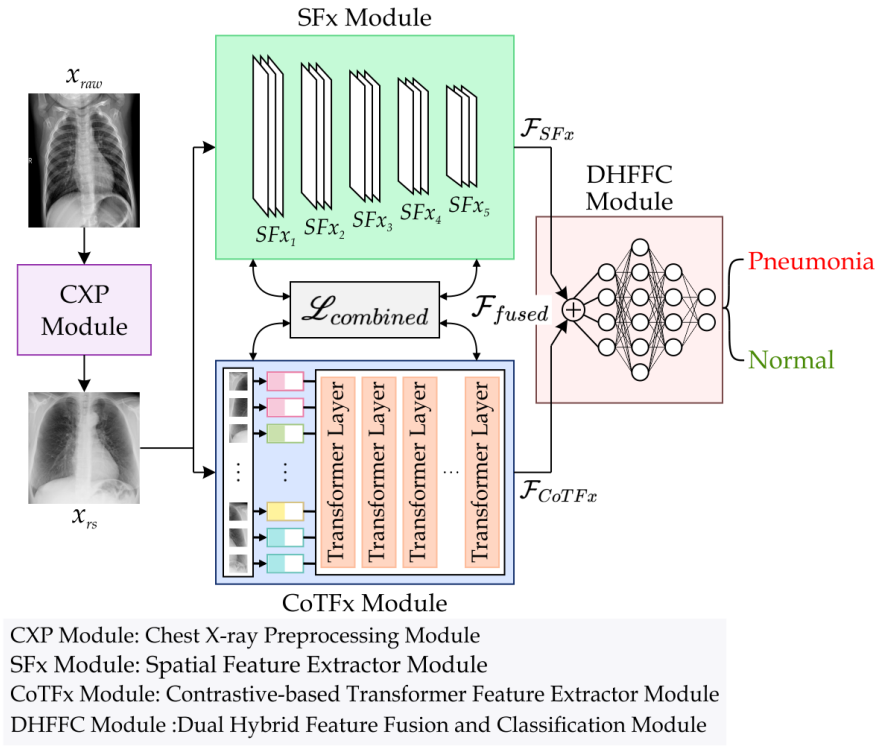


Figure 4.1: The workflow of our proposed XCCNet framework for detecting pediatric pneumonia in chest X-rays.

as a critical gateway, transforming raw CXRs through a series of targeted refinements: (a) Resizing, (b) Synthetic Data Generation, (c) Image Enhancing, (d) Segmentation, and (e) Rib Suppression. These combined processing steps work synergistically to enhance low-level features within the CXRs. This ultimately facilitates improved extraction and maximizes the information captured from each sample, leading to more accurate and robust analysis. Figure 4.2 outlines the pipeline of the CXP module.

(a) Resizer: We resize the collected CXR to spatial dimensions of 256×256 while preserving image quality. The resizing mechanism utilizes bicubic interpolation. Bicubic interpolation stands out as a favored method for image resizing due to its efficient balance between computational cost and quality enhancement. The bicubic interpolation can be mathematically expressed in Equation (4.1):

$$interpolate_{bicubic}\left(x_{raw}, \frac{x_n}{\alpha_x}, \frac{y_n}{\alpha_y}\right) = \sum_x \sum_y B(x-u)\, B(y-v)\, I(x_o+u, y_o+v) \quad (4.1)$$

where, $x_o = \left\lfloor \frac{x_n}{\alpha_x} \right\rfloor$ and $y_o = \left\lfloor \frac{y_n}{\alpha_v} \right\rfloor$ are the corresponding integer coordinates in the original

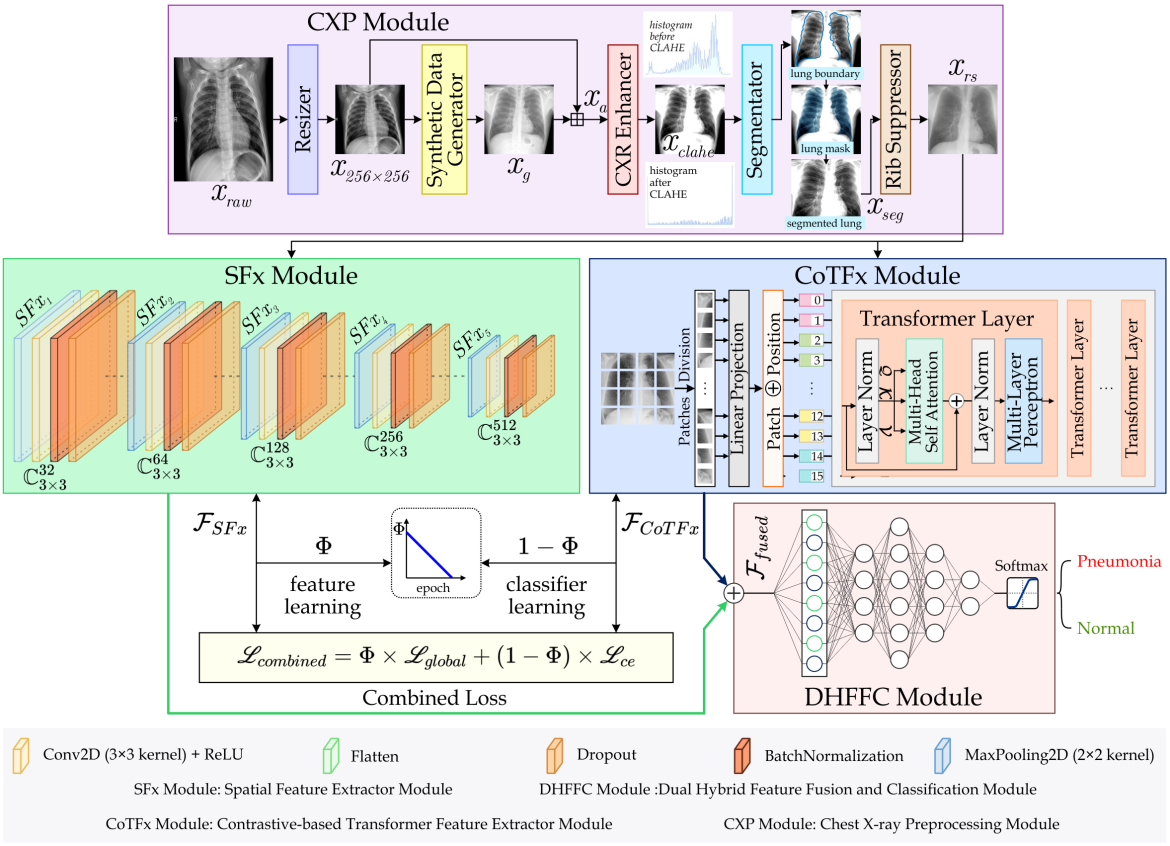


Figure 4.2: Proposed architecture of Explainable Contrastive-based Dilated Convolutional Network with Transformer (XCCNet) for pediatric pneumonia detection in chest X-rays.

image x_{raw}. Variables u and v iterate over the neighboring pixel offsets (-1, 0, 1). $B(\cdot)$ is the bicubic basis function evaluated at $(\cdot)$ and $I(x_o + u, y_o + v)$ is the intensity value at the neighboring pixel.

(b) Synthetic Data Generator: In medical imaging, the scarcity of images poses a significant challenge to effective model training, often hindering convergence. Data augmentation techniques address this challenge by generating synthetic data mimicking real-world samples. Classical data augmentation techniques include geometric transformations such as flipping, rotation, scaling, and skewing. While conventional data augmentation relies on simple replication techniques, these methods may not adequately capture the full spectrum of clinical characteristics in complex medical images, even though they enhance dataset diversity to some extent. However, for medical images with complex textures, the samples obtained by this method may not accurately reflect the images' clinical characterization. This requires an alternative augmentation technique that generates synthetic medical samples with textures

and patterns similar to those observed in medical scans. In response to this, we devised a novel Adversarial-based Data Augmentation Adversarial-based Data Augmentation (ADA) unit to generate high-resolution synthetic CXRs. Recent research in medical imaging demonstrates a significant increase in the utilization of data augmentation methods [104, 105].

Adversarial-based Data Augmentation Unit: We aim to design and implement an Adversarial-based Data Augmentation (ADA) Unit for medical image analysis tasks. The ADA unit will leverage adversarial learning principles to synthetically generate additional data, mitigating the impact of data imbalance issues in medical datasets. Figure 4.3 illustrates the architecture of the ADA unit to generate synthetic CXR.

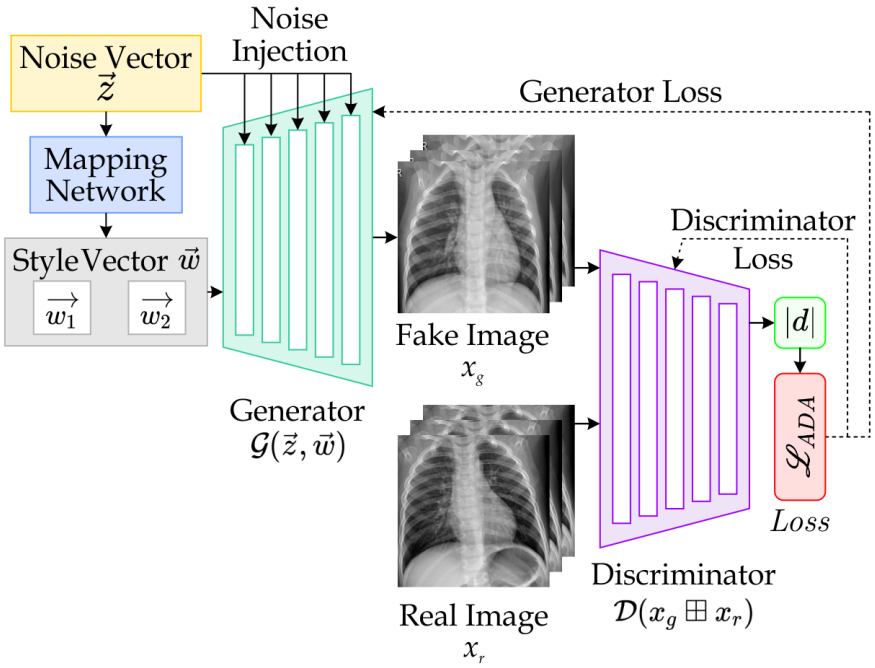


Figure 4.3: Proposed architecture of Adversarial-based Data Augmentation (ADA) unit.

In Figure 4.3, a noise vector $\overrightarrow{z}$, where $\overrightarrow{z} \in \{\overrightarrow{z_1}, \ldots, \overrightarrow{z_n}\}$ sampled from a normal distribution, serves as input to the mapping network. The network employs a Multi-Layer Perceptron (MLP) with numerous hidden units to transform $\overrightarrow{z}$ into a style vector $\overrightarrow{w}$. For synthesis, the style vector bifurcates into $\overrightarrow{w_1}$ and $\overrightarrow{w_2}$ to inject noise into separate upper and lower portions of the synthetic image. The generator block $\mathcal{G}(\overrightarrow{z}, \overrightarrow{w})$ then takes these vectors as inputs to produce fake images x_g, given in Equation (4.2).

$$x_g = \mathcal{G}(\overrightarrow{z}, \overrightarrow{w}) \tag{4.2}$$

The generator block utilizes stacked Adaptive Instance Normalization (AdaIN) to normalize the feature map of synthetic images with $\overrightarrow{w}$, as mathematically presented in Equation (4.3):

$$AdaIN\Big(x_g, \big(\overrightarrow{w}\big)\Big) = \sigma(\overrightarrow{w}) \cdot \frac{\Big(x_g - \mu\big(x_g\big)\Big)}{\sigma(x)} + \mu(\overrightarrow{w}) \tag{4.3}$$

where $\sigma(\overrightarrow{w})$ and $\mu(\overrightarrow{w})$ are the standard deviation and mean of $\overrightarrow{w}$, respectively, and $\mu(x_g)$ and $\sigma(x_g)$ represent the mean and standard deviation of the feature maps of x_g. The generated x_g and real image $x_r \in x_{raw}$ are simultaneously permuted (denoted as $\boxplus$) and fed into the discriminator $\mathcal{D}(x_g \boxplus x_r)$, which outputs a scalar value $|d|$ indicating whether the input is real or fake. The discriminator's and generator's objective functions are given in Equation (4.4) and Equation (4.5), respectively.

$$\mathscr{L}_{\mathcal{D}} = -[log(\mathcal{D}(x_g \uplus x_r))] - E[log(1 - \mathcal{D}(G(\overrightarrow{z}, \overrightarrow{w})))] \tag{4.4}$$

$$\mathscr{L}_{\mathcal{G}} = E[log(\mathcal{D}(\mathcal{G}(\overrightarrow{z}, \overrightarrow{w}))] \tag{4.5}$$

The ADA unit learns parameters from both the generator and discriminator, minimizing the total loss given in Equation (4.6).

$$\mathscr{L}_{ADA} = \mathscr{L}_{\mathcal{D}} + \mathscr{L}_{\mathcal{G}} + \lambda \cdot \mathscr{L}_{reg} \tag{4.6}$$

This loss function comprises three components: $\mathscr{L}_{\mathcal{D}}$ measures differences between fake and real images, $\mathscr{L}_{\mathcal{G}}$ encourages realistic x_g, and $\mathscr{L}_{reg}$ enforces a unit vector constraint on $\overrightarrow{w}$), where λ is a constant. Balancing these components leads to the generation of more realistic synthetic images, x_g. Furthermore, the generated synthetic x_g is combined with real CXR x_r, leading to an augmented dataset $\mathbf{X_a}$, where $x_a \in \mathbf{X_a}$.

Implementation Details: We provided images with dimensions of 512×512 pixels and three color channels as input to the ADA unit. The model underwent training with a batch size of 8 for 30 epochs, generating 128 new images in 11.5 hours using an Nvidia Tesla P100 GPU. Following a manual review, we discarded any incorrect images and repeated this process until achieving a balanced dataset.

(c) CXR Enhancer: To enhance image contrast, we employed Contrast Limited Adaptive

Histogram Equalization (CLAHE). CLAHE demonstrates advantages over Adaptive Histogram Equalization (AHE), particularly for improving low-contrast images in the medical domain [106, 107]. In our experiments, we empirically selected a window size of 8×8 to employ the HE in each local region (window) of the image. Additionally, a clipping scale of 2.0 (represented as $CLAHE2.0$) was chosen, resulting in more excellent contrast enhancement compared to $CLAHE1.0$. Compared to $CLAHE1.0$ with fixed clipping, $CLAHE2.0$ utilizes adaptive clipping limits, mitigating artifacts and preserving image fidelity. To validate this, we computed the correlation coefficient between the enhanced CXR images and the original CXR images in the dataset for various CLAHE scales. Our findings reveal that the correlation coefficient for CLAHE 2.0 surpasses other scales, leading us to select it for contrast enhancement with an 8×8 window and generate the final enhanced image x_{clahe}.

(d) Segmentator: Lung segmentation is a vital step in pneumonia classification, facilitating the isolation of affected regions for enhanced analysis. This refined Region Of Interest (ROI) enables more precise feature extraction and subsequent classification, leading to superior results compared to analyzing the entire chest X-ray. However, achieving robust segmentation remains challenging due to factors such as image variability, overlapping structures, and subtle intensity differences. Therefore, developing a robust lung segmentation technique is essential for efficiently extracting lung regions from Chest X-rays (CXRs) and enabling accurate subsequent analysis tasks. To this end, we propose a modified ResUNet++ architecture for lung segmentation. Figure 4.4 represents the architecture of modified ResUNet++. ResUNet++ augments the UNet architecture with skip connections, enabling the decoder to refine segmentation maps by incorporating information from earlier encoder layers. Additionally, ResUNet++ incorporates dense connections between all subsequent layers, facilitating information flow and feature reuse for enhanced performance.

Figure 4.4 depicts the depicts the encoder block, which utilizes ResNet50 as its backbone network. This backbone employs convolutional layers with residual connections for enhanced information flow. The encoder function, represented by $\mathscr{E}_{seg}$, takes an input-enhanced CXR and generates a feature map $\vec{\mathcal{F}}_{seg}$ containing spatial features. This $\vec{\mathcal{F}}_{seg}$ information is shared across all encoder and decoder layers through skip connections.

The decoder block function, $\mathscr{D}_{seg}\left(\vec{\mathcal{F}}_{seg}\right)$, leverages the feature map $\vec{\mathcal{F}}_{seg}$ generated by the encoder block to construct a segmentation map. This map identifies the lung

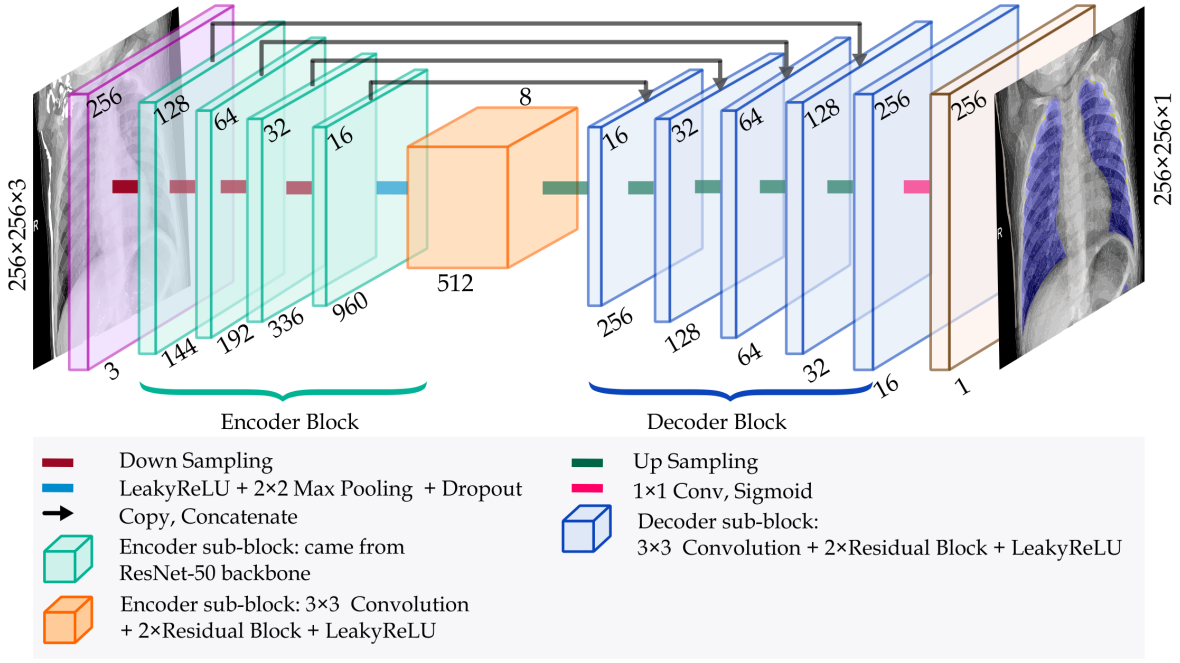


Figure 4.4: Proposed architecture of ResUNet++ for lung segmentation.

boundary and creates the binary mask $\mathcal{B}_{mask}$ for input CXR, assigning a value of 1 to pixels on the boundary and 0 otherwise. To minimize the error between the predicted $\mathcal{B}_{mask}$ and the ground truth segmentation map $\mathcal{S}_{map}$, the binary cross-entropy loss function, $\mathscr{L}_{seg}\Big(\mathcal{S}_{map}, \mathcal{B}_{mask}\Big)$, is employed as defined in Equation (4.7).

$$\mathscr{L}_{seg}\Big(\mathcal{S}_{map}, \mathcal{B}_{mask}\Big) = -\Big[\mathcal{S}_{map} \times log\big(\mathcal{B}_{mask}\big) + \big(1 - \mathcal{S}_{map}\big) \times log\big(1 - \mathcal{B}_{mask}\big)\Big] \tag{4.7}$$

This loss function serves as a quantitative measure of the overall segmentation prediction accuracy. Finally, the generated mask $\mathcal{B}_{mask}$ is overlaid onto the input image using a logical AND ($\wedge$) operation. This yields the final lung mask $\mathcal{L}_{mask}$, which is then cropped based on mask pixels to obtain the segmented CXR image x_{seg}.

(e) Rib Suppressor: Chest X-rays (CXRs) provide valuable diagnostic information about both soft tissues and bones, which is crucial for radiologists and clinicians in making informed decisions regarding diagnosis and treatment. However, overlapping structures like clavicles and ribs can cover soft tissue details. Rib suppression techniques address this challenge by reducing the visibility of bony structures, particularly ribs, while preserving internal soft tissue details.

We utilize the RS_{ResNet} unit, a modified ResNet-based rib suppression model inspired by Rajaraman *et al.* [108], for effective bone removal while preserving soft tissue details in CXR images. This sequential model comprises five stacked ResNet blocks $B^{(i)}_{ResNet}$ for $i = 1$ to 5, each employing 16, 32, 64, 128, and 256 filters, respectively. These blocks utilize 3×3 kernels with zero padding to maintain image dimensions and incorporate skip connections to mitigate vanishing gradient issues and enable the reuse of activations. The final rib-suppressed image x_{rs} is generated by the deepest convolutional layer C_{deep}, which features a single filter with zero padding and a sigmoid activation function. Figure 4.5 present the The architecture of RS_{ResNet} unit. Algorithm 4.1 provides a step-by-step explanation of the Chest X-ray Preprocessing (CXP) Module.

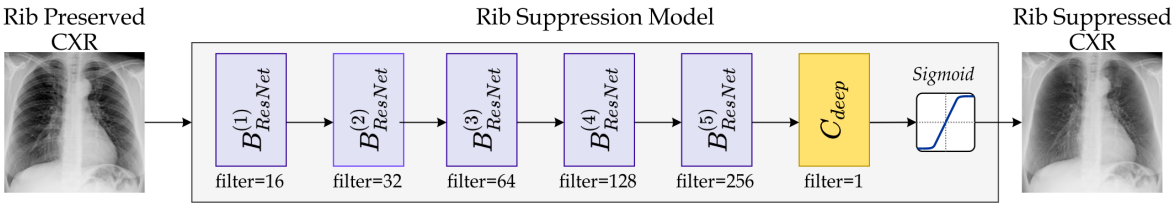


Figure 4.5: Proposed architecture of ResNet-based rib suppression model.

4.2.3 Proposed Framework

Explainable Contrastive-based Dilated Convolutional Network with Transformer (XCCNet) is composed of four primary modules, namely the Chest X-ray Preprocessing (CXP) module, Spatial Feature Extraction Spatial Feature Extraction (SFx) Module, Contrastive-based Transformer Feature Extractor Contrastive-based Transformer Feature Extractor (CoTFx) module, and Dual Hybrid Feature Fusion and Classification Dual Hybrid Feature Fusion and Classification (DHFFC) module. These modules are the foundation pillars of XCCNet, each having its unique role, including CXR preprocessing to final pneumonia classification. In the next sections, we will explore the internal architectural details of each module. Figure 4.2 illustrates the detailed architecture of the XCCNet framework.

4.2.3.1 Spatial Feature Extractor Module

The Spatial Feature Extractor (SFx) module leverages a dilated convolution network to extract fine-grained spatial features like shape, target color, texture, and hierarchical informa-

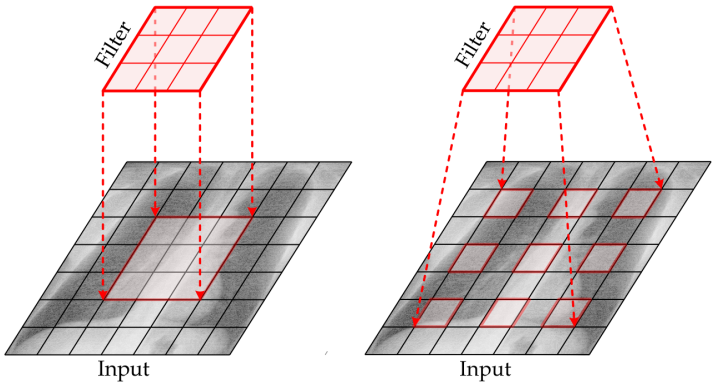
Algorithm 4.1 CXP MODULE

Input: $\langle x_{raw} \rangle \in$ input CXR dimensions.
Output: $\langle x_{rs} \rangle \in$ Rib Suppressed CXR

function RESIZE$_{bicubic}(x_{raw}, x_{256\times256})$
    Calculate scaling factor: $\alpha_x, \alpha_y \leftarrow \frac{256}{I_w}, \frac{256}{I_h}$
    Initialize an empty resized image $h_n \times w_n$
    for each output pixel (x_n, y_n) **do**
        $x_{256\times256}(x_n, y_n) \leftarrow interpolate_{bicubic}(x_{raw}, \frac{x_n}{\alpha_x}, \frac{y_n}{\alpha_y})$
    end for
    return $x_{256\times256}$
end function
function ADA UNIT$(x_{256\times256}, x_a)$
    Sample noise vector $(\vec{z})$, generate style vector $\vec{w}$ using MLP on $(\vec{z})$
    Generate synthetic image x_g using generator $\mathcal{G}$ with $(\vec{z})$ and $(\vec{w})$
    Calculate $\mathscr{L}_{\mathcal{D}}$ and $\mathscr{L}_{\mathcal{G}}$; update combined loss $\mathscr{L}_{\mathcal{ADA}}$ with regularization
    $x_a \leftarrow$ combine generated synthetic x_g with real CXR x_r
    return x_a
end function
function CXR$_{enhance}(x_g, x_{clahe})$
    Apply CLAHE (8×8 window, clipping 2.0) to enhance contrast
    return x_{clahe}
end function
function RESUNET++(x_{clahe}, x_{seg})
    Encoder $\mathscr{E}_{seg}(x_{clahe})$ generates feature map $\vec{\mathcal{F}}_{seg}(x_{clahe})$
    Decoder $\mathscr{D}_{seg}(\vec{\mathcal{F}}_{seg})$ generates binary mask $\mathcal{B}_{mask}$
    $\mathcal{L}_{mask} \leftarrow \mathcal{B}_{mask}$ AND original CXR
    Extract pixels within mask area to obtain segmented CXR
    return x_{seg}
end function
function RS$_{ResNet}(x_{seg}, x_{rs})$
    Feature Extraction: $\mathcal{F}_{rs} = \prod_{i=1}^{5} B^{(i)}_{ResNet}(x_{seg})$
    Final rib suppression: $x_{rs} = C_{deep}(\mathcal{F}_{rs})$
    return x_{rs}
end function
$x_{256\times256} \leftarrow$ RESIZE$_{bicubic}(x_{raw})$
$x_a \leftarrow$ ADA UNIT$(x_{256\times256})$
$x_{clahe} \leftarrow$ CXR$_{enhance}(x_a)$
$x_{seg} \leftarrow$ RESUNET++(x_{clahe})
$x_{rs} \leftarrow$ RS$_{ResNet}(x_{seg})$

tion from CXR images effectively. Unlike standard convolutions, dilated convolutions allow for a larger receptive field while maintaining output size by inserting gaps between kernel elements. The functionality of the SFx module is outlined in Algorithm 4.2. In the SFx

module, a dilation rate of 2 is applied to a 3×3 kernel, resulting in a larger receptive field that captures more data while maintaining computational efficiency. Figure 4.6 illustrates the difference between standard and dilated convolution. Both have the same kernel size, but dilated convolution utilizes a larger receptive area, preserving resolution in feature maps.



(a) Standard Convolution (3×3)

(b) Dilated Convolution $(3 \times 3,\ d = 2)$

Figure 4.6: Comparison between (a) standard 3×3 convolutions and (b) dilated convolutions (dilation rate 2).

As shown in the Figure 4.6, a dilated convolution with a dilation rate of 2 and a 3×3 kernel is functionally similar to a standard convolution with a 5×5 kernel. Equation (4.8) formalizes the Dilated Convolution operation.

$$y[i,j] = \sum_{m=0}^{k-1} \sum_{n=0}^{k-1} x[i - d \cdot m, j - d \cdot n] \cdot f[m,n] \tag{4.8}$$

Here, $y[i,j]$ signifies the output value at the spatial location (i,j) within the feature map. The input feature map itself is denoted by x while f represents the filter kernel. Notably, the dimensions of the filter kernel are captured by k. The dilation rate, denoted by d, controls the spacing between elements within the filter kernel, influencing the receptive field of the convolution operation.

Figure 4.2 illustrates the architecture of the SFx module for feature extraction. Additionally, it incorporates pixel distribution within the image space as part of its feature set. Internally, the SFx module comprises five feature extractor blocks, denoted as SFx_i, $i \in [1, 2 \ldots 5]$ each containing a sequence of convolution operations denoted as $\mathbb{C}^f_{k \times k}$, where $\mathbb{C}$ represents the Conv2D operation, k represents the kernel size, and $f \in [32, 64, 128, 256, 512]$

represents the number of filters employed by each block. To enhance feature robustness, each block incorporates a dropout layer [109] $(\delta_{0.1})$ with a dropout rate of 0.1, followed by batch normalization [110] (β_n) and max-pooling (ρ_{max}). Finally, all extracted features are flattened using the flatten layer $(\digamma)$.

The grid-structured input layer directly feeds the rib-suppressed CXR (x_{rs}) into a sequence of five Spatial Feature Extractor (SFx_i) blocks, where i ranges from 1 to 5. Each block employs $\mathbb{C}^{f}_{k\times k}$ convolutional operations with ReLU activation to actively extract dominant local features. Notably, ReLU is chosen over Sigmoid and Tanh for its reduced computational cost and prevention of vanishing gradients. Simultaneously, a batch normalization layer [110] (β_n) calculates and normalizes the mean and variance of the mini-batch, facilitating faster learning. The normalized output is downsampled by half using a max-pooling layer $(\underset{2\times2}{\rho_{max}})$ with a window of 2×2 anda stride of 2, retaining key features. To prevent overfitting, a dropout layer [109] $\delta_{0.1}$ layer randomly drops neurons during training, encouraging the network to learn more robust features and reduce dependence on individual neurons. Finally, the flatten layer transforms the multidimensional feature map into a one-dimensional vector, denoted as $\mathcal{F}_{SFx} \in \mathbb{N}^{2304}$ with dimensions $N \times 2304$, where N represents the number of input samples. This $\mathcal{F}_{SFx}$ vector is then fed into the Dual Hybrid Feature Fusion and Classification (DHFFC) module to combine local fine-grained features with long-range global contrastive features.

4.2.3.2 Contrastive-based Transformer Feature Extractor Module

Contrastive-based Transformer Feature Extractor (CoTFx) module combines the Vision Transformer Vision Transformer (ViT) and contrastive learning. The architecture of the ViT is schematically illustrated in Figure 4.2. The network is composed of Patch Division, Position Embedding, and Transformer blocks that focus on extracting global information from the input images. Contrastive learning is introduced as a loss function between features from the SFx module and CoTFx module to boost the CXR representations. Contrastive loss improves feature learning by focusing on intra-class similarity and inter-class dissimilarity. Additionally, a cross-entropy loss is incorporated after contrastive loss to drive the network toward accurate classification based on learned features. Leverages both aspects, potentially leading to better performance than using either loss alone. Algorithm 4.2 outlines the work-

ing of the CoTFx module.

(a) Transformer-based Feature Extractor: Transformer-based Feature Extractor leverages the power of Contrastive Language-Image Pretraining Contrastive Language-Image Pretraining (CLIP) framework [111]. The CLIP model utilizes both image and text encoders to generate semantically similar and dissimilar text-image pairs. In our experimentation, we leverage the visual encoder of the CLIP as a backbone to extract global features. Specifically, the ViT-B/32 variant, with its robust feature representation, is pre-trained by vision-language contrastive learning. The ViT/B-32 captures the long-range dependencies between different CXR regions, potentially leading to improved performance on the classification of normal or pneumonia. The ViT-B/32 image encoder operates in three key stages:

(i) Patch Embedding and Positional Encoding: The input rib-suppressed CXRs are partitioned into non-overlapping square patches of 16×16 pixels. Each patch is flattened into a vector and embedded into a lower-dimensional space using a linear layer. This creates a patch embedding $e_i \in \mathbb{R}^{d_p}$ for each patch i, where d_p is the embedding dimension. Additionally, positional encoding, represented by $E_{pos}(i)$ is added to each patch embedding to account for their relative location within the image, $\tilde{p}_i = p_i + E_{pos}(i)$.

(ii) Transformer Encoder: The core of the ViT-B/32 backbone lies in its 12 stacked transformer encoder blocks, each containing multi-head attention, feed-forward networks, layer normalization, and residual connections. Multi-head attention allows the model to learn long-range global relationships between different patches by attending to their individual embeddings and aggregating relevant information. This can be expressed in Equation (4.9) and Equation (4.10).

$$MultiHead = Concat\left(head_1, \ldots, head_n\right) \tag{4.9}$$

$$head_i = Softmax\left(\frac{\mathcal{Q}\mathcal{K}_i^T}{\sqrt{d_{\mathcal{K}}}}\right)\mathcal{V}_i \tag{4.10}$$

Here, $\mathcal{Q}$, $\mathcal{K}$, $\mathcal{V}$ represent the query, key, and value matrices, n signifies the number of attention heads respectively, and $d_{\mathcal{K}}$ is the dimension of the key vectors. Feed-forward networks provide non-linearity and enhance the network's expressive power. Layer normalization and residual connections ensure stable training and information flow through the network effec-

tively. Layer normalization scales and centers the activations of each layer, while residual connections add the original patch embedding to the transformed output, facilitating gradient flow.

(iii) Multi-Layer Perceptron: Finally, the output represents the extracted global visual features of the rib-suppressed CXR. The extracted features undergo dimensionality reduction via a Multi-Layer Perceptron (MLP) head, resulting in a compressed 512-dimensional embedding $\mathcal{F}_{CoTFx} \in \mathbb{N}^{512}$ that captures the salient lung patterns, suggesting its potential for improved CXR analysis.

(b) Contrastive Learning: Contrastive learning focuses on a distance metric, aiming to minimize the distance between similar images (positive pairs) and maximize the distance between dissimilar ones(negative pairs) [112]. The goal is to narrow the distance between the set of normal and pneumonia samples using the cosine similarity metric. The distance between similar pairs and dissimilar pairs focuses on key distinguishing features generated from the SFx module and the CoTFx module. Let these features be denoted as $\langle \overrightarrow{x_{SFx}}, \overrightarrow{x_{CoTFx}} \rangle$. The goal is to increase the similarity index for similar features and decrease the similarity index for dissimilar features. The similarity index $\ddot{s}$ is given in Equation (4.11), which is the dot product (i.e., cosine similarity) for the vector representation of $\overrightarrow{x_n}$ and $\overrightarrow{x_p}$, and can be mathematically calculated as:

$$\ddot{s}\left(\overrightarrow{x_{SFx}}, \overrightarrow{x_{CoTFx}}\right) = \frac{\overrightarrow{x_{SFx}} \cdot \overrightarrow{x_{CoTFx}}}{||\overrightarrow{x_{SFx}}||\,||\overrightarrow{x_{CoTFx}}||} \tag{4.11}$$

where $\overrightarrow{x_{SFx}}$ and $\overrightarrow{x_{CoTFx}}$ represent the feature representation for the SFx module and the CoTFx module. $(\cdot)$ represents the dot product between two vectors. A value of 1 represents more similar vectors, while a value closer to 0 represent more dissimilar vectors. Next, the features with a high similarity index $\ddot{s}$ are taken, and the pair-wise contrastive loss between them is computed as given in Equation (4.12).

$$\mathscr{L}_{CL}(x, y) = -\log\left(\frac{\exp\left(\frac{\ddot{s}\left(\overrightarrow{x_{SFx}}, \overrightarrow{x_{CoTFx}}\right)}{\tau}\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq x]} \exp\left(\frac{\ddot{s}\left(\left(\overrightarrow{x_{SFx}}, \overrightarrow{x_{CoTFx}}\right)\right)}{\tau}\right)}\right) \tag{4.12}$$

Here, τ is the temperature parameter that determines the smoothness of the distribution, N

is the total number of samples, and $\mathbb{1}_{[k \neq x]} \in \{0, 1\}$ is an indicator function that results in 1 iff $k \neq x$. The $\mathscr{L}_{global}$ supports the model for adjacent mapping of similar feature vectors and distant mapping for dissimilar feature vectors. The global contrastive loss $\mathscr{L}_{global}$ is the average of the $\mathscr{L}_{CL}(x, y)$ given in Equation (4.13).

$$\mathscr{L}_{global} = \frac{1}{N(N-1)} \sum_{x=1}^{N} \sum_{y=1, y \neq x}^{N} \mathscr{L}_{CL}(x, y) \tag{4.13}$$

After obtaining the feature representations from the contrastive loss step, we use these features to compute the cross-entropy loss $\mathscr{L}_{CE}$ for pneumonia detection, formulated in Equation (4.14):

$$\mathscr{L}_{ce} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \tag{4.14}$$

where, $\mathscr{L}_{ce}$ represents the cross-entropy loss, N is the number of training samples, and i iterates over each sample. The division by N normalizes the loss across the entire dataset. y_i is the true label for sample i (0 for normal, 1 for pneumonia) and p_i is the predicted probability of pneumonia for sample i. The negative sign ensures the loss is minimized when the predicted probability aligns with the true label. Due to the different natures of the two loss functions, feature learning and classifier learning branches have different data sampling strategies. By combining contrastive loss and cross-entropy loss, the network learns discriminative representations through contrastive learning while optimizing for accurate classification of positive pneumonia samples using cross-entropy loss. The combined loss $\mathscr{L}_{combined}$ is the weighted combination of the contrastive loss and the cross-entropy loss, as formulated in Equation (4.15):

$$\mathscr{L}_{combined} = \Phi \times \mathscr{L}_{global} + (1 - \Phi) \times \mathscr{L}_{ce} \tag{4.15}$$

The variable Φ is a weighted hyperparameter inversely proportional to the epoch number. Φ controls the trade-off between the contrastive loss and the cross-entropy loss. Next, the Dual Hybrid Feature Fusion and Classification (DHFFC) Unit fuses the features obtained from the SFx module and CoTFx module and classifies them as normal or pneumonia.

4.2.3.3 Dual Hybrid Feature Fusion and Classification Module

Previous studies indicate that a combination of transformer and CNN-based techniques extracts richer feature information for medical image classification [33]. Motivated by the combination of transformer-based techniques with convolutional networks, we present a Dual Hybrid Feature Fusion and Classification (DHFFC) Unit. DHFFC efficiently utilizes the local fine-grained features of the SFx module and the long-range global features from the CoTFx module and fuses them all together to focus on the global context information along with the local detail information. The operation of the DHFFC module is illustrated in Algorithm 4.2.

Algorithm 4.2 FEATURE EXTRACTOR AND CLASSIFICATION MODULE

Input: $\langle x_{rs} \rangle \in$ input Rib Suppressed CXR.
Output: $\langle \mathcal{P} \rangle \in$ [normal, pneumonia]

function SFx($x_{rs}, \mathcal{F}_{SFx}$)
  Initialize input: $x_{SFx} \leftarrow x_{rs}$
  for $i = 1$ to 5 **do**
    $\mathcal{F}_{SFx} \leftarrow \delta_{0.1}\left(\underset{2\times2}{\rho_{max}}\left(\beta_n\left(\mathbb{C}^{f}_{3\times3}\right)\right)\right)$
  end for
  $\mathcal{F}_{SFx} \leftarrow \textit{Flatten}(\mathcal{F}_{SFx})$
  return $\mathcal{F}_{SFx} \in \mathbb{N}^{2304}$
end function
function CoTFx($x_{rs}, \mathcal{F}_{CoTFx}$)
  Divide x_{rs} into 16×16 non-overlapped patches + $\mathbf{E}_{pos}$
  Process embeddings through 12 transformer blocks using Multi-Head attention
  $\mathcal{F}_{CoTFx} \leftarrow MLP(\textit{Transformer Output})$
  Compute $\mathscr{L}_{global}$ and $\mathscr{L}_{ce}$
  $\mathscr{L}_{combined} \Leftarrow \Phi \cdot \mathscr{L}_{global} + (1 - \Phi) \cdot \mathscr{L}_{ce}$
  Minimize $\mathscr{L}_{combined}$ via network training
  return $\mathcal{F}_{CoTFx} \in \mathbb{N}^{2304}$
end function
function DHFFC($\mathcal{F}_{SFx}, \mathcal{F}_{CoTFx}$)
  $\mathcal{F}_{fused} \leftarrow \textit{Concat}(\mathcal{F}_{SFx}, \mathcal{F}_{CoTFx})$
  $\mathcal{P} \leftarrow \textit{DeepClassifier}(\mathcal{F}_{fused})$
  return $\mathcal{P} \in$ [normal, pneumonia]
end function
$\mathcal{F}_{SFx} \leftarrow$ SFx(x_{rs})
$\mathcal{F}_{CoTFx} \leftarrow$ CoTFx(x_{rs})
$\mathcal{P} \leftarrow$ DHFFC($\mathcal{F}_{SFx}, \mathcal{F}_{CoTFx}$)

DHFFC utilizes spatial features $\mathcal{F}^{2304}_{SFx}$ and global features $\mathcal{F}^{512}_{CoTFx}$ with the same spatial

dimensions to fuse the features along the channel dimension. Feature fusion is mathematically outlined in the following Equation (4.16), where $\mathcal{F}_{fused} \in \mathbb{N}^{2816}$ and $\oplus$ represents the final fused features and channel-wise concatenation.

$$\mathcal{F}_{fused} = [\mathcal{F}_{SFx}^{2304} \oplus \mathcal{F}_{CoTFx}^{512}] \tag{4.16}$$

Finally, the fused features are fed into the deep classifier with two fully connected layers followed by a sigmoid activation function. The final classification outcome determines whether the input CXR is normal or pneumonia.

4.2.3.4 Visual Explainable Algorithms

To investigate its behavior and feature focus further, we employed four vision-based explainability algorithms: GradCAM [81], GradCAM++ [85], ScoreCAM [86], and Local Interpretable Model Agnostic Explanation (LIME) [87]. These algorithms pinpoint areas of potential pneumonia on chest X-rays.

(i) GradCAM [81]: A class-discriminatory localization technique that generates visual explanations from deep learning architectures. Grad-CAM utilizes the gradient information flowing into the final convolutional layer to determine the importance of each neuron for a specific decision and provide insights into their decision-making processes. GradCAM localization map (GC_L^c) is calculated as:

$$GC_L^c(i,j) = \underset{ReLU}{\alpha} \left(\sum_k w_k^c A_{(i,j)}^k \right) \tag{4.17}$$

where, $\underset{ReLU}{\alpha}$ is the activation function. $A_{(i,j)}^k$ represents the activation of the $k-$th feature map at location (i,j), and w_k^c is the gradient-based weight corresponding to the $k-$th feature map for class (c), obtained from deep learning network.

(ii) GradCAM++ [85]: An improved version of Grad-CAM that provides visual explanations of deep learning predictions for object localization and detection of multiple object instances within an image. It utilizes a weighted combination of the positive partial derivatives of the final convolutional layer feature maps with respect to a specific class score to generate a visual explanation for the target class label. The final class-discriminatory saliency map

$(GC++_L^c)$ is calculated as:

$$GC++_L^c(i,j) = \sum_k w_k^c A_{(i,j)}^k \tag{4.18}$$

where all parameters are the same as Equation (4.17). However, Grad-CAM++ refines the weighting scheme of Grad-CAM, allowing for better localization and identification of multiple object instances.

(iii) ScoreCAM [86]: Score-CAM addresses potential shortcomings in Grad-CAM and GradCAM++, where class-specific weights might not fully capture the impact of feature maps on the final class score. Score-CAM utilizes the forward pass of the network to obtain the rectified score of the entire network rather than relying solely on gradients from the final convolutional layer. Mathematically, the Score-CAM localization map $(SC++_L^c)$ is defined as:

$$SC_L^c(i,j) = \underset{ReLU}{\alpha}\left(\sum_k w_k^c A_{(i,j)}^k\right) \tag{4.19}$$

where all parameters are the same as in Equation (4.17) except for the class-specific weights for the feature map, which are computed by summation over all succeeding layers.

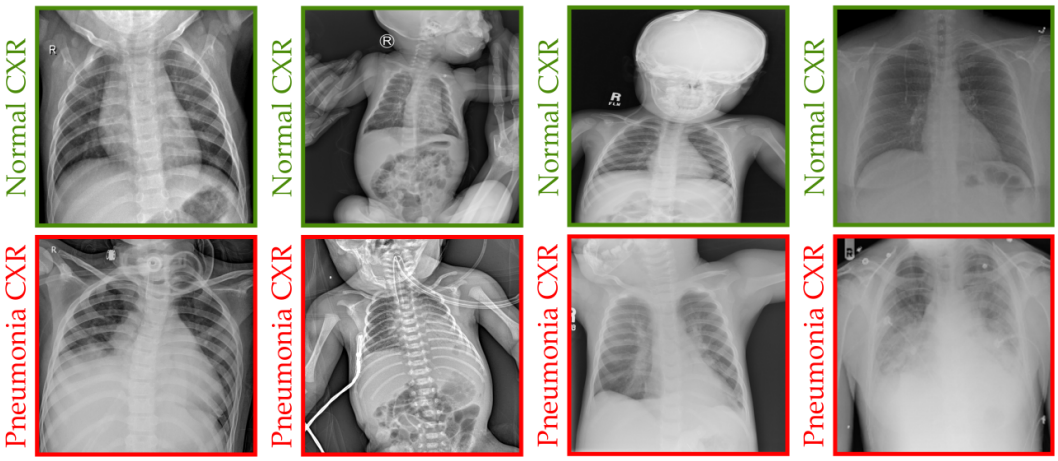
(iv) LIME [87]: Local Interpretable Model-agnostic Explanations (LIME) take a different approach to explaining model predictions. It operates as a model-agnostic technique, meaning it can be applied to explain any model, not just deep learning architectures. LIME focuses on explaining individual predictions by approximating the complex model locally around the specific data point of interest. By analyzing the weights, LIME highlights the corresponding changes in the model's predictions to understand the importance of each feature.

4.3 Experimental Settings

This section describes the datasets used to evaluate the proposed XCCNet framework for pediatric pneumonia classification. It provides details of the benchmark dataset and outlines the evaluation metrics used to assess model performance.

4.3.1 Benchmark Datasets

The proposed XCCNet framework was tested using four publicly available datasets, namely, Kermany Dataset [113], VinDr-PCXR [114], NIH-Pediatric [115], and Trivedi Dataset [20]. Figure 4.7 depicts a selection of chest radiographs utilized for the effectiveness comparison of the proposed XCCNet framework. We briefly summarize the dataset as follows:



(a) Kermany Dataset (b) VinDR-PCXR Dataset (c) NIH-Pediatric Dataset (d) Trivedi Dataset

Figure 4.7: Sample dataset images for pediatric pneumonia classification.

Table 4.1: An overview of benchmark datasets for pediatric pneumonia detection.

Dataset	**Original CXR**$^{\#}$		**Synthetic CXR***		**Total CXR**	
	Normal	**Pneumonia**	**Normal**	**Pneumonia**	**Normal**	**Pneumonia**
Kermany Dataset [113]	1583	4273	2624	0	4207	4273
VinDr-PCXR Dataset [114]	6050	1110	0	4860	6050	5970
NIH-Pediatric Dataset [115]	278	147	0	124	278	271
Trivedi Dataset [20]	1750	4400	2610	0	4360	4400

$^{\#}$ CXR from the original dataset; * Synthetic CXR synthesized using adversarial-based data augmentation unit.

(a) Kermany Dataset: The dataset provided by Kermany [113] serves as the primary source for this study, comprising 5,856 pediatric radiographs. The dataset comprises 1,583 radiographs with "normal" labels and 4,273 with "pneumonia" labels. These chest X-ray images originate from routine screenings of pediatric patients aged 1-5 years at the Guangzhou Women and Children's Medical Center. The dataset details are explored in Table 4.1.

(b) VinDr-PCXR Dataset: Pham *et al.* [114] introduce a new pediatric dataset, VinDr-PCXR, compiled from a major Vietnamese hospital between 2020 and 2021. This dataset

encompasses 9,125 studies, each meticulously annotated by a pediatric radiologist with over ten years of experience. Annotations encompass 36 critical findings and 15 diseases. For our experiments, we extracted CXR labeled as "normal" and "pneumonia." Details of the utilized dataset are presented in Table 4.1.

(c) NIH-Pediatric Dataset: Wang *et al.* [115] present a chest X-ray dataset containing 112,120 frontal-view images of 30,805 patients. Each image has text-mined labels for 14 diseases (allowing for multi-labeling), and demographic information like age and ethnicity is also provided. The National Institutes of Health (NIH) Health Center curated the dataset, which was used in this study. We specifically selected 425 images from children aged 1-5 years (matching the age range in the Guangzhou dataset). Among these, 278 were labeled as "pneumonia", and 147 showed "normal" chest findings. We refer to this subset as the "NIH-Pediatric Dataset" within our work.

(d) Trivedi Dataset: To demonstrate the robustness of our proposed framework, we employed an external dataset utilized by Trivedi *et al.* [20]. Due to the limited number of publicly available pediatric pneumonia datasets, we opted for this adult pneumonia dataset to assess our framework's efficacy and generalizability. It is important to note that Trivedi *et al.* [20] employed two datasets: the Kermany dataset and an external dataset comprising pneumonia and normal chest X-rays. We specifically selected this second, unnamed external dataset for our evaluation, henceforth referred to as the "Trivedi Dataset" in our work. This dataset comprises 6,150 chest X-rays, of which 1,750 are labeled "normal" and 4,400 are labeled "pneumonia."

4.3.2 Evaluation Metrics

The XCCNet framework, designed to detect pneumonia in chest radiographs and categorize each image as either pneumonia or normal, requires thorough evaluation to assess its feasibility. We employ four key metrics: Accuracy (Acc), Precision (Pre), Recall (Rec), and F1-score (F1). As our dataset avoids class imbalance, focusing on accuracy and F1-score provides a robust assessment of XCCNet's overall performance. Additionally, we consider precision and recall to gain insights into the model's ability to identify pneumonia cases and avoid misclassifying normal individuals accurately. Table 4.2 shows the mathematical

representations of the metrics.

Table 4.2: Evaluation metrics of the XCCNet framework for pediatric pneumonia detection in chest X-rays.

Evaluation Metrics	Mathematical Formulation
Accuracy (Acc)	$\frac{TP_{pneu} + TN_{nor}}{TP_{pneu} + FP_{pneu} + TN_{nor} + FN_{nor}}$
Precision (Pre)	$\frac{TP_{pneu}}{TP_{pneu} + FP_{pneu}}$
Recall (Rec)	$\frac{TP_{pneu}}{TP_{pneu} + FN_{nor}}$
F1 - Score (F1)	$2 \times \frac{Precision \times Recall}{Precision + Recall}$

Here, TP_{penu} is True Positive, a pneumonia CXR correctly predicted as pneumonia. TN_{nor} is True Negative, a normal CXR correctly predicted as normal. FP_{penu} is a False Positive, a normal CXR incorrectly predicted as pneumonia. FN_{nor} is False Negative, a pneumonia CXR incorrectly predicted as normal.

4.4 Experimental Results

This section presents the experimental results of the proposed method. It includes quantitative evaluation against existing approaches, followed by qualitative analysis to visualize model performance. Ablation analysis is conducted to assess the contribution of individual components. Cross-dataset evaluation is also performed to examine generalization, along with model complexity analysis.

4.4.1 Quantitative Analysis

In this section, we compare the performance of our XCCNet framework with fourteen state-of the-art (SOTA) pneumonia detection, including Trivedi *et al.* [20], Kiliçarslan *et al.* [24], Yi *et al.* [116], Usman *et al.* [27], Okolo *et al.* [29], Singh *et al.* [28], Ayan *et al.* [117], Chattopadhyay *et al.* [118], Prakash *et al.* [119], Kaya *et al.* [120], Liu *et al.* [33], Ukwuoma *et al.* [79], Yang *et al.* [80] and Hroub *et al.* [82]. We test the learning ability across four diverse datasets (Kermany [113], VinDr-PCXR [114], NIH-Pediatric [115], and Trivedi [20]), presenting both quantitative and qualitative results to comprehensively assess its capabilities.

(a) Performance on Kermany Dataset: This section evaluates the proposed XCCNet against established baseline methods on the Kermany dataset, with results presented in Table 4.3. Our XCCNet achieves outstanding performance, demonstrating an accuracy of 99.76%, precision of 99.75%, recall of 99.75%, and F1-score of 99.75%. XCCNet significantly surpasses the second-best method by Ukwuoma *et al.* [79] in terms of accuracy, recall, and F1-score, with margins of 0.55%, 0.54%, and 0.54%, respectively. While comparable precision is observed with Okolo *et al.* [29], it demonstrates the effectiveness of transformer-based methods for achieving high precision. While many existing baselines focused their evaluation on the Kermany dataset, we conducted a broader evaluation by reproducing their results from scratch and applying them to additional diverse datasets. This comprehensive comparison across different data distributions demonstrates the true generalizability and effectiveness of the XCCNet framework compared to existing methods. Notably, explainability-based techniques also exhibit promising performance. However, XCCNet's unique integration of transformers and convolution-based extraction enables it to capture both fine-grained details and long-range features for superior feature extraction, highlighting its effectiveness compared to existing methods.

Table 4.3: Quantitative analysis of XCCNet framework with existing baselines on the Kermany dataset. **Bold**: best and underlined: second-best scores.

	Methods	Acc	Pre	Rec	F1
I	Trivedi *et al.* [20]	94.23	95.00	96.00	95.00
	Kiliçarslan *et al.* [24]	95.37	96.31	94.15	95.22
	Yi *et al.* [116]	96.06	95.17	93.58	94.36
II	Usman *et al.* [27]	88.46	83.75	85.90	84.81
	Okolo *et al.* [29]	98.08	99.74	97.25	98.48
	Singh *et al.* [28]	97.61	97.12	92.97	95.00
III	Ayan *et al.* [117]	95.83	96.02	92.74	94.35
	Chattopadhyay *et al.* [118]	98.36	98.98	98.79	98.80
	Prakash *et al.* [119]	96.15	97.91	95.90	96.24
	Kaya *et al.* [120]	98.94	99.12	99.12	99.12
	Liu *et al.* [33]	97.03	97.94	97.32	97.63
IV	Ukwuoma *et al.* [79]	99.21	99.21	99.21	99.21
	Yang *et al.* [80]	95.60	95.23	94.51	94.87
	Hroub *et al.* [82]	95.55	94.54	94.55	94.54
Ours	XCCNet	**99.76**	**99.75**	**99.75**	**99.75**

I: Convolutional-based, II: Transformer-based, III: Hybrid Model-based, IV: Explainable-based

(b) Performance on VinDR-CXR Dataset: Table 4.4 showcases the comparison results on the VinDr-PCXR dataset. While the method proposed by Okolo *et al.* [29] achieves the high-

est accuracy (0.01% margin), it shares similarities with XCCNet, including the integration of vision transformers and CNNs and the use of skip-connections for feature fusion. Notably, XCCNet surpasses Ayan *et al.* [117], Kaya *et al.* [120], and even Okolo *et al.* [29] in terms of precision (by 0.30%), recall (by 0.06%), and F1-Score (by 0.64%). This demonstrates XCCNet's effectiveness even against strong contenders. While Ayan *et al.* [117] and Kaya *et al.* [120] leverage hybrid approaches for pneumonia feature learning, XCCNet's integrated adversarial-based data augmentation unit effectively tackles data skewness. This capability leads to more accurate identification of pneumonia cases and minimizes the risk of misclassifying normal CXR images.

Table 4.4: Quantitative analysis of XCCNet framework with existing baselines on the VinDR-PCXR dataset. **Bold**: best and underlined: second-best scores.

	Methods	Acc	Pre	Rec	F1
I	Trivedi *et al.* [20]	87.28	91.83	84.56	88.05
	Kiliçarslan *et al.* [24]	89.43	90.78	86.37	88.52
	Yi *et al.* [116]	89.72	90.21	87.57	88.87
II	Usman *et al.* [27]	86.34	90.79	85.81	88.23
	Okolo *et al.* [29]	**91.57**	91.38	89.48	90.42
	Singh *et al.* [28]	89.57	89.40	86.54	87.94
III	Ayan *et al.* [117]	86.72	91.88	84.79	88.19
	Chattopadhyay *et al.* [118]	91.52	89.73	88.96	89.34
	Prakash *et al.* [119]	87.45	91.05	85.34	88.10
	Kaya *et al.* [120]	91.38	90.77	89.91	90.34
	Liu *et al.* [33]	89.18	91.03	87.52	89.24
IV	Ukwuoma *et al.* [79]	91.27	90.61	89.19	89.89
	Yang *et al.* [80]	87.39	90.95	85.81	88.30
	Hroub *et al.* [82]	88.54	91.28	83.25	87.08
Ours	XCCNet	91.56	**92.18**	**89.97**	**91.06**

I: Convolutional-based, II: Transformer-based, III: Hybrid Model-based, IV: Explainable-based

(c) Performance on NIH-Pediatric Dataset: As shown in Table 4.5, the proposed XCCNet framework surpasses all other state-of-the-art (SOTA) techniques on the NIH-Pediatric dataset in terms of accuracy, precision, and F1-score. It achieves impressive values of 92.87%, 91.72%, and 91.65%, respectively. While the hybrid feature extraction technique by Kaya *et al.* [120] exhibits comparable performance, securing the second-best results across all metrics, its approach relies on an ensemble strategy with majority voting. This contrasts with XCCNet's robust preprocessing module, which effectively amplifies low-radiation CXRs into high-quality images. This enhanced input allows XCCNet to learn deeper features and extract valuable insights, ultimately leading to superior performance.

Table 4.5: Quantitative analysis of XCCNet framework with existing baselines on the NIH-Pediatric dataset. **Bold**: best and underlined: second-best scores.

	Methods	Acc	Pre	Rec	F1
I	Trivedi *et al.* [20]	85.45	86.66	87.78	87.22
	Kiliçarslan *et al.* [24]	87.71	87.23	89.14	88.17
	Yi *et al.* [116]	84.22	86.33	87.18	86.76
II	Usman *et al.* [27]	86.67	86.98	88.91	87.93
	Okolo *et al.* [29]	89.54	88.91	90.26	89.58
	Singh *et al.* [28]	88.19	87.66	89.27	88.46
III	Ayan *et al.* [117]	91.08	89.67	91.23	90.44
	Chattopadhyay *et al.* [118]	91.34	88.99	90.82	89.89
	Prakash *et al.* [119]	88.48	88.14	89.37	88.75
	Kaya *et al.* [120]	92.68	90.80	**92.07**	91.43
	Liu *et al.* [33]	91.09	88.93	91.43	90.16
IV	Ukwuoma *et al.* [79]	91.71	90.20	91.58	90.89
	Yang *et al.* [80]	88.36	88.10	89.00	88.55
	Hroub *et al.* [82]	87.05	87.16	89.28	88.20
Ours	XCCNet	**92.87**	**91.72**	91.59	**91.65**

I: Convolutional-based, II: Transformer-based, III: Hybrid Model-based, IV: Explainable-based

(d) Performance on Trivedi Dataset: This section presents the proposed XCCNet's performance on the Trivedi dataset against established baselines (refer Table 4.6). XCCNet achieves the highest accuracy, precision, and F1-score, demonstrating exceptional results with values of 97.19%, 96.72%, and 97.00%, respectively. While Okolo *et al.* [29] edge out a slightly higher recall of 97.85%, their approach heavily relies on integrating convolutions with transformers, which may not generalize as effectively across diverse datasets. XCCNet surpasses competitors in other metrics: Liu *et al.* [33] for accuracy and F1-score, and Ukwuoma *et al.* [79] for precision. Notably, the margins between the best and second-best results are minimal (0.83% for accuracy, 0.17% for precision, and 0.69% for F1-Score), showcasing the competitive performance. Compared to existing techniques, XCCNet utilizes contrastive learning with binary cross-entropy. This approach effectively enhances feature learning by considering both inter-class similarities and dissimilarities, ultimately leading to superior classification performance and accurate differentiation between pneumonia and normal cases.

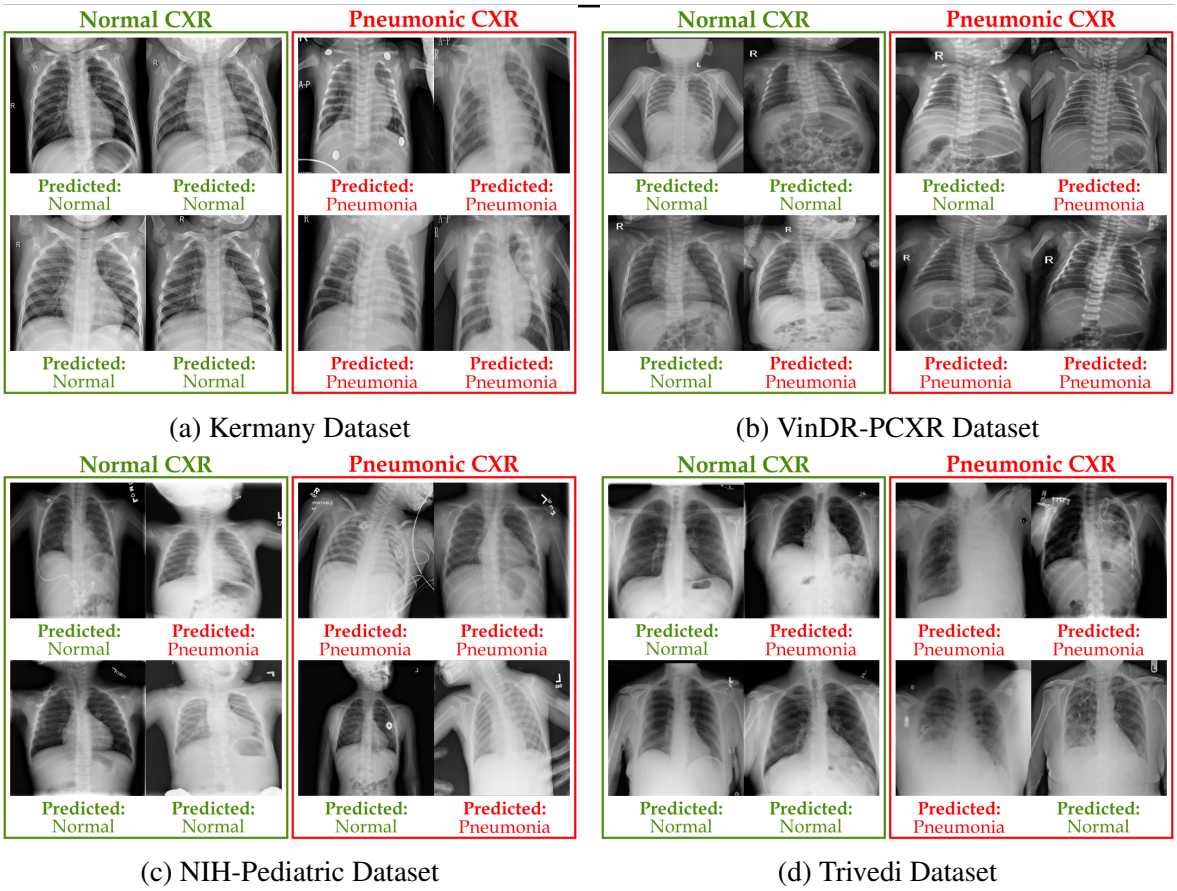
Table 4.6: Quantitative analysis of XCCNet framework with existing baselines on the Trivedi dataset. **Bold**: best and underlined: second-best scores.

	Methods	Acc	Pre	Rec	F1
I	Trivedi *et al.* [20]	87.46	90.64	89.71	90.17
	Kiliçarslan *et al.* [24]	94.43	94.78	96.68	95.72
	Yi *et al.* [116]	90.46	94.79	95.48	95.13
II	Usman *et al.* [27]	94.43	93.13	96.70	94.88
	Okolo *et al.* [29]	95.86	94.42	**97.85**	96.10
	Singh *et al.* [28]	88.52	92.91	92.41	92.66
III	Ayan *et al.* [117]	95.73	92.41	96.05	94.19
	Chattopadhyay *et al.* [118]	96.33	95.22	95.52	95.37
	Prakash *et al.* [119]	94.77	92.49	93.53	93.01
	Kaya *et al.* [120]	96.08	95.08	96.56	95.81
	Liu *et al.* [33]	96.36	96.22	96.41	96.31
IV	Ukwuoma *et al.* [79]	95.61	96.55	94.36	95.44
	Yang *et al.* [80]	92.49	94.91	91.54	93.19
	Hroub *et al.* [82]	93.14	93.71	95.67	94.68
Ours	XCCNet	**97.19**	**96.72**	97.28	**97.00**

I: Convolutional-based, II: Transformer-based, III: Hybrid Model-based, IV: Explainable-based

4.4.2 Qualitative Analysis

Figure 4.8 presents the qualitative results of diagnosing pediatric pneumonia using our proposed framework across four diverse datasets (Kermany [113], VinDr-PCXR [114], NIH-Pediatric [115], and Trivedi [20]). The framework demonstrates a robust ability to recognize healthy lung patterns and complex pneumonic lung patterns, successfully detecting pneumonia cases across multiple datasets, particularly in the Kermany and NIH-Pediatric datasets. We categorize the failure cases into two scenarios. In the first scenario, the CXR regions exhibit high similarity to the surrounding normal regions, resulting in low contrast. This low contrast makes distinguishing between CXR and normal regions challenging, hindering accurate lung area extraction. In the second scenario, anatomical variability and external noise artifacts introduce challenges. These factors make it difficult to retain feature responses, compromising the diagnostic results. Overall, the model exhibits high accuracy in correctly identifying normal and pneumonic cases across different datasets, showcasing its generalizability and diagnostic strength.



(a) Kermany Dataset

(b) VinDR-PCXR Dataset

(c) NIH-Pediatric Dataset

(d) Trivedi Dataset

Figure 4.8: Qualitative analysis of XCCNet framework.

4.4.3 Ablation Analysis

In this section, we systematically evaluate the impact of individual components within the XCCNet framework on the Kermany Dataset [113] by selectively removing or modifying them while maintaining the remaining framework. We denote ablated versions as XCCNet (technique/module), where the indicated module is the component under analysis, and "w/o" signifies the absence of a specific component.

(a) Effect of Adversarial-based Data Augmentation: To analyze the contributions of adversarial-based data augmentation, we conducted an ablation study with three variants: no augmentation, traditional augmentation, and our proposed adversarial-based data augmentation (ADA) within the XCCNet framework. Results in Table 4.7 demonstrate that the absence of augmentation leads to a 3.47% degradation in accuracy, indicating overfitting. While traditional augmentation techniques like scaling, cropping, translation, flipping, and rotation improve performance, they still lag 2.58% behind XCCNet. Notably, incorporating

the ADA unit yields superior performance. These findings confirm the efficacy of the ADA unit in enhancing XCCNet's pneumonia detection capabilities. The ADA unit leverages adversarial learning principles to synthetically generate realistic CXRs that preserve essential features, effectively mitigating data skewness and boosting XCCNet's performance.

Table 4.7: Ablation analysis of XCCNet framework with different adversarial-based data augmentation techniques on the Kermany dataset.

Methods	Acc	Pre	Rec	F1
$\text{XCCNet}_{\text{No Augmentation}}$	96.29	94.31	97.48	95.87
$\text{XCCNet}_{\text{Traditional Augmentation}}$	97.18	94.52	98.65	96.54
$\text{XCCNet}_{\text{Adversarial-based Data Augmentation}}$	**99.76**	**99.75**	**99.75**	**99.75**

(b) Effect of CXR Enhancer: Table 4.8 summarizes the impact of CXR enhancer, which enhances the quality of CXR utilizing Contrast Limited Adaptive Histogram Equalization CLAHE. CLAHE enhances the contrast and improves the visibility of subtle features, leading to improved feature extraction and diagnostic accuracy. Results of CLAHE are compared with raw CXRs and other image enhancement techniques, namely contrastive enhancement, histogram equalization, and sharpening. Notably, leveraging the sharpen yields inferior results with an accuracy drop of 5.22%. Histogram equalization yields the second-best performance in terms of accuracy. The result demonstrated that the utilization of CLAHE in CXR yields the best performance in the detection of pneumonia.

Table 4.8: Ablation analysis of XCCNet framework with different CXR enhancer techniques on the Kermany dataset.

Methods	Acc	Pre	Rec	F1
$\text{XCCNet}_{\text{Raw CXR}}$	96.88	94.57	94.32	94.44
$\text{XCCNet}_{\text{Contrastive Enhancement}}$	97.01	94.22	95.64	94.92
$\text{XCCNet}_{\text{Histogram Equalization}}$	97.88	96.83	97.52	97.17
$\text{XCCNet}_{\text{Sharpen}}$	94.54	96.66	96.09	96.37
$\text{XCCNet}_{\text{CLAHE}}$	**99.76**	**99.75**	**99.75**	**99.75**

(c) Effect of CLAHE: Table 4.9 demonstrates that applying CLAHE significantly enhances the performance of XCCNet in detecting pediatric pneumonia in CXR images. Increasing CLAHE parameters (clip limit and grid size) progressively improves the framework Acc with an absolute improvement of 4.61%. A clip limit of 2.0 and a grid size of 8×8 yield the highest performance, and we utilize this configuration in our proposed framework. Figure 4.9

illustrates the variations in normal and pneumonic CXR images using different CLAHE parameters, aligning with the results presented in Table 4.9. These findings highlight the importance of optimizing CLAHE parameters to enhance the diagnostic performance of XCCNet for pediatric pneumonia detection in CXR images.

Table 4.9: Ablation analysis of XCCNet framework with different CLAHE parameters on the Kermany dataset.

Methods	Acc	Pre	Rec	F1
$\text{XCCNet}_{\text{CLAHE: CL(0.5)+GS(2×2)}}$	95.15	95.48	94.53	95.00
$\text{XCCNet}_{\text{CLAHE: CL(1.0)+GS(4×4)}}$	97.09	97.74	96.29	97.01
$\text{XCCNet}_{\text{CLAHE: CL(1.5)+GS(6×6)}}$	98.06	98.99	97.04	98.01
$\text{XCCNet}_{\text{CLAHE: CL(2.0)+GS(8×8)}}$	**99.76**	**99.75**	**99.75**	**99.75**

CL: clipLimit, GS: tileGridSize

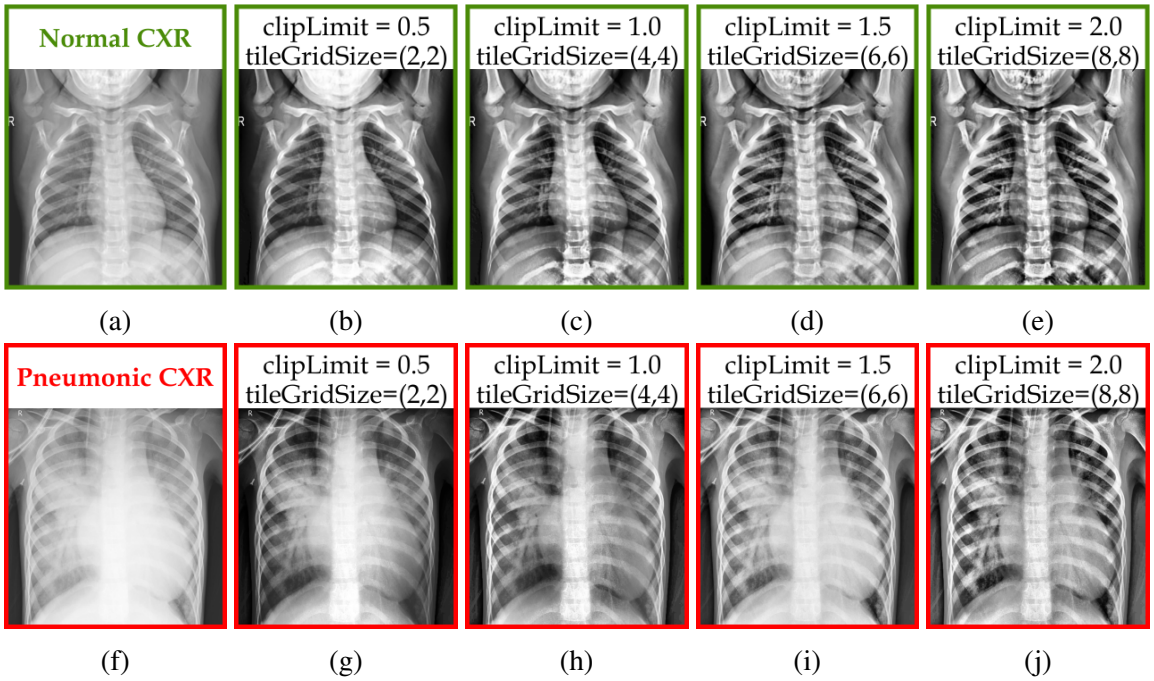


Figure 4.9: Impact of varying CLAHE parameters on chest X-rays: (a)–(e) Normal case and (f)–(j) Pneumonic case

(d) Effect of Segmentator: As detailed in Table 4.11, our proposed framework, $\text{XCCNet}_{\text{ResUNet++}}$, achieves superior performance. Replacing ResUNet++ with UNet [46] and UNet++ [121] significantly reduces accuracy by 1.92% and 1.74%, respectively. Utilizing ResUNet++ facilitates isolating affected regions for enhanced analysis and accurately identifying the Region of Interest (ROI). Extracting the segmented lung area enables the

framework to learn more targeted features, improving diagnostic effectiveness. By concentrating on segmented CXRs, our framework captures more precise and discriminative features for accurately detecting pediatric pneumonia.

Table 4.10: Ablation analysis of XCCNet framework with different segmentators on the Kermany dataset.

Methods	Acc	Pre	Rec	F1
$\text{XCCNet}_{\text{UNet}}$	97.84	96.94	98.55	97.74
$\text{XCCNet}_{\text{UNet++}}$	98.02	98.35	97.15	97.75
$\text{XCCNet}_{\text{ResUNet++}}$	**99.76**	**99.75**	**99.75**	**99.75**

(e) Effect of Rib Suppressor: Table 4.11 presents the performance comparison of XCCNet with and without rib suppression. Integrating a rib suppressor module into the XCCNet framework significantly improves the Acc, Pre, Rec, and F1 by 1.71%, 2.92%, 2.27%, and 2.60% compared to the model without rib suppression. The rib suppressor module enhances lung visibility by removing rib structures, allowing the XCCNet model to focus on lung parenchyma and other relevant anatomical features. Furthermore, rib suppression enhances the CXRs with more discriminative features for pneumonia detection. By suppressing the ribs, the framework effectively learns complex patterns and variations associated with pneumonia, resulting in improved diagnostic accuracy to identify both positive and negative cases of pneumonia accurately.

Table 4.11: Ablation analysis of XCCNet framework with different rib suppressor techniques on the Kermany dataset.

Methods	Acc	Pre	Rec	F1
$\text{XCCNet}_{\text{w/o Rib Suppressor}}$	98.59	96.83	97.48	97.15
$\text{XCCNet}_{\text{with Rib Suppressor}}$	**99.76**	**99.75**	**99.75**	**99.75**

(f) Effect of Feature Extractor: Table 4.12 presents an ablation study investigating the impact of feature extractors on XCCNet performance. The combination of SFx($d_r = 2$) and CoTFx (ViT-B/32) achieves the highest performance across all metrics, demonstrating a minimal performance drop of 3.64% and 2.54% compared to other configurations. The choice of dilation convolution within the XCCNet framework expands the receptive field without increasing the parameters and captures fine-grained details within the CXRs. We specifically

chose the ViT-B/32 variant to capture global features due to its superior performance compared to other ViT variants (ViT-B/16 and ViT-L/14). Integrating SFx and CoTFx is crucial for the superior performance of XCCNet in pediatric pneumonia detection.

Table 4.12: Ablation analysis of XCCNet framework with different feature extraction techniques on the Kermany dataset.

	Methods	Acc	Pre	Rec	F1
I	$\text{XCCNet}_{\text{SFx}(d_r = 1)}$	94.91	93.72	95.64	94.67
	$\text{XCCNet}_{\text{SFx}(d_r = 2)}$	96.12	96.23	95.75	95.99
	$\text{XCCNet}_{\text{SFx}(d_r = 3)}$	95.15	94.72	95.20	94.96
II	$\text{XCCNet}_{\text{CoTFx(ViT-B/16)}}$	95.03	94.22	95.42	94.82
	$\text{XCCNet}_{\text{CoTFx(ViT-B/32)}}$	<u>97.22</u>	<u>96.50</u>	<u>97.72</u>	<u>97.11</u>
	$\text{XCCNet}_{\text{CoTFx(ViT-L/14)}}$	95.52	94.72	95.93	95.32
Ours	**$\text{XCCNet}_{\text{Hybrid Feature Extractor}}$**	**99.76**	**99.75**	**99.75**	**99.75**

I: Spatial Feature Extractor (SFx), II: Coupled Transformer (CoTFx)

(g) Effect of Feature Fusion: The ablation study investigates the impact of serial and parallel feature fusion approaches on XCCNet performance. The parallel fusion technique actively incorporates contrastive loss within samples, while the serial fusion technique does not. Results in Table 4.13 indicate that parallel fusion of full features from both the Spatial Feature Extractor (SFx) and Contextual Transformer Feature Extractor (CoTFx) yields the highest accuracy of 99.76%. In serial fusion, applying CoTFx before SFx shows competitive performance with a marginal performance drop of 0.85%, suggesting that capturing global features before local features improves results. To assess computational efficiency, partial features (75% of total features) were also considered, but their utilization resulted in performance degradation. Overall, the study indicates that parallel fusion of full features is the most effective strategy for enhancing XCCNet performance, outperforming partial feature utilization and serial fusion.

4.4.4 Cross-Dataset Analysis

We evaluated the generalizability of XCCNet through cross-dataset training experiments on four publicly available pediatric pneumonia X-ray datasets: Kermany [113], VinDr-PCXR [114], NIH-Pediatric [115], and Trivedi [20]. Table 4.14 comprehensively summarizes the performance of XCCNet when trained on each dataset and tested on all others. The results indicate that training on a dataset similar to the testing dataset generally yields the best

Table 4.13: Ablation analysis of XCCNet framework with different feature fusion techniques on the Kermany dataset.

	Methods	Acc[a]	Pre[b]	Rec[c]	F1[d]
I	$XCCNet_{SFx + CoTFx}$	97.58	96.75	98.22	97.48
	$XCCNet_{CoTFx + SFx}$	98.91	98.99	98.75	98.87
II	$XCCNet_{SFx(P) + CoTFx(P)}$	95.65	96.75	94.39	95.56
	$XCCNet_{SFx(F) + CoTFx(P)}$	97.94	98.25	97.52	97.88
	$XCCNet_{SFx(P) + CoTFx(F)}$	98.79	99.00	98.51	98.76
Ours	**$XCCNet_{SFx(F) + CoTFx(F)}$**	**99.76**	**99.75**	**99.75**	**99.75**

I: Serial Fusion, II: Parallel Fusion (Contrastive). P: Partial, F: Full

performance. For example, training and testing on the Kermany dataset achieves the highest accuracy of 99.76%. When trained and tested on the same data, similar patterns are observed for the VinDr-PCXR, NIH-Pediatric, and Trivedi datasets. Interestingly, XCCNet trained on the Kermany dataset demonstrates a high accuracy of 91.64% on the NIH-Pediatric dataset. The high performance is attributed to the framework's ability to learn generalizable features from the Kermany dataset. These findings suggest that XCCNet exhibits robust performance even when variations exist in the diverse training and testing datasets.

Table 4.14: Cross-Dataset analysis of XCCNet framework with different training and testing datasets.

Training Dataset	**Testing Dataset**							
	Kermany Dataset				**VinDr-PCXR Dataset**			
	Acc	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**
Kermany Dataset [113]	**99.76**	**99.75**	**99.75**	**99.75**	88.93	90.45	87.80	89.11
VinDr-PCXR Dataset [114]	89.45	91.21	87.47	89.30	**91.56**	**92.18**	**89.97**	**91.06**
NIH-Pediatric Dataset [115]	80.97	83.67	78.35	80.92	76.00	80.40	72.73	76.37
Trivedi Dataset [20]	96.48	96.48	96.24	96.36	89.94	91.71	87.95	89.79
Training Dataset	**Testing Dataset**							
	NIH-Pediatric Dataset:				**Trivedi Dataset**			
	Acc	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**
Kermany Dataset [113]	91.64	**93.97**	89.26	91.55	93.21	93.47	92.54	93.00
VinDr-PCXR Dataset [114]	90.42	91.46	89.00	90.21	92.00	90.95	92.35	91.65
NIH-Pediatric Dataset [115]	**92.87**	91.72	**91.59**	**91.65**	78.55	83.67	74.83	79.00
Trivedi Dataset [20]	90.79	90.95	90.05	90.50	**97.19**	**96.72**	**97.28**	**97.00**

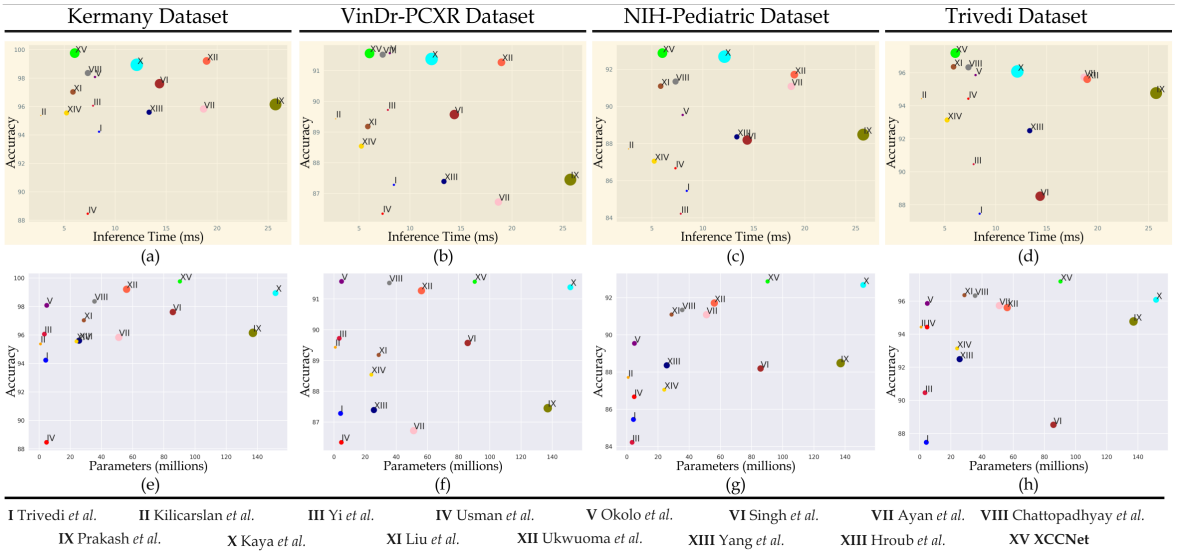


Figure 4.10: Comparative analysis of model parameters, inference time, and accuracy for XCCNet framework against existing baselines on benchmark datasets.

4.4.5 Model Complexity Analysis

Table 4.15 compares the complexity of the XCCNet framework with existing baselines for pneumonia detection, considering inference time and the number of parameters. Convolutional-based models like those proposed by Trivedi *et al.* [20] and Kiliçarslan *et al.* [24] exhibit fast inference times (under ten milliseconds) with a moderate number of parameters (under 5 million). In contrast, transformer-based models such as the one proposed by Singh *et al.* [28] demonstrate significantly higher complexity, with over 85 million parameters and inference times exceeding 14 milliseconds. Our proposed XCCNet framework strikes a balance between computational efficiency and model complexity. While possessing a moderate number of parameters (around 90 million), XCCNet achieves a competitive inference time of approximately six milliseconds. Only 2 million of XCCNet's total 90 million parameters are trainable. Overall, XCCNet demonstrates a promising approach by achieving good accuracy with a balanced level of complexity, making it suitable for real-world deployment scenarios where computational resources may be limited. Figure 4.10 illustrates a comparative analysis of inference time and model parameters for the XCCNet framework and existing baselines in pneumonia detection.

Table 4.15: Model Complexity Analysis of the XCCNet framework comparing parameters and inference times with existing baselines.

	Methods	Inference Time (ms)	Parameters (M)
I	Trivedi *et al.* [20]	8.42	4.25
	Kiliçarslan *et al.* [24]	2.71	0.88
	Yi *et al.* [116]	7.83	3.41
II	Usman *et al.* [27]	7.31	4.75
	Okolo *et al.* [29]	8.02	4.89
	Singh *et al.* [28]	14.35	85.80
III	Ayan *et al.* [117]	18.67	51.12
	Chattopadhyay *et al.* [118]	7.32	35.54
	Prakash *et al.* [119]	25.73	137.16
	Kaya *et al.* [120]	12.12	151.65
	Liu *et al.* [33]	5.86	24.50
IV	Ukwuoma *et al.* [79]	18.97	56.13
	Yang *et al.* [80]	13.33	25.63
	Hroub *et al.* [82]	5.23	24.36
Ours	XCCNet	6.03	90.31

I: Convolutional-based, II: Transformer-based, III: Hybrid Model-based, IV: Explainable-based, M: million, ms: milliseconds.

4.5 Conclusion

This chapter presents an Explainable Contrastive-based Dilated Convolutional Network (XCCNet) designed for pediatric pneumonia detection in chest X-rays (CXR). We leverage the synergy of contrastive learning, transformers, and convolutional networks to extract global and fine-grained spatial features from CXRs effectively. XCCNet's preprocessing module enhances CXRs with low-radiation intensity and segments the lung area for targeted analysis. Additionally, a novel adversarial-based data augmentation technique mitigates skewed data distribution by generating high-quality synthetic CXRs. By incorporating explainability into XCCNet, we provide valuable insights into lung pathology for medical experts and radiographers. Extensive evaluation across four benchmark datasets demonstrates the superiority of XCCNet. Its potential clinical utility lies in accurately diagnosing normal and pneumonic CXRs. Despite its promising performance, the XCCNet framework lacks pneumonia severity prediction capabilities. Exploring semi-supervised learning within XCCNet will enhance its potential to assist radiologists in accurate disease prediction.

Chapter 5

Hierarchical Fusion of Semi and Self-Supervised Learning for Knee Osteoarthritis Severity Grading

CONTENTS

5.1 Introduction . . . 89
5.2 Methodology . . . 90
5.2.1 Framework Overview . . . 90
5.2.2 Preprocessing . . . 91
5.2.3 Proposed Framework . . . 91
5.3 Experimental Settings . . . 106
5.3.1 Benchmark Datasets . . . 106
5.3.2 Evaluation Metrics . . . 108
5.4 Experimental Results . . . 108
5.4.1 Quantitative Analysis . . . 109
5.4.2 Qualitative Analysis . . . 111
5.4.3 Ablation Analysis . . . 112
5.4.4 Cross-Dataset Analysis . . . 116
5.4.5 Model Complexity Analysis . . . 119
5.5 Conclusion . . . 120

5.1 Introduction

This chapter focuses on the classification of the knee osteoarthritis (KOA) severity and addresses the objective of data-efficient learning for medical image analysis. The aim is to develop a method for a reliable grading performance under limited labeled samples. This preserves the sensitivity to disease progression in KOA. Knee osteoarthritis is a common degenerative joint disorder that causes pain, including stiffness, and reduces mobility in older adults. Early and accurate severity assessment is important for treatment planning and management. Knee X-rays (KXR) are widely used for clinical grading of KOA. Still, the final interpretation depends on the expert judgment. In addition, the knee radiographs exhibit low contrast, blurred joints, overlapping anatomical structures, and variability in acquisition settings. These factors make feature extraction more complex, especially when grade differences are very close. Existing supervised approaches require large datasets; on the other hand, the semi-supervised methods rely on noisy pseudo-labels. These limitations motivate the frameworks to leverage labeled and unlabeled data effectively under limited supervision.

To address these limitations, this chapter presents a Hierarchical Semi-Supervised framework with Self-Supervision (H-SemiS) for grading knee osteoarthritis severity. The framework leverages self- and semi-supervised learning with reconstructed proxy samples to support classification with limited labeled data. A hierarchical mechanism decomposes the multi-class grading task into binary decisions, which promotes balanced learning across different severity levels. An adversarial-inspired masked image reconstruction module generates reconstructed samples that improve feature coverage. A Similarity-aware Reconstructed Image Labeler assigns labels to unlabeled samples while filtering noisy predictions. In addition, quantum-inspired learning principles are incorporated to model complex anatomical relationships within knee radiographs.

The remainder of this chapter is organized as follows. Section 5.2 describes the methodology, covering masked image reconstruction, similarity-aware labeling, and the quantum-infused teacher-student architecture. Section 5.3 provides details of the benchmark datasets and evaluation metrics. Section 5.4 presents the quantitative and qualitative findings, including ablation studies and model complexity analysis. Finally, Section 5.5 summarizes the findings and discusses the clinical utility of the H-SemiS framework.

5.2 Methodology

This section presents the proposed framework for grading knee osteoarthritis severity. It describes the overall pipeline, including preprocessing and framework design. The framework captures the key features as a learnable and meaningful tool, as detailed in the following subsections.

5.2.1 Framework Overview

As illustrated in Figure 5.1, the framework operates in a three-stage pipeline to address data scarcity and class imbalance. STAGE 1 (Masked Image Reconstruction) generates proxy data by utilizing unlabeled samples from $\mathcal{D}_u$ and partitioned into non-overlapping patches $x_u^k = \{x_u^{(1)}, \ldots, x_u^{(p)}\}$, where each patch $x^k \in \mathbb{R}^{p^2 \times ch}$. The positional information is added, and 75% of patches are randomly masked, after which the model reconstructs the missing regions by minimizing the reconstruction loss ($\mathcal{L}_{rec}$), producing reconstructed samples x_{rec}^u. Next, STAGE 2 (Similarity-aware Reconstructed Image Labeler) assigns proxy labels y_{rec} to these samples by maximizing a combined cosine and Euclidean similarity metric $\ddot{\mathbb{S}}_i(x_{rec}^u, \mathcal{T}_i)$ against class-specific templates $\mathcal{T}_i$. Finally, in STAGE 3 (Hierarchical Quantum-infused Teacher-Student Framework), the proxy-labeled reconstructed samples and weakly and strongly augmented views of unlabeled data, $x_w = \mathcal{A}_w(x_u)$ and $x_s = \mathcal{A}_s(x_u)$, are jointly used within a hierarchical quantum-infused teacher–student framework. The teacher parameters θ_{Te} are updated using an exponential moving average of the student parameters θ_{St}, and the trained teacher model is finally used to predict KOA severity.

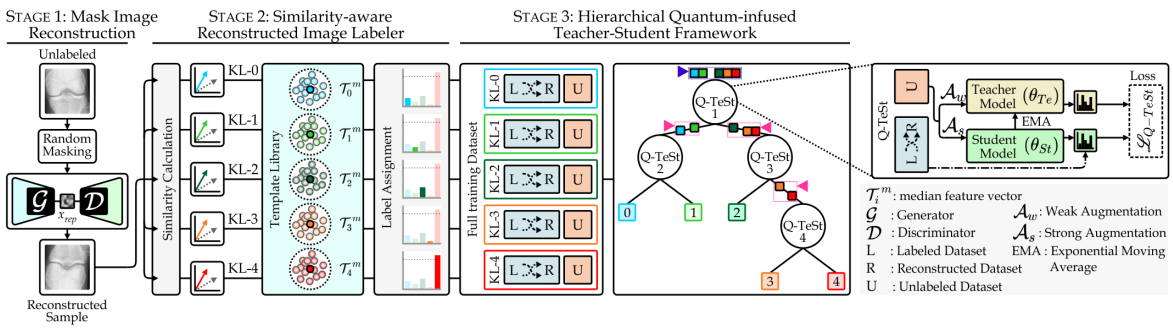


Figure 5.1: The workflow of our proposed H-SemiS framework for knee osteoarthritis severity grading in knee X-rays.

5.2.2 Preprocessing

To ensure uniform inputs across the dataset, which includes KXR capturing bilateral knees at varying resolutions, we employ YOLOv2 [122]. This detection step ensures that bilateral scans are accurately separated, yielding single-knee radiograph samples. Furthermore, to isolate the specific area of interest prior to model training, any images encompassing extended surrounding bone structures, such as the femur, tibia, or fibula, are cropped to focus on the targeted knee joint. Following this extraction, the single-knee images are standardized to the resolution of 224×224. To address variations in image quality and exposure across the datasets, we applied Contrast Limited Adaptive Histogram Equalization (CLAHE) [15]. This enhancement step improves the visibility of low-level structures and the overall image contrast before the radiographs are provided for framework learning. Furthermore, we employ BoneFinder [123] to detect key anatomical landmarks of the knee and preserve joint space narrowing Joint Space Narrowing (JSN), which plays an important role in accurately capturing the structural characteristics associated with knee osteoarthritis.

Augmentations: We adopt two augmentation strategies for consistency training in the teacher–student framework (see Section 5.2.3.3): weak $(\mathcal{A}_w)$ and strong $(\mathcal{A}_s)$. For weak augmentation $(\mathcal{A}_w)$, applied to the teacher, we flip KXR samples vertically and/or horizontally with probability 0.5 and randomly translate them by up to 0.2 of the image dimensions along each axis. For strong augmentation $(\mathcal{A}_s)$, applied to the student, we use RandAugment [124], which applies a sequence of transformations including pixel inversion, X/Y shearing in [-0.3, 0.3], scaling in [0.51, 0.60], horizontal and vertical translation in [-0.3, 0.3], brightness adjustment in [0.05, 0.95], rotation in [-30°, 30°], HSV ratio modification in [-1, 1], and Gaussian blurring with kernel sizes in [3,5].

5.2.3 Proposed Framework

5.2.3.1 Stage 1: Masked Image Reconstruction

Inspired by recent advances in self-supervised masked image reconstruction [125–127], we employ self-supervised learning within a semi-supervised framework for grading osteoarthritis severity. The proposed Masked Image Reconstruction Masked Image Recon-

struction (MI-Rec) module builds on Masked Autoencoders [127] and Generative Adversarial Networks [128]. Through adversarial reconstruction of masked images, MI-Rec learns robust anatomical features from unlabeled data and generates high-fidelity proxy samples that (i) alleviate data scarcity, (ii) balance class distributions, and (iii) inject diverse discriminative features into the training set. As shown in Figure 5.2, MI-Rec comprises a reconstruction generator ($\mathcal{G}_{MI\text{-}Rec}$) and a reconstruction discriminator ($\mathcal{D}_{MI\text{-}Rec}$), where the generator reconstructs masked regions, and the discriminator distinguishes real patches from reconstructed ones, ensuring anatomical correctness and structural consistency.

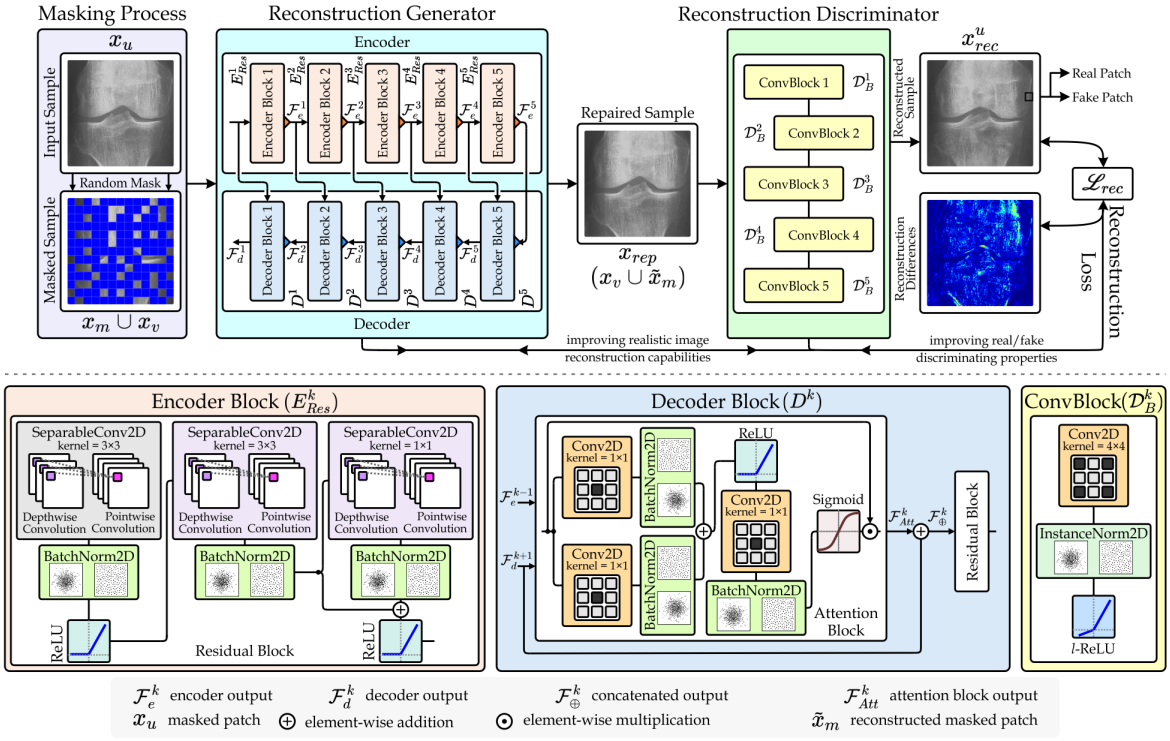


Figure 5.2: Proposed architecture of Masked Image Reconstruction (MI-Rec).

(a) Masking Process: Given an input sample $x_u \in \mathbb{R}^{h\times w\times ch}$, where h, w, and ch denote the height, width, and number of channels. We partition the input sample into N_p non-overlapping patches of size $p \times p$, with $N_p = \frac{h\times w}{p^2}$ patches. Each patch $p_i \in \mathbb{R}^{p\times p\times ch}$ is flatten into a vector $z_i \in \mathbb{R}^{p^2\cdot ch}$ forming the patch sequence $\mathbf{z} = \{z_1, \ldots, z_{N_p}\}$. We add positional encoding PE_i to each patch z_i to preserve the spatial information, where $\mathbf{PE} = \{PE_1, \ldots, PE_{N_p}\}$. The resulting patch representations with positional encoding are

defined in Equation (5.1).

$$\tilde{z}_i = z_i + PE_i; \qquad \tilde{\mathbf{z}} = \{\tilde{z}_1, \ldots, \tilde{z}_{N_p}\} \tag{5.1}$$

After patching and adding positional encoding, we follow a uniform distribution to apply random masking to create a uniformly distributed random set $\mathbf{z_m}$. A masking ratio of 75% to partition the sequence of patches into the masked subset x_m and visible subset x_v, which are defined in Equation (5.2a) and Equation (5.2b).

$$x_m = \{z_i \,|\, i \in \mathbf{z_m}\}, \quad \mathbf{z_m} = RandomSample\left(\{1, \ldots, N_p\}, N_m\right) \tag{5.2a}$$

$$x_v = \{z_i \,|\, i \notin \mathbf{z_m}\} \tag{5.2b}$$

Here, $N_m = 0.75N_p$ represents the number of masked patches, and $x_m \cup x_v = \{1, \ldots, N_p\}$ and $x_m \cap x_v = \varnothing$.

Why high masking ratio? When masking is high, only a few patches remain visible, forcing the model to learn deeper patterns instead of relying on random predictions from neighboring pixels. Uniform random masking ensures the model learns to reconstruct missing patches from all parts of the image equally, preventing any bias towards specific regions like edges or the center. Since most patches are masked, the input to the reconstruction generator is primarily sparse. This sparsity encourages the reconstruction generator (described next) to process information from the visible patches efficiently.

(b) Reconstruction Generator: Our reconstruction generator $\mathcal{G}_{MI\text{-}Rec}$ follows an encoder–decoder architecture inspired by Masked Autoencoders (MAE). Its primary objective is to reconstruct the masked patches $\tilde{x}m$ from the visible patches x_v. The encoder comprises five residual blocks $E^k_{Res}, \{k = 1 \ldots 5\}$, where each block applies separable convolutions $(\mathbb{C}_{Sep})$, batch normalization (β_n), and ReLU activation (α_{ReLU}), followed by downsampling to reduce spatial resolution. The encoder progressively reduces the spatial resolution while increasing feature dimensionality, producing encoded features $\mathcal{F}^k_e$ as defined in Equation (5.3). The decoder then uses the encoded representations to reconstruct the masked

patches $\tilde{x}_m$, which are validated by the reconstruction discriminator.

$$\mathcal{F}_e^k = E_{Res}^k \left(\mathcal{F}_e^{k-1}\right), \quad k \in \{1, \ldots, 5\} \tag{5.3}$$

Here, the initialization $\mathcal{F}_e^0 = x_v$ and E_{Res}^k takes the encoded representation as input from the previous block's output. $\mathcal{F}_e^5$ is the final output of the last encoder block E_5.

After encoding, the highest-level encoded feature maps are passed to the decoder to reconstruct the masked patches. The decoder mirrors the encoder and consists of five decoder blocks D^k, $k \in 1, \ldots, 5$. Each block applies upsampling, an attention unit (D_{Att}^k), and a residual block (D_{Res}^k) to progressively recover spatial resolution. A transpose convolution upsamples the features from the previous block. Next, an attention unit fuses the upsampled features with the corresponding encoder output $\mathcal{F}_e^{k-1}$ to produce an attention-refined representation $\mathcal{F}_{Att}^k$. The attention output is concatenated with the upsampled features to form $\mathcal{F}_{\oplus}^k$, which is then refined by D_{Res}^k to generate the decoded features $\mathcal{F}_d^k$, as defined in Equation (5.4):

$$\mathcal{F}_d^k = D^k \left(\mathcal{F}_d^{k+1}\right), \quad k \in \{1, \ldots, 5\} \tag{5.4}$$

For the final decoding stage, D^5 receives the encoder output $\mathcal{F}_e^5$ (i.e., $\mathcal{F}_d^6 = \mathcal{F}_e^5$), and the output $\mathcal{F}_d^1$ represents the reconstructed masked image $\tilde{x}_m$.

(c) Reconstruction Discriminator: Next, the reconstruction discriminator $\mathcal{D}_{MI\text{-}Rec}$ predicts whether each reconstructed patch is real or synthetic. It takes a repaired image x_{rep} as input, formed by replacing the masked patches x_m in the visible image x_v with reconstructed patches $\tilde{x}_m$, such that $x_{rep} = x_v \cup \tilde{x}_m$ and $x_m \cap x_v = \varnothing$. Given the ground-truth label sequence $\tilde{\mathbf{y}} = \{\tilde{y}_1, \ldots, \tilde{y}_{N_p}\}$, where $\tilde{y}_k \in \{0, 1\}$ denotes real (1) or synthetic (0) patches, the discriminator learns to predict these labels from x_{rep}. The discriminator comprises five convolutional blocks $\mathcal{D}_B^k$, $k \in \{1, \ldots, 5\}$, each applying convolution $\mathbb{C}$, instance normalization (β_{in}), and leaky ReLU activation $(\underset{lReLU}{\alpha})$, followed by progressive downsampling. The network outputs a scalar confidence score indicating the likelihood of each patch being real, as defined in Equation (5.5):

$$\mathcal{D}_{MI-Rec} = prob_{\mathcal{D}} \left(\tilde{y}_k | x_{rep}, k\right), \quad k \in \{1, \ldots, N_p\} \tag{5.5}$$

where $prob_{\mathcal{D}}$ denotes the discriminator's confidence for classifying each patch as real or synthetic. When the confidence score is high, the repaired image is denoted as x^u_{rec}, representing the final unlabeled reconstructed sample.

(d) Loss Strategy: The MI-Rec framework adopts a dual loss strategy to optimize the reconstruction generator $\mathcal{G}_{MI\text{-}Rec}$ and discriminator $\mathcal{D}_{MI\text{-}Rec}$. The generator loss $\mathscr{L}_{gen}$ combines adversarial loss (Binary Cross-Entropy Binary Cross-Entropy (BCE)) and reconstruction loss (L1). The adversarial term $\mathscr{L}_{adv}$ encourages $\mathcal{G}_{MI\text{-}Rec}$ to produce patches classified as real by $\mathcal{D}_{MI\text{-}Rec}$, while the reconstruction term $\mathscr{L}_{rec}$ preserves consistency with masked regions. The generator objective is defined in Equation (5.6):

$$\mathscr{L}_{gen} = \underbrace{BCE\big(\mathcal{D}_{MI-Rec}(\tilde{x}_m), 1\big)}_{\mathscr{L}_{adv}} + \alpha \cdot \underbrace{\|\tilde{x}_m - x_m\|_1}_{\mathscr{L}_{rec}} \tag{5.6}$$

where, $\tilde{x}_m$ denotes the reconstructed patch, x_m the original masked patch, $\mathcal{D}_{MI\text{-}Rec}$ the discriminator prediction for the reconstructed patch, and α a weighting factor that balances the two loss terms. The reconstruction discriminator loss $\mathscr{L}_{dis}$ measures its ability to classify real patches as real and reconstructed patches as fake using BCE loss. The discriminator objective is defined in Equation (5.7).

$$\mathscr{L}_{dis} = \underbrace{\frac{1}{2}\big(BCE(\mathcal{D}_{MI-Rec}(x_m), 1)\big)}_{\mathscr{L}_{real}} + \alpha \cdot \underbrace{\frac{1}{2}\big(BCE(\mathcal{D}_{MI-Rec}(\tilde{x}_m), 0)\big)}_{\mathscr{L}_{fake}} \tag{5.7}$$

where, $\mathcal{D}_{MI\text{-}Rec}(x_m)$ and $\mathcal{D}_{MI\text{-}Rec}(\tilde{x}_m)$ denote the discriminator predictions for real and generated patches. The real patch is assigned a label of 1, while the fake patch is labeled 0. The loss averages the two BCE terms to balance real and fake classifications. The final objective combines the generator and discriminator losses, as defined in Equation (5.8):

$$\mathscr{L}_{MI-Rec} = \mathscr{L}_{gen} + \mathscr{L}_{dis} \tag{5.8}$$

where, $\mathscr{L}_{gen}$ promotes realistic and accurate patch reconstruction, while $\mathscr{L}dis$ enforces effective discrimination between real and reconstructed patches.

5.2.3.2 Stage 2: Similarity-aware Reconstructed Image Labeler

The Similarity-aware Reconstructed Image Labeler (SiRL) assigns proxy labels to unlabeled reconstructed samples by measuring their similarity to predefined templates. SiRL performs three steps: (a) template library construction by extracting and storing features, (b) similarity computation between a sample and the median feature vector of each template, and (c) similarity ranking and labeling by selecting the template with the highest similarity score. An overview is shown in Figure 5.3.

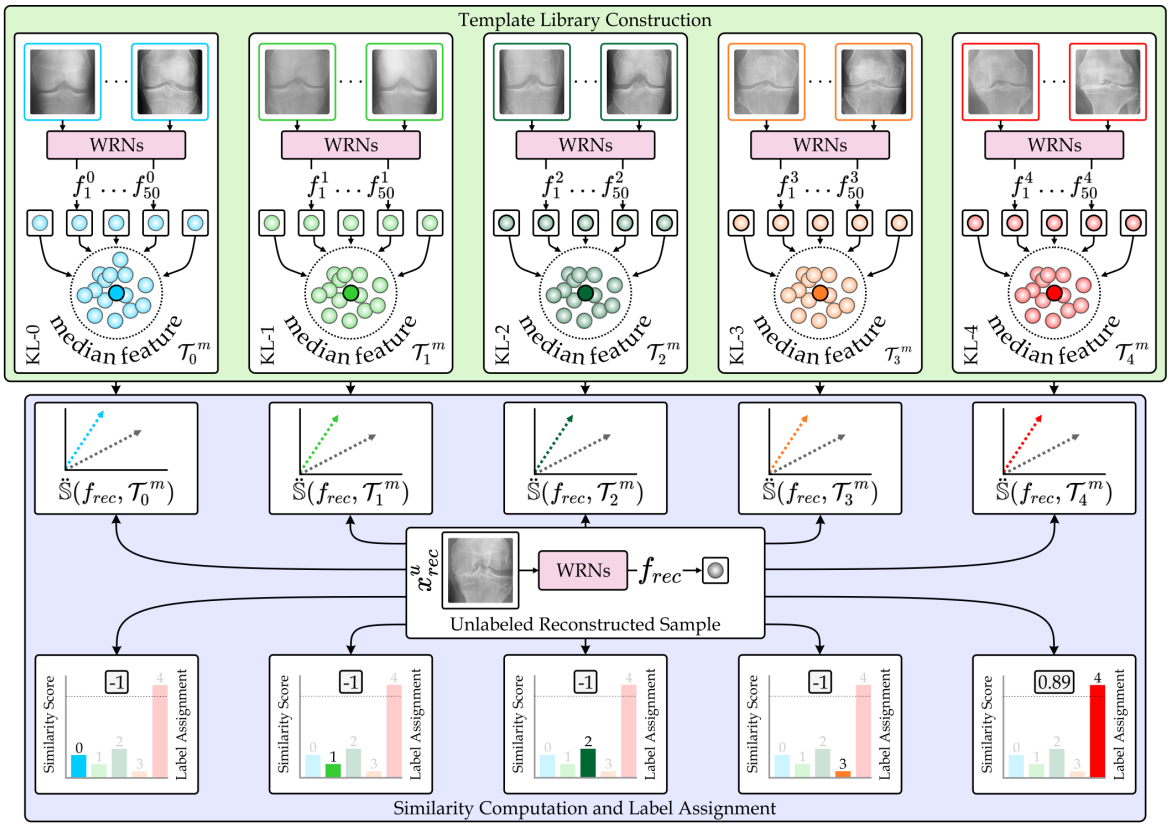


Figure 5.3: Proposed architecture of Similarity-aware Reconstructed Image Labeler (SiRL).

(a) Template Library Construction: The process begins by constructing a template library $\mathcal{T} = \{\mathcal{T}_0, \ldots, \mathcal{T}_4\}$ for the five osteoarthritis severity classes. For each class, we randomly select 50 labeled samples and extract super-pixel level features f_s^i using a pre-trained Wide Residual Network [129]. These features capture the essential visual patterns of each class. For each template, we compute a representative feature vector $\mathcal{T}_i^m$ by aggregating the extracted features from the selected samples, which serves as the reference for subsequent similarity matching, where $i \in \{0, \ldots, 4\}$.

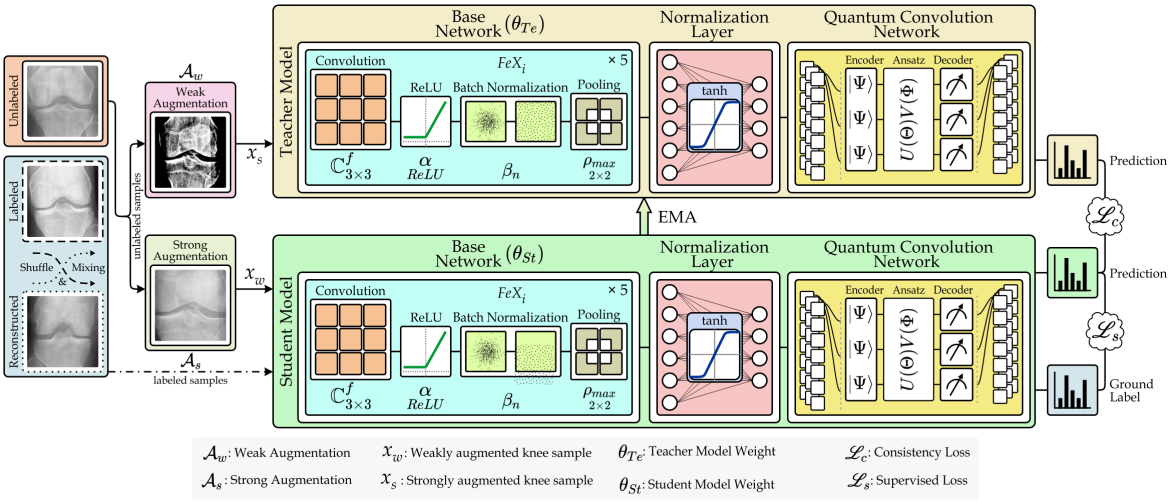


Figure 5.4: Proposed architecture of Quantum-infused Teacher-Student (Q-TeSt) framework.

(b) Similarity Computation and Label Assignment: For each unlabeled reconstructed sample x^u_{rec}, SiRL compares its feature representation f_{rec} with the median feature vectors in the template library. It computes similarity using cosine similarity and Euclidean distance, which are combined into a unified score $\ddot{\mathbb{S}}$ to provide a balanced similarity measure, as defined in Equation (5.9):

$$\ddot{\mathbb{S}}\left(f_{rec}, \mathcal{T}_i^m\right) = \alpha \cdot \underbrace{\frac{f_{rec} \cdot \mathcal{T}_i^m}{\|f_{rec}\| \cdot \|\mathcal{T}_i^m\|}}_{cosine-similarity} + (1-\alpha) \cdot \underbrace{\frac{1}{\sqrt{(f_{rec} - \mathcal{T}_i^m)(f_{rec} - \mathcal{T}_i^m)^{\mathrm{T}}}}}_{euclidean-distance} \tag{5.9}$$

where α is a tunable parameter that balances the contributions of cosine similarity and Euclidean distance in the overall score. To ensure reliable template matching and avoid incorrect proxy label assignment, we apply the following labeling rules, formalized in Equation (5.10):

$$y_{rec} = \begin{cases} \mathcal{C} & \text{if} \quad \arg\max\limits_{i \in \mathcal{C}} \left\{\ddot{\mathbb{S}}\left(f_{rec}, \mathcal{T}_i^m\right) \geq \tau\right\} \\ -1 & \text{if} \quad \arg\max\limits_{i \in \mathcal{C}} \left\{\ddot{\mathbb{S}}\left(f_{rec}, \mathcal{T}_i^m\right) < \tau\right\} \end{cases} \tag{5.10}$$

where $\mathcal{C} = \{KL_0, \ldots, KL_4\}$ denotes the osteoarthritis severity classes, -1 indicates an unassigned sample that is excluded from further processing, and τ is a threshold set to 0.80 (τ_{80}) to filter unreliable proxy labels. The operator $\arg\max\limits_i \ddot{\mathbb{S}}_i(f_{rec}, \mathcal{T}_i)$ selects the template $\mathcal{T}_i$ with the highest similarity score and assigns the corresponding label y_{rec} to the reconstructed sample. The labeled sample is then forwarded to the Hierarchical Quantum-infused

Teacher–Student framework for final severity grading.

5.2.3.3 Stage 3: Hierarchical Quantum-infused Teacher-Student Framework

The proposed Hierarchical Quantum-infused Teacher–Student Framework Hierarchical Quantum-infused Teacher–Student (HQ-TeSt) is a semi-supervised teacher–student model that leverages self-supervised, reconstructed samples for grading the severity of osteoarthritis. HQ-TeSt enhances representation learning by integrating reconstructed samples into training, while the student model continuously updates the teacher via an exponential moving average. The framework comprises two stages: (i) semi-supervised training of the quantum-infused teacher–student (Q-TeSt) model and (ii) hierarchical multi-classification of osteoarthritis severity. We first describe the Q-TeSt architecture, then present its consistency training strategy, and finally detail the hierarchical multi-classifier.

Quantum-infused Teacher-Student Framework: The proposed Quantum-infused Teacher–Student Quantum-infused Teacher-Student (Q-TeSt) framework is inspired by the Mean-Teacher (MT) architecture [37]. Unlike MT, which utilizes a ResNet backbone, Q-TeSt employs a standard convolutional neural network infused with Quantum Convolutional Neural Networks, thereby forming a hybrid semi-supervised framework [130]. As shown in Figure 5.4, Q-TeSt extracts visual features from different augmented views of the same sample in both classical and Hilbert spaces, thereby capturing fine-grained spatial patterns and nonlinear relationships. The framework addresses data scarcity in osteoarthritis severity grading by learning from limited labeled data while leveraging unlabeled samples. Q-TeSt comprises three components: (i) a base network, (ii) a normalization layer, and (iii) a quantum convolution network Quantum Convolutional Network (QCN).

(i) Base Network: The base network in Q-TeSt captures spatial information using a stack of five feature extraction blocks FeX_i, $i \in \{1, \ldots, 5\}$. Each block applies a 3×3 Conv2D operation $\mathbb{C}^{f}_{3\times3}$ with ReLU activation, where the number of filters increases as $f \in \{32, 64, 128, 256, 512\}$. ReLU mitigates vanishing gradients with low computational cost, while batch normalization β_n stabilizes training by normalizing mini-batch statistics. A 2×2 max-pooling layer $\rho^{2\times2}_{max}$ with stride 2 then downsamples the feature maps. After the five blocks, the output of FeX_5 is flattened and passed through two fully connected layers,

FC_{1024} and FC_{512}. The operations in the i-th block are defined in Equation (5.11):

$$FeX_i(x) = \rho_{max}^{2\times2}\left(\beta_n\left(\alpha_{ReLU}\left(\mathbb{C}_{3\times3}^{f}(x)\right)\right)\right) \tag{5.11}$$

where x is the augmented input sample. Finally, global average pooling is applied to the FC_{512} output to produce the base feature vector $\mathcal{F}_{base} \in \mathbb{R}^{512}$.

To address the limitations of current quantum systems, which support small qubit circuits (eight in our case), we first map classical features into a quantum-compatible space. We project the high-dimensional feature vector into a 256-dimensional representation $\mathcal{F}_t \in \mathbb{R}^{256}$, where $256 = 2^8$ corresponds to the number of qubits, as defined in Equation (5.12):

$$\mathcal{F}_t = \mathcal{W}^{(t)} \cdot \mathcal{F}_{base} + b^{(t)} \tag{5.12}$$

where, $\mathcal{W}^{(t)} \in \mathbb{R}^{256\times512}$ and $b^{(t)}$ denote the projection weights and bias. The transformed feature $\mathcal{F}_t$ is then fed into the quantum convolutional network. This layer bridges the classical and quantum components of the Q-TeSt framework, enabling efficient quantum processing of osteoarthritis samples.

(ii) Normalization Layer: The transformed feature vector $\mathcal{F}_t$ is fed into QCN, which requires normalized input within a bounded range for stable learning [131]. As $\mathcal{F}_t$ is not normalized, previous work uses $\tanh$ to restrict feature values to the $[-1, 1]$ range [130]. However, $\tanh$ saturates near ±1, leading to vanishing gradients as $\nabla \tanh(x) = 1 - \tanh^2(x)$. Following [132] we adopt the $L_2-\tanh$ normalization to improve gradient flow while preserving feature magnitude and nonlinearity. The normalized output $\mathcal{F}_{norm}$ is defined in Equation (5.13):

$$\mathcal{F}_{norm} = \tanh\left(\omega \cdot \frac{\mathcal{F}_t}{\|\mathcal{F}_t\|_2}\right) \tag{5.13}$$

where $\mathcal{F}_t$ denotes the projected feature embedding and ω controls the saturation range, ensuring that the normalized features lie within a Bloch sphere of radius ω.

(iii) Quantum Convolution Network: The quantum convolution network (QCN) operates directly in the quantum domain. First, the quantum encoder $\mathcal{U}_e(\mathcal{F}_{norm})$ maps the normalized feature vector $\mathcal{F}_{norm}$ to a quantum state. Next, the ansatz $\mathcal{U}_a(\Theta)$ transforms the encoded state using learnable parameters Θ. Finally, the decoder applies quantum state measurement to

project the qubit-encoded information back into the classical domain for further processing. During training, the decoder updates Θ by minimizing the cost function.

Quantum Encoder: The quantum algorithm operates on quantum states, and normalized classical features cannot be used directly by the quantum convolution network. Therefore, the encoder maps the input features into a quantum state before ansatz processing. Among common encoding strategies, computation basis encoding converts inputs into binary states, whereas amplitude encoding maps data to quantum state amplitudes. We adopt amplitude encoding, which represents the normalized feature vector $\mathcal{F}_{norm}$ as the amplitudes of a quantum state, ensuring that the sum of squared amplitudes equals one. This encoding applies a unitary transformation, defined in Equation (5.14):

$$|\Psi\rangle = \sum_{i=1}^{N} a_i |i\rangle \tag{5.14}$$

where, $|\Psi\rangle$ denotes the quantum state, a_i are the components of $\mathcal{F}_{norm}$ satisfying $\sum_{i=1}^{N} |a_i|^2 = 1$, and $|i\rangle$ represents the computational basis states. Amplitude encoding is qubit-efficient, encoding N features with $\lceil \log_2 N \rceil$ qubits, thereby limiting circuit width and improving scalability.

Ansatz: An ansatz is a parameterized quantum circuit composed of tunable quantum gates whose parameters are optimized to approximate a target quantum state. By adjusting these parameters, the ansatz explores a broad space of quantum states, capturing correlations and

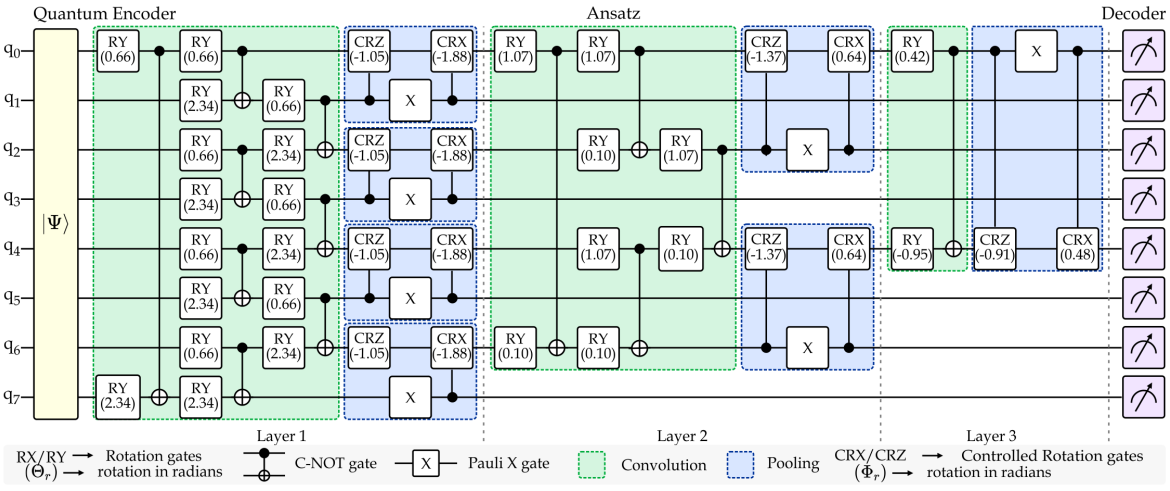


Figure 5.5: Proposed architecture of Quantum Convolutional Network (QCN).

hidden patterns that enable efficient learning of complex relationships. In this work, the QCN circuit comprises three layers, as shown in Figure 5.5, where each layer applies unitary operations analogous to classical convolution and pooling.

The convolutional layers apply unitary transformations $U(\Theta)$ implemented as parameterized quantum gates acting on qubits. In our setting, each transformation consists of rotation gates $R_y(\Theta_0)$ and $R_y(\Theta_1)$ followed by a controlled-NOT (CNOT) gate, as defined in Equation (5.15):

$$U(\Theta) = CNOT \cdot R_y(\Theta_0) \otimes R_y(\Theta_1) \tag{5.15}$$

where the CNOT gate flips the target qubit when the control qubit is in the $|1\rangle$ state, and $R_y(\Theta_q)$ rotates qubit $q \in 0, 1$ around the Y-axis of the Bloch sphere by Θ_q. The tensor product $\otimes$ combines the single-qubit rotations into a two-qubit operation. The pooling operation $V(\Phi)$ reduces quantum state dimensionality using Controlled Rotation-Z $(CR_z(\Phi_q))$, Pauli-X (X), and Controlled Rotation-X $(CR_x(\Phi_q))$ gates to selectively reduce entanglement between qubit pairs. This downsampling limits circuit complexity while preserving salient information, playing a role analogous to pooling in classical CNNs while exploiting quantum entanglement and superposition for feature extraction.

Decoder: The decoder maps the quantum state produced by the ansatz to classical features through quantum measurement. It performs computational basis measurements on the processed quantum state and estimates the associated outcome probabilities, which form the classical output vector. In our implementation, measurement is applied to wire 4, yielding the probability distribution over computational basis states. For a single qubit, the probability of observing state $|0\rangle$ or $|1\rangle$ is defined in Equation (5.16):

$$p(i) = |\langle q_i|\ddot{\Psi}\rangle|^2 \tag{5.16}$$

where $|q_i\rangle$ denotes the computational basis state with $q_i \in \{0, 1\}$, and $\ddot{\Psi}$ is the final quantum state after the convolution and pooling operations. The ansatz parameters are optimized using a hybrid quantum–classical approach. The quantum circuit produces measurement outcomes for current parameters and then used to evaluate an objective function. We employ cross-entropy loss $(\mathscr{L}_{CE})$ to iteratively update the parameters and minimize this objective.

Semi-Supervised Training: Semi-supervised training forms the first stage of the Hierar-

chical Quantum-infused Teacher–Student (HQ-TeSt) framework, leveraging labeled, reconstructed labeled, and unlabeled samples. The objective is to minimize supervised loss on labeled data while reducing consistency loss on unlabeled samples. During training, the teacher predicts on strongly augmented unlabeled samples, and the student enforces consistency on weakly augmented views of the same inputs. This process promotes invariant feature learning and enhances robustness to perturbations, thereby improving generalization. The semi-supervised training of Q-TeSt proceeds in three steps: (a) model initialization, (b) supervised training on labeled data, and (c) consistency training on unlabeled data.

(a) Model Initialization: The quantum-infused teacher–student (Q-TeSt) framework consists of a teacher and a student model with identical architectures, each comprising a convolutional network followed by quantum convolution (see Section 5.2.3.3). The convolutional weights of the student and teacher are initialized independently and updated using EMA, while the quantum convolution weights are kept identical across both models. This design stabilizes the classical feature-extraction stage, yielding more consistent inputs to the QCN. We denote the convolutional weights of the teacher and student by θ_{Te} and θ_{St}, respectively, and the shared quantum weights by θ_{QCN}.

(b) Supervised Training on Labeled data: In supervised training, we use labeled samples to train the student model. Given the scarcity of labeled data, the student alone may not learn sufficiently discriminative representations. To address this limitation, we incorporate self-supervised learning (see Section 5.2.3.1) to generate reconstructed samples, which serves a dual purpose: (i) increasing the volume of labeled data strengthens the student model's ability to capture local patterns within the samples, and (ii) training the student on enriched samples subsequently improves the teacher model's performance when processing unlabeled data. The student model $(\mathcal{M}St)$ is optimized using supervised loss $(\mathscr{L}sup)$, implemented as cross-entropy to align predictions with ground-truth labels. The supervised objective is defined in Equation (5.17).

$$\mathscr{L}_{sup} = \mathsf{CE}\left(\mathcal{M}_{St}(x_i; \theta_{St}, y_i)\right) \tag{5.17}$$

where x_i and y_i denote the labeled input and its ground-truth, respectively, and CE represents the cross-entropy loss.

(c) Consistency Training on Unlabeled data: Consistency training uses a teacher–student framework to improve predictions on unlabeled data. The student model $\mathcal{M}_{St}$ is first trained on labeled samples in a supervised manner, and the teacher model $\mathcal{M}_{Te}$ is updated via EMA of the student's weights. The approach encourages both models to produce consistent predictions under different augmentations of the same unlabeled sample. We apply weak augmentations $\mathcal{A}_w$, such as flipping and translation, to generate inputs for $\mathcal{M}_{Te}$, and strong augmentations $\mathcal{A}_s$ using RandAugment [124], including invert, shear, and scale operations, for $\mathcal{M}_{St}$ (see Section 5.2.2). For an unlabeled sample x_u, the teacher predicts a pseudo-label from the weakly augmented input $x_w = \mathcal{A}_w(x_u)$, while the student predicts from the strongly augmented counterpart $x_s = \mathcal{A}_s(x_u)$. The corresponding predictions are given by $y_{Te} = \mathcal{M}_{Te}(x_w; \theta_{Te})$ and $y_{St} = \mathcal{M}_{St}(x_s; \theta_{St})$. The unsupervised loss $\mathscr{L}_{unsup}$ enforces prediction consistency between the teacher and student models under different augmentations of the same input, as defined in Equation (5.18). Here, $\mathcal{M}(\cdot)$ denotes the Q-TeSt classification framework, while θ_{Te} and θ_{St} represent the teacher and student model parameters, respectively.

$$\mathscr{L}_{unsup} = \mathcal{M}(\cdot)(\theta_{Te}, \mathcal{A}_w)(\theta_{St}, \mathcal{A}_s) \tag{5.18}$$

An individual consistency mechanism [37] is employed in our framework to ensure consistency between the teacher and student model outputs. As defined in Equation (5.19), the loss penalizes per-sample prediction discrepancies under weak and strong augmentations:

$$\mathscr{L}_{con} = \sum_{i=1}^{n+m} \mathbb{E}_{\mathcal{A}_w, \mathcal{A}_s} \| \mathcal{M}_{Te}(x_i, \theta_{Te}, \mathcal{A}_w) - \mathcal{M}_{St}(x_i, \theta_{St}, \mathcal{A}_s) \|_2^2 \tag{5.19}$$

where $\mathcal{A}_w$ and $\mathcal{A}_s$ denote weak and strong augmentations; n and $m, (m \gg n)$ are labeled and unlabeled samples; x_i is a training sample; and θ_{Te}, θ_{St} are model parameters. Following the Mean Teacher framework [37], we update the teacher via EMA applied only to the student's base network, while keeping the QCN fixed to stabilize classical feature extraction and provide consistent inputs to the quantum module, as defined in Equation (5.20).

$$\theta_{Te}^{(t+1)} \leftarrow \mu \cdot \theta_{Te}^{(t)} + (1 - \mu) \cdot \theta_{St}^{(t)} \tag{5.20}$$

where t represents the training iteration and μ is the EMA smoothing coefficient, set to 0.99

following [37], which controls the update rate of the teacher parameters. This consistency strategy gradually transfers knowledge from the student to the teacher, aligns their predictions, and improves performance on both labeled and unlabeled data.

Hierarchical Multi-classifier: The Hierarchical Multi-classifier Hierarchical Multi-classifier (HiM) classifies knee X-rays into five osteoarthritis severity grades using hierarchical binary classification. Motivated by [133], HiM mitigates class imbalance by decomposing the multi-class task into a sequence of binary decisions. The key distinction lies at each binary decision node, where reconstructed samples are incorporated and the classifier is trained in a semi-supervised manner using the Q-TeSt framework (refer to Section 5.2.3.3).
(i) Hierarchical Decomposition: In the hierarchical decomposition stage, we convert the multiclass osteoarthritis grading problem $\mathcal{C} = KL_0, \ldots, KL_4$ into a sequence of binary classification tasks using a dual rule-based strategy. At the root node, classes are grouped by severity into normal $\mathcal{C}_n = \{KL_0, KL_1\}$ and osteoarthritis $\mathcal{C}_o = \{KL_2, KL_3, KL_4\}$. Subsequent levels leverage majority and minority class distributions to form balanced binary splits. This process builds a hierarchical binary tree, where each decision node is implemented with the Q-TeSt framework Q-TeSt$_i$ and leaf nodes correspond to specific severity grades. Hierarchical decomposition improves classification accuracy and mitigates multi-class imbalance.
(ii) Classification: At each decision node (Q-TeSt$_i$), we train a dedicated Q-TeSt framework in a semi-supervised manner and incorporate reconstructed samples to mitigate class imbalance. As shown in Figure 5.6, the hierarchy performs binary decisions at successive levels. The root node Q-TeSt$_1$ separates combined normal (KL$_0$, KL$_1$) from osteoarthritis (KL$_2$, KL$_3$, KL4). The left branch Q-TeSt$2L$ further distinguishes KL$_0$ from KL$_1$, while the right branch Q-TeSt$_{2R}$ splits KL$_2$ from (KL$_3$, KL$_4$). At the final level, Q-TeSt$_3$ discriminates between KL$_3$ and KL$_4$. After training all nodes, we aggregate their outputs to compute the final classification accuracy, providing a structured solution to the multi-class grading task.
(iii) Aggregation: In the aggregation stage, we compute the final classification performance by combining the outputs of all Q-TeSt frameworks using a weighted scheme. Each framework Q-TeSt$_i$ is assigned a weight w_i proportional to its depth l_i in the hierarchy, defined as $w_i = \ln(l_i+1)$. The weighting scheme assigns greater importance to classifiers at deeper levels of the hierarchy, which capture finer-grained distinctions, than to higher-level classifiers that model coarse separations. The aggregated accuracy A_{agg} is computed as the normalized

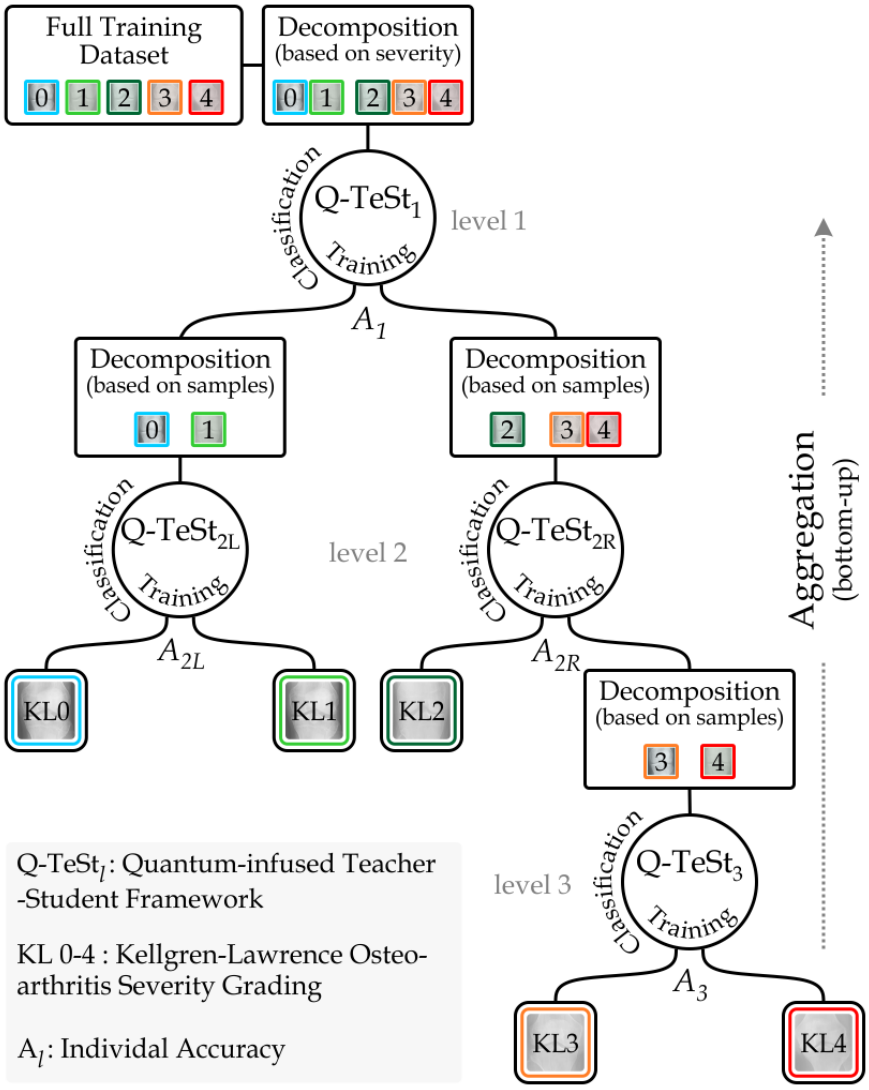


Figure 5.6: Proposed architecture of Hierarchical Multi-classifier (HiM) module.

weighted sum of individual accuracies A_i, as shown in Equation (5.21).

$$A_{agg} = \frac{\sum_i w_i \cdot A_i}{\sum_i w_i} \tag{5.21}$$

where $i \in \{\text{Q-TeSt}_1, \text{Q-TeSt}_{2L}, \text{Q-TeSt}_{2R}, \text{Q-TeSt}_3\}$. This normalization keeps A_{agg} within $[0, 1]$ and preserves the hierarchical structure, thereby providing a balanced overall performance measure.

5.3 Experimental Settings

This section describes the datasets used to evaluate the proposed H-Semi framework for grading of knee osteoarthritis severity. It provides the details of the benchmark datasets and outlines the evaluation metrics used to assess model performance.

5.3.1 Benchmark Datasets

We evaluate H-SemiS using two multiclass benchmark datasets to examine its generalizability on two publicly available binary classification datasets. In all experiments, we use single-joint anteroposterior (AP) knee X-ray radiographs, which align with the standard clinical imaging perspective used for Kellgren-Lawrence (KL) severity grading in KOA datasets. Table 5.1 provides a summary of the original and reconstructed samples employed during training, while Figure 5.7 presents representative examples of knee osteoarthritis cases drawn from each dataset.

Class	**Original Samples**	**Reconstructed Samples**	**Total Samples**	**Original Samples**	**Reconstructed Samples**	**Total Samples**
Multi-class Dataset	**OAI Dataset**			**DKXI Dataset**		
Class 0	3857	0	3857	576	0	576
Class 1	1770	1982	3752	514	0	514
Class 2	2578	1243	3821	273	243	516
Class 3	1286	608	1894	267	0	267
Class 4	295	1567	1862	206	54	260
Binary Dataset	**OP Dataset**			**KO Dataset**		
Class 0	1651	478	2129	1434	2089	3523
Class 1	2155	0	2155	3554	0	3554

Table 5.1: An overview of benchmark datasets for knee osteoarthritis classification.

(a) Multiclass Dataset

(i) Osteoarthritis Initiative Dataset: The Osteoarthritis Initiative (OAI) [134] is a publicly available, multi-center dataset designed to study the development and progression of knee osteoarthritis. It contains 9,786 single-knee X-ray images from 4,796 subjects aged 45 to 79, each with a resolution of 299×299 pixels. To address multiclass imbalance in a hierarchical binary setting, we augment the dataset with 5,400 reconstructed KXR samples. Images are

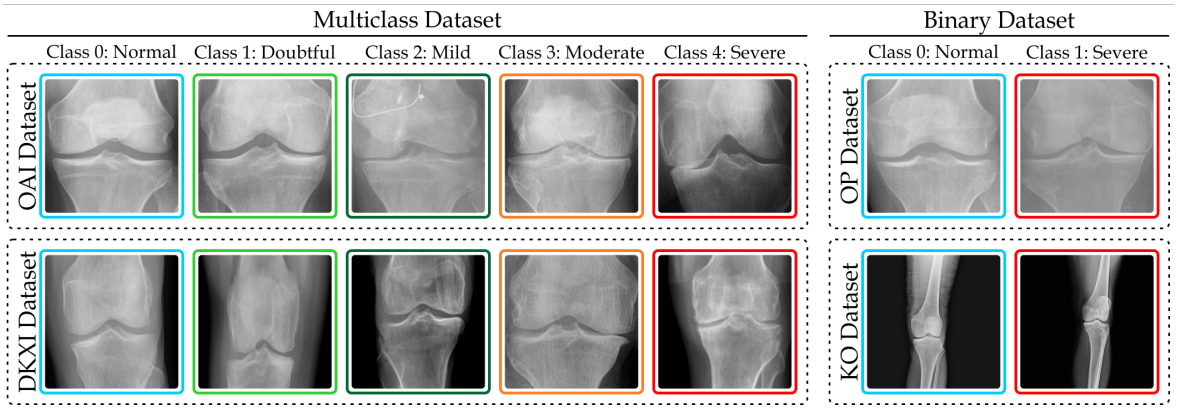


Figure 5.7: Sample dataset images for knee osteoarthritis classification.

annotated using SiRL (see Section 5.2.3.2), which follows the five-grade KL system: KL-0 (normal), KL-1 (doubtful), KL-2 (mild), KL-3 (moderate), and KL-4 (severe).

(ii) Digital Knee X-ray Images: The Digital Knee X-ray Images (DKXI) dataset [135] comprises 8-bit grayscale images acquired using a PROTEC PRS 500E X-ray system from multiple hospitals and diagnostic centers. Two medical experts annotated all samples using the KL grading system. The dataset includes both single-knee images and scans containing both knees; we crop the latter into individual knees for consistency, resulting in 1,836 knee X-ray samples. All images are provided in PNG format, with resolutions ranging from 300×162 to 640×481 pixels. To mitigate class imbalance, we further augment the dataset with 297 reconstructed KXR samples.

(b) Binary Dataset

(i) Osteoarthritis Prediction Dataset: The Osteoarthritis Prediction (OP) dataset [136], released by the University of Florida and the OAI organization, contains anteroposterior, lateral, and oblique knee X-rays from diverse patients. In this study, we use only anteroposterior. After cropping images that contain both knees into single-knee samples, the dataset includes 1,651 normal (class 0) and 2,155 severe (class 1) X-rays. To mitigate class imbalance, we augment the dataset with 478 reconstructed normal KXR samples.

(ii) Knee Osteoarthritis Dataset: The Knee Osteoarthritis (KO) dataset [137] contains X-rays with and without KL grading across multiple joints, including knees, wrists, elbows, necks, and hips. We use knee X-rays without KL grading to evaluate the generalizability of the proposed framework in a binary classification setting. Samples are grouped into nor-

mal (class 0) and severe (class 1). For consistency, we crop images containing both knees into single-knee views, yielding 4,988 knee X-ray samples. To address class imbalance, we augment the dataset with 2,089 reconstructed normal KXR samples.

5.3.2 Evaluation Metrics

We evaluate H-SemiS using standard metrics: Accuracy (Acc), Precision (Pre), Recall (Rec), and F1-score (F1) for grading KOA severity. For multiclass evaluation, Precision, Recall, and F1-score are macro-averaged across classes. Table 5.2 summarizes the mathematical definitions of these metrics. Here, TP_{KOA} denotes KOA samples correctly classified as KOA, TN_{Nor} denotes normal samples correctly classified as normal, FP_{KOA} denotes normal samples misclassified as KOA, and FN_{Nor} denotes KOA samples misclassified as normal.

Table 5.2: Evaluation metrics of H-SemiS framework osteoarthritis severity grading.

Evaluation Metrics	**Mathematical Formulation**
Accuracy (Acc)	$\frac{\text{TP}_{\text{KOA}} + TN_{nor}}{\text{TP}_{\text{KOA}} + \text{FP}_{\text{KOA}} + TN_{nor} + \text{FN}_{\text{Nor}}}$
Precision (Pre)	$\frac{\text{TP}_{\text{KOA}}}{\text{TP}_{\text{KOA}} + \text{FP}_{\text{KOA}}}$
Recall (Rec)	$\frac{\text{TP}_{\text{KOA}}}{\text{TP}_{\text{KOA}} + \text{FN}_{\text{Nor}}}$
F1-Score (F1)	$2 \times \frac{Precision \times Recall}{Precision + Recall}$

5.4 Experimental Results

This section presents the experimental results of the proposed method. It includes quantitative evaluation against existing approaches, followed by qualitative analysis to visualize model performance. Ablation analysis is conducted to assess the contribution of individual components. Cross-dataset evaluation is also performed to examine generalization, along with model complexity analysis.

5.4.1 Quantitative Analysis

We report quantitative results across three training settings: supervised, self-supervised, and semi-supervised. In the supervised setting, models are trained using the full labeled dataset. In the self-supervised setting, models learn representations from unlabeled data without manual annotations. In the semi-supervised setting, all baselines are trained with 20% labeled samples and no additional data, matching the H-SemiS protocol.

Table 5.3: Quantitative analysis of H-SemiS framework with existing baselines on the OAI and DKXI datasets. **Bold**: best and underlined: second-best scores.

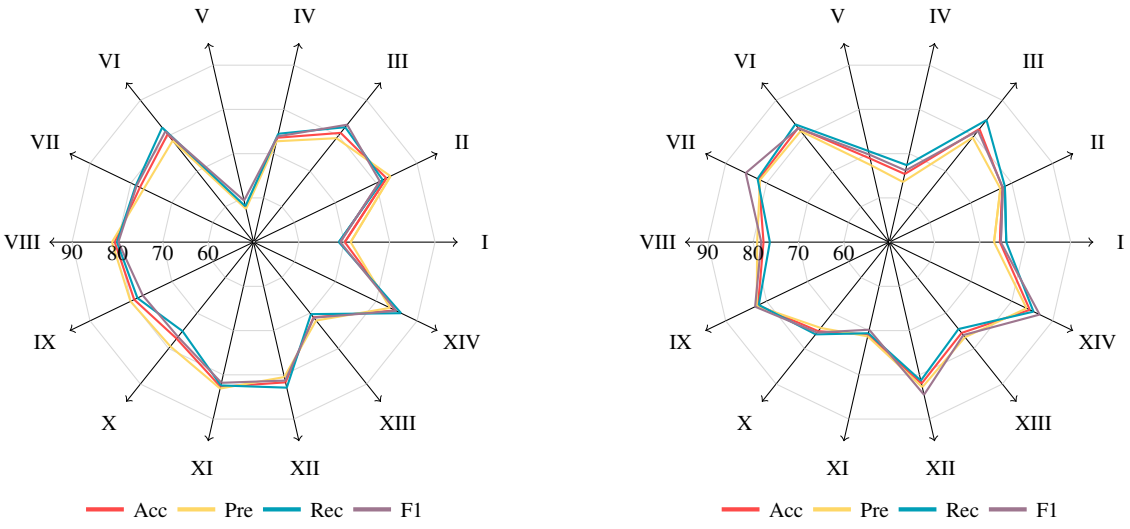
	Methods	OAI Dataset				DKXI Dataset			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
I	Nguyen *et al.* [138]	69.1	68.8	71.4	70.1	74.6	75.8	73.1	74.4
	Lo *et al.* [139]	80.9	81.6	**83.5**	82.5	77.8	78.3	77.2	77.7
	Teoh *et al.* [140]	83.1	82.4	79.3	80.8	81.4	84.4	79.5	81.9
	Hu *et al.* [141]	73.9	74.5	72.8	73.6	66.2	67.4	63.6	65.4
II	Wu *et al.* [59]	59.4	58.1	57.5	57.8	69.8	70.6	67.3	68.9
	Azizi *et al.* [142]	81.2	82.3	78.6	80.4	82.1	83.2	81.3	82.2
	Cai *et al.* [143]	78.6	78.8	76.9	77.8	85.1	82.2	81.6	81.9
	Azizi *et al.* [144]	79.8	80.1	81.2	80.6	78.2	76.3	79.1	77.7
III	Berrimi *et al.* [41]	77.1	78.4	80.1	79.2	82.8	81.9	82.4	82.1
	Farooq *et al.* [42]	76.7	75.1	79.4	77.2	75.5	76.0	74.2	75.1
	Huo *et al.* [145]	81.8	82.4	83.1	82.7	69.8	70.6	71.3	70.9
	Burton *et al.* [146]	82.9	82.9	80.5	81.7	84.5	81.3	82.8	82.0
	Nguyen *et al.* [147]	71.3	70.3	72.1	71.2	76.2	74.5	76.8	75.6
Ours	H-SemiS	**84.9**	**86.1**	83.4	**84.7**	**86.8**	**85.3**	**83.6**	**84.4**

I: Supervised-based Techniques; II: Self-Supervised-based Techniques; III: Semi-Supervised-based Techniques.

(i) Analysis on OAI Dataset: We report quantitative results in Table 5.3, showing that H-SemiS demonstrates superior performance across nearly all metrics. Among supervised methods, H-SemiS achieves the strongest overall performance. Only [139] achieves a marginally higher recall by 0.1, while the best competing supervised approach, [140], still lags by 1.8. These gaps indicate that H-SemiS exhibits stronger feature learning under limited supervision. H-SemiS also surpasses all self-supervised methods with [59], yielding the lowest accuracy, highlighting the limitations of learning from unlabeled data alone. Among semi-supervised approaches [145] achieves the closest performance with an F1-score of 82.7 whereas H-SemiS maintains a clear advantage with an F1-score of 84.7. Although [146] attains the highest precision among existing methods, it still falls short of H-SemiS by 3.2. Overall, these results demonstrate the effectiveness of the hierarchical design and self-

supervised components in enhancing feature learning. This maintains the balanced performance and establishes the H-SemiS as a strong framework for grading osteoarthritis severity with limited labeled data.

(ii) Analysis on DKXI Dataset: For the DKXI dataset, H-SemiS outperforms all competing baselines across every evaluation metric, as shown in Table 5.3, confirming its effectiveness in grading osteoarthritis severity. Among supervised methods, [140] achieves the closest performance with an F1-score of 81.9. Although a self-supervised method attains the second-highest accuracy, H-SemiS remains superior by 1.7, reflecting stronger discriminative feature learning. [143] reports the highest accuracy among self-supervised approaches, followed by H-SemiS. [142] achieves the best self-supervised F1-score of 82.2 but still trails H-SemiS by 2.2. This gap reflects the advantage of combining self-supervision with hierarchical learning. Among semi-supervised methods, [146] and [41] show competitive results. However, despite achieving 84.5 accuracy and 82.8 recall, [146] remains below H-SemiS. Overall, these results show that H-SemiS effectively addresses limitations in prior work. The framework leverages quantum principles to capture complex patterns in knee X-ray data. In addition, the use of proxy labels filtered from noisy and ambiguous samples supports learning more



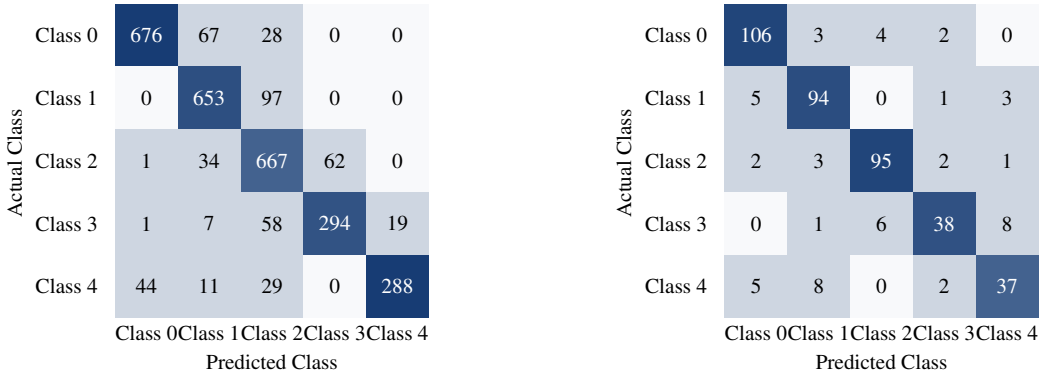
(a) OAI Dataset **(b)** DKXI Dataset

I: Nguyen *et al.* [138] II: Lo *et al.* [139] III: Teoh *et al.* [140] IV: Hu *et al.* [141] V: Wu *et al.* [59]
VI: Azizi *et al.* [142] VII: Cai *et al.* [143] VIII: Azizi *et al.* [144] IX: Berrimi *et al.* [41] X: Farooq *et al.* [42]
XI: Huo *et al.* [145] XII: Burton *et al.* [146] XIII: Nguyen *et al.* [147] XIV: **H-SemiS**

Figure 5.8: Quantitative comparison of the proposed H-SemiS framework against existing baselines on (a) OAI dataset and (b) DKXI dataset.

discriminative features for KOA severity grading.

We utilize radar plots to visualize the quantitative results, as shown in Figure 5.8. These plots provide a comparative view of Acc, Pre, Rec, and F1-score achieved by the proposed framework and competing baselines on the OAI [134] and DKXI [135] datasets. We further report confusion matrices for both datasets in Figure 5.9 to illustrate class-wise prediction behavior.



(a) OAI Dataset Evaluation

(b) DKXI Dataset Evaluation

Figure 5.9: Confusion matrices for classification on the OAI and DKXI datasets.

5.4.2 Qualitative Analysis

We evaluate qualitative performance of the H-SemiS using Explainable Artificial Intelligence (XAI) methods including Grad-CAM [81], Grad-CAM++ [85] and Score-CAM [86]. To assess the reliability and effectiveness of resulting explanation maps we use the ADCC score [148] to identify critical regions for KOA severity. The ADCC score (χ_{ADCC}) is defined as the harmonic mean of Average Drop (χ_{AvD}), Coherency (χ_{Co}) and Complexity (χ_{Cx}), computed as follows:

$$\chi_{ADCC} = 3\left(\frac{1}{1-\chi_{AvD}} + \frac{1}{\chi_{Co}} + \frac{1}{1-\chi_{Cx}}\right)^{-1} \tag{5.22}$$

Figure 5.10 visualizes class activation maps with ADCC scores demonstrating the ability of H-SemiS to localize clinically relevant features. Higher ADCC values indicate more reliable and interpretable explanations aligned with the framework's predictions. A lower χ_{AvD} reflects reduced confidence when highlighted regions are removed, while χ_{Co} mea-

sures the consistency of activation maps with expert interpretation. The χ_{Cx} term ensures that highlighted regions remain compact and meaningful. Together, these components provide quantitative validation of the interpretability of H-SemiS and its clinical relevance for assessing KOA.

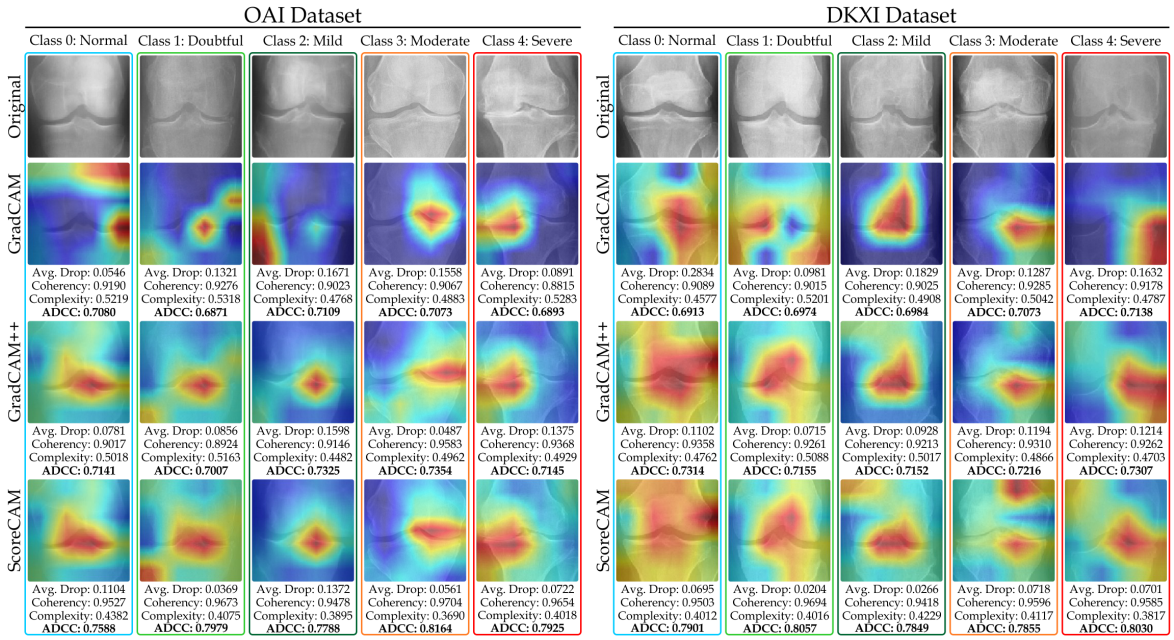


Figure 5.10: Qualitative analysis of H-SemiS framework.

5.4.3 Ablation Analysis

We conduct an ablation study on the OAI [134] dataset to assess the contribution of individual components in the H-SemiS framework. By varying one component at a time while keeping all other settings fixed, we quantify their impact on overall performance. The following subsections summarize the results.

(a) Effect of Image Reconstruction: We examine the contribution of Masked Image Reconstruction (MI-Rec) (see Section 5.2.3.1) by analyzing the effects of the masking ratio and the choice of reconstruction technique. Table 5.4 reports performance across masking ratios (N_m), showing optimal results at 75%. Performance drops by 1.1 and 1.8 when the masking ratio is set to 70% and 80%, respectively. Figure 5.11 visualizes reconstructed samples with high masking (80%), which yield blurry reconstructions with poorly defined edges, and degrade performance. Figure 5.12 further shows a peak accuracy of 84.9 at 75%, with

accuracy decreasing as the masking ratio moves away from the value. This trend indicates that the MI-Rec performs best at 75%, where missing patches are reconstructed uniformly without bias toward specific regions.

Table 5.4: Ablation analysis of H-SemiS framework with different masking ratios and reconstruction techniques on the OAI dataset.

	Methods	Acc	Pre	Rec	F1
I	$N_m = 0.70$	83.8	82.5	83.1	82.8
	$N_m = 0.80$	83.1	84.3	82.6	83.4
II	MAE	79.5	80.1	77.2	78.6
	CMAE	80.4	82.6	81.9	82.2
	ViT-AE++	78.5	79.7	82.6	81.1
	GAN-MAE	83.1	80.4	83.2	81.8
Ours	**$N_m = 0.75$ + MI-Rec**	**84.9**	**86.1**	**83.4**	**84.7**

I: Effect of Masking Ratio; II: Effect of Reconstruction Technique.

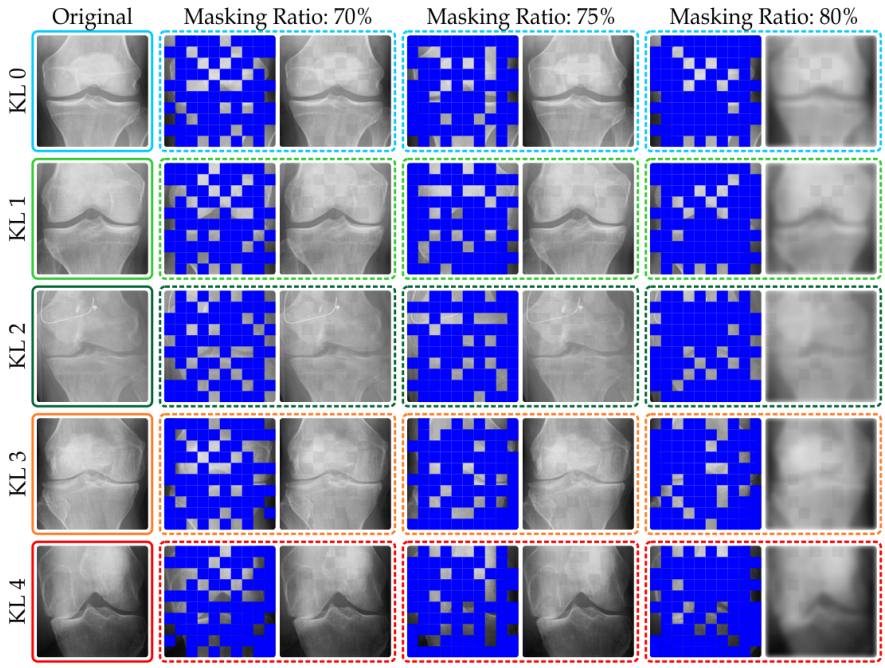


Figure 5.11: KXR reconstruction examples across different KL grades.

We next compare reconstruction techniques at a fixed masking ratio of 75%. As summarized in Table 5.4, we evaluate MAE [127], CMAE [149], ViT-AE++ [150], and GAN-MAE [126] against the proposed MI-Rec. GAN-MAE achieves the second-best performance, while MI-Rec consistently outperforms all alternatives across metrics. These results suggest that MI-Rec better preserves structural consistency and key KOA markers, including

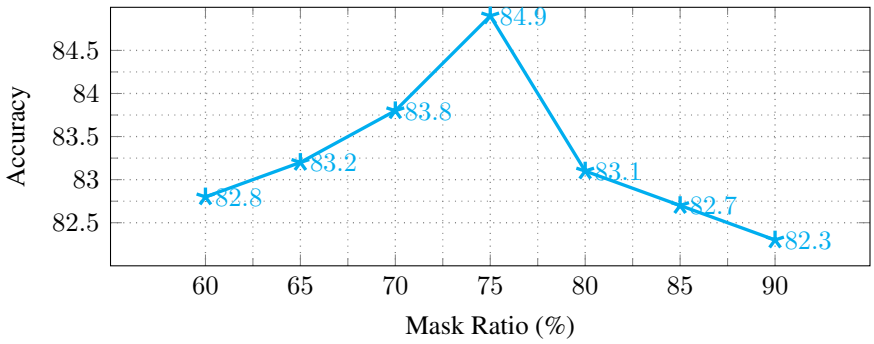


Figure 5.12: Mask ratio ablation study on H-SemiS performance.

cartilage thickness and joint space narrowing. The combined choice of an optimal masking ratio and reconstruction strategy prevents spatial feature loss and enables high-fidelity reconstructions, leading to more accurate KOA classification.

(b) Effect of Template Matching: Table 5.5 analyzes the role of the Similarity-aware Reconstructed Image Labeler (SiRL), described in Section 5.2.3.2, within the H-SemiS framework. SiRL assigns proxy labels to reconstructed unlabeled samples based on similarity to predefined templates. We evaluate thresholds from τ_{60} to τ_{90} and identify τ_{80} as optimal, achieving the best performance across all metrics. Lower thresholds (τ_{60}, τ_{70}) admit more reconstructed samples but introduce noise, reducing performance by 6.3% and 1.1%, respectively. A higher threshold (τ_{90}) admits too few samples, limiting class imbalance mitigation and causing a 5% drop. The strong performance at τ_{80} indicates effective selection of reliable reconstructed samples.

Table 5.5: Ablation analysis of H-SemiS framework with different thresholds for labeling reconstructed samples on the OAI dataset.

Methods	**Acc**	**Pre**	**Rec**	**F1**
H-SemiS (τ_{60})	78.6	75.2	76.7	75.9
H-SemiS (τ_{70})	83.8	84.1	83.6	83.8
H-SemiS (τ_{90})	79.9	81.2	80.3	80.7
H-SemiS (τ_{80})	**84.9**	**86.1**	**83.4**	**84.7**

(c) Effect of Labeling in Teacher-Student Framework: We evaluate H-SemiS using varying labeled data ratios (1%, 5%, 10%, and 20%) and different architectural setups. As shown in Figure 5.13, accuracy improves across all configurations as the labeled data increases. To isolate the specific contributions of the quantum-infused teacher–student framework (Q-

TeSt) (see Section 5.2.3.3), we ablate the architecture into its core components: the Base Network (BN), Normalization Layer (NL), and Quantum Convolutional Neural Network QCN. To verify that our improvements from quantum operations and not just a larger model size, we also test a classical baseline (BN+NL+MLP). This setup replaces the QCN with a Multi-Layer Perceptron containing a similar number of parameters. As detailed in Table 5.6, the base configurations maintain stable accuracy even under severe data scarcity (1% labels). However, adding the QCN consistently yields better results than classical Multi-Layer Perceptron (MLP) projection at every supervision level. At the 20% label ratio, the full BN+NL+QCN architecture reaches 84.9% accuracy and beats the equivalent-capacity MLP by a 0.9% absolute margin. The findings show that QCN effectively maps complex and non-linear knee pathologies in quantum feature space to explain the accuracy gain that exceeds typical run-to-run variance and is independent of expanded parameter size.

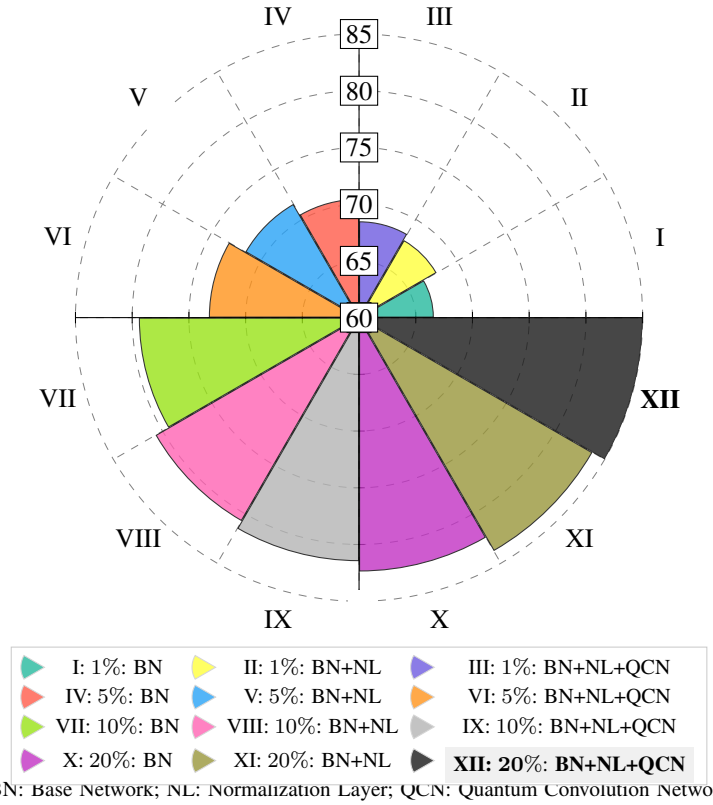


Figure 5.13: Impact of labeled data percentage on Teacher-Student framework accuracy.

(d) Effect of Decomposition: We analyze the impact of decomposition strategies in the hierarchical multi-classifier on H-SemiS performance for KOA severity grading. As shown in Figure 5.14, we compare three strategies: no decomposition, random decomposition, and hi-

Table 5.6: Ablation analysis of H-SemiS framework with different percentages of labeled samples and architectural components on the OAI dataset.

%	BN	NL	MLP	QCN	Acc	Pre	Rec	F1
1%	✓	×	×	×	66.4	65.5	63.2	64.3
	✓	×	×	✓	67.7	67.2	67.1	67.1
	✓	✓	✓	×	67.8	68.6	68.1	68.4
	✓	✓	×	✓	68.3	69.4	69.4	69.4
5%	✓	×	×	×	70.3	71.3	70.6	70.9
	✓	×	×	✓	71.4	71.7	72.8	72.2
	✓	✓	✓	×	72.1	73.2	72.8	73.0
	✓	✓	×	✓	73.1	74.7	73.9	74.3
10%	✓	×	×	×	79.3	80.2	78.3	79.2
	✓	×	×	✓	80.6	81.8	78.9	80.3
	✓	✓	✓	×	80.9	81.7	79.0	80.5
	✓	✓	×	✓	81.4	82.2	79.4	80.8
20%	✓	×	×	×	82.3	83.5	80.7	82.1
	✓	×	×	✓	83.7	84.9	81.6	83.2
	✓	✓	✓	×	84.0	85.1	82.3	83.8
	✓	✓	×	✓	**84.9**	**86.1**	**83.4**	**84.7**

BN: Base Network, NL: Normalization Layer, MLP: Multi-Layer Perceptron, QCN: Quantum Convolutional Neural Network.

erarchical decomposition, and examine the use of original samples (OriS) and reconstructed samples (RecS). Without decomposition, the framework performs direct multiclass classification, achieving an accuracy of 75.4% with OriS. Random decomposition, which splits classes into binary tasks without accounting for class imbalance, yields inferior results, with performance drops of 12.4% using OriS and 10.8% using OriS+RecS relative to hierarchical decomposition. Incorporating reconstructed samples consistently improves performance across all decomposition strategies, demonstrating their effectiveness in mitigating multiclass imbalance. Hierarchical decomposition achieves the best performance, reaching an accuracy of 84.9% with OriS+RecS, corresponding to a 4.5% gain over using OriS alone. These results show that hierarchical decomposition effectively integrates information from both sample types. In addition, the proposed hierarchical strategy employs a dual rule-based design (see Section 5.2.3.3) that supports balanced learning under multiclass imbalance.

5.4.4 Cross-Dataset Analysis

We assess the generalizability of H-SemiS through cross-dataset evaluation by training the framework on one dataset and testing it on another, including concatenated datasets. To

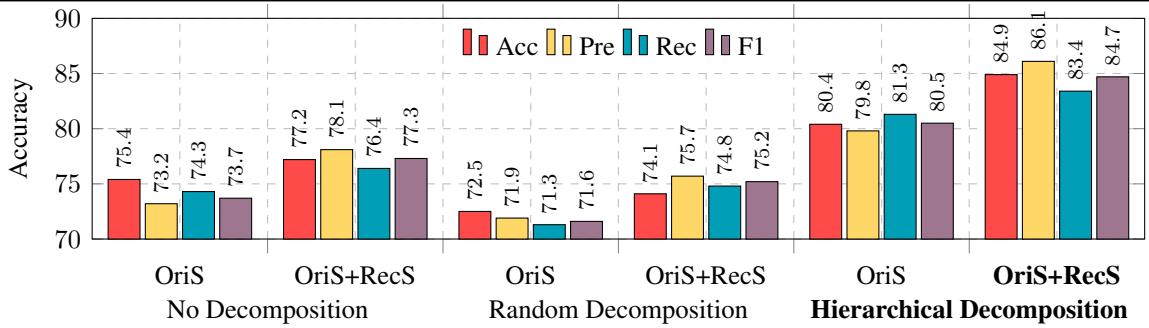


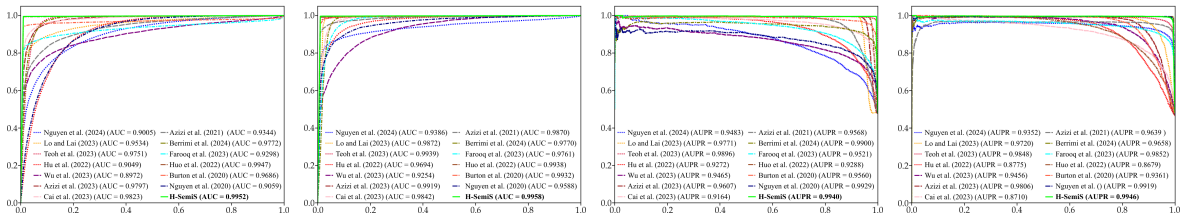
Figure 5.14: Performance analysis of H-SemiS under different decomposition techniques.

strengthen this analysis, we also consider additional binary datasets (see Section 5.3.1) alongside the multi-class datasets. Table 5.7 reports performance across different training and testing distributions for both multi-class and binary settings. H-SemiS performs best when trained and evaluated on the same dataset, whereas models trained on smaller datasets perform worse when tested on larger or combined datasets. For multi-class evaluation, training and testing on OAI [134] yields the highest accuracy of 84.9, which drops by 7.7 and 3.7 when tested on DKXI [135] and the combined OAI+DKXI dataset, respectively. A similar trend is observed in the binary setting, where H-SemiS achieves accuracies of 90.8, 92.3, and 91.7 when trained and tested on OP, KO, and OP+KO datasets, respectively. In contrast, training on OP [136] and testing on KO [137] and OP+KO leads to accuracy reductions of 11.3 and 18.9. We further report ROC and PR curves with AUC and AUPR for the bi-

Table 5.7: Evaluation of H-SemiS generalizability across varied training and testing samples.

Training	**Testing**											
	Category-1: Multi-class Dataset											
	OAI				**DKXI**				**OAI + DKXI**			
	Acc	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**
OAI	84.9	86.1	83.4	84.7	77.3	73.6	79.1	76.3	81.2	79.3	80.8	80.0
DKXI	68.2	67.9	67.3	67.6	86.8	85.3	83.6	84.4	73.6	69.8	72.1	70.9
OAI+DKXI	83.8	84.5	83.1	83.8	80.5	79.6	80.3	79.9	82.1	80.5	81.8	81.1
	Category-2: Binary Dataset											
	OP				**KO**				**OP + KO**			
	Acc	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**
OP	90.8	92.5	91.7	92.1	79.5	77.1	78.2	77.6	71.9	70.3	72.8	71.5
KO	88.2	89.3	90.2	89.7	92.3	94.6	92.8	93.7	85.3	83.2	84.9	84.0
OP+KO	87.6	85.1	86.3	85.7	89.5	92.7	91.3	92.0	91.7	90.8	89.4	90.1

nary datasets in Figure 5.15, demonstrating the robustness of H-SemiS under cross-dataset shifts. Overall, the cross-dataset results confirm that H-SemiS generalizes reliably across both multi-class and binary settings and remains competitive on unseen datasets.



(a) ROC Curve on OP Dataset (b) ROC Curve on KO Dataset (c) PR Curve on OP Dataset (d) PR Curve on KO Dataset

Figure 5.15: Performance comparison: ROC and PR curves for OP and KO datasets.

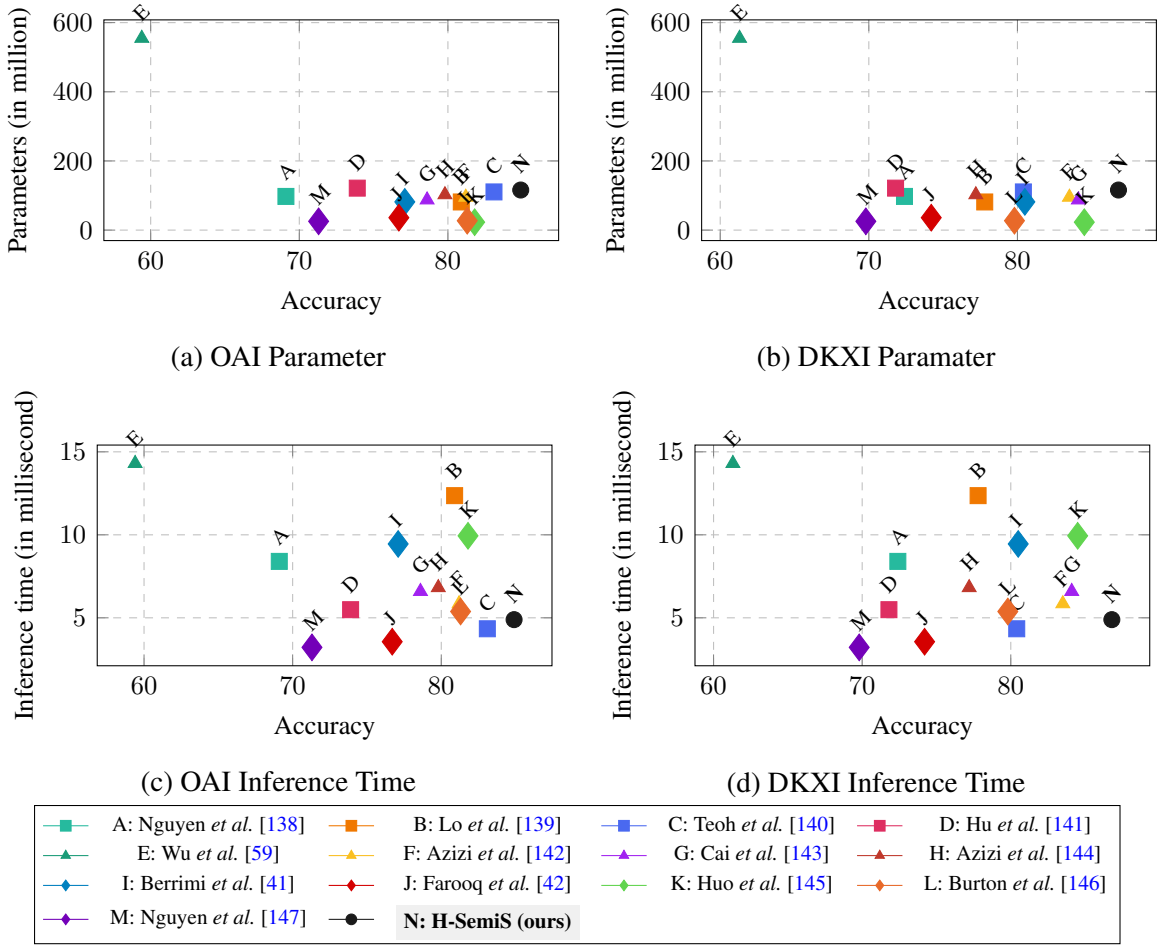


Figure 5.16: Comparative analysis of model parameters, inference time, and accuracy for the H-SemiS framework against existing baselines on benchmark datasets.

5.4.5 Model Complexity Analysis

We evaluate the computational cost of H-SemiS by comparing trainable parameters and inference time against supervised, self-supervised, and semi-supervised baselines. Among supervised methods, [140] reports the lowest inference time at 4.34 milliseconds (ms) with 110.40 million (M) parameters, whereas [141] has the largest model size at 121.35M. In the self-supervised setting, [59] exhibits the highest complexity with 554.75M parameters and a 14.30 ms inference time, while [142] is more efficient with 95.15M parameters and 5.85 ms inference. Among semi-supervised approaches, [147] achieves the fastest inference at 3.22 ms with 25.35M parameters, whereas [145] shows higher latency at 9.94 ms despite fewer parameters. In comparison, H-SemiS integrates hierarchical semi-supervised learning with self-supervision, maintaining a parameter count of 115.89M and an inference time of 4.09 ms, among the lowest for high-performing models. To ensure a fair comparison, the inference times for all competing baselines are measured under identical hardware conditions (see the Training Setting section for detailed specifications). As summarized in Table 5.8 and visualized in Figure 5.16, H-SemiS achieves state-of-the-art performance while balancing accuracy, model size, and inference efficiency.

Table 5.8: Model Complexity Analysis of the H-SemiS framework, comparing parameters and inference times with existing baselines.

	Methods	**Parameters (M)**	**Inference time (ms)**
I	Nguyen *et al.* [138]	97.40	8.40
	Lo *et al.* [139]	81.36	12.36
	Teoh *et al.* [140]	110.40	4.34
	Hu *et al.* [141]	121.35	5.49
II	Wu *et al.* [59]	554.75	14.30
	Azizi *et al.* [142]	95.15	5.85
	Cai *et al.* [143]	87.32	6.59
	Azizi *et al.* [144]	102.35	6.82
III	Berrimi *et al.* [41]	81.58	9.45
	Farooq *et al.* [42]	35.87	3.56
	Huo *et al.* [145]	22.89	9.94
	Burton *et al.* [146]	27.11	5.38
	Nguyen *et al.* [147]	25.35	3.22
Ours	H-SemiS	**115.89**	**4.09**

I: Supervised-based Techniques, II: Self-Supervised-based Techniques, III: Semi-Supervised-based Techniques; **M**: million, **ms**: millisecond, All methods evaluated on identical hardware (details in Training Setting).

5.5 Conclusion

In this work, we propose H-SemiS, a hierarchical semi-supervised framework with self-supervision for grading knee osteoarthritis (KOA) severity from knee X-ray (KXR) images. The framework addresses data scarcity and multiclass imbalance by combining limited labeled data with reconstructed samples. To handle imbalance, H-SemiS adopts a dual strategy. First, it decomposes the multiclass problem into a hierarchy of binary tasks, using a rule-based approach to promote balanced learning across severity levels. Second, it introduces an adversarial-inspired self-supervised module to generate reconstructed samples, increasing feature diversity and improving data distribution. We further integrate quantum learning within a teacher-student framework to capture complex nonlinear anatomical relationships and enhance fine-grained feature discrimination. Extensive experiments on multiclass imbalanced datasets demonstrate that the H-SemiS outperforms competing methods across key evaluation metrics. Ablation studies confirm the contribution of each component. These results establish H-SemiS is a reliable and effective approach for grading KOA severity, particularly in resource-constrained clinical settings.

Chapter 6

Transformer-aware Multiscale Progressive Encoder-Decoder Network for Tumor and Liver Segmentation

CONTENTS

6.1 Introduction . . . 123
6.2 Methodology . . . 124
6.2.1 Framework Overview . . . 124
6.2.2 Preprocessing . . . 125
6.2.3 Proposed Framework . . . 125
6.3 Experimental Settings . . . 138
6.3.1 Benchmark Dataset . . . 138
6.3.2 Evaluation Metrics . . . 140
6.4 Experimental Results . . . 141
6.4.1 Quantitative Analysis . . . 141
6.4.2 Qualitative Analysis . . . 146
6.4.3 Ablation Analysis . . . 147
6.4.4 Cross-Dataset Analysis . . . 149
6.4.5 Model Complexity Analysis . . . 152
6.5 Conclusion . . . 152

6.1 Introduction

This chapter focuses on the liver and tumor segmentation and addresses the objective of precise boundary refinement for medical image analysis. The aim is to develop a method for accurately localizing pathological regions while preserving structural boundaries in complex imaging. Accurate segmentation of the liver and tumors in CT/MRI is important in treatment planning, including disease monitoring and surgical guidance. On the other hand, manual identification is labor-intensive and subject to observer variability. Automated segmentation remains difficult because liver tissue shares similar intensity patterns with surrounding organs. Also, tumors have similar texture variation in shape and size. In many cases, lesions appear small and low-contrast, making the boundaries difficult to locate. Existing convolution-based methods capture local details but overlook the long-range context. On the other hand, transformer-based approaches capture global context but miss fine spatial structures. These limitations motivate the need for a framework for jointly modeling contextual and spatial information with precise boundary improvement.

To address these limitations, this chapter presents a Transformer-aware Multiscale Progressive Encoder-Decoder Network (T-MPEDNet) for liver and tumor segmentation. The framework employs a progressive encoder-decoder with adaptive feature extraction and channel recalibration to preserve spatial consistency during feature fusion. A Dynamic Contextual Attention module combines attention learning with dynamic convolutions to capture global dependencies and suppress irrelevant background. A Multi-Scale Atrous Spatial module extracts discriminative patterns with multiple receptive fields of varying lesion sizes and textures. In addition, a Morphological Boundary Refinement module sharpens the boundary and reduces the ambiguity between the adjacent tissues. The proposed framework provides the accurate localization of the complex hepatic structures.

The remainder of this chapter is organized as follows. Section 6.2 presents the proposed methodology, including the progressive encoder-decoder architecture, multi-scale learning, and boundary refinement modules. Section 6.3 describes the benchmark datasets and evaluation metrics. Section 6.4 discusses the quantitative and qualitative findings, including ablation studies and model complexity analysis. Finally, Section 6.5 summarizes the main outcomes and clinical relevance of the proposed framework.

6.2 Methodology

This section presents the proposed framework for liver and tumor segmentation. It describes the overall pipeline, including preprocessing and framework design. The framework learns from the key components for capturing meaningful representations, as detailed in the following subsections.

6.2.1 Framework Overview

Figure 6.1 presents the architectural blueprint of the T-MPEDNet framework, which enhances segmentation performance for both liver and tumor regions. The framework begins with a preprocessing block, followed by a progressive feature encoder-decoder and a morphological boundary refinement module. The pre-processing block assists the progressive feature encoder-decoder structure by eliminating irrelevant details from CT and MRI scans. The preprocessing step employs normalization techniques to improve sample quality, which optimizes the network's learning efficiency and learning rate for deep feature extraction. The progressive feature encoder-decoder extracts deep feature maps by integrating the Dynamic Contextual Attention (δ_{DCA}) module and the Multi-Scale Atrous Spatial (μ_{MSAS}) module. The δ_{DCA} module globally analyzes deep features from a semantic perspective to highlight effective channels, while μ_{MSAS} enhances valuable features across both global and local scales. Finally, the morphological boundary refinement module (β_{MBR}) sharpens unclear boundaries using morphological erosion, thereby refining blurred edges and improving the anatomical accuracy of liver and tumor contours. Section 6.2.3 provides a detailed explana-

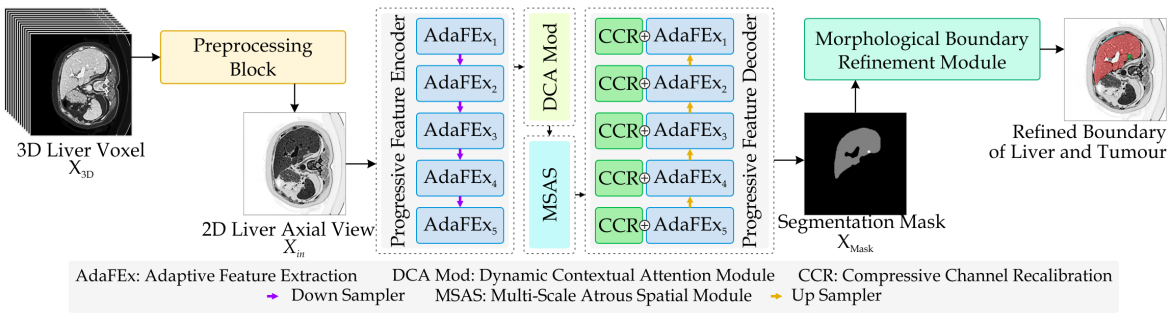


Figure 6.1: The workflow of our proposed T-MPEDNet framework for segmenting liver and tumors in computed tomography and magnetic resonance imaging.

tion of the proposed framework.

6.2.2 Preprocessing

The preprocessing block systematically performs a sequence of steps to prepare medical imaging data for accurate liver area segmentation, as illustrated in Figure 6.2. We design this block to handle both CT and MRI scans using modality-specific preprocessing techniques. For both modalities, we first convert the 3D voxel data X_{3D} into 2D slices of 512×512 pixels, then resize them to 256×256 to optimize computational efficiency. For CT scans, we apply Hounsfield Unit (HU) windowing with a window size of $(-250, 200)$ to exclude irrelevant anatomical structures. HU windowing is a standardized method for quantifying tissue density, with liver tissue typically falling between 40 and 50 HU. For MRI scans, which do not use HU values, we use min-max normalization to scale pixel intensities to 0-1, ensuring a consistent intensity range across acquisitions. Next, we apply Contrast Limited Adaptive Histogram Equalization (CLAHE) to both CT and MRI images to enhance local contrast, suppress over-amplification of noise, and improve visibility of anatomical boundaries. Finally, we use z-score normalization to standardize intensity values and reduce inter-scan variability, making the data more robust for framework training across both modalities.

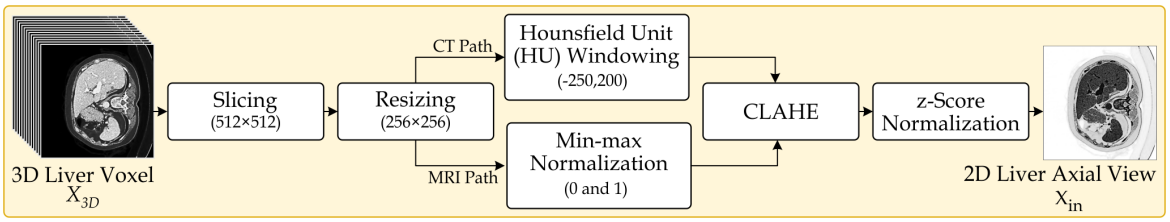


Figure 6.2: Visualization of the processing pipeline within the preprocessing block.

6.2.3 Proposed Framework

6.2.3.1 Progressive Feature Encoder Decoder

Figure 6.3 outlines the architecture of the progressive feature encoder-decoder block integrated into the proposed T-MPEDNet for liver and tumor segmentation in CT and MRI scans. The proposed encoder-decoder module aims to (a) efficiently extract features from input CT and MRI scans, (b) adaptively recalibrate fused features on a channel-wise basis,

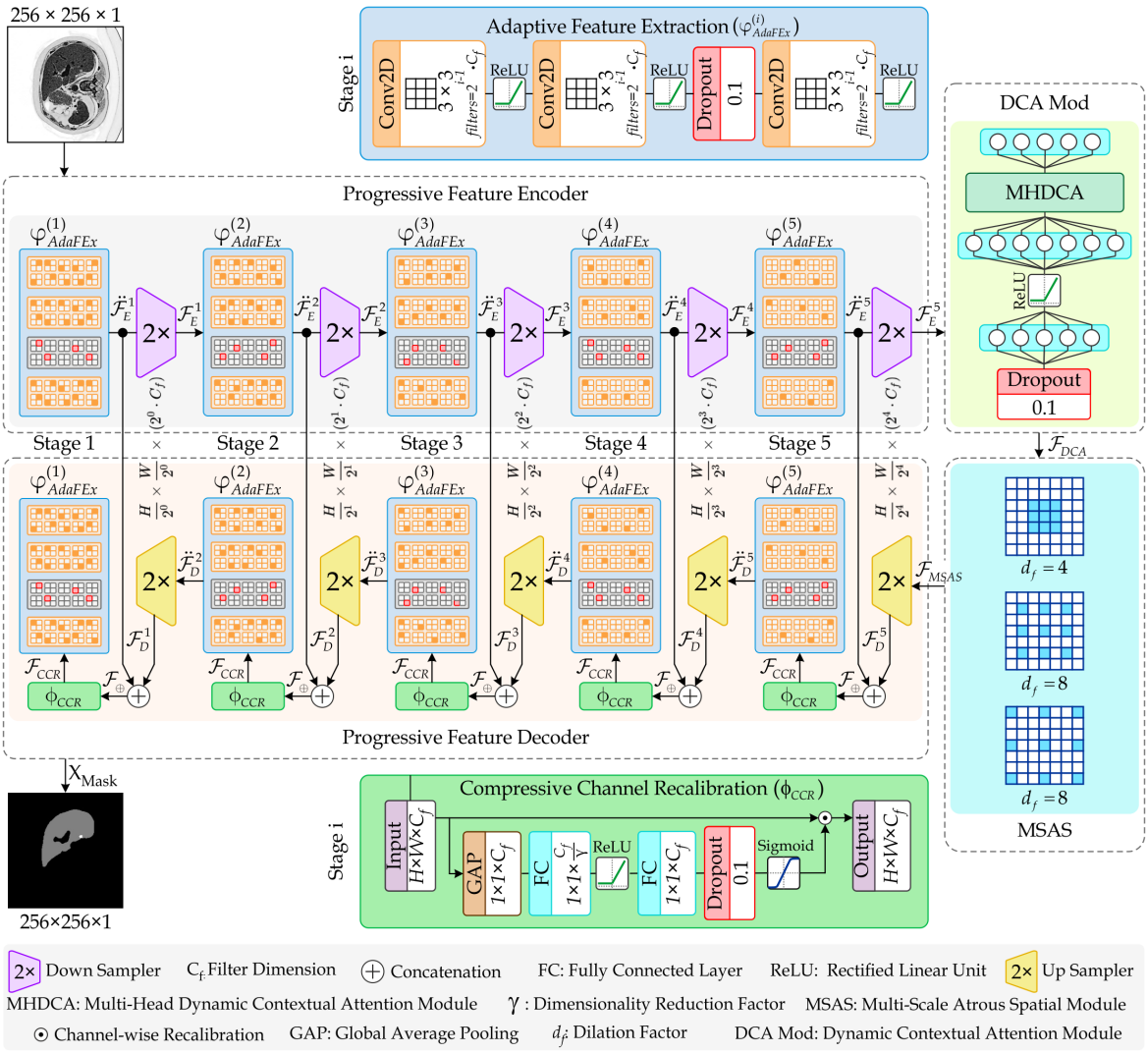


Figure 6.3: Proposed architecture of the Progressive Feature Encoder-Decoder (PED) block.

(c) preserve spatial integrity to extract higher-level global features, and (d) capture diverse multi-scale features with varying receptive fields to enhance low-level feature representation with intricate spatial details. To achieve these objectives, the encoder-decoder network integrates four key components: the Adaptive Feature Extraction (AdaFEx) module, the Compressive Channel Recalibration (CCR) module used in the decoder network, the Dynamic Contextual Attention (DCA) module, and the Multi-Scale Atrous Spatial (MSAS) module.

The progressive encoder is composed of five AdaFEx modules, labeled as $\varphi^{(i)}_{AdaFEx}$, $i \in [1, 2, \ldots, 5]$, as illustrated in Figure 6.3. The φ_{AdaFEx} takes an input scan X_{in} with dimensions $H \times W \times 1$ as input, where H is the height, W is the width, and 1 is the grayscale channel. Adaptive deep features, denoted as $\ddot{\mathcal{F}}^{i}_{E}$, are extracted for each AdaFEx module,

followed by a down-sampling operation. The down-sampler, denoted as $\bigvee$, incorporates a max pooling layer with a 2×2 pooling window, reducing the spatial dimensions by half while preserving the most salient features. The output after down-sampling are labeled as $\mathcal{F}_E^1, \mathcal{F}_E^2, \ldots, \mathcal{F}_E^5$. The progressive encoder block computes the final feature map by employing a down-sampler, as formally expressed in Equation (6.1):

$$\mathcal{F}_E^i = \bigvee \left(\varphi_{AdaFEx}^{(i)}(\mathcal{F}_E^{i-1}; \theta_E^i) \right), \quad i \in \{1, 2, \ldots, 5\} \tag{6.1}$$

Here, θ_E^i denotes the parameters of the i-th AdaFEx module $\varphi_{AdaFEx}^{(i)}$, $\mathcal{F}_E^i$ is the feature map generated by the down-sampler integrated with $\varphi_{AdaFEx}^{(i)}$, and $\mathcal{F}_E^0 = X_{in}$ is the input image.

The progressive encoder generates spatially rich feature maps, which are fed into the Dynamic Contextual Attention (DCA) module. The progressive encoder generates static contextual information through extracted deep feature maps. At the same time, the DCA module dynamically learns attention weights to adaptively recalibrate these features, capturing dynamic contextual information. This fusion of static and dynamic contextual information enhances spatial representation and global reasoning, resulting in the enriched feature map $\mathcal{F}_{DCA}$. The resulting $\mathcal{F}_{DCA}$ is then fed into the Multi-Scale Atrous Spatial (MSAS) module, which actively extracts multi-scale features with diverse receptive fields, effectively capturing the rich spatial details within low-level features. The resulting multi-scaled output, represented as $\mathcal{F}_{MSAS}$, serves as the input to the progressive decoder. The progressive feature decoder refines the extracted $\mathcal{F}_{MSAS}$ by upsampling it and generating segmentation masks. The progressive feature decoder uses encoder features to guide upsampling via five mirrored AdaFEx modules, each replicating the encoder's block architecture.

As input, each AdaFEx module receives recalibrated features, denoted as $\phi_{CCR}^{(i)} \in [1, 2, \ldots, 5]$ from the Compressive Channel Recalibration (CCR) module. The CCR module combines up-sampled reconstructed features (denoted as $\mathcal{F}_D^i$) from the previous $\varphi_{AdaFEx}^{(i+1)}$ module with the output of the corresponding AdaFEx module from the mirrored encoder. A deconvolution layer with a 2×2 kernel size acts as an up-sampler ($\bigwedge$), increasing the spatial dimensionality of the feature maps and enabling the network to generate high-resolution segmentation masks. Each AdaFEx module in the decoder outputs a refined feature map, denoted as $\ddot{\mathcal{F}}_D^1, \ddot{\mathcal{F}}_D^2, \ldots, \ddot{\mathcal{F}}_D^5$. The final feature map is generated by the last AdaFEx module,

and its computation is formally expressed in Equation (6.2). Finally, the progressive decoder outputs a 2D binary mask of the liver and the corresponding tumor.

$$\ddot{\mathcal{F}}_D^i = \varphi_{AdaFEx}^{(i)}\Big(\phi_{CCR}^{(i)}\big(\ddot{\mathcal{F}}_E^i \oplus \bigwedge(\ddot{\mathcal{F}}_D^{i+1})\big); \theta_D^i\Big)\; i \in \{1, 2, \ldots, 5\} \tag{6.2}$$

Here, $\oplus$ representing the concatenation, θ_D^i represents the parameters of the i-th AdaFEx module $\varphi_{AdaFEx}^{(i)}$, and $\ddot{\mathcal{F}}_D^i$ is the feature map produced by the AdaFEx integrated with $\phi_{CCR}^{(i)}$. Also, $\ddot{\mathcal{F}}_D^5 = \mathcal{F}_E^5$ and $\ddot{\mathcal{F}}_D^1 = X_{Mask}$ is the final 2D-mask of the liver and corresponding tumor.

The progressive encoder-decoder architecture employs skip connections between corresponding AdaFEx modules. These connections bridge the gap between feature maps with lower and higher spatial resolution, alleviating information loss during downsampling in the encoder. During upsampling in the decoder, skip connections provide a rich source of contextual information to higher-resolution layers, enabling them to generate finer details in the final segmentation mask.

(a) Adaptive Feature Extraction Module: Neural networks leverage the feature-extraction capabilities of convolutional layers, making them a powerful approach for learning data characteristics. However, medical image analysis remains a significant challenge despite advancements in the field, primarily because diverse medical image types require different feature extraction strategies. This work introduces Adaptive Feature Extraction (AdaFEx), a novel feature extraction module denoted as φ_{AdaFEx}. AdaFEx efficiently learns salient features of liver and tumor tissues from CT and MRI scans, empowering the framework to locate and segment these structures with high precision.

AdaFEx progressively extracts deep context from low-level details, gradually transitioning to higher-level semantic representations. As illustrated in Figure 6.3, AdaFEx takes X_{in} as input and feeds through five stages $\varphi_{AdaFEx}^{(i)}\Big|_{i=1}^{5}$. Each stage comprises three Conv layers with f filters of 3×3 kernel, where f progressively increases at each stage according to the formula $2^{i-1} \cdot C_f$, where the value of C_f is 16. ReLU activation functions after each Conv layer introduce nonlinearity, enabling the framework to extract and learn complex hidden connections and relationships within the data. A Dropout layer is integrated after the second Conv layer in each stage to enhance regularization and reduce neuron interdependencies.

(b) Compressive Channel Recalibration Module: The Compressive Channel Recalibra-

tion (CCR) module, denoted as ϕ_{CCR}, effectively leverages channel-wise dependencies in feature maps. It automatically learns weights and recalibrates the original feature channels, thereby enhancing feature representation. The detailed internal structure of the CCR module is visualized in Figure 6.4. It first compresses the concatenated feature map $\mathcal{F}_{\oplus} \in \mathbb{R}^{H \times W \times C_f}$ via channel aggregation, producing a compressed channel representation. Subsequently, a recalibration operation generates modulation weights for channels, which are applied to $\mathcal{F}_{\oplus}$ to yield the output $\mathcal{F}_{CCR}$. This output then serves as input to the AdaFEx module of the progressive decoder block.

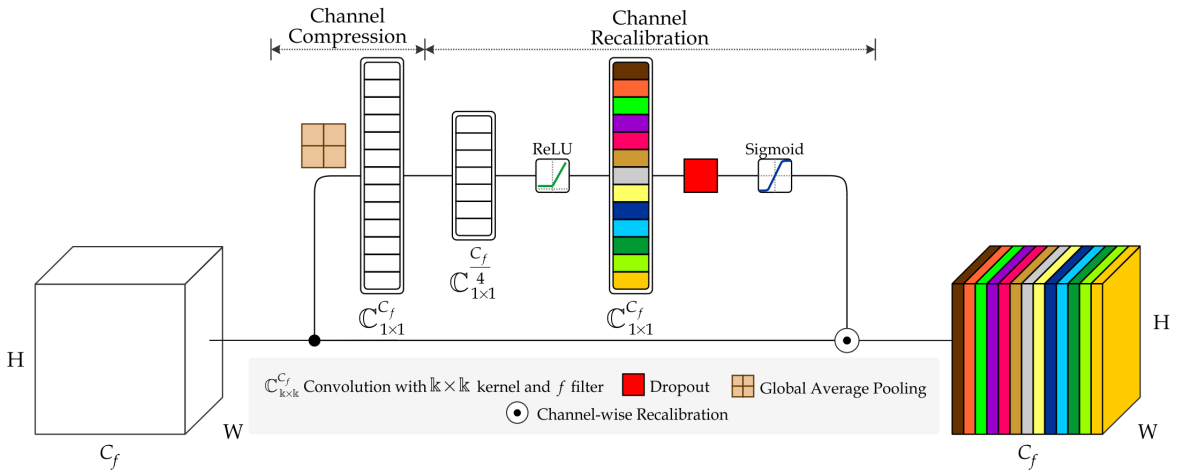


Figure 6.4: Proposed architecture of Compressive Channel Recalibration (CCR) module.

For an input $\mathcal{F}_{\oplus} = [f^1_{\oplus}, f^2_{\oplus}, \ldots, f^{C_f}_{\oplus}]$, where $f^k_{\oplus} \in \mathbb{R}^{H \times W \times C_f}$, we leverage global average pooling to reduce the spatial dimensions of each channel while preserving channel-specific information. This produces a static vector $\mathcal{Z} \in \mathbb{R}^{1 \times 1 \times C_f}$ by by shrinking $\mathcal{F}_{\oplus}$ along its spatial dimensions $H \times W$. The k-th element of $\mathcal{Z}$ is calculated using Equation (6.3):

$$z_k = \mathcal{F}_{com}(f^k_{\oplus}) = \sum_{i=1}^{H \times C} f^k_{\oplus}(j) \tag{6.3}$$

The above procedure embeds global spatial information within a vector $z_k \in \mathcal{Z}$. Following the compression operation, the recalibration operation ($\mathcal{F}_{rec}$) is calculated to capture channel-wise dependencies comprehensively. Recalibration involves transforming $\mathcal{Z}$ to $\hat{\mathcal{Z}}$ by integrating two Fully Connected (FC) layers as outlined in Equation (6.4):

$$\mathcal{A} = \mathcal{F}_{rec}(\mathcal{Z}, \mathcal{W}) = \underset{Sigmoid}{\sigma}\Big(\big(\mathcal{W}_2 \underset{ReLU}{\alpha}(\mathcal{W}_1 \mathcal{Z})\big)\Big) \tag{6.4}$$

Here, $\mathcal{W}_1 \in \mathbb{R}^{C_f \times \frac{C_f}{\gamma}}$, $\mathcal{W}_2 \in \mathbb{R}^{\frac{C_f}{\gamma} \times C_f}$ denotes the weight matrices of the first and second fully connected layers, and γ determines the dimensionality reduction factor controlling the compression ratio. We empirically found that $\gamma = 4$ yields optimal performance. Also, $\underset{Sigmoid}{\sigma}$ and $\underset{ReLU}{\alpha}(\cdot)$ are the sigmoid function and ReLU operator. To retain the non-linear mapping capabilities of the sigmoid function a dropout layer with a 0.1 probability of dropping units are employed before the sigmoid function. The final step involves rescaling the original feature channel $\mathcal{F}_{\oplus}$ with the earned modulation weights $\mathcal{A}$ as given in Equation (6.5). This adaptive scaling emphasize on the informative channels and suppresses less relevant.

$$f_{CCR}^{j} = a_k \odot z_k \tag{6.5}$$

The CCR module generates the final output $\mathcal{F}_{CCR} = [f_{CCR}^{1}, f_{CCR}^{2}, \ldots, f_{CCR}^{C_f}]$, a tensor containing rescaled feature maps. Each element f_{CCR}^{i} in $\mathcal{F}_{CCR}$ is obtained by adaptively modulating the corresponding channel $f_{\oplus}^{k} \in \mathbb{R}^{H \times W \times C_f}$ of the input feature map using the learned activation $a_k \in \mathcal{A}$. This channel-wise modulation, denoted by $\odot$, effectively amplifies informative channels while suppressing less relevant ones. As the network trains, it dynamically tunes these activations, prioritizing features that contribute most to the framework's representation power.

(c) Dynamic Contextual Attention Module: The Dynamic Contextual Attention (DCA) module employs a novel Transformer-inspired building block, the Multi-Head Dynamic Contextual Attention (MHDCA). MHDCA enhances the representational properties of deep networks by leveraging contextual information among input keys and facilitates self-attention learning. The primary objective of the DCA module is to capture extensive dependencies within the adaptive features ($\mathcal{F}_E^5$) derived from a progressive encoder block while preserving the integrity of spatial information. We utilize a transformer block that includes a Fully Connected (FC) layer, MHDCA, feed-forward layers, and a dropout layer to achieve this. Given the crucial role of the attention mechanism in the DCA module, we provide a detailed introduction to MHDCA before delving into the remaining components.

Multi-Head Dynamic Contextual Attention (MHDCA): The Multi-Head Dynamic Contextual Attention (MHDCA) module, inspired by the Transformer's success in capturing long-range relationships in sentences, enhances the framework's representational power in two

key ways. Firstly, MHDCA extends the framework's capacity to focus on diverse positions, as each head's encoding incorporates knowledge from the other heads' encodings. Secondly, the partitioning of input features broadens the representation subspaces, leading to more indicative attention weights for each partition. Concatenating the enriched representation yields a superior overall representation that boosts classification accuracy. Within the Dynamic Contextual Attention (DCA) module, MHDCA leverages dynamic convolutions to adapt convolutional kernel parameters to specific input samples, thereby enhancing its representational flexibility.

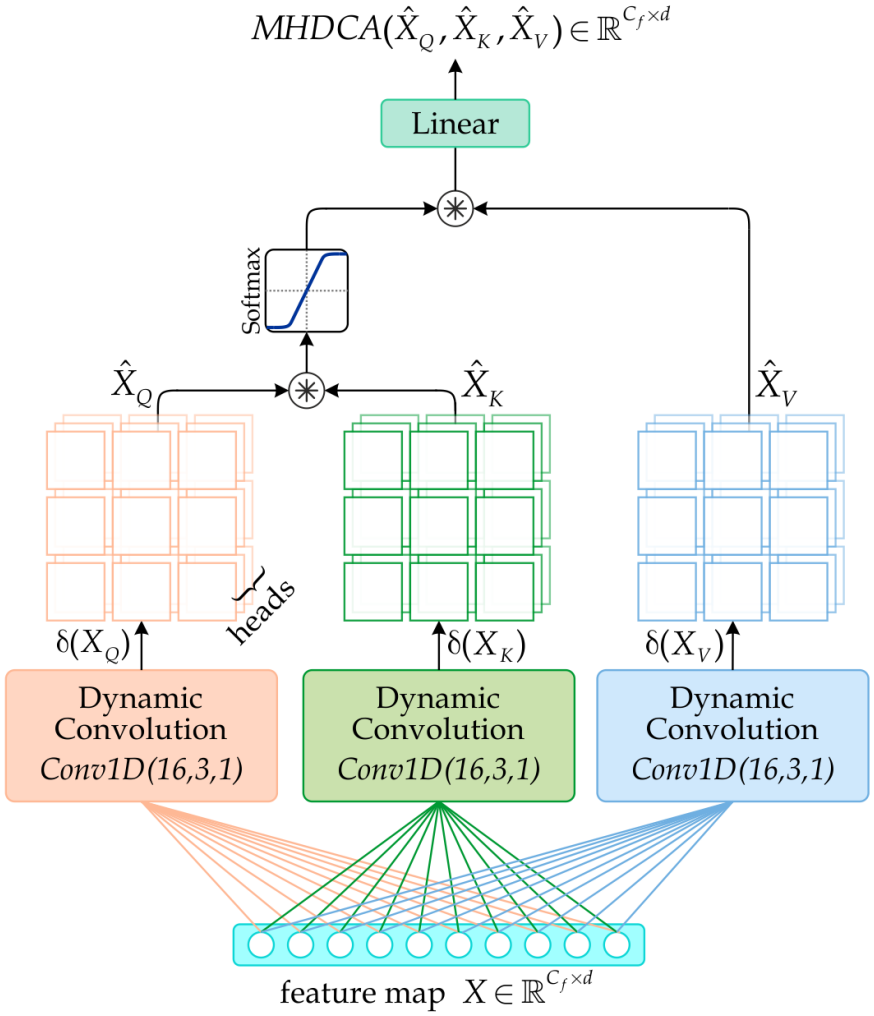


Figure 6.5: Proposed architecture of Multi-Head Dynamic Contextual Attention (MHDCA) module.

The FC layer of the DCA module outputs $\mathcal{F}^{FC_1}$, a tensor of dimension $C_f \times d$, where C_f signifies the number of convolution filters and d represents the length of a tensor. This

key input, visualized in Figure 6.5, feeds into the MHDCA block, extracting the discriminative features. Notably, MHDCA requires three identical copies of $\mathcal{F}^{FC_1}$ as inputs. First, it dynamically adjusts the internal representation of $\mathcal{F}^{FC_1}$ through a dynamic convolution (represented by Ω), generating $\widehat{\mathcal{F}}^{FC_1} = \Omega(\mathcal{F}^{FC_1})$. Subsequently, these three transformed matrices, query, key, and value of $\widehat{\mathcal{F}}^{FC_1}$ are fed into the attention mechanism, as described in Equation (6.6).

$$ATT(\widehat{\mathcal{F}}^{FC_1}_{\mathcal{Q}}, \widehat{\mathcal{F}}^{FC_1}_{\mathcal{K}}, \widehat{\mathcal{F}}^{FC_1}_{\mathcal{V}}) = \underset{Softmax}{\sigma}\left(\frac{(\widehat{\mathcal{F}}^{FC_1}_{\mathcal{Q}})(\widehat{\mathcal{F}}^{FC_1}_{\mathcal{K}})^T}{\sqrt{d}}\right) \cdot \widehat{\mathcal{F}}^{FC_1}_{\mathcal{V}} \tag{6.6}$$

Here, $(\cdot)$ is the multiplication operation, and $\underset{Softmax}{\sigma}$ represents the softmax function. We expand the attention mechanism across $\mathcal{H}$ heads for each of the three duplicated matrices derived from the FC layer output to further enrich the representation. Specifically, we partitioned each matrix $\widehat{\mathcal{F}}^{FC_1}$ into $\mathcal{H}$ sub-spaces, denoted as $\widehat{\mathcal{F}}^{FC_1}_1, \ldots, \widehat{\mathcal{F}}^{FC_1}_{\mathcal{H}}$, where $\widehat{\mathcal{F}}^{FC_1}_h \in \mathbb{R}^{C_f \times \frac{d}{\mathcal{H}}}$, $1 \leq h \leq \mathcal{H}$. Within particular subspaces h, an attention similarity score A^h is computed, and concatenating all $\mathcal{H}$ representations together produces the final result, as expressed in Equation (6.7a) and Equation (6.7b).

$$A^h = ATT(\widehat{\mathcal{F}}^{FC_1}_{\mathcal{Q}_h}, \widehat{\mathcal{F}}^{FC_1}_{\mathcal{K}_h}, \widehat{\mathcal{F}}^{FC_1}_{\mathcal{V}_h}) \in \mathbb{R}^{C_f \times \frac{d}{\mathcal{H}}} \tag{6.7a}$$

$$\mathcal{F}_{MHDCA} = MHDCA(\widehat{\mathcal{F}}^{FC_1}_{\mathcal{Q}}, \widehat{\mathcal{F}}^{FC_1}_{\mathcal{K}}, \widehat{\mathcal{F}}^{FC_1}_{\mathcal{V}}) = Concat(A^1, \ldots, A^{\mathcal{H}}) \tag{6.7b}$$

Following the MHDCA layer, the DCA module employs a dedicated Feed-Forward (FF) neural network to further refine features. This network uses a stacked configuration of two FC layers, where each layer employs the non-linear ReLU activation function to introduce non-linearity, enabling the network to learn complex relationships in the data. Additionally, the FF network incorporates an L1/L2 regularizer to prevent excessive weight values and achieve robust generalization. Formally, this operation can be expressed as $\mathcal{F}_{DCA} = \psi_{FC_4}\Big(r_{L_1L_2}\big(\underset{ReLU}{\sigma}\psi_{FC_3}(\mathcal{F}_{MHDCA})\big)\Big)$, where ψ_{FC_3} and ψ_{FC_4} represent the two FC layers, $\underset{ReLU}{\sigma}$ represents the ReLU activation function, and $r_{L_1L_2}$ represent the L1/L2 regularizer in the DCA module. The resulting $\mathcal{F}_{DCA}$ captures a broader spectrum of contextual

relationships, preserving spatial information to capture local and long-range dependencies within the input data.

(d) Multi-Scale Atrous Spatial Module: In computer vision tasks, extracting detailed features while maintaining global context presents a significant challenge. Balancing feature resolution and receptive field size is crucial as larger receptive fields capture broader context but sacrifice detail, while smaller ones preserve fine-grained information but lack context. We address this challenge by proposing a novel Multi-Scale Atrous Spatial (MSAS) module that efficiently captures fine and coarse-grained features.

MSAS leverages multiple parallel atrous depthwise convolutions with varying dilation rates, followed by pointwise convolutions (1×1 kernel), to extract multi-scale features from diverse receptive fields. In addition to these atrous convolutions, MSAS incorporates a parallel 2×2 max pooling operation to capture low-level local features. The max pool branch provides downsampled spatial features, which complement the multi-dilated convolution outputs and enhance the framework's ability to segment structures of varying sizes. The different branches (dilated convolutions with dilation rates $1, 4, 8, 12$ and max pooling) are then concatenated along the channel dimension, as shown in Figure 6.6. A 1×1 convolution followed by batch normalization, dropout, and ReLU activation further fuses and regularizes the extracted multi-scale features.

In Figure 6.6, $\mathcal{F}_{DCA}$ represents the input feature map comprising contextual features from the previous Dynamic Contextual Attention (DCA) module, while $\mathcal{F}_{MSAS}$ denotes the output enriched with multi-scale spatial features. The atrous depthwise convolution operation is mathematically defined in Equation (6.8).

$$\mathcal{F}_{K,d_f}[l] = \sum_{\kappa=1}^{K} \mathcal{F}_{DCA}[l + d_f * \kappa] * \omega_\kappa \tag{6.8}$$

Here l is the location on the feature map, d_f denotes the dilation factor, ω_κ is the κ-th filter weight, and K represents the filter size. The final MSAS feature map, incorporating dilated convolutions and max pooling, is formulated in Equation (6.9).

$$\mathcal{F}_{MSAS} = \mathcal{F}_{3,1}\langle\mathcal{F}_{DCA}\rangle + \mathcal{F}_{3,4}\langle\mathcal{F}_{DCA}\rangle + \mathcal{F}_{3,8}\langle\mathcal{F}_{DCA}\rangle + \mathcal{F}_{3,12}\langle\mathcal{F}_{DCA}\rangle + \mathcal{F}_{\text{MaxPool}}\langle\mathcal{F}_{DCA}\rangle \tag{6.9}$$

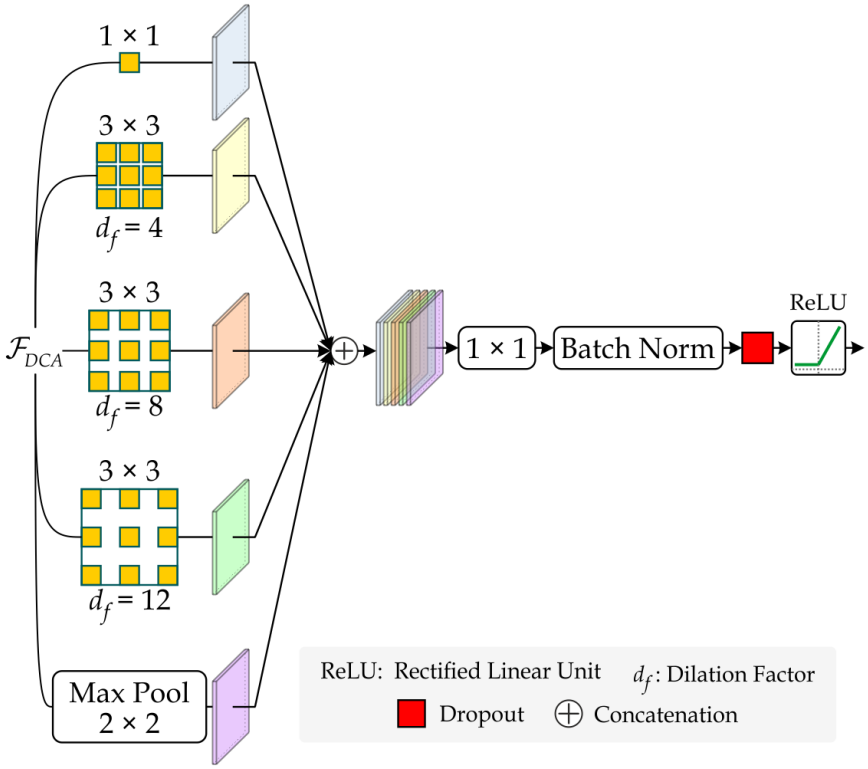


Figure 6.6: Proposed architecture of Multi-Scale Atrous Spatial (MSAS) module.

Here, $\mathcal{F}_{3,d_f}$ denotes the output of 3×3 atrous convolutions at different dilation rates, and $\mathcal{F}_{\text{MaxPool}}$ denotes the max pooling feature map. By concatenating the multi-scale features, including a max-pooled feature and an MSAS module, a rich contextual representation is generated. This representation is progressively upsampled and fused with recalibrated features in the decoder to further enhance the segmentation performance across diverse anatomical structures and imaging conditions.

6.2.3.2 Morphological Boundary Refinement Module

Segmenting liver and tumor regions is challenging due to indistinct boundaries with neighboring organs, leading to weaker boundary pixels. This limitation is addressed by incorporating edge priors that provide precise boundary information, thereby improving segmentation accuracy by guiding the framework to capture finer details. The Morphological Boundary Refinement (MBR) module is integrated after the Progressive Encoder-Decoder architecture to enhance focus on the liver and tumor boundaries by actively extracting additional boundary information, thereby improving segmentation accuracy.

The MBR module operates as a post-processing step, refining the initial segmentation mask X_{Mask}, as illustrated in Figure 6.7, by applying binary morphological erosion followed by mask subtraction to isolate boundary pixels. This process occurs in two main steps:

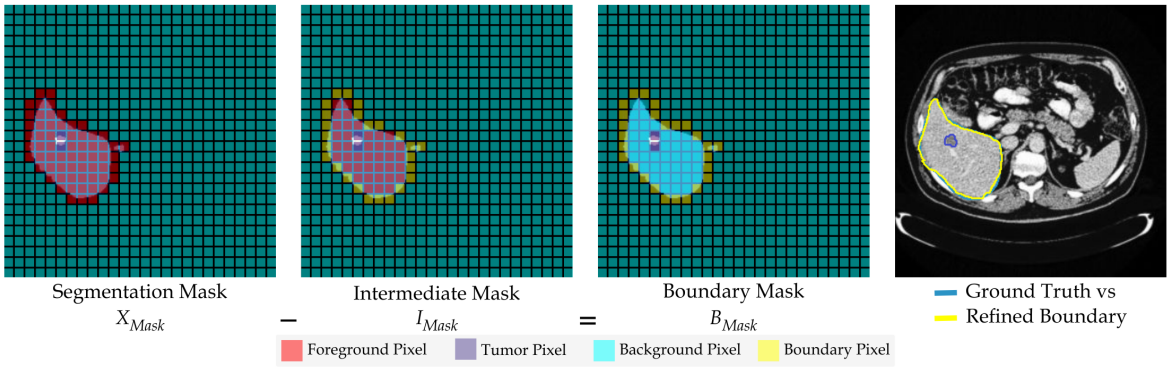


Figure 6.7: Proposed architecture of Morphological Boundary Refinement (MBR) module.

(i) Morphological Erosion: First, morphological erosion is applied to the input mask X_{Mask} using a predefined structuring element, S. Erosion effectively shrinks the foreground regions (red and purple areas in Figure 6.7) in the mask by removing pixels from their boundaries. The resulting eroded mask, I_{Mask}, is computed pixel-by-pixel. A pixel (p, q) in I_{Mask} is set to the foreground value 1 if and only if the structuring element S, when centered at (p, q), is entirely contained within the foreground region of X_{Mask}. Otherwise, the pixel (p, q) is set to the background value 0. The operation is formally defined in Equation (6.10).

$$I_{Mask}(p, q) = (X_{Mask} \ominus S)(p, q) = \begin{cases} 1, & \text{if } S_{(p,q)} \subseteq X_{Mask} \\ 0, & \text{otherwise} \end{cases} \tag{6.10}$$

Here, X_{Mask} represents the input binary mask, where 1 indicates the foreground and 0 represents the background. S is the structuring element (a 3×3 square in our case) that defines a neighborhood shape. I_{Mask} is the output eroded mask, (p, q) are the pixel coordinates, $\ominus$ denotes the erosion operator, and $S_{(p,q)}$ signifies the set of coordinates within the structuring element S translated to be centered at (p, q). This erosion step generates the intermediate eroded mask I_{Mask}. This mask represents the shrunken foreground, conceptually depicted as the red and purple regions within the yellow boundary in Figure 6.7.

(ii) Boundary Extraction: Second, the boundary mask, B_{Mask}, is extracted by performing a pixel-wise subtraction of the eroded mask I_{Mask} from the original input mask X_{Mask}. This

relationship is visually indicated between the panels in Figure 6.7. The calculation is defined in Equation (6.11):

$$B_{Mask} = X_{Mask} - I_{Mask} \tag{6.11}$$

This subtraction operation, where $1-0=1$, $1-1=0$, and $0-0=0$, effectively isolates the pixels that were part of the original mask foreground ($X_{Mask}(p,q)=1$) but were removed during the erosion process ($I_{Mask}(p,q)=0$). Therefore, B_{Mask} contains only the pixels corresponding to the boundary layer that was stripped away by the erosion, representing the desired refined boundary information, shown as the light yellow pixels in Figure 6.7. These boundary pixels are assigned a value of 1, while all other pixels are 0.

The resulting B_{Mask} explicitly identifies the boundary pixels, providing refined edge information that is subsequently incorporated to improve the overall accuracy of the segmentation process directly. Table 6.1 displays the overall configuration of the proposed T-MPEDNet in terms of feature dimensions along with internal layers configuration, respectively. Here, CCR denotes the Compressive Channel Recalibration module

6.2.3.3 Loss Calculation

Given the complex liver anatomy and the presence of small tumors in CT and MRI samples, relying solely on dice loss may result in unstable training. To address this challenge, we employ a combined loss function ($\mathscr{L}_{\text{C}}$), inspired by [151], which effectively manages the variations in liver structures. This combined loss ($\mathscr{L}_{\text{C}}$) consists of a weighted combination of dice loss ($\mathscr{L}_{\text{DSC}}$) and binary cross-entropy (BCE) loss ($\mathscr{L}_{\text{BCE}}$). Both loss functions mitigate the class imbalance between background and foreground pixels and accelerate convergence. We define the dice loss and BCE loss in Equation (6.12) and Equation (6.13) respectively.

$$\mathscr{L}_{\text{DSC}} = 1 - \frac{2\sum(\mathsf{G} \cap \mathsf{P}) + \varphi}{\sum(\mathsf{G} + \mathsf{P}) + \varphi} \tag{6.12}$$

$$\mathscr{L}_{\text{BCE}} = -[\mathsf{G} \cdot \log(\mathsf{P}) + (1 - \mathsf{G}) \cdot \log(1 - \mathsf{P})] \tag{6.13}$$

Here G and P represent a ground-truth and predicted labels, respectively, and φ represents a smoothing factor to prevent division by zero. We define the final combined loss function is

Table 6.1: Progressive Encoder-Decoder Network architecture of T-MPEDNet with feature dimensions and layer configurations.

Module	Input	Output	Network	CCR
Progressive Encoder Block				
$AdaFEx_1$	$256 \times 256 \times 1$	$256 \times 256 \times 16$	Conv2D + Conv2D + Drop + Conv2D	$\times$
Down-Sampler	$256 \times 256 \times 16$	$128 \times 128 \times 16$	Max Pooling	
$AdaFEx_2$	$128 \times 128 \times 16$	$128 \times 128 \times 32$	Conv2D + Conv2D + Drop + Conv2D	$\times$
Down-Sampler	$128 \times 128 \times 32$	$64 \times 64 \times 32$	Max Pooling	
$AdaFEx_3$	$64 \times 64 \times 32$	$64 \times 64 \times 64$	Conv2D + Conv2D + Drop + Conv2D	$\times$
Down-Sampler	$64 \times 64 \times 64$	$32 \times 32 \times 64$	Max Pooling	
$AdaFEx_4$	$32 \times 32 \times 64$	$32 \times 32 \times 128$	Conv2D + Conv2D + Drop + Conv2D	$\times$
Down-Sampler	$32 \times 32 \times 128$	$16 \times 16 \times 128$	Max Pooling	
$AdaFEx_5$	$16 \times 16 \times 128$	$16 \times 16 \times 256$	Conv2D + Conv2D + Drop + Conv2D	$\times$
Down-Sampler	$16 \times 16 \times 256$	$8 \times 8 \times 256$	Max Pooling	
Dynamic Contextual Attention Module				
DCA	$8 \times 8 \times 256$	$8 \times 8 \times 256$	FC+MHDCA+FC+Drop	$\times$
Multi-Scale Atrous Spatial Module				
MSAS	$8 \times 8 \times 256$	$8 \times 8 \times 512$	$d_f(4 \oplus 8 \oplus 12)$ + BN + Drop	$\times$
Progressive Decoder Block				
Up-Sampler	$8 \times 8 \times 512$	$16 \times 16 \times 256$	Conv2DTranspose	✓
$AdaFEx_5$	$16 \times 16 \times 256$	$16 \times 16 \times 256$	Conv2D + Conv2D + Drop + Conv2D	
Up-Sampler	$16 \times 16 \times 256$	$32 \times 32 \times 128$	Conv2DTranspose	✓
$AdaFEx_4$	$32 \times 32 \times 128$	$32 \times 32 \times 128$	Conv2D + Conv2D + Drop + Conv2D	
Up-Sampler	$32 \times 32 \times 128$	$64 \times 64 \times 64$	Conv2DTranspose	✓
$AdaFEx_3$	$64 \times 64 \times 64$	$64 \times 64 \times 64$	Conv2D + Conv2D + Drop + Conv2D	
Up-Sampler	$64 \times 64 \times 64$	$128 \times 128 \times 32$	Conv2DTranspose	✓
$AdaFEx_2$	$128 \times 128 \times 32$	$128 \times 128 \times 32$	Conv2D + Conv2D + Drop + Conv2D	
Up-Sampler	$128 \times 128 \times 32$	$256 \times 256 \times 16$	Conv2DTranspose	✓
$AdaFEx_1$	$256 \times 256 \times 16$	$256 \times 256 \times 1$	Conv2D + Conv2D + Drop + Conv2D	

CCR included: ✓; CCR not-included: $\times$.

expressed in Equation (6.14):

$$\mathscr{L}_{\text{C}} = \omega_1 \cdot \mathscr{L}_{\text{DSC}} + \omega_2 \cdot \mathscr{L}_{BCE} \tag{6.14}$$

where ω_1 and ω_2 control the weights of the dice loss and BCE loss, respectively. In this study, we set ω_1 and ω_2 to 0.5 to ensure balanced contributions from both loss terms during training.

6.3 Experimental Settings

This section describes the datasets used to evaluate the proposed T-MPEDNet framework for liver and tumor segmentation. It provides details of the benchmark dataset and outlines the evaluation metrics used to assess model performance.

6.3.1 Benchmark Dataset

We evaluated the proposed T-MPEDNet framework using three publicly available datasets: the Liver Tumor Segmentation Benchmark (LiTS) [152], 3D Image Reconstruction for Comparison of Algorithm Database (3DIRCADb) [153], and the A Tumor and Liver Automatic Segmentation (ATLAS) dataset [154]. Figure 6.8 illustrates a selection of CT and MRI scans used for both quantitative and qualitative analysis. For a comprehensive evaluation, we randomly divided each dataset into 80% for training, 10% for validation, and 10% for independent testing, as detailed in Table 6.2. The following section provides a summary of the datasets.

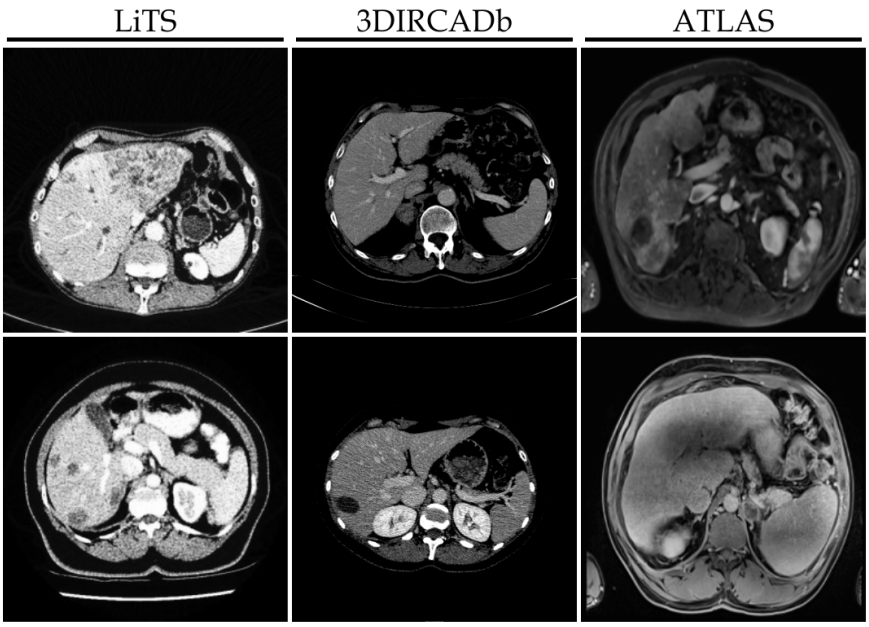


Figure 6.8: Sample dataset images for liver and tumor segmentation.

(a) LiTS: The Liver Tumor Segmentation Challenge (LiTS) dataset [152] is a popular benchmark for evaluating automated liver and tumor segmentation techniques in contrast-

enhanced abdominal CT scans. Launched in conjunction with ISBI 2017 and MICCAI 2017, LiTS [152] comprises 131 anonymized labeled CT scans from various clinical sites worldwide. In a single volume, the slice count exhibits significant variability, ranging from 75 to 841. The voxel spacing within the scans varies along the X, Y, and Z axes, with values spanning from 0.6 to 0.98 mm, 0.6 to 0.96 mm, and 0.4 to 0.45 mm, respectively. Table 6.2 presents a detailed overview of the dataset characteristics.

(b) 3DIRCADb: The 3D Image Reconstruction for Computer-Assisted Diagnosis Database (3DIRCADb) dataset [153] features 20 anonymous CT volumes (10 male, 10 female) acquired across European hospitals from different patients. These volumes were manually annotated by expert radiologists to highlight various structures of interest. 75 percent of patients possess tumors with varying sizes, shapes, and locations. The remaining 25 percent comprises non-tumor reference cases, essential for assessing the segmentation's generalizability across diverse clinical techniques. Encompassing a wide range of tumor characteristics within the dataset allows for a comprehensive assessment of the segmentation technique's robustness and generalizability. The single volume of CT scan exhibits non-uniform voxel spacing, with values ranging from 0.56-0.87 mm in the X axis, 0.56 - 0.87 mm in the Y axis, and 1.60 - 4.05 mm in the Z axis. Table 6.2 summarizes the key characteristics of the dataset.

(b) ATLAS: A Tumor and Liver Automatic Segmentation (ATLAS) dataset [154] is a publicly available dataset in automatic liver and tumor segmentation, particularly for primary liver cancer. ATLAS [154] comprises contrast-enhanced magnetic resonance imaging (CE-MRI) scans from 90 patients with unresectable this form of liver cancer, each accompanied by expert-annotated segmentation masks for both liver and liver tumors. The dataset is organized into a training set of 60 cases (acquired between 2012 and 2020) and a test set of 30 cases (acquired between 2020 and 2023). Each 3D CE-MRI scan contains between 44 and 136 axial slices, covering the entire liver and tumor regions, with in-plane resolutions ranging from 0.68×0.68 mm^2 to 1.41×1.41 mm^2, while slice thicknesses vary between 2 mm and 4 mm. Key characteristics of the dataset are summarized in Table 6.2.

Table 6.2: An overview of benchmark datasets for liver and tumor segmentation.

Dataset	Train	Val	Test	Resolution	Slices	X	Y	Z
LiTS [152]	54065	5001	4573	512×512	75–841	0.60–0.98	0.60–0.98	0.45–0.50
3DIRCADb [153]	2258	282	283	512×512	74–260	0.56–0.87	0.56–0.87	1.60–4.05
ATLAS [154]	3795	475	474	256×256	44–136	0.68–1.41	0.68–1.41	2.00–4.00

Train/Val/Test denotes dataset splits, Slices: number of slices per volume, X, Y, Z: voxel spacing (mm) along each axis

6.3.2 Evaluation Metrics

To quantitatively evaluate the framework performance, we utilize four established metrics: Dice Similarity Coefficient (DSC), Intersection over Union (IoU)/Jaccard Index, Accuracy (Acc), and the 95th percentile Hausdorff distance (HD95) that assess the segmentation quality in terms of overlap, including correctness and boundary precision. Specifically, DSC measures the normalized overlap between predicted and ground-truth masks, calculated as twice the intersection divided by the sum of their areas. IoU calculates the ratio of the intersection to the union of the predicted and ground-truth masks, representing the extent of overlap relative to the total area covered by both. Accuracy represents the overall proportion of correctly classified pixels across the region of interest, including both foreground and background. Lastly, HD95 measures the distance between the boundaries of predicted and ground truth segmentations using the 95th percentile of distances, which provides robustness to small segmentation outliers and indicates boundary agreement (lower values are better). The mathematical formulations are as follows:

$$\text{DSC} = \frac{2 \times |\mathsf{P} \cap \mathsf{G}|}{|\mathsf{P}| + |\mathsf{G}|} \tag{6.15}$$

$$\text{IoU} = \frac{|\mathsf{P} \cap \mathsf{G}|}{|\mathsf{P}| + |\mathsf{G}| - |\mathsf{P} \cap \mathsf{G}|} \tag{6.16}$$

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.17}$$

$$\text{HD95} = \max \left\{ \underset{x \in \mathsf{P}}{HD_{95}} \min_{y \in \mathsf{G}} \|x - y\|, \ \underset{y \in \mathsf{G}}{HD_{95}} \min_{x \in \mathsf{P}} \|y - x\| \right\} \tag{6.18}$$

In the above equations, P represents the set of pixels in the predicted segmentation mask (or the set of boundary points for HD95 calculation), while G denotes the corresponding set for the ground truth mask (or its boundary points for HD95). The operator $|\cdot|$ signifies the

cardinality, indicating the number of pixels in a set. The symbols $\cap$ and $\cup$ represent the intersection (common pixels) and union of sets, respectively. For the Accuracy calculation, TP denotes True Positives ($|\mathsf{P} \cap \mathsf{G}|$), TN represents True Negatives (correctly classified background pixels), FP signifies False Positives ($|\mathsf{P}/\mathsf{G}|$), and FN indicates False Negatives ($|\mathsf{G}/\mathsf{P}|$), where $/$ is the set difference operator. In the HD95 formula, $||x - y||$ calculates the Euclidean distance between a boundary point $x \in \mathsf{P}$ and a boundary point $y \in \mathsf{G}$, $\min$ finds the minimum distance from a point in one set to the points in the other set, HD_{95} computes the 95th percentile of these minimum distances across all points in the source boundary set, and $\max$ takes the larger of the two directed 95th percentile Hausdorff distances.

6.4 Experimental Results

This section presents the experimental results of the proposed method. It includes quantitative evaluation against existing approaches, followed by qualitative analysis to visualize model performance. Ablation analysis is conducted to assess the contribution of individual components. Cross-dataset evaluation is also performed to examine generalization, along with model complexity analysis.

6.4.1 Quantitative Analysis

We compare the performance of various liver and tumor segmentation methods utilizing three publicly available datasets, LiTS [152], 3DIRCADb [153], and ATLAS [154]. To evaluate framework performance, we use the Dice Similarity Coefficient (DSC), Intersection over Union (IoU), Accuracy (Acc), and the 95th percentile Hausdorff Distance (HD95) to quantify overlap between predicted and ground-truth regions. For a structured comparison, we group existing state-of-the-art (SOTA) methods into three categories: Encoder-Decoder-based Segmentation, Attention-based Segmentation, and Multi-Scale Feature Extractor-based Segmentation.

(a) Performance on LiTS Dataset: Table 6.3 presents a quantitative comparison of T-MPEDNet with various state-of-the-art segmentation methods for liver and tumor segmentation on the LiTS [152] dataset. For both liver and tumor segmentation, MSED-Net [52]

Table 6.3: Quantitative analysis of T-MPEDNet framework with existing baselines on the LiTS dataset. **Bold**: best and underlined: second-best scores.

	Methods	Liver Segmentation				Tumor Segmentation			
		DSC	IoU	Acc	HD95	DSC	IoU	Acc	HD95
I	MSED-Net [53]	97.5	94.7	97.7	5.5	87.1	85.4	91.6	15.1
	EG-UNETR [54]	93.5	91.8	91.3	11.8	85.8	82.1	90.7	23.9
	MSCF-Net [155]	96.4	92.4	**98.2**	6.4	87.3	82.4	91.1	22.3
	DRAUNet [55]	96.9	91.8	95.0	12.7	86.6	79.9	88.9	16.7
	Slice-Fusion [56]	95.6	94.2	96.2	4.8	86.4	80.2	92.6	14.4
	AutoLT [57]	93.8	90.3	93.3	15.8	84.9	77.2	89.9	15.3
II	UgCmA-Net [61]	96.4	93.6	95.8	4.2	86.3	84.8	93.4	15.5
	SBCNet [62]	94.2	94.9	93.4	10.1	85.7	**86.1**	90.9	23.5
	TD-Net [63]	97.1	93.1	94.9	7.7	85.2	78.4	89.2	18.2
	AC-E Net [156]	94.6	91.7	96.1	14.8	84.5	75.1	86.1	24.8
	HFRU-Net [157]	92.3	92.2	94.4	5.8	83.4	82.4	90.2	21.7
	LiTS-Net [158]	93.6	93.5	96.4	7.4	82.1	77.0	87.2	22.3
III	MSML-AttUNet [70]	94.1	95.1	95.7	9.4	84.9	83.6	92.1	18.6
	RMAU-Net [71]	95.1	93.7	96.6	8.9	83.1	82.1	91.8	20.1
	FRA-UNet [74]	96.8	92.4	97.8	5.1	85.5	73.6	93.5	14.6
	DefED-Net [73]	97.0	94.8	96.2	5.9	88.1	79.5	88.3	19.1
	KiU-Net [75]	93.5	90.6	94.3	11.0	82.6	76.8	90.8	19.2
Ours	T-MPEDNet	**97.6**	**95.6**	**98.2**	**3.1**	**89.1**	85.7	**94.5**	**13.7**

I: Encoder Decoder-based Segmentation; II: Attention-based Segmentation; III: Multi-Scale Feature Extractor-based Segmentation.

and DefED-Net [73] deliver performance comparable to T-MPEDNet, each achieving the second-best DSC scores of 97.5% and 88.1%, with declines of 0.1% and 1.0%, respectively. In terms of IoU, MSML-AttUNet [70] ranks second for liver segmentation, showing a performance gap of 0.5%. For tumor segmentation, SBCNet [62] achieves the next-best IoU, lagging by 0.4%. Although FRA-UNet [74] surpasses other baselines in overall Acc, it still underperforms compared to T-MPEDNet, with a margin of 0.4% for liver and 1.0% for tumor segmentation. T-MPEDNet demonstrates precise boundary localization, achieving HD95 values of 3.1 and 13.7 for liver and tumor segmentation, respectively. These results highlight the effectiveness of its progressive encoder-decoder architecture in extracting discriminative features. The framework successfully captures fine-grained anatomical details and enhances representational capacity while maintaining spatial consistency by adaptively recalibrating fused channel-wise features through skip connections.

(b) Performance on 3DIRCADb Dataset: Table 6.4 summarizes the quantitative evaluation of the T-MPEDNet framework on the 3DIRCADb [153] dataset. T-MPEDNet outperforms all existing baselines in terms of Dice Similarity Coefficient (DSC), Intersection

Table 6.4: Quantitative analysis of T-MPEDNet framework with existing baselines on the 3DIRCADb dataset. **Bold**: best and underlined: second-best scores.

	Methods	Liver Segmentation				Tumor Segmentation			
		DSC	IoU	Acc	HD95	DSC	IoU	Acc	HD95
I	MSED-Net [53]	97.9	92.4	94.1	3.8	82.9	76.0	85.7	**16.2**
	EG-UNETR [54]	96.7	88.4	91.3	5.2	83.3	75.7	86.1	17.6
	MSCF-Net [155]	96.5	90.2	93.5	6.6	78.2	73.2	82.8	17.9
	DRAUNet [55]	97.2	92.3	94.2	6.2	77.4	69.9	83.4	22.7
	Slice-Fusion [56]	95.4	88.1	90.6	7.7	77.1	70.5	88.7	17.1
	AutoLT [57]	96.3	88.0	93.5	8.7	75.7	68.4	79.6	22.4
II	UgCmA-Net [61]	95.1	94.2	93.8	9.5	82.6	73.8	85.2	17.8
	SBCNet [62]	96.3	92.0	95.4	**3.6**	78.2	71.0	84.1	17.4
	TD-Net [63]	97.8	91.7	95.5	6.5	79.4	73.2	80.9	19.7
	AC-E Net [156]	97.4	90.7	92.1	7.0	76.5	68.3	82.9	16.9
	HFRU-Net [157]	97.9	88.6	93.1	10.1	79.2	73.1	86.8	17.0
	LiTS-Net [158]	96.7	86.9	94.3	6.4	80.1	72.5	81.4	17.2
III	MSML-AttUNet [70]	96.9	92.3	95.8	7.3	79.6	76.3	85.0	18.6
	RMAU-Net [71]	97.5	89.8	96.4	5.5	82.1	75.6	88.6	20.3
	FRA-UNet [74]	97.7	90.2	93.6	8.0	75.1	68.2	81.6	23.0
	DefED-Net [73]	96.8	91.2	95.5	5.0	74.2	71.0	81.2	22.8
	KiU-Net [75]	97.1	93.3	96.0	9.2	76.3	69.0	79.4	20.4
Ours	T-MPEDNet	**98.3**	**94.8**	**97.1**	3.7	**83.6**	**77.6**	**89.2**	**16.2**

I: Encoder Decoder-based Segmentation; II: Attention-based Segmentation; III: Multi-Scale Feature Extractor-based Segmentation.

over Union (IoU), and Accuracy (Acc) for both liver and tumor segmentation, demonstrating robust, consistent performance across metrics. While T-MPEDNet leads in most metrics, SBCNet [62] slightly surpasses T-MPEDNet's HD95 liver segmentation score of 3.7 by a negligible margin of 0.1, achieving the best HD95 for the liver. For tumor segmentation, T-MPEDNet achieves a strong HD95 score, with AC-E Net [156] falling short by 0.7, reflecting its precision in boundary identification. MSED-Net [52] and HFRU-Net [157] exhibit comparable DSC scores for liver segmentation but do not surpass T-MPEDNet. Although EG-UNETR [54] achieves strong performance in tumor segmentation, it falls short by 0.3% in DSC. UgCmA-Net [61] and MSML-AttUNet [70] surpass many existing baselines in IoU for both liver and tumor segmentation. However, T-MPEDNet achieves superior IoU, exceeding UgCmA-Net by 0.6% and MSML-AttUNet by 1.3%. RMAU-Net [71] and Slice-Fusion [56] achieve the second-best results among all existing methods for liver and tumor segmentation but fall short by 0.7% and 0.5%, respectively, compared to T-MPEDNet. T-MPEDNet's superior performance highlights the effectiveness of its core architectural components, particularly the transformer-inspired attention module. This module efficiently captures deep global

features and learns complex contextual relationships among spatial information within the samples. This capability enables the framework to achieve accurate, robust segmentation, surpassing methods that rely primarily on local features, thereby contributing to its state-of-the-art results.

Table 6.5: Quantitative analysis of T-MPEDNet framework with existing baselines on the ATLAS dataset. **Bold**: best and underlined: second-best scores.

	Methods	Liver Segmentation				Tumor Segmentation			
		DSC	IoU	Acc	HD95	DSC	IoU	Acc	HD95
I	MSED-Net [53]	90.4	89.0	94.3	11.8	84.9	80.1	89.9	17.8
	EG-UNETR [54]	87.8	86.2	90.2	15.8	82.1	79.5	90.1	19.5
	MSCF-Net [155]	89.3	85.7	91.3	10.2	82.9	82.4	88.4	21.8
	DRAUNet [55]	85.4	84.5	90.5	19.8	79.9	78.0	88.2	23.5
	Slice-Fusion [56]	86.1	85.3	89.8	17.3	85.1	79.5	87.1	18.6
	AutoLT [57]	84.9	83.7	87.0	21.5	80.9	78.3	86.9	24.1
II	UgCmA-Net [61]	89.0	88.0	92.2	9.4	83.6	80.7	89.3	17.6
	SBCNet [62]	90.5	89.3	94.2	8.7	82.5	82.1	89.2	**14.6**
	TD-Net [63]	87.7	86.4	92.4	16.9	82.1	79.7	87.0	18.7
	AC-E Net [156]	86.6	85.9	90.6	14.4	80.9	77.5	86.2	21.8
	HFRU-Net [157]	84.2	83.8	88.7	20.6	80.5	78.6	87.7	19.9
	LiTS-Net [158]	85.3	84.2	90.0	18.8	82.0	81.8	89.5	18.1
III	MSML-AttUNet [70]	91.7	88.9	93.4	**7.7**	84.1	81.3	88.6	18.5
	RMAU-Net [71]	89.3	87.6	92.5	11.2	81.8	79.5	90.2	20.2
	FRA-UNet [74]	86.2	87.9	90.8	15.6	80.8	78.2	87.2	18.8
	DefED-Net [73]	85.7	84.8	89.3	17.3	78.6	77.0	86.5	22.6
	KiU-Net [75]	87.1	86.5	89.9	12.5	81.5	80.5	89.4	19.2
Ours	T-MPEDNet	**92.3**	**90.4**	**95.7**	7.8	**85.3**	**83.7**	**91.3**	14.6

I: Encoder Decoder-based Segmentation; II: Attention-based Segmentation; III: Multi-Scale Feature Extractor-based Segmentation.

(c) Performance on ATLAS Dataset: Table 6.5 highlights the performance of T-MPEDNet on the ATLAS [154] dataset for liver and tumor segmentation compared to existing state-of-the-art methods. Among encoder-decoder-based segmentation methods, MSED-Net [52] outperforms all other baselines for liver segmentation but shows a decline of 1.4% in accuracy compared to T-MPEDNet. For tumor segmentation, MSCF-Net [155] and Slice-Fusion [56] achieve the second-best IoU and DSC but fall short of T-MPEDNet by 1.3% and 0.2%, respectively. Considering attention-based methods, UgCmA-Net [61] accurately segments tumor regions but shows a decline of 3.0 in HD95 compared to T-MPEDNet. Additionally, SBCNet [62] achieves an IoU of 89.3%, lagging by 1.1% compared to T-MPEDNet, and demonstrates similar performance to T-MPEDNet in tumor segmentation with an HD95 of 14.6. Among multi-scale feature extractor-based methods, MSML-AttUNet [70] and

RMAU-Net [71] deliver the second-best performance with 91.7% DSC and 90.2% Acc for liver and tumor segmentation, yet still trail behind T-MPEDNet. The superior performance of T-MPEDNet demonstrates its robust feature learning capabilities in boundary preservation to enhance the localization performance. This advantage becomes especially critical in challenging cases involving noisy scans and small tumors. Integrating transformer-aware dynamic contextual attention, multi-scale atrous spatial modules, and morphological boundary refinement enables T-MPEDNet to process complex spatial relationships and fine details, resulting in accurate segmentation even in challenging clinical scenarios. Figure 6.9 presents a visual comparison of T-MPEDNet with various existing baseline methods on the LiTS [152], 3DIRCADb [153], and ATLAS [154] datasets.

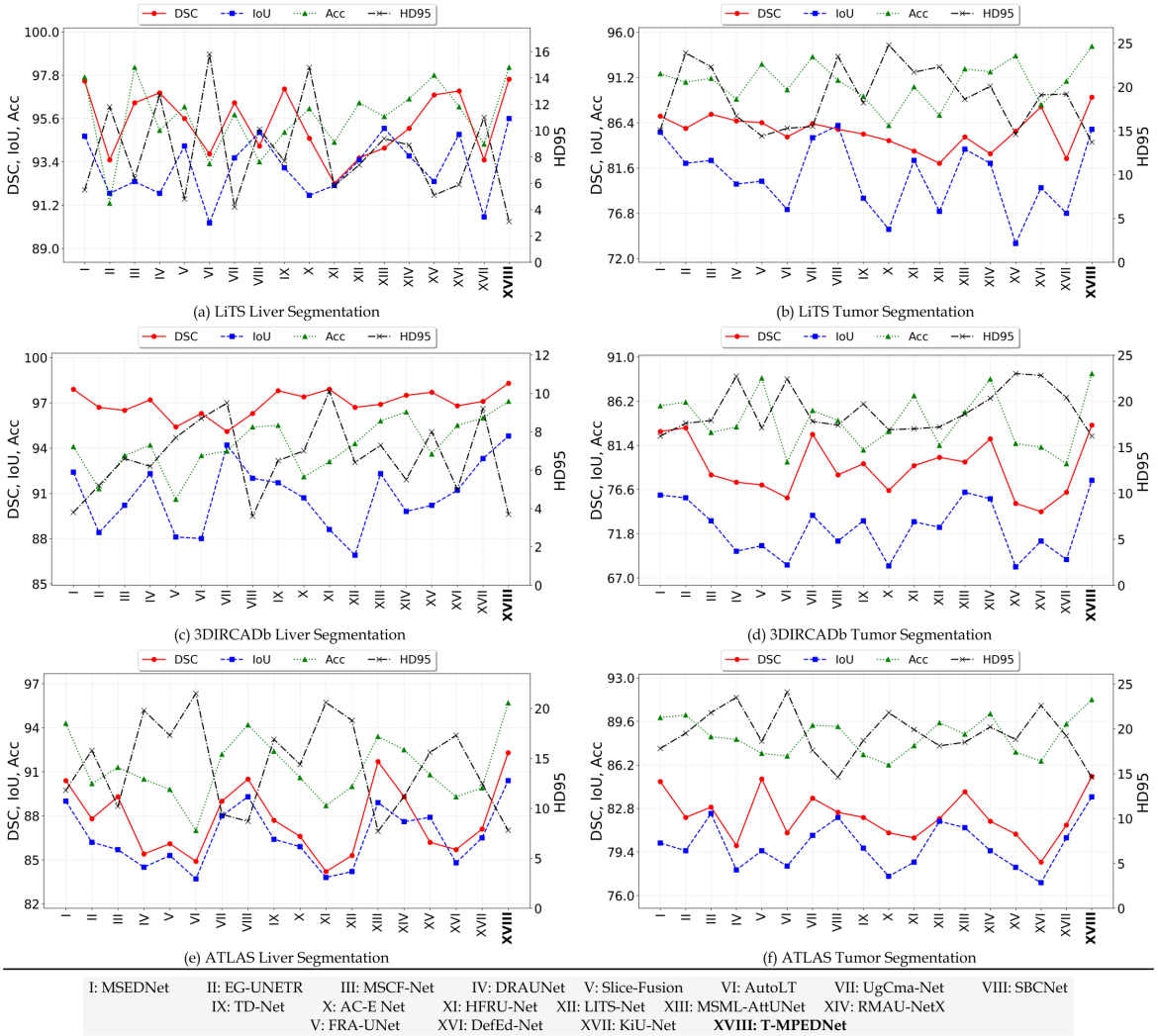


Figure 6.9: Comparative analysis of T-MPEDNet against existing baselines across three datasets using DSC, IoU, Accuracy, and HD95.

6.4.2 Qualitative Analysis

The proposed T-MPEDNet framework demonstrates superior performance in liver and tumor segmentation, particularly in challenging scenarios involving noisy scans and small tumors. In Figure 6.10, we present the top three best-performing methods in terms of Dice Similarity Coefficient within each category for the CT-based datasets LiTS [152] and 3DIRCADb [153], as well as the MRI-based ATLAS [154] dataset. In the visual analysis, gray masks represent liver segmentation, while white masks indicate tumor regions. T-MPEDNet accurately segments complex liver anatomy with small tumors and effectively suppresses non-region-of-interest (NROI) areas, including surrounding organs. On the LiTS [152] dataset, baseline methods such as AC-E Net [156] and MSML-AttUNet [70] struggle to capture liver and tumor regions due to noise interference. In contrast, T-MPEDNet accurately segments the liver and tumor while maintaining clear anatomical structures. This superior performance demonstrates the effectiveness of its integrated Adaptive Feature Extractor (AdaFEx) and Compressive Channel Recalibration (CCR) modules. These modules ensure the capture of fine-grained details and preserve spatial information that existing methods overlook. Similarly, on the 3DIRCADb [153] dataset, methods such as DRAUNet [55] and TD-Net [63] fail to segment small tumors in noisy conditions. T-MPEDNet generates precise segmentation maps by employing the Dynamic Contextual Attention (DCA) module, which fuses global and local features to suppress background noise. On the ATLAS [154] dataset, MRI-based segmentation presents additional challenges due to texture variations. Baseline methods such as KiU-Net [75] and SBCNet [62] generate fragmented masks. T-MPEDNet, however, adapts to texture variations, producing accurate segmentation masks of liver and tumors despite irregular texture. A Multi-Scale Atrous Spatial (MSAS) module contributes to this performance by aggregating multi-scale features to capture the fine details and variations of liver and tumor structures at different scales. Furthermore, T-MPEDNet enhances boundary precision in blurry scans with small tumors through its Morphological Boundary Refinement (MBR) module. This module employs morphological operations to refine binary masks, reduce boundary ambiguity, and improve overall segmentation quality. In summary, the qualitative analysis highlights the reliability of T-MPEDNet in generating precise liver and tumor segmentation masks across diverse

imaging modalities.

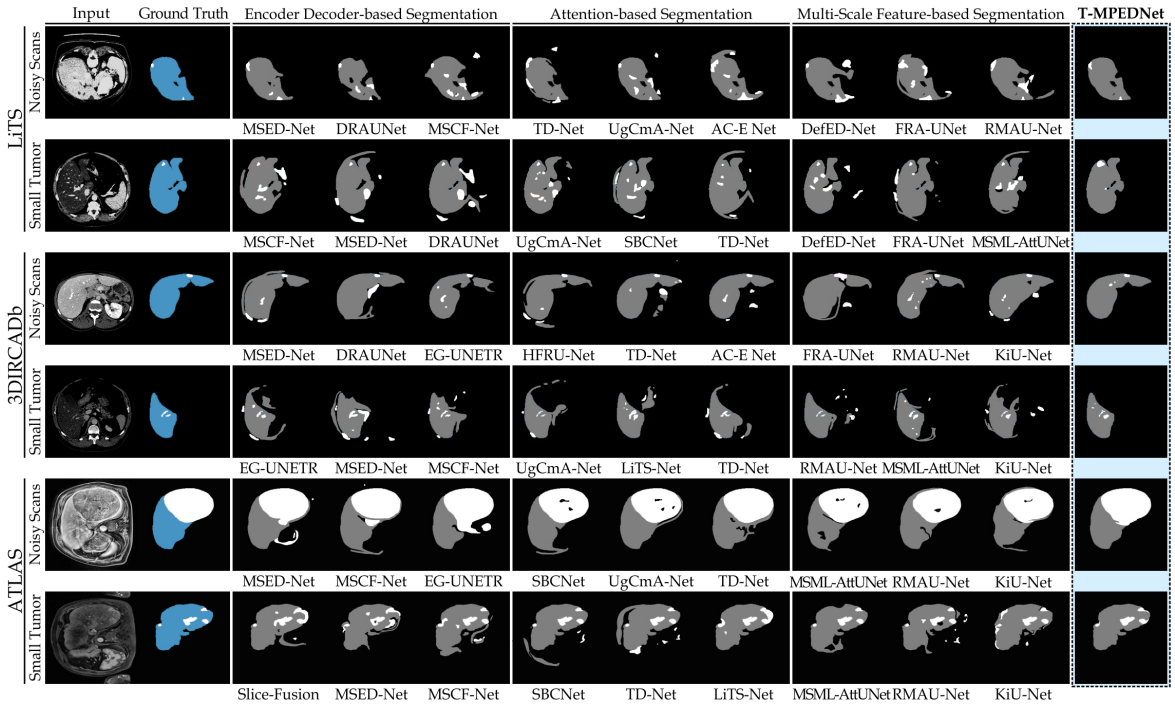


Figure 6.10: Qualitative analysis of T-MPEDNet framework.

Table 6.6: Ablation analysis of the T-MPEDNet framework with the contribution of each T-MPEDNet component on the LiTS dataset.

PED	CCR	DCA	MSAS	MBR	Liver Segmentation				Tumor Segmentation			
					DSC	IoU	Acc	HD95	DSC	IoU	Acc	HD95
✓	×	×	×	×	93.5	90.4	94.1	7.3	81.4	78.6	87.5	19.3
✓	✓	×	×	×	95.4	92.3	95.0	6.5	83.9	80.1	88.3	17.6
✓	✓	✓	×	×	96.7	93.8	96.5	5.1	86.2	82.5	91.6	15.5
✓	✓	✓	✓	×	97.4	94.1	97.7	4.2	88.6	84.1	93.2	14.8
✓	**✓**	**✓**	**✓**	**✓**	**97.6**	**95.6**	**98.2**	**3.1**	**89.1**	**85.7**	**94.5**	**13.7**

6.4.3 Ablation Analysis

The ablation study in Table 6.6 evaluates the individual contribution of each of the components in the T-MPEDNet architecture for liver and tumor segmentation on the LiTS [152] dataset. In this table, a checkmark (✓) indicates a component is included, while a cross mark (×) signifies its exclusion. The baseline configuration, which incorporates only Progressive Encoder-Decoder Block (PED) to yield the lowest segmentation performance with Dice Similarity Coefficients (DSC) of 93.5% for the liver and 81.4% for the tumor, and relatively high

HD95 values of 7.3 and 19.3, respectively. These results highlight limited boundary localization and segmentation accuracy in the absence of the additional module. As components are added incrementally, performance consistently improves. Introducing the Compressive Channel Recalibration Module (CCR) increases the liver DSC to 95.4% and reduces HD95 to 6.5, indicating enhanced spatial representation. Adding the Dynamic Contextual Attention Module (DCA) further refines contextual feature learning, improving both liver and tumor segmentation metrics. The Multi-Scale Atrous Spatial Module (MSAS) strengthens multi-scale feature extraction, raising liver segmentation accuracy to 97.4% and tumor DSC to 88.6%. Incorporating the Morphological Boundary Refinement Module (MBR) sharpens boundary precision. The complete T-MPEDNet framework which integrates all five modules (PED, CCR, DCA, MSAS, and MBR) achieves the overall performance with DSC values of 97.6% for liver and 89.1% for tumor segmentation and the lowest HD95 scores of 3.1 and 13.7. This ablation study demonstrates that each component plays a critical role in enhancing feature representation and boundary accuracy, validating the effectiveness of the T-MPEDNet design for segmenting complex anatomical structures.

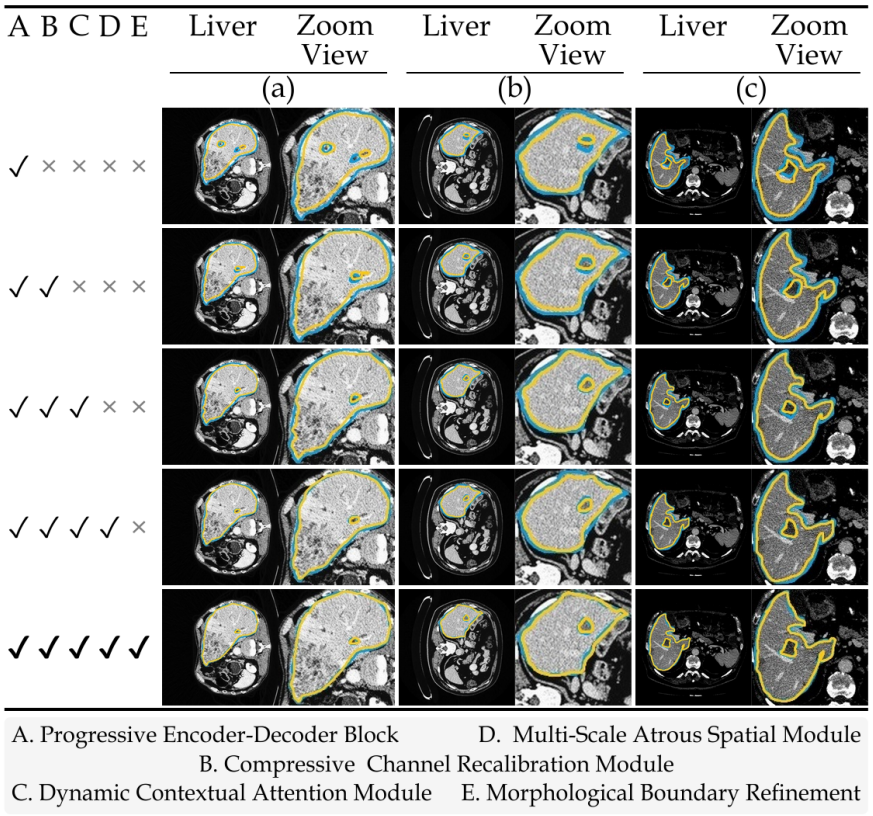


Figure 6.11: Visual results showing the contribution of key components in T-MPEDNet.

Figure 6.11 visualizes the segmentation result on three CT scans using various configurations of the T-MPEDNet framework. The first row depicts the T-MPEDNet without additional modules. The significant deviations between the predicted segmentations (yellow boundary) and the ground truth (blue boundary) indicate the framework's inability to capture discriminative features for accurate segmentation. As more modules are incorporated in subsequent rows, the predicted segmentations align closely with the ground truth. The progressive improvement is shown in Figure 6.12, suggesting that each module contributes to enhanced feature extraction. The integration of these learned features leads to superior segmentation performance.

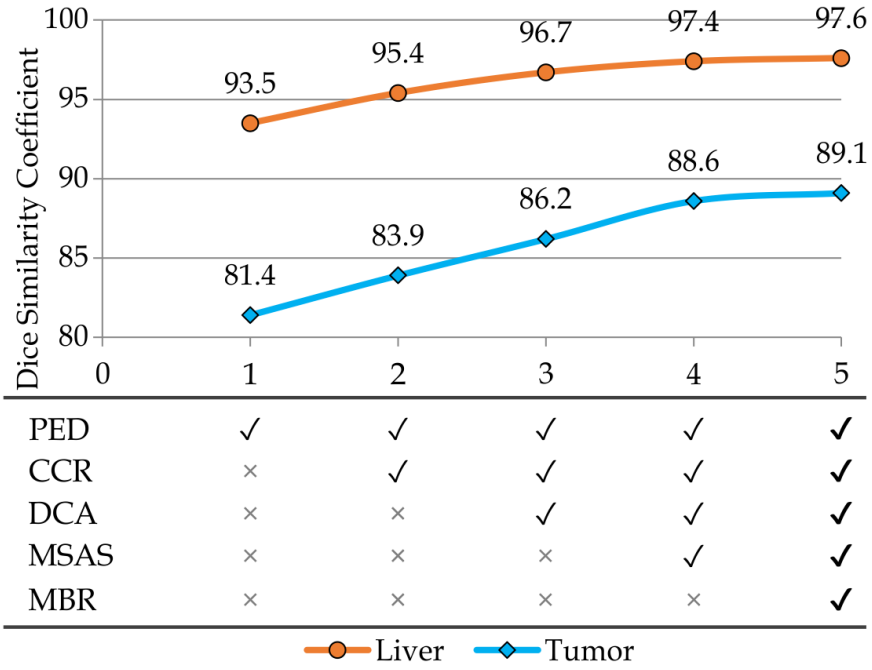


Figure 6.12: Impact analysis curves of T-MPEDNet components for liver and tumor segmentation on the LiTS dataset using Dice similarity coefficient.

6.4.4 Cross-Dataset Analysis

We evaluate the generalizability of T-MPEDNet in Table 6.7 across CT-based (LiTS [152], 3DIRCADb [153]) and MRI-based (ATLAS [154]) datasets, demonstrating its strengths and limitations in both cross-dataset and cross-modality liver and tumor segmentation scenarios. T-MPEDNet performs best when trained and tested on the same dataset, benefiting from consistent data distribution and imaging modality. This consistency enables the framework

Table 6.7: Cross-Dataset analysis of T-MPEDNet framework with different training and testing datasets.

Training	Testing											
	Category-1: Liver Segmentation											
	LiTS				**3DIRCADb**				**ATLAS**			
	DSC	**IoU**	**Acc**	**HD95**	**DSC**	**IoU**	**Acc**	**HD95**	**DSC**	**IoU**	**Acc**	**HD95**
LiTS	97.6	95.6	98.2	3.1	69.7	68.4	63.3	30.1	58.6	57.2	60.4	42.5
3DIRCADb	65.3	64.1	68.9	32.6	98.3	94.8	97.1	3.7	49.8	50.5	53.9	45.1
ATLAS	45.8	44.1	49.6	42.8	48.6	47.7	52.1	45.1	92.3	90.4	95.7	7.8
	Category-2: Tumor Segmentation											
	LiTS				**3DIRCADb**				**ATLAS**			
	DSC	**IoU**	**Acc**	**HD95**	**DSC**	**IoU**	**Acc**	**HD95**	**DSC**	**IoU**	**Acc**	**HD95**
LiTS	89.1	85.7	94.5	13.7	64.8	63.3	67.4	31.1	54.7	53.0	56.9	39.4
3DIRCADb	60.2	59.5	64.2	34.8	83.6	77.6	89.2	16.2	45.7	44.3	49.4	40.3
ATLAS	43.8	42.1	45.7	41.6	46.3	44.7	50.1	39.8	85.3	83.7	91.3	14.6

to learn dataset-specific features and deliver accurate predictions effectively. However, when we train on LiTS [152] and test on 3DIRCADb [153] for liver segmentation, T-MPEDNet achieves a DSC of 69.7% and an IoU of 68.4%, while maintaining competitive accuracy of 63.3% and an HD95 of 30.1. For tumor segmentation, the DSC drops to 64.8% and the IoU to 63.3%, indicating that even within the same imaging modality, a reduced volume of training data leads to decreased performance. These results reveal that the framework's generalizability is sensitive to the amount of training data, even within similar modalities. The significant challenge arises in cross-modality settings, where T-MPEDNet is trained on CT datasets and tested on MRI data, or vice versa: segmentation performance declines regardless of the number of training samples. When trained on ATLAS [154] an MRI-based dataset T-MPEDNet exhibits limited generalizability to CT-based datasets with lower metrics observed when tested on LiTS [152] such as a DSC of 45.8% for the liver and 43.8% for the tumor segmentation. Conversely, testing on the ATLAS [154] after training on LiTS [152] yields moderate performance with a DSC of 58.6% for the liver and 54.7% for tumors, suggesting limited cross-modality. However, a significant performance gap remains compared to within-modality evaluation. These findings highlight the challenge of achieving robust cross-modality generalization, as T-MPEDNet struggles to adapt to the distinct imaging characteristics of CT and MRI. Our findings emphasize the importance of maintaining modality

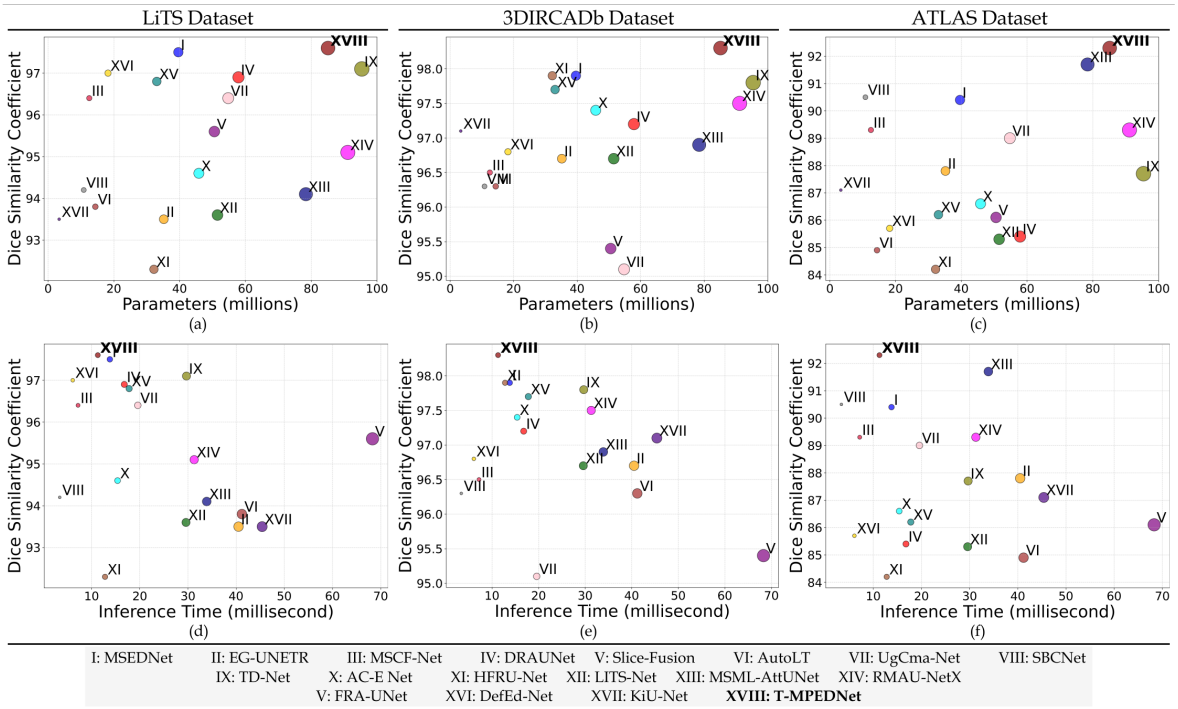


Figure 6.13: Comparative analysis of model parameters, inference time, and accuracy for the T-MPEDNet framework against existing baselines on benchmark datasets.

consistency and of using sufficient training data to develop a robust, generalized segmentation framework. The results also highlight the real-world applicability of our approach.

Table 6.8: Model Complexity Analysis of the T-MPEDNet framework comparing parameters, inference times, and memory with existing baselines.

	Methods	**Params (M)**	**Inf. Time (ms)**	**Memory (GB)**
I	MSED-Net [53]	39.62	13.8	381.4
	EG-UNETR [54]	35.19	40.5	338.7
	MSCF-Net [155]	12.53	7.2	120.6
	DRAUNet [55]	57.89	16.8	557.2
	Slice-Fusion [56]	50.57	68.3	486.8
	AutoLT [57]	14.38	41.2	125.7
II	UgCmA-Net [61]	54.78	19.6	527.3
	SBCNet [62]	10.85	3.4	104.4
	TD-Net [63]	95.41	29.7	918.3
	AC-E Net [156]	45.86	15.4	441.4
	HFRU-Net [157]	32.19	12.8	309.8
	LiTS-Net [158]	51.52	29.6	495.9
III	MSML-AttUNet [70]	78.41	33.9	754.7
	RMAU-Net [71]	91.15	31.3	877.3
	FRA-UNet [74]	33.06	17.8	318.2
	DefED-Net [73]	18.23	6.1	175.5
	KiU-Net [75]	3.38	45.4	32.5
Ours	T-MPEDNet	**85.13**	**11.3**	**819.4**

I: Encoder Decoder-based Segmentation; II: Attention-based Segmentation; III: Multi-Scale Feature Extractor-based Segmentation.

6.4.5 Model Complexity Analysis

We analyze the computational complexity of the proposed T-MPEDNet framework as presented in Table 6.8 by considering the number of parameters, inference time, and memory. T-MPEDNet maintains a reasonable parameter count of 85.13 million, comparable to other high-performing models such as MSML-AttUNet [70] and RMAU-Net [71], while achieving a significantly lower inference time of 11.3 milliseconds. Despite its size T-MPEDNet delivers this fast inference outperforming methods such as Slice-Fusion [56] at 68.3 milliseconds and MSML-AttUNet [70] at 33.9 milliseconds. However, the framework requires a significant amount of memory, using 819.4 MB. This memory usage aligns closely with other high-capacity networks such as TD-Net [63], which consumes 918.3 MB and RMAU-Net [71] that need 877.3 MB show extensive multi-path feature aggregation in encoder-decoder design. This computational investment yields a Dice Similarity Coefficient (DSC) of 97.6%, the highest among the compared methods, demonstrating a balance between resource demands, superior segmentation performance, and real-time inference. Figure 6.13 visually illustrates the relationship between the DSC, the number of parameters, and inference time for existing baselines in liver segmentation on the LiTS [152] dataset. Overall, T-MPEDNet achieves an optimal balance among segmentation performance, computational cost, and inference speed, making it a highly efficient and scalable solution for medical image segmentation tasks.

6.5 Conclusion

In this study, we present a novel **T**ransformer-aware **M**ultiscale **P**rogressive **E**ncoder-**D**ecoder **Net**work (T-MPEDNet) for automated liver and tumor segmentation from CT and MRI scans. T-MPEDNet harnesses the power of a Transformer-inspired dynamic attention mechanism coupled with a multi-scale feature extractor. This synergistic approach enables the capture of fine-grained features utilizing a larger receptive field while preserving global features. The dynamic attention mechanism learns contextual dependencies within hepatic tumor and non-tumor regions by dynamically adjusting the convolutional kernel, enabling the mapping of spatial relations. Subsequently, the multi-scale feature extractor enhances

the global feature map's contextual understanding of liver and tumor anatomy. Additionally, morphological operations are applied to refine liver and tumor boundaries, ensuring precise segmentation. Extensive evaluations on the LiTS, 3DIRCADb, and ATLAS datasets demonstrate the superiority of T-MPEDNet over 17 state-of-the-art methods. The proposed framework effectively automates the segmentation of intricate and small liver and tumor boundaries in CT and MRI scans, presenting valuable applications in clinical procedures such as treatment planning for hepatic diseases and tumors.

Chapter 7

A Multiscale Network with Spatial-enhanced Attention for Segmentation of Polyps in Colonoscopy

CONTENTS

7.1 Introduction . . . 156
7.2 Methodology . . . 157
7.2.1 Framework Overview . . . 157
7.2.2 Preprocessing . . . 157
7.2.3 Proposed Framework . . . 159
7.3 Experimental Settings . . . 167
7.3.1 Benchmark Dataset . . . 168
7.3.2 Evaluation Metrics . . . 168
7.4 Experimental Results . . . 168
7.4.1 Quantitative Analysis . . . 169
7.4.2 Qualitative Analysis . . . 171
7.4.3 Ablation Analysis . . . 172
7.4.4 Cross-Dataset Analysis . . . 175
7.4.5 Model Complexity Analysis . . . 177
7.5 Conclusion . . . 178

7.1 Introduction

This chapter investigates automated polyp segmentation and addresses the primary objective of the thesis: multi-scale feature fusion for advanced medical image analysis. The aim is to develop robust computational methods that accurately identify the colorectal polyps across a diverse range of shapes and sizes, including similar texture patterns within colonoscopy frames. Early detection and precise identification of polyps are important for preventing colorectal cancer and improving long-term patient outcomes. While colonoscopy is the gold standard for clinical screening, manual inspection remains a labor-intensive process and depends on the clinician's expertise. Automated segmentation faces challenges due to uneven illumination, reflections, low contrast, and blurred anatomical boundaries. Furthermore, polyps exhibit morphological variations where complex background pixels make segmentation challenging. Existing deep learning methods struggle to balance the fine local details with broader contextual relationships across multiple scales. These challenges motivate the need for a framework that integrates the multi-scale feature learning with spatially enhanced attention mechanisms for the pathological segmentation.

To address limitations, this chapter presents the Multiscale Network with Spatial-enhanced Attention (MNet-SAt) for polyp segmentation. The framework employs an encoder-decoder architecture with the Edge-Guided Feature Enrichment units to preserve boundary information and capture fine-grained patterns. A Multi-Scale Feature Aggregator extracts contextual features across variable receptive fields to handle diverse polyp sizes. A Spatial-Enhanced Attention unit learns the global dependencies and highlights the relevant region. In addition, a Channel-Enhanced Atrous Spatial Pyramid Pooling module refines multi-scale representations via atrous convolutions and channel interactions. The proposed framework demonstrates superior performance on challenging benchmark datasets.

The remainder of this chapter is organized as follows. Section 7.2 presents the proposed methodology, including edge-guided learning, multi-scale aggregation, and attention modules. Section 7.3 describes the benchmark datasets and evaluation metrics. Section 7.4 discusses the quantitative and qualitative findings, including ablation studies and model complexity analysis. Finally, Section 7.5 summarizes the major outcomes and clinical relevance of the proposed framework.

7.2 Methodology

This section presents the proposed framework for polyp segmentation. It describes the overall pipeline, including preprocessing and framework design. The framework comprises key components for learning meaningful representations, as detailed in the following subsections.

7.2.1 Framework Overview

In this section, we present a brief overview of our proposed MNet-SAt framework for polyp segmentation. Figure 7.1 illustrates the workflow of our methodology. For ease of understanding, we have broken down the MNet-SAt framework into three distinct stages, sequentially described in the following sections. First, the input image undergoes preprocessing, which includes resizing and contrast enhancement, followed by a series of data augmentation steps. Next, the encoder downsamples the input data while extracting edge-aware features using Edge-Guided Feature Enrichment (EGFE) units. The Hybrid Multi-Scale Attention (HMAtt) module, consisting of the Multi-Scale Feature Aggregator (MSFA), Spatial-Enhanced Attention (SEAt), and Channel-Enhanced Atrous Spatial Pyramid Pooling (CE-ASPP), extracts and refines multi-scale features. The decoder then restores the data to its original dimensions, utilizing skip connections. Finally, the aggregated features are processed to produce the segmentation probabilities. We compute the segmentation loss by combining Binary Cross-Entropy Loss and Dice Loss to optimize boundary accuracy and overall segmentation quality.

7.2.2 Preprocessing

Medical image quality is crucial for accurate diagnosis and treatment planning. However, colonoscopy images often exhibit variations in resolution, contrast, and illumination due to the challenges of capturing images within the complex anatomy of the human colon. The performance limits of automated polyp segmentation [159, 160]. To address the challenges, we employ a preprocessing pipeline that resizes images to a standardized size and enhances contrast to improve the visibility and feature extraction. We next apply data augmentation

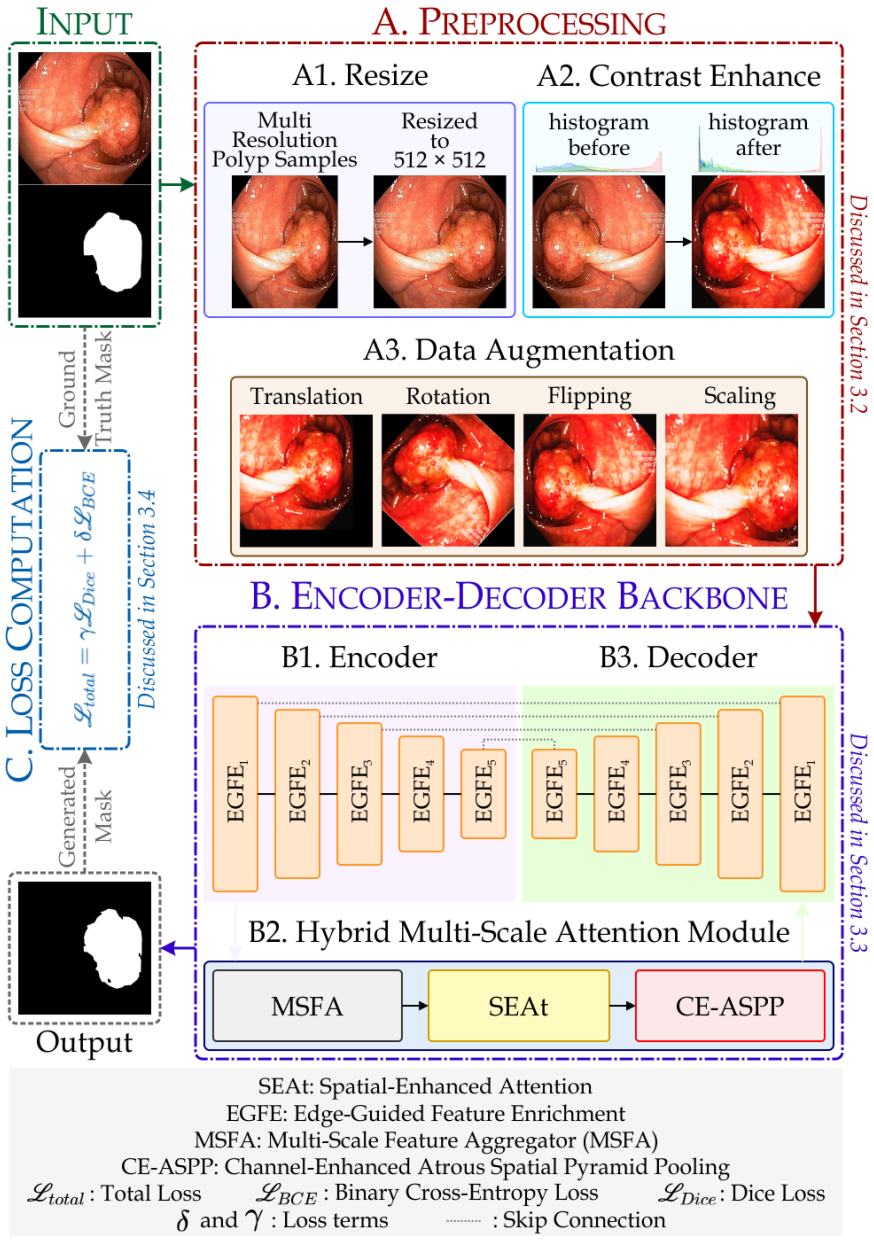


Figure 7.1: The workflow of our proposed MNet-SAt framework for polyp segmentation in colonoscopy images.

techniques to increase the data diversity and prevent overfitting.

(a) Resizing The preprocessing pipeline begins by resizing all polyp images to 512×512 pixels using bicubic interpolation [15]. This adjustment addresses the multi-resolution inconsistency present in datasets such as Kvasir-SEG [161], which spans 332×487 to 1920×1072 pixels, and CVC-ClinicDB [162], which has a fixed resolution of 384×288 pixels. Resizing to a fixed resolution ensures that each sample contributes equally to the training process,

avoiding bias introduced by image sizes. Using the original high resolutions demands significant memory and processing power, making efficient training impractical on standard hardware. The choice of 512×512 pixels also aligns with practice in medical image segmentation [163], balancing detail and computational efficiency for polyp detection.

(b) Contrast Enhancement Clinicians face difficulties in accurately identifying potential lesion areas due to the low contrast between polyps and the surrounding healthy mucosa in colonoscopy images [164]. This low contrast also results in poor feature extraction as polyp and surrounding tissue appear similar [165]. To address this, we apply Contrast Limited Adaptive Histogram Equalization (CLAHE) [14] to amplify contrast and polyp visibility. CLAHE divides the image into small regions and applies histogram equalization to each, improving the visibility of hidden features. Using CLAHE enhances polyp visibility, distinguishing them more clearly from the background.

(c) Data Augmentation Data augmentation is essential for improving automated image segmentation performance and class imbalance with deep learning networks [166]. To achieve robust and generalizable performance, a deep learning model requires data to learn rich representations and adapt to variations in data distribution. Data augmentation increases the size of the training dataset by applying a series of transformations to existing images. In this study, we augment the data by applying transformations such as random translations (up to 10 pixels in each direction), rotations (turning the image by a random angle between -15 and +15 degrees), flipping (horizontal and vertical), and scaling (with a scale factor of 0.8 to 1.2) to each original image. This increases the diversity of the training data and mitigates the class imbalance, leading to improved segmentation performance.

7.2.3 Proposed Framework

The MNet-SAt framework, illustrated in Figure 7.2, adopts a U-shaped architecture comprising an encoder and a decoder, each with five stages and Edge-Guided Feature Enrichment (EGFE) units. The encoder progressively downsamples the input data, with each stage i ($i \in [1, 2 \dots 5]$) reducing the feature map resolution to $\frac{h}{2^{i-1}} \times \frac{w}{2^{i-1}}$. The EGFE units integrate edge-aware feature extraction and enhance boundary-aware feature representation, generating feature maps rich in edge information. The features from the final encoder stage

are fed into the Hybrid Multi-Scale Attention (HMAtt) module, which facilitates communication between the encoder and decoder by employing multi-scale receptive fields to enrich feature representations and enhance the framework's adaptability. The HMAtt module enables the framework to focus on critical regions, achieving precise anomaly localization and segmentation.

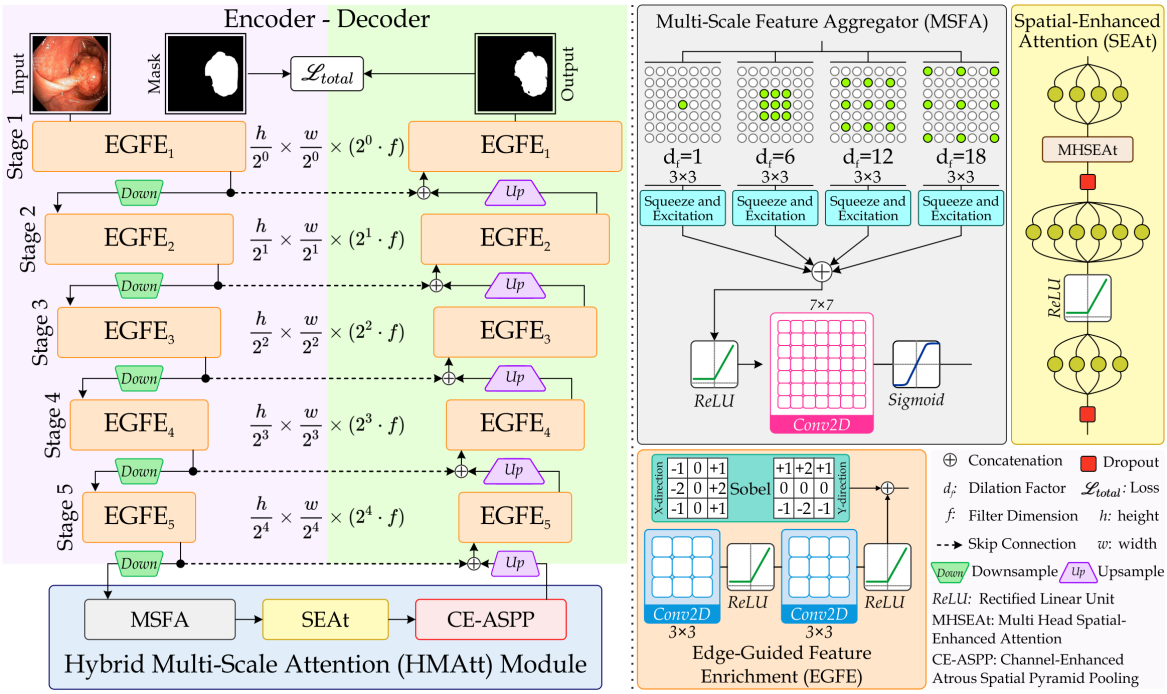


Figure 7.2: Proposed architecture of Multiscale Network with Spatial-enhanced Attention (MNet-SAt) for polyp segmentation in colonoscopy images.

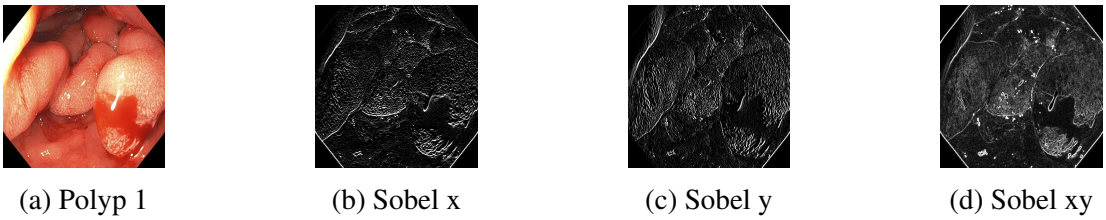
The decoder path begins with the HMAtt module's output and uses skip connections to incorporate features from the encoder's EGFE units to generate segmentation masks. Each decoder stage upsamples its output by a factor of 2×2 and concatenates it with the corresponding encoder feature map, effectively bridging the encoder and decoder pathways. These skip connections mitigate information loss from downsampling and help the decoder produce finer details in the final segmentation mask. Finally, a convolutional layer with a sigmoid activation function processes the aggregated features from the decoder's last stage to compute segmentation probabilities. Table 7.1 details the feature map dimensions for each stage in the encoder and decoder.

Table 7.1: Feature map dimensions across different stages of the MNet-SAt architecture.

Stage	Encoder Blocks	Encoder Output	Skip	Decoder Blocks	Decoder Output
Stage 1	$EGFE_1$ Down	$512 \times 512 \times 64$ $256 \times 256 \times 64$	$512 \times 512 \times 64$	$EGFE_1$ Up	$512 \times 512 \times 3$ $512 \times 512 \times 64$
Stage 2	$EGFE_2$ Down	$256 \times 256 \times 128$ $128 \times 128 \times 128$	$256 \times 256 \times 128$	$EGFE_2$ Up	$256 \times 256 \times 64$ $256 \times 256 \times 128$
Stage 3	$EGFE_3$ Down	$128 \times 128 \times 256$ $64 \times 64 \times 256$	$128 \times 128 \times 256$	$EGFE_3$ Up	$128 \times 128 \times 128$ $128 \times 128 \times 256$
Stage 4	$EGFE_4$ Down	$64 \times 64 \times 512$ $32 \times 32 \times 512$	$64 \times 64 \times 512$	$EGFE_4$ Up	$64 \times 64 \times 256$ $64 \times 64 \times 512$
Stage 5	$EGFE_5$ Down	$32 \times 32 \times 1024$ $16 \times 16 \times 1024$	$32 \times 32 \times 1024$	$EGFE_5$ Up	$32 \times 32 \times 512$ $32 \times 32 \times 1024$

7.2.3.1 Edge-Guided Feature Enrichment Unit

The scarcity of annotated data in clinical practice often limits the ability of existing polyp segmentation methods to capture polyp boundaries accurately. To address this, we propose the Edge-Guided Feature Enrichment (EGFE) unit, designed to enhance boundary quality in segmentation masks. To improve boundary detail, each EGFE unit applies the 2D Sobel operator [62] before the convolution operation, preserving essential edge information. The EGFE unit effectively maintains edge details, mitigates weak boundaries, bridges the semantic gap between low-level and high-level features, and directly captures spatial features of polyps from input images.



(a) Polyp 1 (b) Sobel x (c) Sobel y (d) Sobel xy

(e) Polyp 2 (f) Sobel x (g) Sobel y (h) Sobel xy

Figure 7.3: Edge maps of the Sobel operator in x-y directions

The Edge-Guided Feature Enrichment (EGFE) unit actively enhances feature representation by integrating edge knowledge into the set of features. It applies a Sobel operator to the input feature set, generating separate gradient magnitude maps for horizontal ($\mathcal{S}_x$) and

vertical ($\mathcal{S}_y$) directions (Figure 7.3). As illustrated in Figure 7.3, the Sobel operation in a single direction captures limited information about the polyp. However, combining these maps (Figure 7.3d and Figure 7.3h) effectively captures polyp boundaries and enriches learned features. The final gradient magnitude map, produced using the Euclidean norm as defined in Equation (7.1), highlights regions with significant intensity changes at polyp boundaries.

$$\mathcal{S} = \sqrt{\mathcal{S}_x^2 + \mathcal{S}_y^2} \tag{7.1}$$

Here, the output of the prior edge detector is denoted by $\mathcal{S}$, representing the gradient magnitude map. This map is then element-wise added to the output of the convolution layers for feature fusion. Each EGFE unit consists of two convolutional layers with f filters of kernel size 3×3 and stride 2. The number of filters, f, progressively increases at each stage according to the formula $64 \cdot 2^{i-1}$, where i represents the stage number. ReLU activation functions are applied after each convolution to induce non-linearity. This allows the framework to learn and extract complex, hidden relationships within the data. Equation (7.2) formalizes this process:

$$\mathcal{F}_{egfe}^{i} = \mathcal{S} \oplus \left(\underset{ReLU}{\alpha} \left(Conv_2 \left(\underset{ReLU}{\alpha} (Conv_1(\mathcal{F}_{egfe}^{i-1})) \right) \right) \right) \tag{7.2}$$

where $\mathcal{S}$, $\oplus$, $\underset{ReLU}{\alpha}$, and $Conv_j$ denote 2D Sobel operation, concatenation operation, ReLU activation function, the number of 2D convolutions with $j = \{1, 2\}$ respectively, and $\mathcal{F}_{egfe}^{i}$ represents the edge-guided features learned at the i-th layer. This element-wise addition facilitates the fusion of edge information extracted by the Sobel operator with the spatial features learned by the convolutional layers. For the first stage $(i = 1)$, the edge-guided feature vector is derived directly from the input image. We apply the Sobel operator to the input image to generate the initial edge maps in the horizontal and vertical directions. These edge maps are then concatenated with the original input image to form the edge-guided feature vector for the first stage. This process effectively incorporates edge information from the raw input image into the encoder's initial stage. This fusion enables the framework to focus on and refine the representation of polyps, ultimately contributing to enhanced segmentation performance.

7.2.3.2 Hybrid Multi-Scale Attention Module

The Hybrid Multi-Scale Attention (HMAtt) module extracts and refines multi-scale features. It consists of three sequentially arranged components: the Multi-Scale Feature Aggregator (MSFA), Spatial-Enhanced Attention (SEAt), and Channel-Enhanced Atrous Spatial Pyramid Pooling (CE-ASPP), which are described in detail below.

(a) Multi-Scale Feature Aggregator Unit: Colonoscopy image analysis often assumes that specific features reside at specific scales [167]. However, existing multi-scale approaches primarily focus on feature extraction [167], neglecting aggregated channel- and spatial-level dimensions. To mitigate this, we introduce a Multi-Scale Feature Aggregator (MSFA) Unit that integrates Squeeze-and-Excitation (SE) blocks [65] after multi-scale feature extraction and aggregates them. This emphasis on salient regions enhances the polyp's feature representation.

The MSFA unit receives the encoder's final EGFE unit output, $\mathcal{F}^{5}_{egfe}$, as input. It then processes this feature map through multiple convolutions with varying dilation rates, as defined mathematically in Equation (7.3).

$$\mathcal{F}_{k,d_r}[l] = \sum_{\kappa=1}^{k} \mathcal{F}^{5}_{egfe}[l + d_r * \kappa] * \varphi_{\kappa} \tag{7.3}$$

Here, we define l as the location on the feature map, $d_r \in \{1, 6, 12, 18\}$ as the dilation rate, φ_{κ} as the $\kappa-$th parameter of the convolution filter, and k signifies the filter size. Equation (7.3) illustrates varying dilation rates (d_r), enabling the acquisition of diverse receptive fields, serving as multiple scales to learn spatial components. Afterward, we recalibrate these multi-scale features through the SE block to prioritize the relevant multi-scale features. By concatenating these multi-scale aggregated features with distinct receptive fields, MSFA generates high-level contextual features. Equation (7.4a) and Equation (7.4b) provide the mathematical formulation of MSFA.

$$\mathcal{F}_{\oplus} = Concat\Big(SE\big(\mathcal{F}_{3,i}(\mathcal{F}^{5}_{egfe})\big)\Big), i \in (1, 6, 12, 18) \tag{7.4a}$$

$$\mathcal{F}_{msfa} = \underset{softmax}{\alpha}\Big(\mathcal{F}_{7,1}\big(\underset{ReLU}{\alpha}(\mathcal{F}_{\oplus})\big)\Big) \tag{7.4b}$$

Here, $\mathcal{F}_{3,i}$ represents the multi-scale features captured at different dilation rates $(1, 6, 12, 18)$ enhanced by SE block, $\mathcal{F}_{\oplus}$ denotes the concatenated multi-scale features which are further processed by $\underset{ReLU}{\alpha}$, ReLU activation function. Finally, all concatenated features are processed through convolution with 7×7 kernel to generate final multi-scale aggregated features, denoted by $\mathcal{F}_{msfa}$.

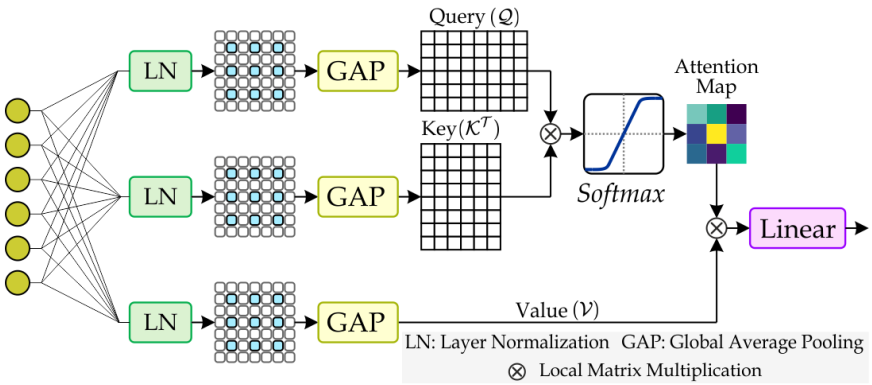


Figure 7.4: Proposed architecture of Multi-Head Spatial-Enhanced Attention (MHSEAt).

(b) Spatial-Enhanced Attention Unit: As illustrated in Figure 7.2, the Spatial-Enhanced Attention (SEAt) unit, a core component of the HMAtt module, captures spatial-aware global dependencies within the multi-scale aggregated features $\mathcal{F}_{msfa}$ derived from the MSFA unit. The SEAt comprises a Fully Connected (FC) layer, MHSEAt, feed-forward layers, and a dropout layer. Due to the pivotal role of the attention mechanism within SEAt, we provide a detailed introduction to MHSEAt below.

Multi-Head Spatial-Enhanced Attention (MHSEAt): The Multi-Head Spatial-Enhanced Attention (MHSEAt) performs the self-attention operation along the spatial dimension, as shown in Figure 7.4. Initially, the three identical copies of feature maps are normalized through Layer Normalization (LN), followed by convolution operation $(Conv)$ with 3×3 kernel having a dilation rate of 2. Next, the spatial-enhanced features are aggregated using Global Average Pooling (GAP) to obtain the Query $(\mathcal{Q})$, Key $(\mathcal{K})$, and Values $(\mathcal{V})$ represen-

tations, as defined in Equation (7.5a), Equation (7.5b) and Equation (7.5c).

$$\mathcal{Q} = LN(\underset{3\times3}{Conv}(GAP(X)))W_{\mathcal{Q}} \tag{7.5a}$$

$$\mathcal{K} = LN(\underset{3\times3}{Conv}(GAP(X)))W_{\mathcal{K}} \tag{7.5b}$$

$$\mathcal{V} = LN(\underset{3\times3}{Conv}(GAP(X)))W_{\mathcal{V}} \tag{7.5c}$$

Here, X represents the input feature map, $W_{\mathcal{Q}}, W_{\mathcal{K}}, W_{\mathcal{V}} \in \mathbb{R}^{c\times c}$ are the projection matrices. The resulting spatial attention is computed using the softmax-scaled dot product of $\mathcal{Q}$ and $\mathcal{K}$, which is then multiplied by $\mathcal{V}$ to form the spatial-enhanced attention output X_{SEAt}, as shown in Equation (7.6).

$$X_{SEAt} = \underset{softmax}{\alpha}\left(\frac{\mathcal{Q}\cdot\mathcal{K}^T}{\sqrt{d_{\mathcal{K}}}}\right)\cdot\mathcal{V} \tag{7.6}$$

X_{SEAt} is the feature map after linear spatial reduction multi-head spatial-enhanced attention. $\underset{softmax}{\alpha}$ and $\sqrt{d_{\mathcal{K}}}$ denotes the softmax function and dimension of the key. This attention mechanism enables the model to selectively focus on spatial features, thereby enhancing the framework's adaptability and improving feature refinement. The SEAt unit follows the MH-SEAt layer with two stacked Fully Connected (FC) layers. Each layer uses the ReLU activation (denoted by $\underset{ReLU}{\alpha}$) to incorporate non-linearity, allowing the network to model complex data relationships. Formally, $\mathcal{F}_{SEAt} = FC_2\Big(\underset{ReLU}{\sigma}\big(FC_1(\mathcal{F}_{MHSEAt})\big)\Big)$, where $\mathcal{F}_{MHSEAt}$ represents the spatial-enhanced attention features. This process effectively captures local and global dependencies while preserving spatial information in the final output, $\mathcal{F}_{SEAt}$.

(c) Channel-Enhanced Atrous Spatial Pyramid Pooling: To improve polyp representation by resampling features at different scales, we propose Channel-Enhanced Atrous Spatial Pyramid Pooling (CE-ASPP). Unlike the standard Atrous Spatial Pyramid Pooling (ASPP) [168], which uses dilated convolutions for multiscale extraction and standard convolutions for fusion, CE-ASPP incorporates a three-step refinement to adaptively recalibrate the importance of various feature channels.

Within CE-ASPP (refer to Figure 7.5), we first compress the input feature channels by a factor of four. This compression is then reversed before the final output is generated (refer

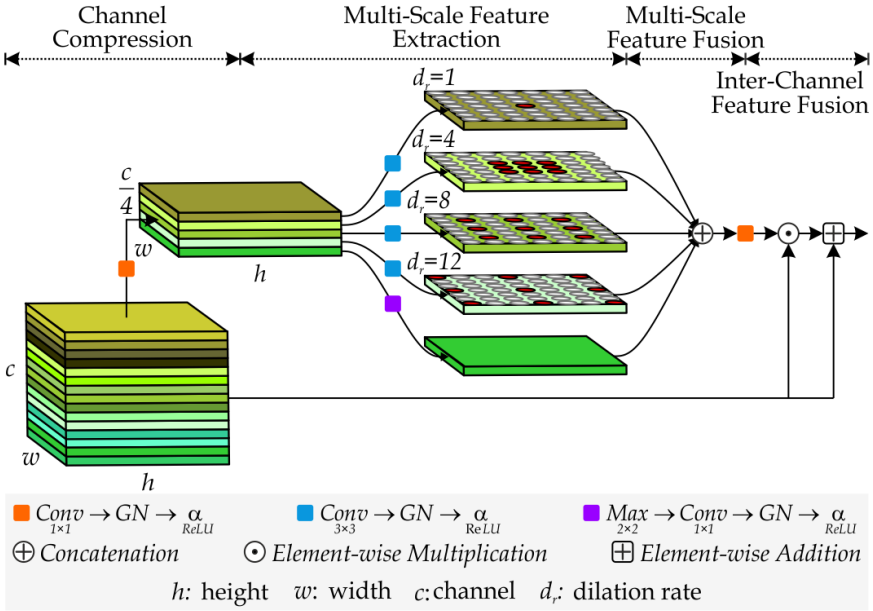


Figure 7.5: Proposed architecture of Channel-Enhanced Atrous Spatial Pyramid Pooling (CE-ASPP).

to Equation (7.7)).

$$\mathcal{F}_c = \underset{ReLU}{\alpha}\Big(GN\big(\underset{1\times1}{Conv}(\mathcal{F}_{SEAt})\big)\Big) \tag{7.7}$$

Here, $\underset{ReLU}{\alpha}$ denotes the ReLU activation function, $\mathcal{F}_c$ is the channel-compressed features, and GN is Group Normalization [169], chosen for its accuracy and stability across batch sizes.

Secondly, $\mathcal{F}_c$ undergoes five parallel routing convolutions: four with 3×3 depth-wise atrous convolutions (dilation rates 1, 4, 8, and 12) and one with max-pooling followed by 1×1 point-wise convolution for salient feature extraction and channel dimension adjustment. This process is expressed below in Equation (7.8a) and Equation (7.8b).

$$\mathcal{F}_c^{d_r} = \underset{ReLU}{\alpha}\Big(GN\big(\underset{3\times3}{Conv}(\mathcal{F}_c)\big)\Big), d_r \in [1,4,8,12] \tag{7.8a}$$

$$\mathcal{F}_c^{Max} = \underset{ReLU}{\alpha}\Big(GN\big(\underset{1\times1}{Conv}(\underset{2\times2}{Max}(\mathcal{F}_c))\big)\Big) \tag{7.8b}$$

where $\mathcal{F}_c^{d_r}$ and $\mathcal{F}_c^{Max}$ denote the compressed multi-scale features obtained from atrous convolutions (varying dilation rates) and max-pooling, respectively. Lastly, Multi-scale and inter-channel fusion are then applied to generate $\mathcal{F}_{ce-aspp}$ as expressed in Equation (7.9)

and Equation (7.10).

$$\mathcal{F}_{\oplus} = Concat(\mathcal{F}_c^1, \mathcal{F}_c^4, \mathcal{F}_c^8, \mathcal{F}_c^{12}, \mathcal{F}_c^{Max}) \tag{7.9}$$

$$\mathcal{F}_{ce-aspp} = \mathcal{F}_{SEAt} \odot \underset{ReLU}{\alpha}\Big(GN\big(\underset{1\times1}{Conv}(\mathcal{F}_{\oplus})\big)\Big) \boxplus \mathcal{F}_{SEAt} \tag{7.10}$$

Here, $\mathcal{F}_{\oplus}$ denotes the concatenated multi-scale features. $\boxplus$ and $\odot$ represent element-wise multiplication and addition, respectively. The CE-ASPP module generates the final feature, $\mathcal{F}_{ce-aspp}$ is fed into the decoder.

7.2.3.3 Loss Computation

Our framework tackles the class imbalance commonly encountered in medical image segmentation by employing a total loss function inspired by [170], as defined in Equation (7.11). $\mathscr{L}_{total}$, defined as a weighted sum of Dice loss ($\mathscr{L}_{Dice}$) and binary cross-entropy loss ($\mathscr{L}_{BCE}$), as expressed in Equation (7.12a) and Equation (7.12b).

$$\mathscr{L}_{total} = \gamma \cdot \mathscr{L}_{Dice} + \delta \cdot \mathscr{L}_{BCE} \tag{7.11}$$

Dice loss, which maximizes the Dice score and Intersection over Union (IoU), is further utilized for parameter optimization. The mathematical expressions for these loss components are provided in Equation (7.12a) and Equation (7.12b).

$$\mathscr{L}_{Dice} = 1 - \frac{2\sum(y_g \cap y_p) + \epsilon}{\sum(y_g + y_p) + \epsilon} \tag{7.12a}$$

$$\mathscr{L}_{BCE} = -[y_g \cdot \log(y_p) + (1 - y_g) \cdot \log(1 - y_p)] \tag{7.12b}$$

Here y_g and y_p represent the ground-truth and predicted labels and ϵ is a smoothing factor. To ensure balanced contributions from both loss terms we set the weights γ and δ to 0.5.

7.3 Experimental Settings

This section describes the datasets used to evaluate the proposed MNet-SAt framework for polyp segmentation. It provides details of the benchmark dataset and outlines the evaluation metrics used to assess model performance.

7.3.1 Benchmark Dataset

(a) Kvasir-SEG [161]: Provided by Vestre Viken Health Trust (VV) Norway it consists of 1000 gastrointestinal polyp image collected with endoscopic equipment and label by experienced gastroenterologists. Image resolutions range between 332×487 pixels to 1920×1072 pixels.

(b) CVC-ClinicDB [162]: A popular biomedical dataset comprise 612 image (384×288 pixels) extract from 31 colonoscopy video collected at the Hospital Clinic in Barcelona Spain. This dataset served as training data for Automatic Polyp Detection in the MICCAI 2015 Sub-Challenge.

7.3.2 Evaluation Metrics

To evaluate the performance of the MNet-SAt framework we employed eight commonly used metrics including Dice Similarity Coefficient (DSC), Intersection over Union (IoU), Precision (Pre), Recall (Rec), Accuracy (Acc), F2 Score (F2), Dice Loss (DL), and the 95th percentile Hausdorff distance (HD95). DSC measure similarity between predicted segmentation and the ground truth while IoU quantifies overlap. Precision evaluate proportion of correctly identified polyp pixels and Recall assess proportion of actual polyp regions correctly identified. Accuracy represent overall correctness of the predictions and F2 provide balanced measure of Recall and Precision. DL quantifies the dissimilarity between the predicted and ground truth segmentations while HD95 calculates the spatial distance between their boundary focusing on the 95th percentile to reduce the impact of outliers. Superior polyp segmentation performance is characterized by maximizing metrics such as DSC, IoU, Precision, Recall, Accuracy and F2 while minimizing Dice Loss and HD95. These metrics collectively evaluate the overlap boundary precision and overall accuracy of the predicted polyp regions compared to the ground truth.

7.4 Experimental Results

This section presents the experimental result of the proposed method. It include quantitative evaluation against existing approaches followed by qualitative analysis to visualize model

performance. Ablation analysis is conducted to assess the contributions of the individual component. Cross-dataset evaluation is also performed to examine generalization, along with model complexity analysis.

7.4.1 Quantitative Analysis

In this section we comprehensively evaluate performance of the MNet-SAt framework on Kvasir-SEG [161] and the CVC-ClinicDB [162] benchmark dataset. We utilize standard segmentation metrics including Dice Similarity Coefficient (DSC), Intersection over Union (IoU), Precision (Pre), Recall (Rec), Accuracy (Acc), F2 Score (F2), Dice Loss (DL), 95th percentile Hausdorff distance (HD95) to analyze effectiveness in polyp segmentation. Figure 7.6 display the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curve in comparison to existing baseline.

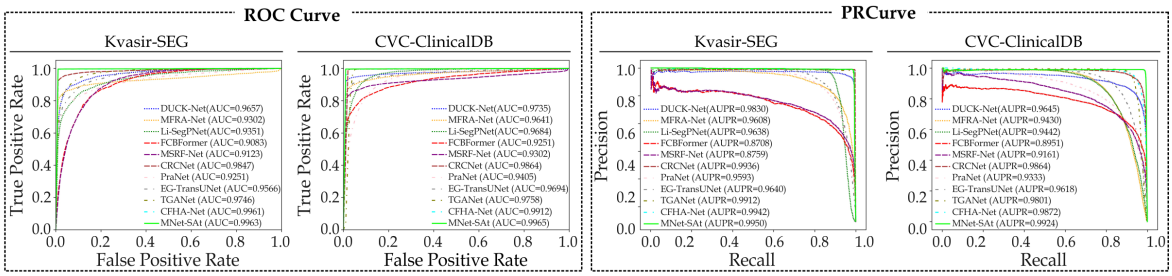


Figure 7.6: Comparative analysis of MNet-SAt against existing baselines across two datasets using ROC curves and PR curves.

(a) Performance on Kvasir-SEG Dataset: Table 7.2 summarizes the quantitative evaluation of the MNet-SAt framework on the Kvasir-SEG [161] dataset. Our framework demonstrates superior performance consistently achieving the highest scores for DSC, IoU, Pre and Rec, highlighting segmenting polyp effectively. While EGTransUNet [171] slightly surpasses MNet-SAt in Acc by 0.51% this difference underscores importance of attention mechanism in capturing global dependencies and emphasizing regions of interest. Similarly MFRA-Net [172] shows a marginal 0.15% improvement in F2 over MNet-SAt demonstrating the effectiveness of multi-scale feature extraction for capturing polyps of varying sizes. The MNet-SAt framework achieves the lowest HD95 value demonstrating that integrating the Enhanced Global Feature Extraction (EGFE) module improves identification of boundary in complex regions with irregular shape.

Table 7.2: Quantitative analysis of MNet-SAt framework with existing baselines on the Kvasir-SEG dataset. **Bold**: best and underlined: second-best scores.

	Methods	DSC	IoU	Pre	Rec	Acc	F2	DL	HD95
I	DUCK-Net [173]	95.56	90.53	95.32	90.76	96.12	91.82	4.44	40.32
	MFRA-Net [172]	94.37	90.76	92.28	96.69	95.47	**95.82**	5.63	36.78
	Li-SegPNet [174]	91.86	89.17	92.79	87.88	93.34	89.17	8.14	29.53
II	FCBFormer [77]	91.08	88.48	94.54	93.63	94.59	94.18	8.92	30.10
	MSRF-Net [78]	93.71	90.71	96.69	92.75	95.38	93.46	6.29	33.43
	CRCNet [175]	90.60	89.38	91.75	89.94	92.74	90.73	9.40	27.55
III	PraNet [68]	88.21	85.16	87.24	89.61	93.46	91.34	11.79	25.84
	EGTransUNet [171]	93.19	90.15	93.02	93.96	**98.34**	95.16	6.81	29.83
	TGANet [176]	88.58	84.22	89.74	90.98	94.54	92.63	11.42	31.58
	CFHA-Net [67]	92.50	90.30	93.60	92.86	95.98	94.75	7.50	28.68
Ours	MNet-SAt	**96.61**	**96.92**	**97.36**	**97.12**	97.83	95.67	**03.39**	**23.91**

I: Encoder Decoder-based Segmentation, II: Multi-Scale Feature-based Segmentation, III: Attention-based Segmentation.

Table 7.3: Quantitative analysis of MNet-SAt framework with existing baselines on the CVC-ClinicDB dataset. **Bold**: best and underlined: second-best scores.

	Methods	DSC	IoU	Pre	Rec	Acc	F2	DL	HD95
I	DUCK-Net [173]	97.15	88.55	94.02	**96.39**	95.65	94.68	2.85	21.54
	MFRA-Net [172]	96.24	93.36	96.83	94.37	94.23	95.62	3.76	16.36
	Li-SegPNet [174]	91.85	93.55	**96.99**	87.09	95.81	92.77	8.15	23.49
II	FCBFormer [77]	95.59	89.83	93.77	94.08	94.64	94.30	4.41	20.05
	MSRF-Net [78]	94.74	91.25	94.61	93.24	96.77	**96.84**	5.26	15.86
	CRCNet [175]	94.83	94.85	94.35	92.93	95.04	92.91	5.17	18.31
III	PraNet [68]	88.19	85.88	87.73	88.32	92.39	91.90	11.81	15.63
	EGTransUNet [171]	95.12	92.36	94.34	96.27	96.91	94.26	4.88	14.88
	TGANet [176]	94.98	91.10	94.81	94.05	93.47	93.73	5.02	23.54
	CFHA-Net [67]	95.53	95.54	95.53	95.01	95.76	95.31	4.47	24.07
Ours	MNet-SAt	**98.60**	**95.89**	96.89	96.13	**97.32**	95.88	**1.40**	**13.57**

I: Encoder Decoder-based Segmentation, II: Multi-Scale Feature-based Segmentation, III: Attention-based Segmentation.

(b) Performance on CVC-ClinicDB Dataset: Table 7.3 presents the quantitative evaluation of the MNet-SAt framework on the CVC-ClinicDB [162] datasets. MNet-SAt framework achieve excellent performance with DSC and IoU scores of 98.60% and 95.89% respectively surpassing DUCK-Net [173] by 1.45% and CFHA-Net [67] by 0.35%. MNet-SAt also exhibits slightly higher Pre of 96.99% and Rec of 96.39% compared to Li-SegPNet [174] and DUCK-Net [173] highlight superiority to capture multi-scale features essential for accurately identifying polyps of varying sizes and complexities. Furthermore MNet-SAt and EGTransUNet [171] demonstrate better accuracy and HD95 than existing baselines emphasizing the importance of learning global features for accurately segmenting polyps with diverse textures and irregular shapes.

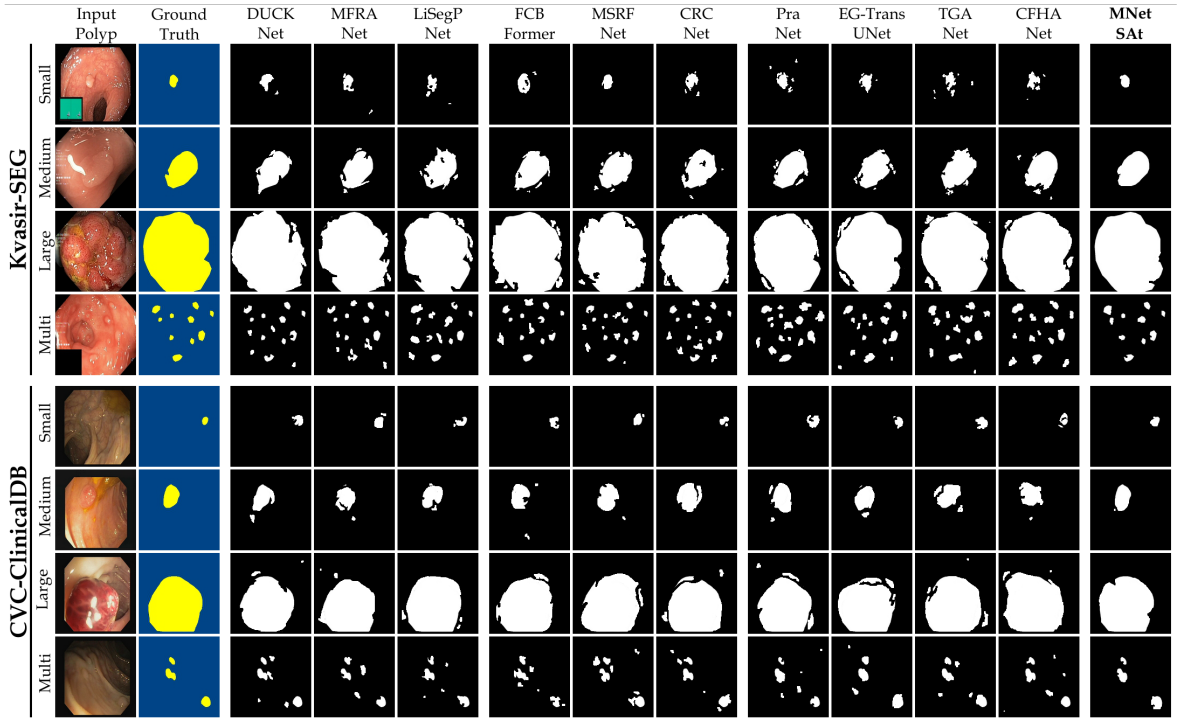


Figure 7.7: Qualitative analysis of MNet-SAt framework.

7.4.2 Qualitative Analysis

Figure 7.7 offers qualitative validation to show MNet-SAt superior polyp mask generation compared to baselines across four cases. This qualitative superiority aligns with the previously established quantitative result. In the case of small polyps (first and fifth rows) all baseline method initially appear to produce satisfactory visual result. However upon closer inspection they fail to maintain the fine boundary detail. Our framework excel preserving polyp anatomy and topology for medium and large polyps significantly outperforming the baseline methods. Notably even in multiple polyps our framework successfully captures more polyps and approximates the ground truth closely. The primary reason for this superior performance is the EGFE module, which effectively eliminates noise in conjunction with the HMAtt module, resulting in a few false positives. Our analysis reinforces MNet-SAt's effectiveness in handling challenging polyp scenarios (small, medium, large-scale, and multi-polyps) while suppressing non-regions of interest.

7.4.3 Ablation Analysis

(a) Component-wise Ablation: In this section, we systematically remove or modify individual components of MNet-SAt, keeping the rest of the framework unchanged. We denote ablated versions as *MNet-SAt*$_{technique}$, where the subscript indicates the component under analysis, and "w/o" indicates the absence of that component.

(i) Effect of Encoder-Decoder Backbone: To analyze the contributions of the encoder-decoder backbone, we conducted an ablation study with three variants: w/o EGFE, with EGFE only, and with both EGFE and Sobel operation (our proposed framework). The results in Table 7.4 demonstrate that removing the EGFE module leads to a significant drop in DSC by 3.59% and an increase in HD95 by 4.65. While including EGFE without Sobel improves performance over the baseline (i.e., without EGFE), it remains inferior to the proposed framework model, which leverages EGFE with Sobel for enhanced boundary capture. The proposed framework with both EGFE and Sobel operation achieves the best performance across all metrics, with a DSC of 96.61%, IoU of 96.92%, and the lowest HD95 of 23.91. These results confirm the role of the EGFE module with the Sobel operation to enhance feature extraction and boundary identification, enabling superior polyp segmentation performance.

Table 7.4: Ablation analysis of MNet-SAt framework with different encoder-decoder backbones on Kvasir-SEG dataset.

Methods	**DSC**	**IoU**	**Pre**	**Rec**	**Acc**	**F2**	**DL**	**HD95**
MNet-SAt$_{\text{w/o EGFE}}$	93.02	92.87	93.55	94.33	94.33	93.26	6.98	28.56
MNet-SAt$_{\text{EGFE + w/o Sobel}}$	95.32	94.19	95.84	95.79	95.84	94.18	4.68	25.92
MNet-SAt$_{\text{EGFE + Sobel}}$	**96.61**	**96.92**	**97.36**	**97.12**	**97.83**	**95.67**	**03.39**	**23.91**

w/o: without, EGFE: Edge-Guided Feature Enrichment.

(ii) Effect of MSFA Module: To gain deeper insights into the extracted features within the MSFA module, the baseline configuration, which employs a simple dilation rate ($d_r \in 1, 6, 12, 18$), demonstrates a lower DSC of 93.52% but achieves a comparable F2 score, with only a marginal difference of 0.24%. Integrating the SE block improves feature weighting, resulting in a DSC of 95.16% and a reduction in HD95 to 32.78. Incorporating dilated convolutions with varying dilation rates further enhances the aggregation of spatial

and contextual features, improving DSC, IoU, Precision, and Recall. The complete MSFA configuration, which includes both SE blocks and dilated convolutions, achieves the best results across all metrics, with a DSC of 96.61%. Additionally, experiments with smaller dilation rates ($\tilde{d}_r \in 1, 6, 12$) demonstrate similar performance comparable to simple dilation rate ($d_r \in 1, 6, 12, 18$), shows the dilation rate of 18 contributes less in scenarios with smaller input sizes. The results in Table 7.5 highlight the importance of SE blocks and dilated convolutions in capturing multi-scale features and diverse polyp appearances for robust segmentation.

Table 7.5: Ablation analysis of MNet-SAt framework with different multi-scale feature aggregator components on Kvasir-SEG dataset.

Methods	DSC	IoU	Pre	Rec	Acc	F2	DL	HD95
MNet-SAt$_{d_r}$	93.52	93.72	93.58	94.03	93.58	95.43	6.48	29.43
MNet-SAt$_{d_r}$ + SE	95.16	95.58	96.14	95.92	94.47	93.59	4.84	32.78
MNet-SAt$_{d_r}$ + SE + Conv$_3$	96.07	95.93	96.74	96.73	96.01	95.21	3.93	27.67
MNet-SAt$_{\tilde{d}_r}$ + SE + Conv$_7$	96.02	95.88	96.60	96.11	95.54	95.22	3.98	30.27
MNet-SAt$_{d_r}$ + SE + Conv$_7$	**96.61**	**96.92**	**97.36**	**97.12**	**97.83**	**95.67**	**03.39**	**23.91**

d_r: dilation rate, SE: Squeeze & Excitation, $Conv_k$: $k \times k$ Conv operation.

(iii) Effect of SEAt Module: Table 7.6 presents an ablation study evaluating the impact of different components within the Spatial Enhancement Attention (SEAt) module on MNet-SAt's performance. The baseline configuration, which excludes Layer Normalization (LN) and employs a dilation rate of 1, exhibits lower performance in DSC, IoU, Pre, and Rec while achieving a comparable accuracy of 96.32%. Incorporating LN with a dilation rate of 2 significantly improves DSC and IoU, demonstrating the benefit of LN in enhancing feature representations. The complete SEAt module achieves the best performance across all metrics with LN, a dilation rate of 2, and Global Average Pooling (GAP). This highlights the contribution of LN, dilated convolutions, and GAP within the SEAt module in capturing rich spatial and contextual information for accurate polyp segmentation.

(iv) Effect of CE-ASPP: The ablation study in Table 7.7 examines the contributions of Multi-Scale Feature Fusion (MSF), Channel Compression (CC), and Inter-Channel Feature Fusion (ICF) within the CE-ASPP module. The integration of CC and MSF achieves a higher Pre of 96.52% compared to utilizing MSF alone, demonstrating the benefits of dimensionality reduction in refining feature representations. However, this configuration still falls short of

Table 7.6: Ablation analysis of MNet-SAt framework with different spatial-enhanced attention components on Kvasir-SEG dataset.

Methods	DSC	IoU	Pre	Rec	Acc	F2	DL	HD95
MNet-SAt$_{\text{w/o LN + } d_r^1}$	94.26	93.51	94.17	94.12	96.32	93.64	5.74	28.91
MNet-SAt$_{\text{LN + } d_r^2}$	95.87	95.79	95.38	96.10	95.73	94.75	4.13	25.07
MNet-SAt $_{\text{LN + } d_r^2 \text{ + GAP}}$	**96.61**	**96.92**	**97.36**	**97.12**	**97.83**	**95.67**	**03.39**	**23.91**

LN: Layer Normalization, d_r: dilation rate, GAP: Global Average Pooling.

the complete MNet-SAt framework. Adding ICF (i.e., w/o CC) to MSF further enhances performance, demonstrating the importance of capturing channel relationships. The complete CE-ASPP module, with CC, MSF, and ICF, achieves the best performance across all metrics. This superior performance highlights the framework's ability to effectively capture complex dependencies between feature channels.

Table 7.7: Ablation analysis of MNet-SAt framework with different channel-enhanced atrous spatial pyramid pooling components on Kvasir-SEG dataset.

Methods	DSC	IoU	Pre	Rec	Acc	F2	DL	HD95
MNet-SAt$_{\text{MSF}}$	94.21	94.76	94.39	95.05	94.77	91.73	5.79	28.96
MNet-SAt$_{\text{CC + MSF}}$	95.32	95.41	96.52	95.87	96.24	93.49	4.68	26.34
MNet-SAt$_{\text{w/o CC + MSF + IF}}$	95.89	96.27	97.08	96.54	95.86	94.41	4.11	25.67
MNet-SAt$_{\text{CC + MSF + ICF}}$	**96.61**	**96.92**	**97.36**	**97.12**	**97.83**	**95.67**	**03.39**	**23.91**

MSF: Multi-Scale Feature Fusion, CC: Channel Compression, ICF: Inter-Channel Feature Fusion.

(v) Effect of Optimizer: Table 7.8 presents the impact of different optimizers on MNet-SAt performance using a consistent learning rate of 1e-4. Among the tested optimizer SGD yield the lowest performance across all metric indicating its limitation in effectively optimizing parameter. RMSProp demonstrate superior performance over SGD reflecting its ability to handle sparse gradient and adapt learning rate to different parameter. N-AdaM shows further enhancements but does not outperform the AdaM optimizer. AdaM achieves the highest performance across all metrics, highlighting its superior capability in addressing non-convex optimization challenges. The results indicate the role of optimizer selection in maximizing the framework performance.

(b) Analytical Ablation: The ablation study in Table 7.9 systematically evaluates the impact of incrementally adding individual components to the MNet-SAt framework and their contribution to segmentation performance. Starting with only the Encoder-Decoder Backbone

Table 7.8: Ablation analysis of MNet-SAt framework with different optimizers on Kvasir-SEG dataset.

Optimizer	DSC	IoU	Pre	Rec	Acc	F2	DL	HD95
MNet-SAt$_{SGD}$	93.27	86.45	95.53	93.75	92.51	92.47	6.73	29.80
MNet-SAt$_{RMSProp}$	95.56	91.73	94.37	92.18	94.82	94.66	4.44	22.58
MNet-SAt$_{N\text{-}AdaM}$	94.46	90.22	93.48	90.38	95.47	93.51	5.54	27.35
MNet-SAt$_{AdaM}$	**96.61**	**96.92**	**97.36**	**97.12**	**97.83**	**95.67**	**03.39**	**23.91**

(EDB) the model achieves a baseline DSC of 93.50% and IoU of 94.12%. Adding the Multi-Scale Feature Aggregator (MSFA) enhances performance, improving the DSC to 94.74% and the IoU to 95.19%, highlighting the effectiveness of multi-scale feature extraction. Incorporating the Spatial-Enhanced Attention (SEAt) module further boosts performance, increasing the DSC to 95.81% and IoU to 96.41%, underscoring the importance of attentive spatial features in refining multi-scale aggregated features. Finally, the complete MNet-SAt framework, including the Channel-Enhanced Atrous Spatial Pyramid Pooling (CE-ASPP) module, achieves the highest DSC of 96.61% and IoU of 96.92%, representing overall gains of 3.11% and 2.80%, respectively. This analysis highlights the cumulative contribution of each module to MNet-SAt's overall performance in accurate polyp segmentation.

Table 7.9: Ablation analysis of MNet-SAt framework with different components on the Kvasir-SEG dataset.

EDB	MSFA	SEAt	CE-ASPP	DSC	IoU	Pre	Rec	Acc	F2	DL	HD95
✓	×	×	×	93.50	94.12	94.07	94.18	94.15	93.07	6.50	26.79
✓	✓	×	×	94.74	95.19	95.48	95.22	95.33	93.83	5.26	25.88
✓	✓	✓	×	95.81	96.41	96.71	96.39	96.87	94.14	4.19	24.67
✓	**✓**	**✓**	**✓**	**96.61**	**96.92**	**97.36**	**97.12**	**97.83**	**95.67**	**3.39**	**23.91**

EDB: Encoder-Decoder Backbone, CE-ASPP: Channel-Enhanced Atrous Spatial Pyramid Pooling, MSFA: Multi-Scale Feature Aggregator, SEAt: Spatial-Enhanced Attention.

7.4.4 Cross-Dataset Analysis

The proposed MNet-SAt framework demonstrates its generalizability through cross-dataset evaluation. Table 7.10 shows MNet-SAt's outstanding performance when trained on the Kvasir-SEG dataset [161] and evaluated on the previously unseen CVC-ClinicDB dataset [162]. The framework achieves a Dice Similarity Coefficient (DSC) of 91.52%, significantly surpassing all existing baselines across segmentation paradigms, including

encoder-decoder, multi-scale feature, and attention-based methods. Moreover, MFRA-Net [172] and Li-SegPNet [174] show marginally better Recall and F2-score (by 1.38% and 1.25%, respectively) compared to our proposed framework, but display less satisfactory HD95 values. MNet-SAt achieves the lowest HD95 value of 29.91, highlighting its superior ability to identify blurred or ambiguous polyp boundaries. Compared to multi-scale feature extraction techniques, MNet-SAt demonstrates better overall performance, clearly showcasing its capability to effectively extract multi-scale features and handle polyps of varying shapes and sizes. Notably, CFHA-Net [67] exhibits slightly higher precision (1.61% difference), highlighting the importance of spatial-enhanced global features.

Table 7.10: Cross-Dataset analysis of MNet-SAt framework with training on Kvasir-SEG and testing on CVC-ClinicDB.

	Methods	**DSC**	**IoU**	**Pre**	**Rec**	**Acc**	**F2**	**DL**	**HD95**
I	DUCK-Net [173]	88.92	86.11	87.65	86.23	85.37	88.85	11.08	45.32
	MFRA-Net [172]	86.61	83.33	84.52	90.72	84.79	90.14	13.39	40.78
	Li-SegPNet [174]	83.78	86.62	86.23	81.17	83.01	91.75	16.22	35.53
II	FCBFormer [77]	85.00	81.41	87.80	86.32	82.42	87.87	15.00	35.10
	MSRF-Net [78]	85.80	84.80	88.40	88.10	84.65	88.35	14.20	38.43
	CRCNet [175]	85.95	83.09	84.88	83.02	83.32	88.06	14.05	32.55
III	PraNet [68]	81.66	78.90	82.07	81.46	78.48	80.07	18.34	30.84
	EGTransUNet [171]	85.99	83.39	84.99	87.65	84.89	87.95	14.01	33.83
	TGANet [176]	83.51	78.05	84.28	84.37	77.32	89.14	16.49	35.58
	CFHA-Net [67]	87.62	83.58	89.37	86.60	84.13	89.85	12.38	34.68
Ours	MNet-SAt	**91.52**	**88.26**	**90.98**	**89.34**	**90.28**	**90.50**	**8.48**	**29.91**

I: Encoder Decoder-based Segmentation; II: Multi-Scale Feature-based Segmentation; III: Attention-based Segmentation.

Table 7.11 illustrates the generalization performance of MNet-SAt when trained on the CVC-ClinicDB dataset [162] and evaluated on the unseen Kvasir-SEG dataset [161]. Achieving a DSC of 92.28% and IoU of 87.62%, MNet-SAt demonstrates its robust ability to adapt to variations in unseen data. Furthermore, the highest precision (91.19%) and competitive recall (89.89%), highlighting MNet-SAt's capacity to balance true polyp regions compared to actual polyp regions, are significant for effective polyp segmentation, minimizing both false positives and false negatives. The framework's performance is validated by achieving the lowest Hausdorff Distance (HD95) of 33.91, demonstrating its ability to accurately extract boundary pixels. These results illustrate the effectiveness of MNet-SAt's multiscale and spatial-enhanced attention mechanism in adapting to unseen data variations while maintaining high segmentation performance. This strong generalization ability, ex-

ceeding many existing baselines, underscores MNet-SAt's potential for real-world application in diverse clinical settings.

Table 7.11: Cross-Dataset analysis of MNet-SAt framework with training on CVC-ClinicDB and testing on Kvasir-SEG.

	Methods	DSC	IoU	Pre	Rec	Acc	F2	DL	HD95
I	DUCK-Net [173]	91.81	80.89	87.91	88.87	84.52	89.26	8.19	46.32
	MFRA-Net [172]	89.04	87.90	88.52	87.06	85.59	89.37	10.96	42.78
	Li-SegPNet [174]	86.78	87.28	89.57	88.32	83.29	84.61	13.22	36.53
II	FCBFormer [77]	89.39	84.61	87.47	87.76	82.10	90.91	10.61	38.10
	MSRF-Net [78]	89.25	83.07	88.62	85.02	81.84	89.26	10.75	39.43
	CRCNet [175]	87.50	87.62	90.02	86.16	83.99	87.19	12.50	34.55
III	PraNet [68]	80.84	77.58	83.23	81.14	79.02	80.45	19.16	32.84
	EGTransUNet [171]	87.54	83.78	89.18	89.04	83.39	89.72	12.46	36.83
	TGANet [176]	87.40	84.76	89.13	86.74	77.34	90.17	12.60	38.58
	CFHA-Net [67]	88.45	86.82	88.52	88.36	85.54	90.70	11.55	34.68
Ours	MNet-SAt	**92.28**	**87.62**	**91.19**	**89.89**	**85.87**	**92.71**	**7.72**	**33.91**

I: Encoder Decoder-based Segmentation; II: Multi-Scale Feature-based Segmentation; III: Attention-based Segmentation.

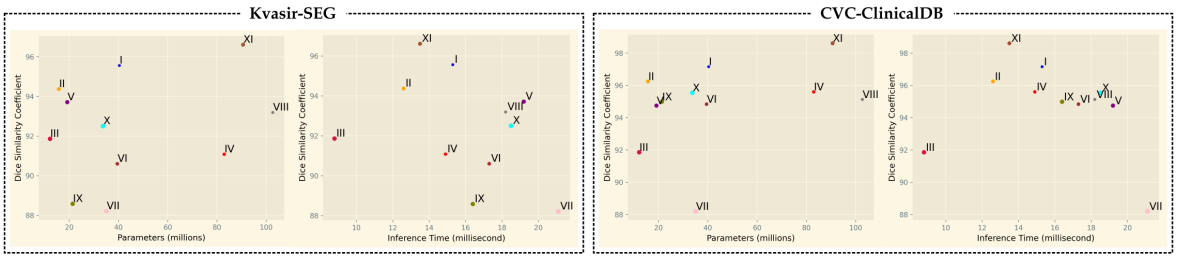


Figure 7.8: Comparative analysis of model parameters, inference time, and accuracy for the MNet-SAt framework against existing baselines on benchmark datasets.

7.4.5 Model Complexity Analysis

Table 7.12 presents a computational analysis of the MNet-SAt framework compared to existing baselines. Among encoder-decoder-based segmentation methods, DUCK-Net [173] features fewer parameters (40.36M) but exhibits a slower inference time of 15.3 ms compared to MNet-SAt. Within multi-scale feature-based segmentation approaches, CRCNet [175] demonstrates efficient parameter utilization with competitive FLOPs (39.53G) but an increased inference time (17.3ms). Attention-based methods, such as EGTransUNet [171], show the highest computational complexity (31.56G MACs). In contrast, the MNet-SAt framework optimally balances computational efficiency and model complexity. Despite

having a moderate number of parameters (approximately 90 million), MNet-SAt delivers a competitive inference time of around 13 milliseconds. Figure 7.8 illustrates a comparative analysis of model parameters and inference time. Overall, MNet-SAt achieves superior performance with a balanced level of computational complexity, making it well-suited for real-world deployment scenarios with limited computational resources.

Table 7.12: Model Complexity Analysis of the MNet-SAt framework comparing parameters, FLOPS, MACs, training time, and inference times with existing baselines.

	Methods	Params (M)	FLOPS (G)	MAC (G)	Train Time (min)	Inf. Time (ms)
I	DUCK-Net [173]	40.36	53.71	25.98	145	15.3
	MFRA-Net [172]	15.75	11.46	5.75	153	12.6
	Li-SegPNet [174]	12.19	18.67	9.14	169	8.8
II	FCBFormer [77]	82.96	32.39	16.88	225	14.9
	MSRF-Net [78]	19.19	48.74	24.57	195	19.2
	CRCNet [175]	39.52	39.53	20.32	101	17.3
III	PraNet [68]	35.08	15.11	7.67	152	21.1
	EGTransUNet [171]	102.70	60.75	31.56	241	18.2
	TGANet [176]	21.38	40.35	11.45	177	16.4
	CFHA-Net [67]	33.71	23.92	12.34	169	18.5
Ours	MNet-SAt	**90.63**	**38.52**	**19.37**	**187**	**13.5**

I: Encoder Decoder-based Segmentation; II: Multi-Scale Feature-based Segmentation; III: Attention-based Segmentation.
M: Million; G: Giga; min: minutes; ms: millisecond.

7.5 Conclusion

In this research, we propose a novel Multiscale Network with Spatial-enhanced Attention (MNet-SAt) for polyp segmentation in colonoscopy. MNet-SAt leverages a unique encoder-decoder architecture incorporating an edge-preserving unit to refine polyp boundaries. We enhance feature learning by channel-wise recalibrating and aggregating multi-scale features, capturing information across all scales. Subsequently, spatial-aware attention prioritizes global dependencies within these aggregated regions, highlighting salient features. The attentive features are then resampled at multiple scales to improve polyp representation. Extensive evaluations on Kvasir-SEG and CVC-ClinicDB datasets show that MNet-SAt outperforms existing methods. Future research will focus on adapting MNet-SAt for broader applicability in medical image segmentation, incorporating techniques such as self-supervised learning and knowledge distillation to enhance model robustness and generalizability.

Chapter 8

Conclusion and Future Directions

CONTENTS

8.1 Summary of the Thesis 181

8.1.1 Coupled Transformer–Convolutional Network with an Adaptive Graph Reconstruction for Leukemia Detection 181

8.1.2 Contrastive–based Dilated Convolutional Network with Transformer for Pediatric Pneumonia Detection 181

8.1.3 Hierarchical Fusion of Semi and Self-Supervised Learning for Knee Osteoarthritis Severity Grading 182

8.1.4 Transformer-aware Multiscale Progressive Encoder-Decoder Network for Tumor and Liver Segmentation 183

8.1.5 A Multiscale Network with Spatial-enhanced Attention for Segmentation of Polyps in Colonoscopy 183

8.2 Future Work 186

In this chapter, we summarize the primary research findings and technical contributions detailed in this thesis, followed by a discussion of prospective directions for future investigations in computer-aided diagnosis.

8.1 Summary of the Thesis

In this thesis, we address the critical challenges in automated pathological image analysis across diverse medical imaging modalities, including blood smears, CXR, CT, MRI, and colonoscopy images. The work is organized around four core objectives: global–local feature synthesis, data-efficient learning, precise boundary refinement, and multi-scale feature fusion. Across varied modalities, the proposed framework demonstrates strong quantitative and qualitative performance, highlighting its usability in CAD systems.

8.1.1 Coupled Transformer–Convolutional Network with an Adaptive Graph Reconstruction for Leukemia Detection

This chapter addresses automated leukemia classification from blood smear images, where strong visual similarity between healthy and malignant cells, with staining variations and overlapping nuclei, hinders reliable diagnosis. To overcome these limitations, the proposed CoTCoNet framework combined transformer-based feature learning with convolutional-based local feature extraction to capture hematological patterns. A graph-based feature reconstruction module enriches hidden feature relationships by integrating leukocyte segmentation, improving focus on diagnostically relevant regions. To mitigate the impact of limited data and class imbalance, the synthetic leukocyte generation was incorporated using an adversarial learning strategy. Feature optimization was employed to retain discriminative representation for the final classification. Experimental evaluation across multiple benchmark datasets shows superior gains in performance and robustness compared to existing methods. The study demonstrates that integrating global context with cellular-level detail can improve leukemia recognition and support early hematological screening.

8.1.2 Contrastive–based Dilated Convolutional Network with Transformer for Pediatric Pneumonia Detection

This chapter focuses on pediatric pneumonia detection from chest radiographs, where low contrast, anatomical overlap, and data scarcity reduce diagnostic reliability. The proposed XCCNet framework was developed to jointly learn global contextual information and local

pathological patterns. The transformer-based model captures broader dependencies using dilated convolutions to extract fine spatial detail associated with the pulmonary infection. Contrastive learning is introduced to align the complementary features and improve class separation. A preprocessing pipeline enhanced chest radiographs by improving contrast during lung region extraction and suppressing ribs to reduce irrelevant structures. In addition, adversarial data generation helped address class imbalance by producing realistic synthetic samples. An explainability module highlights the image region that influences the prediction, thereby improving transparency. Extensive experiments across benchmark datasets showed superior performance compared with the competitive baselines. The findings confirm that combining feature synthesis, including preprocessing and interpretability, strengthens the automated pediatric pneumonia diagnosis in resource-constrained clinical settings.

8.1.3 Hierarchical Fusion of Semi and Self-Supervised Learning for Knee Osteoarthritis Severity Grading

This chapter investigates the severity grading of knee osteoarthritis on radiographs under limited-label settings. Accurate KOA grading is challenging because adjacent classes appear similar and also requires expert effort. To address this issue, the proposed H-SemiS framework combines self-supervised and semi-supervised learning in a hierarchical structure. A masked image reconstruction stage generates a proxy sample from the unlabeled data to improve feature diversity. A similarity-aware labeling strategy assigns labels to generated pseudo-labels while filtering uncertain samples. The multi-class grading task was decomposed into binary decisions to reduce the class imbalance and support balanced learning. Quantum-inspired modeling further incorporates to capture complex structural relationships in knee anatomy. Results across benchmark datasets show that the framework achieved reliable grading performance with a small number of data labels. This chapter demonstrates that the annotation-efficient learning technique reduces dependence on the expert labels while maintaining superior performance.

8.1.4 Transformer-aware Multiscale Progressive Encoder-Decoder Network for Tumor and Liver Segmentation

This chapter addresses the liver and tumor segmentation from CT and MRI scans of indistinct boundaries and heterogeneous lesions with similarities in surrounding organs. This complicates the automated analysis. To overcome these issues, the T-MPEDNet framework introduced the progressive encoder–decoder architecture for accurate identification and structural preservation. Adaptive feature extraction and channel recalibration improve feature transfer during encoding and decoding. A dynamic contextual attention module captures the long-range dependencies and suppresses the irrelevant background responses. Multi-scale atrous feature learning enabled the representation of lesions with varying sizes and textures. In addition, a morphological boundary-refinement unit sharpened the segmentation mask and reduced the ambiguity at the tissue interface. An extensive experiment on the benchmark dataset demonstrates improvement over the existing segmentation method for both the liver and tumor regions. The findings confirm that combining contextual reasoning with multi-scale learning, including boundary refinement, improves segmentation precision for clinical tasks.

8.1.5 A Multiscale Network with Spatial-enhanced Attention for Segmentation of Polyps in Colonoscopy

This chapter focuses on automated polyp segmentation in colonoscopy images, an important task for early colorectal cancer detection. Colonoscopy frames contain low contrast reflections with blurred boundaries and large variations in polyp size and appearance. These multiple variations make segmentation challenging. The proposed MNet-SAt framework addressed these challenges through the multi-scale encoder–decoder design. The Edge-guided feature enrichment preserved boundary information and captured the fine structural information. A multi-scale feature aggregation unit extracts contextual patterns across different receptive fields to handle diverse polyps. A spatially enhanced attention mechanism highlights the relevant region with broader dependencies. Channel-enhanced atrous pyramid pooling refines fused representations for accurate mask generation. An experimental evaluation on

the benchmark dataset demonstrates superior performance over competing baselines in both quantitative and qualitative analyses. This chapter shows that effective multi-scale feature fusion with attention enhances the polyp identification within endoscopic image samples.

The work presented in the thesis can be further extended along several promising research directions outlined below.

(i) **Leukemia Classification** Future work will prioritize integrating blood smear histology with cytochemical developmental-state information. By combining visual patterns with biochemical markers, the semi-supervised model can distinguish more complex subtypes and specific developmental cell stages with higher precision, enabling comprehensive hematological assessment. In parallel, reducing CoTCoNet's parameter count through knowledge distillation, where a lightweight student network is trained to replicate the predictions of the current model, or through structured pruning of the graph attention layers, would lower the computational footprint of the framework, making it more suitable for deployment in resource-constrained clinical settings.

(ii) **Pneumonia Detection**: For pediatric pneumonia detection, future research should integrate Electronic Health Records (EHR) with multimodal data, including patient history and text-based report generation. Such integration will improve predictions of disease progression over time, enabling clinicians to implement early diagnostic interventions. Furthermore, incorporating human feedback into the learning curve will help the model identify fine-grained patterns more accurately. In addition, extending the architecture to multi-task learning, where a severity-grading head is trained jointly with the detection task using a combined classification and ordinal regression loss, would allow the model to predict both the presence of pneumonia and its clinical severity level, using severity-labeled data derived from established radiological grading scales.

(iii) **Knee Osteoarthritis Severity Grading**: To enhance knee osteoarthritis grading, we aim to extend the H-SemiS framework into 3D imaging, like MRIs and CTs, for capturing internal structural variations. The future area includes developing confidence-based error-correction mechanisms to mitigate the risk of misclassification in the hierarchical decision framework. These mechanisms allow the later stages of the hierarchy

to re-evaluate and refine earlier routing decisions, ensuring stable diagnostic prediction. In addition, training efficiency can be improved by parallelizing node training across sibling nodes at the same hierarchy level, which do not share gradient dependencies, along with weight sharing across nodes with similar quantum-feature extraction requirements to reduce computation and improve scalability for larger hierarchies.

(iv) **Tumor and Liver Segmentation**: Future developments for liver and tumor segmentation will focus on 3D volumetric data analysis to capture the inter-slice dependency in abdominal scans. Research will also aim to expand the framework's explainability and scope to support simultaneous multi-organ segmentation. Improving boundary precision for point-sized tumors that resemble neighboring tissues remains a priority to ensure the framework can handle challenging cases. In addition, replacing the computationally intensive morphological post-processing with a differentiable boundary-based loss, such as a boundary IoU or Hausdorff-distance-based loss, integrated directly into end-to-end training, is a promising direction to reduce inference-time computation while preserving boundary precision.

(v) **Polyp Segmentation**: Future investigation for polyp segmentation focuses on attention-guided filtering and adversarial training to effectively differentiate surgical instruments from pathological tissue. Additionally, exploring model compression and knowledge distillation techniques is essential to reduce computational complexity. These optimizations will enable real-time deployment of the MNet-SAt framework in clinical endoscopy suites, providing immediate support to gastroenterologists during live procedures. As the current framework is trained and evaluated on static polyp images, extending it to process video sequences by incorporating a temporal attention mechanism across consecutive frames is another key direction, allowing the model to propagate contextual and boundary information over time and maintain segmentation stability under motion blur or illumination variation encountered during live endoscopic procedures.

8.2 Future Work

The work presented in the thesis can be further extended along several promising research directions outlined below.

(i) **Leukemia Classification** Future work will prioritize integrating blood smear histology with cytochemical developmental-state information. By combining visual patterns with biochemical markers, the semi-supervised model can distinguish more complex subtypes and specific developmental cell stages with higher precision, enabling comprehensive hematological assessment. In parallel, reducing CoTCoNet's parameter count through knowledge distillation, where a lightweight student network is trained to replicate the predictions of the current model, or through structured pruning of convolutional and transformer components, could lower the computational footprint of the framework, making it more suitable for deployment in resource-constrained clinical settings.

(ii) **Pneumonia Detection** For pediatric pneumonia detection, future research should integrate Electronic Health Records (EHR) with multimodal data, including patient history and text-based report generation. Such integration will improve predictions of disease progression over time, enabling clinicians to implement early diagnostic interventions. Furthermore, incorporating clinician feedback into the learning process will help the model identify fine-grained patterns more accurately. In addition, extending the architecture to multi-task learning, where a severity-grading head is trained jointly with the detection task using an appropriate severity-aware objective, such as ordinal classification or ordinal regression, would allow the model to predict both the presence of pneumonia and its clinical severity level using appropriately severity-labeled data.

(iii) **Knee Osteoarthritis Severity Grading** To enhance knee osteoarthritis grading, we aim to extend the H-SemiS framework into 3D imaging, such as MRI and CT, for capturing internal structural variations. The future area includes developing confidence-based error-correction mechanisms to mitigate the risk of misclassification in the hierarchical decision framework. These mechanisms would allow later stages of the

hierarchy to re-evaluate and refine earlier routing decisions, ensuring stable diagnostic prediction. In addition, training efficiency could be improved by parallelizing the training of independent sibling nodes at the same hierarchy level and exploring parameter sharing across selected hierarchical nodes to reduce redundant computation and improve scalability for larger hierarchies.

(iv) **Tumor and Liver Segmentation** Future developments for liver and tumor segmentation will focus on 3D volumetric data analysis to capture inter-slice dependencies in abdominal scans. Research will also aim to expand the framework's explainability and scope to support simultaneous multi-organ segmentation. Improving boundary precision for point-sized tumors that resemble neighboring tissues remains a priority to ensure that the framework can handle challenging cases. In addition, replacing the computationally intensive morphological post-processing with differentiable boundary-aware or distance-transform-based losses integrated directly into end-to-end training is a promising direction that could reduce the dependence on post-processing while preserving boundary precision.

(v) **Polyp Segmentation** Future investigation for polyp segmentation focuses on attention-guided filtering and adversarial training to effectively differentiate surgical instruments from pathological tissue. Additionally, exploring model compression and knowledge distillation techniques is essential to reduce computational complexity. These optimizations could facilitate real-time deployment of the MNet-SAt framework in clinical endoscopy suites, providing immediate support to gastroenterologists during live procedures. Since the current framework is trained and evaluated on static polyp images, extending it to process video sequences by incorporating a temporal attention mechanism over consecutive frames is another key direction, allowing the model to propagate contextual and boundary information over time and maintain segmentation stability under motion blur or illumination variation encountered during live endoscopic procedures.

Bibliography

[1] C. A. Leitão, G. L. de Oliveira Salvador, B. M. Idowu, and F. Dako, "Drivers of global health care worker migration," *Journal of the American College of Radiology*, vol. 21, no. 8, pp. 1188–1193, 2024, doi:10.1016/j.jacr.2024.03.005.

[2] E. Sokolovskaya, T. Shinde, R. B. Ruchman, A. J. Kwak, S. Lu, Y. K. Shariff, E. F. Wiggins, and L. Talangbayan, "The effect of faster reporting speed for imaging studies on the number of misses and interpretation errors: a pilot study," *Journal of the American College of Radiology*, vol. 12, no. 7, pp. 683–688, 2015, doi:10.1016/j.jacr.2015.03.040.

[3] N. Achour, T. Zapata, Y. Saleh, B. Pierscionek, N. Azzopardi-Muscat, D. Novillo-Ortiz, C. Morgan, and M. Chaouali, "The role of ai in mitigating the impact of radiologist shortages: a systematised review," *Health and Technology*, vol. 15, no. 3, pp. 489–501, 2025, doi:10.1007/s12553-025-00970-y.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi:10.1038/nature14539.

[5] G. Litjens, 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, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017, doi:10.1016/j.media.2017.07.005.

[6] R. Kumar, P. Kumbharkar, S. Vanam, and S. Sharma, "Medical images classification using deep learning: a survey," *Multimedia Tools and Applications*, vol. 83, no. 7, pp. 19 683–19 728, 2024, doi:10.1007/s11042-023-15576-7.

[7] S. U. R. Khan, S. Asif, M. Zhao, W. Zou, Y. Li, and X. Li, "Optimized deep learning model for comprehensive medical image analysis across multiple modalities," *Neurocomputing*, vol. 619, p. 129182, 2025, doi:10.1016/j.neucom.2024.129182.

[8] M. Li, Y. Jiang, Y. Zhang, and H. Zhu, "Medical image analysis using deep learning algorithms," *Frontiers in public health*, vol. 11, p. 1273253, 2023, doi:10.3389/fpubh.2023.1273253.

[9] H. Wang, Q. Jin, S. Li, S. Liu, M. Wang, and Z. Song, "A comprehensive survey on deep active learning in medical image analysis," *Medical Image Analysis*, vol. 95, p. 103201, 2024, doi:10.1016/j.media.2024.103201.

[10] C. Chen, N. A. M. Isa, and X. Liu, "A review of convolutional neural network based methods for medical image classification," *Computers in biology and medicine*, vol. 185, p. 109507, 2025, doi:10.1016/j.compbiomed.2024.109507.

[11] R. Azad, E. K. Aghdam, A. Rauland, Y. Jia, A. H. Avval, A. Bozorgpour, S. Karimijafarbigloo, J. P. Cohen, E. Adeli, and D. Merhof, "Medical image segmentation review: The success of u-net," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10 076–10 095, 2024, doi:10.1109/TPAMI.2024.3435571.

[12] T. Dissanayake, Y. George, D. Mahapatra, S. Sridharan, C. Fookes, and Z. Ge, "Few-shot learning for medical image segmentation: A review and comparative study," *ACM Computing Surveys*, vol. 58, no. 1, pp. 1–36, 2025, doi:10.1145/3746224.

[13] J. Zhang, X. Chen, B. Yang, Q. Guan, Q. Chen, J. Chen, Q. Wu, Y. Xie, and Y. Xia, "Advances in attention mechanisms for medical image segmentation," *Computer Science Review*, vol. 56, p. 100721, 2025, doi:10.1016/j.cosrev.2024.100721.

[14] C. S. Raghaw, A. Sharma, S. Bansal, M. Z. U. Rehman, and N. Kumar, "Cotconet: An optimized coupled transformer-convolutional network with an adaptive graph reconstruction for leukemia detection," *Computers in Biology and Medicine*, vol. 179, p. 108821, 2024, doi:10.1016/j.compbiomed.2024.108821.

[15] C. S. Raghaw, P. S. Bhore, M. Z. U. Rehman, and N. Kumar, "An explainable contrastive-based dilated convolutional network with transformer for pediatric pneumonia detection," *Applied Soft Computing*, vol. 167, 2024, doi:10.1016/j.asoc.2024.112258.

[16] C. S. Raghaw, A. Parwal, S. S. Dar, P. Darade, and N. Kumar, "H-semis: Hierarchical fusion of semi and self-supervised learning for knee osteoarthritis severity grading," *Expert Systems with Applications*, p. 132279, 2026, doi:10.1016/j.eswa.2026.132279.

[17] C. S. Raghaw, J. S. Sanjotra, M. Z. U. Rehman, S. Bansal, S. S. Dar, and N. Kumar, "T-mpednet: Unveiling the synergy of transformer-aware multiscale progressive encoder-decoder network with feature recalibration for tumor and liver segmentation," *Biomedical Signal Processing and Control*, vol. 110, p. 108225, 2025, doi:10.1016/j.bspc.2025.108225.

[18] C. S. Raghaw, A. Yadav, J. S. Sanjotra, S. Dangi, and N. Kumar, "Mnet-sat: A multiscale network with spatial-enhanced attention for segmentation of polyps in colonoscopy," *Biomedical Signal Processing and Control*, vol. 102, p. 107363, 2025, doi:10.1016/j.bspc.2024.107363.

[19] B. Leng, H. Jiang, B. Wang, J. Wang, and G. Luo, "A lightweight deep learning model for acute myeloid leukemia-related blast cell identification," *The Journal of Supercomputing*, 2024, doi:10.1007/s11227-024-06063-3.

[20] M. Trivedi and A. Gupta, "A lightweight deep learning architecture for the automatic detection of pneumonia using chest x-ray images," *Multimedia Tools and Applications*, vol. 81, 2022, doi:10.1007/s11042-021-11807-x.

[21] Y. Chen, C. Zhang, B. Chen, Y. Huang, Y. Sun, C. Wang, X. Fu, Y. Dai, F. Qin, Y. Peng, and Y. Gao, "Accurate leukocyte detection based on deformable-detr and multi-level feature fusion for aiding diagnosis of blood diseases," *Computers in Biology and Medicine*, vol. 170, p. 107917, 2024, doi:10.1016/j.compbiomed.2024.107917.

[22] S. Dhalla, A. Mittal, and S. Gupta, "Leukocapsnet: a resource-efficient modified capsnet model to identify leukemia from blood smear images," *Neural Computing and Applications*, vol. 36, pp. 2507–2524, 2024, doi:10.1007/s00521-023-09157-w.

[23] K. K. Mohammed, A. E. Hassanien, and H. M. Afify, "Refinement of ensemble strategy for acute lymphoblastic leukemia microscopic images using hybrid cnn-grubilstm and msvm classifier," *Neural Computing and Applications*, vol. 35, 2023, doi:10.1007/s00521-023-08607-9.

[24] S. Kiliçarslan, C. Közkurt, S. Baş, and A. Elen, "Detection and classification of pneumonia using novel superior exponential (supex) activation function in convolutional neural networks," *Expert Systems with Applications*, vol. 217, 2023, doi:10.1016/j.eswa.2023.119503.

[25] Q. Han, X. Qian, H. Xu, K. Wu, L. Meng, Z. Qiu, T. Weng, B. Zhou, and X. Gao, "Dm-cnn: Dynamic multi-scale convolutional neural network with uncertainty quantification for medical image classification," *Computers in Biology and Medicine*, vol. 168, p. 107758, 2024, doi:10.1016/j.compbiomed.2023.107758.

[26] I. Adjei-Mensah, X. Zhang, I. O. Agyemang, S. B. Yussif, A. A. Baffour, B. M. Cobbinah, C. Sey, L. D. Fiasam, I. A. Chikwendu, and J. R. Arhin, "Covfed: Federated learning-based framework for covid-19 diagnosis using chest x-ray scans," *Engineering Applications of Artificial Intelligence*, vol. 128, 2024, doi:10.1016/j.engappai.2023.107448.

[27] M. Usman, T. Zia, and A. Tariq, "Analyzing transfer learning of vision transformers for interpreting chest radiography," *Journal of Digital Imaging*, vol. 35, 2022, doi:10.1007/s10278-022-00666-z.

[28] S. Singh, M. Kumar, A. Kumar, B. K. Verma, K. Abhishek, and S. Selvarajan, "Efficient pneumonia detection using vision transformers on chest x-rays," *Scientific Reports*, vol. 14, 2024, doi:10.1038/s41598-024-52703-2.

[29] G. I. Okolo, S. Katsigiannis, and N. Ramzan, "Ievit: An enhanced vision transformer architecture for chest x-ray image classification," *Computer Methods and Programs in Biomedicine*, vol. 226, 2022, doi:10.1016/j.cmpb.2022.107141.

[30] O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, and A. Ayatollahi, "Medvit: A robust vision transformer for generalized medical image classification," *Computers in Biology and Medicine*, vol. 157, 2023, doi:10.1016/j.compbiomed.2023.106791.

[31] C. S. Raghaw, T. J. Sahu, P. Darade, S. S. Dar, and N. Kumar, "Quantum-guided textural–spatial fusion network for thyroid nodule classification in ultrasound imaging," *International Journal of Imaging Systems and Technology*, vol. 36, 2026, doi:10.1002/ima.70428.

[32] M. Fang, M. Fu, B. Liao, X. Lei, and F.-X. Wu, "Deep integrated fusion of local and global features for cervical cell classification," *Computers in Biology and Medicine*, vol. 171, p. 108153, 2024, doi:10.1016/j.compbiomed.2024.108153.

[33] S. Liu, L. Wang, and W. Yue, "An efficient medical image classification network based on multi-branch cnn, token grouping transformer and mixer mlp," *Applied Soft Computing*, vol. 153, p. 111323, 2024, doi:10.1016/j.asoc.2024.111323.

[34] F. Hörst, M. Rempe, L. Heine, C. Seibold, J. Keyl, G. Baldini, S. Ugurel, J. Siveke, B. Grünwald, J. Egger, and J. Kleesiek, "Cellvit: Vision transformers for precise cell segmentation and classification," *Medical Image Analysis*, vol. 94, p. 103143, 2024, doi:10.1016/j.media.2024.103143.

[35] P. S. Mann, S. D. Panchal, S. Singh, G. S. Saggu, and K. Gupta, "A hybrid deep convolutional neural network model for improved diagnosis of pneumonia," *Neural Computing and Applications*, vol. 36, pp. 1791–1804, 2024, doi:10.1007/s00521-023-09147-y.

[36] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 2017, pp. 1–17. [Online]. Available: https://openreview.net/forum?id=BJ6oOfqge

[37] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, vol. 2017-December, 2017, p. 1195–1204. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/68053af2923e00204c3ca7c6a3150cf7-Paper.pdf

[38] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020. [Online]. Available: https://papers.nips.cc/paper/2020/file/06964dce9addb1c5cb5d6e3d9838f733-Paper.pdf

[39] J. Kim, Y. Hur, S. Park, E. Yang, S. J. Hwang, and J. Shin, "Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 14 567–14 579, 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/a7968b4339a1b85b7dbdb362dc44f9c4-Paper.pdf

[40] H. Lee, S. Shin, and H. Kim, "Abc: Auxiliary balanced classifier for class-imbalanced semi-supervised learning," *Advances in Neural Information Processing Systems*,

vol. 34, pp. 7082–7094, 2021. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/3953630da28e5181cffca1278517e3cf-Paper.pdf

[41] M. Berrimi, D. Hans, and R. Jennane, "A semi-supervised multiview-mri network for the detection of knee osteoarthritis," *Computerized Medical Imaging and Graphics*, vol. 114, p. 102371, Jun. 2024, doi:10.1016/j.compmedimag.2024.102371.

[42] M. U. Farooq, Z. Ullah, A. Khan, and J. Gwak, "Dc-aae: Dual channel adversarial autoencoder with multitask learning for kl-grade classification in knee radiographs," *Computers in Biology and Medicine*, vol. 167, p. 107570, Dec. 2023, doi:10.1016/j.compbiomed.2023.107570.

[43] S. Khan, M. Sajjad, N. Abbas, J. Escorcia-Gutierrez, M. Gamarra, and K. Muhammad, "Efficient leukocytes detection and classification in microscopic blood images using convolutional neural network coupled with a dual attention network," *Computers in Biology and Medicine*, p. 108146, 2024, doi:10.1016/j.compbiomed.2024.108146.

[44] N. A. Atasoy and A. F. A. A. Rahhawi, "Examining the classification performance of pre-trained capsule networks on imbalanced bone marrow cell dataset," *International Journal of Imaging Systems and Technology*, vol. 34, 2024, doi:10.1002/ima.23067.

[45] B. Chen, F. Qin, Y. Shao, J. Cao, Y. Peng, and R. Ge, "Fine-grained imbalanced leukocyte classification with global-local attention transformer," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, 2023, doi:10.1016/j.jksuci.2023.101661.

[46] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, vol. 9351, 2015, pp. 234–241, doi:10.1007/978-3-319-24574-4_28.

[47] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11045 LNCS, 2018, pp. 3–11, doi:10.1007/978-3-030-00889-5_1.

[48] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, 2020, doi:10.1016/j.isprsjprs.2020.01.013.

[49] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. D. Lange, P. Halvorsen, and H. D. Johansen, "Resunet++: An advanced architecture for medical image segmentation," in *Proceedings - 2019 IEEE International Symposium on Multimedia, ISM 2019*, 2019, pp. 225–2255, doi:10.1109/ISM46123.2019.00049.

[50] Z. Zhu, K. Yu, G. Qi, B. Cong, Y. Li, Z. Li, and X. Gao, "Lightweight medical image segmentation network with multi-scale feature-guided fusion," *Computers in Biology and Medicine*, vol. 182, 2024, doi:10.1016/j.compbiomed.2024.109204.

[51] Y. Li, Z. Wang, L. Yin, Z. Zhu, G. Qi, and Y. Liu, "X-net: a dual encoding–decoding method in medical image segmentation," *Visual Computer*, vol. 39, 2023, doi:10.1007/s00371-021-02328-7.

[52] R. Agarwal, P. Ghosal, A. K. Sadhu, N. Murmu, and D. Nandi, "Multi-scale dual-channel feature embedding decoder for biomedical image segmentation," *Computer Methods and Programs in Biomedicine*, vol. 257, 2024, doi:10.1016/j.cmpb.2024.108464.

[53] R. Agarwal, A. Chowdhury, R. K. Chatterjee, H. Chel, C. Murmu, N. Murmu, and D. Nandi, "Deep quasi-recurrent self-attention with dual encoder-decoder in biomedical ct image segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, 2024, doi:10.1109/jbhi.2024.3447689.

[54] D. Cheng, Z. Zhou, and J. Zhang, "Eg-unetr: An edge-guided liver tumor segmentation network based on cross-level interactive transformer," *Biomedical Signal Processing and Control*, vol. 97, 2024, doi:10.1016/j.bspc.2024.106739.

[55] Y. Chen, C. Zheng, T. Zhou, L. Feng, L. Liu, Q. Zeng, and G. Wang, "A deep residual attention-based u-net with a biplane joint method for liver segmentation from ct scans," *Computers in Biology and Medicine*, vol. 152, 2023, doi:10.1016/j.compbiomed.2022.106421.

[56] D. Y. Tu, P. C. Lin, H. H. Chou, M. R. Shen, and S. Y. Hsieh, "Slice-fusion: Reducing false positives in liver tumor detection for mask r-cnn," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 20, 2023, doi:10.1109/TCBB.2023.3265394.

[57] S. Di, Y. Zhao, M. Liao, Z. Yang, and Y. Zeng, "Automatic liver tumor segmentation from ct images using hierarchical iterative superpixels and local statistical features," *Expert Systems with Applications*, vol. 203, 2022, doi:10.1016/j.eswa.2022.117347.

[58] J. Zhao, X. Hou, M. Pan, and H. Zhang, "Attention-based generative adversarial network in medical imaging: A narrative review," *Computers in Biology and Medicine*, vol. 149, 2022, doi:10.1016/j.compbiomed.2022.105948.

[59] W. Wu, K. Hu, W. Yue, W. Li, M. Simic, C. Li, W. Xiang, and Z. Wang, "Self-supervised multimodal fusion network for knee osteoarthritis severity grading," in *2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, vol. 8. IEEE, Nov. 2023, p. 57–64, doi:10.1109/dicta60407.2023.00017.

[60] N. K. Tomar, D. Jha, M. A. Riegler, H. D. Johansen, D. Johansen, J. Rittscher, P. Halvorsen, and S. Ali, "Fanet: A feedback attention network for improved biomedical image segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, 2023, doi:10.1109/TNNLS.2022.3159394.

[61] J. Zhao and S. Li, "Uncertainty-guided and cross-modality attention network for liver tumor segmentation and quantification via integrating dynamic mri," *Knowledge-Based Systems*, vol. 310, 2025, doi:10.1016/j.knosys.2025.113021.

[62] K.-N. Wang, S.-X. Li, Z. Bu, F.-X. Zhao, G.-Q. Zhou, S.-J. Zhou, and Y. Chen, "Sbcnet: Scale and boundary context attention dual-branch network for liver tumor segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, 2024, doi:10.1109/JBHI.2024.3370864.

[63] S. Di, Y. Q. Zhao, M. Liao, F. Zhang, and X. Li, "Td-net: A hybrid end-to-end network for automatic liver tumor segmentation from ct images," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, 2023, doi:10.1109/JBHI.2022.3181974.

[64] H. Liu, Y. Fu, S. Zhang, J. Liu, Y. Wang, G. Wang, and J. Fang, "Gcha-net: Global context and hybrid attention network for automatic liver segmentation," *Computers in Biology and Medicine*, vol. 152, 2023, doi:10.1016/j.compbiomed.2022.106352.

[65] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 7132–7141, 2020, doi:10.1109/TPAMI.2019.2913372.

[66] Y. Su, J. Cheng, C. Zhong, Y. Zhang, J. Ye, J. He, and J. Liu, "Fednet: Feature decoupled network for polyp segmentation from endoscopy images," *Biomedical Signal Processing and Control*, vol. 83, p. 104699, 2023, doi:10.1016/j.bspc.2023.104699.

[67] L. Yang, C. Zhai, Y. Liu, and H. Yu, "Cfha-net: A polyp segmentation method with cross-scale fusion strategy and hybrid attention," *Computers in Biology and Medicine*, vol. 164, p. 107301, 2023, doi:10.1016/j.compbiomed.2023.107301.

[68] D. P. Fan, G. P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Pranet: Parallel reverse attention network for polyp segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, vol. 12266 LNCS, 2020, pp. 263–273, doi:10.1007/978-3-030-59725-2_26.

[69] C. Kaul, S. Manandhar, and N. Pears, "Focusnet: An attention-based fully convolutional network for medical image segmentation," in *Proceedings - International Symposium on Biomedical Imaging*, vol. 2019-April, 2019, pp. 455–458, doi:10.1109/ISBI.2019.8759477.

[70] Z. Hu, H. Chen, L. Hua, X. Ren, and W. Mei, "Msml-attunet: A hierarchical attention network with multi-scale and multi-task for precision liver tumor segmentation," *Biomedical Signal Processing and Control*, vol. 99, 2025, doi:10.1016/j.bspc.2024.106861.

[71] L. Jiang, J. Ou, R. Liu, Y. Zou, T. Xie, H. Xiao, and T. Bai, "Rmau-net: Residual multi-scale attention u-net for liver and tumor segmentation in ct images," *Computers in Biology and Medicine*, vol. 158, 2023, doi:10.1016/j.compbiomed.2023.106838.

[72] Y. Chen, C. Zheng, W. Zhang, H. Lin, W. Chen, G. Zhang, G. Xu, and F. Wu, "Msfanet: Multi-scale feature attention network for liver tumor segmentation," *Computers in Biology and Medicine*, vol. 163, 2023, doi:10.1016/j.compbiomed.2023.107208.

[73] T. Lei, R. Wang, Y. Zhang, Y. Wan, C. Liu, and A. K. Nandi, "Defed-net: Deformable encoder-decoder network for liver and liver tumor segmentation," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 6, 2022, doi:10.1109/TRPMS.2021.3059780.

[74] Y. Chen, C. Zheng, F. Hu, T. Zhou, L. Feng, G. Xu, Z. Yi, and X. Zhang, "Efficient two-step liver and tumour segmentation on abdominal ct via deep learning and a conditional random field," *Computers in Biology and Medicine*, vol. 150, 2022, doi:10.1016/j.compbiomed.2022.106076.

[75] J. M. J. Valanarasu, V. A. Sindagi, I. Hacihaliloglu, and V. M. Patel, "Kiu-net: Overcomplete convolutional architectures for biomedical image and volumetric segmentation," *IEEE Transactions on Medical Imaging*, vol. 41, 2022, doi:10.1109/TMI.2021.3130469.

[76] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 3349–3364, 2021, doi:10.1109/TPAMI.2020.2983686.

[77] E. Sanderson and B. J. Matuszewski, "Fcn-transformer feature fusion for polyp segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13413 LNCS, 2022, pp. 892–907, doi:10.1007/978-3-031-12053-4_65.

[78] A. Srivastava, D. Jha, S. Chanda, U. Pal, H. Johansen, D. Johansen, M. Riegler, S. Ali, and P. Halvorsen, "Msrf-net: A multi-scale residual fusion network for biomedical image segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, pp. 2252–2263, 2022, doi:10.1109/JBHI.2021.3138024.

[79] C. C. Ukwuoma, Z. Qin, M. B. B. Heyat, F. Akhtar, O. Bamisile, A. Y. Muaad, D. Addo, and M. A. Al-antari, "A hybrid explainable ensemble transformer encoder for pneumonia identification from chest x-ray images," *Journal of Advanced Research*, vol. 48, 2023, doi:10.1016/j.jare.2022.08.021.

[80] Y. Yang, G. Mei, and F. Piccialli, "A deep learning approach considering image background for pneumonia identification using explainable ai (xai)," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2022, doi:10.1109/TCBB.2022.3190265.

[81] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, p. 336–359, Oct. 2019, doi:10.1007/s11263-019-01228-7.

[82] N. A. Hroub, A. N. Alsannaa, M. Alowaifeer, M. Alfarraj, and E. Okafor, "Explainable deep learning diagnostic system for prediction of lung disease from medical images," *Computers in Biology and Medicine*, vol. 170, p. 108012, 2024, doi:10.1016/j.compbiomed.2024.108012.

[83] S. Desai and H. G. Ramaswamy, "Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization," in *Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020*, 2020, pp. 983–991, doi:10.1109/WACV45572.2020.9093360.

[84] M. B. Muhammad and M. Yeasin, "Eigen-cam: Class activation map using principal components," in *Proceedings of the International Joint Conference on Neural Networks*, 2020, pp. 1–7, doi:10.1109/IJCNN48605.2020.9206626.

[85] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018*, vol. 2018-January, 2018, pp. 839–847, doi:10.1109/WACV.2018.00097.

[86] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 2020-June, 2020, pp. 24–25, doi:10.1109/CVPRW50498.2020.00020.

[87] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, vol. 13-17-August-2016, 2016, pp. 1135–1144, doi:10.1145/2939672.2939778.

[88] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10 2023, pp. 3992–4003, doi:10.1109/ICCV51070.2023.00371.

[89] S. Mirjalili, "Sca: A sine cosine algorithm for solving optimization problems," *Knowledge-Based Systems*, vol. 96, 2016, doi:10.1016/j.knosys.2015.12.022.

[90] S. Gupta and A. Panwar, "Deep models for analysis of pneumonia infection using chest radiographs," in *Proceedings of International Conference on Information Technology and Applications: ICITA 2021*, vol. 350, 2022, pp. 263–272, doi:10.1007/978-981-16-7618-5_23.

[91] R. D. Labati, V. Piuri, and F. Scotti, "All-idb: The acute lymphoblastic leukemia image database for image processing," in *2011 18th IEEE international conference on image processing*, 2011, pp. 2045–2048, doi:10.1109/ICIP.2011.6115881.

[92] A. Gupta, R. Duggal, S. Gehlot, R. Gupta, A. Mangal, L. Kumar, N. Thakkar, and D. Satpathy, "Gcti-sn: Geometry-inspired chemical and tissue invariant stain normalization of microscopic medical images," *Medical Image Analysis*, vol. 65, 2020, doi:10.1016/j.media.2020.101788.

[93] R. Khandekar, P. Shastry, S. Jaishankar, O. Faust, and N. Sampathila, "Automated blast cell detection for acute lymphoblastic leukemia diagnosis," *Biomedical Signal Processing and Control*, vol. 68, 2021, doi:10.1016/j.bspc.2021.102690.

[94] K. Gokulkannan, T. A. Mohanaprakash, J. DafniRose, and B. Sriman, "Multiscale adaptive and attention-dilated convolutional neural network for efficient leukemia detection model with multiscale trans-res-unet3+ -based segmentation network," *Biomedical Signal Processing and Control*, vol. 90, p. 105847, 2024, doi:10.1016/j.bspc.2023.105847.

[95] A. K. Al-Bashir, R. E. Khnouf, and L. R. B. Issa, "Leukemia classification using different cnn-based algorithms-comparative study," *Neural Computing and Applications*, pp. 1–16, 2024, doi:10.1007/s00521-024-09554-9.

[96] D. S. Depto, M. M. Rizvee, A. Rahman, H. Zunair, M. S. Rahman, and M. R. Mahdy, "Quantifying imbalanced classification methods for leukemia detection," *Computers in Biology and Medicine*, vol. 152, 2023, doi:10.1016/j.compbiomed.2022.106372.

[97] P. K. Das, B. Sahoo, and S. Meher, "An efficient detection and classification of acute leukemia using transfer learning and orthogonal softmax layer-based model," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2022, doi:10.1109/TCBB.2022.3218590.

[98] M. Jawahar, S. H, J. A. L, and A. H. Gandomi, "Alnett: A cluster layer deep convolutional neural network for acute lymphoblastic leukemia classification," *Computers in Biology and Medicine*, vol. 148, 2022, doi:10.1016/j.compbiomed.2022.105894.

[99] S. Gehlot, A. Gupta, and R. Gupta, "Sdct-auxnetθ: Dct augmented stain deconvolutional cnn with auxiliary classifier for cancer diagnosis," *Medical Image Analysis*, vol. 61, 2020, doi:10.1016/j.media.2020.101661.

[100] U. Kadak, "Fractional sampling operators of multivariate fuzzy functions and applications to image processing," *Applied Soft Computing*, vol. 132, 2023, doi:10.1016/j.asoc.2022.109901.

[101] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M. H. Yang, and L. Shao, "Learning enriched features for fast image restoration and enhancement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, 2023, doi:10.1109/TPAMI.2022.3167175.

[102] H. Rahman and G. C. Paul, "Tripartite sub-image histogram equalization for slightly low contrast gray-tone image enhancement: Slightly low contrast image enhancement with tsihe," *Pattern Recognition*, vol. 134, 2023, doi:10.1016/j.patcog.2022.109043.

[103] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, vol. 11045 LNCS, 2018, pp. 3–11, doi:10.1007/978-3-030-00889-5_1.

[104] Y. Chen, X. H. Yang, Z. Wei, A. A. Heidari, N. Zheng, Z. Li, H. Chen, H. Hu, Q. Zhou, and Q. Guan, "Generative adversarial networks in medical image augmentation: A review," *Computers in Biology and Medicine*, vol. 144, 2022, doi:10.1016/j.compbiomed.2022.105382.

[105] Z. Xu, J. Tang, C. Qi, D. Yao, C. Liu, Y. Zhan, and T. Lukasiewicz, "Cross-domain attention-guided generative data augmentation for medical image analysis with limited data," *Computers in Biology and Medicine*, vol. 168, p. 107744, 2024, doi:10.1016/j.compbiomed.2023.107744.

[106] F. J. M. Shamrat, S. Azam, A. Karim, K. Ahmed, F. M. Bui, and F. D. Boer, "High-precision multiclass classification of lung disease through customized mobilenetv2 from chest x-ray images," *Computers in Biology and Medicine*, vol. 155, 2023, doi:10.1016/j.compbiomed.2023.106646.

[107] A. M. Rifai, S. Raharjo, E. Utami, and D. Ariatmanto, "Analysis for diagnosis of pneumonia symptoms using chest x-ray based on mobilenetv2 models with image enhancement using white balance and contrast limited adaptive histogram equalization (clahe)," *Biomedical Signal Processing and Control*, vol. 90, p. 105857, 2024, doi:10.1016/j.bspc.2023.105857.

[108] S. Rajaraman, G. Zamzmi, L. Folio, P. Alderson, and S. Antani, "Chest x-ray bone suppression for improving classification of tuberculosis-consistent findings," *Diagnostics*, vol. 11, 2021, doi:10.3390/diagnostics11050840.

[109] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, 2014. [Online]. Available: https://www.jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf

[110] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *32nd International Conference on Machine Learning, ICML 2015*, vol. 1, 2015, pp. 448–456. [Online]. Available: https://proceedings.mlr.press/v37/ioffe15.pdf

[111] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, vol. 139, 2021, pp. 8748–8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a/radford21a.pdf

[112] X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang, W. Yang, J. Huang, and X. Han, "Transformer-based unsupervised contrastive learning for histopathological image classification," *Medical Image Analysis*, vol. 81, 2022, doi:10.1016/j.media.2022.102559.

[113] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M. K. Prasadha, J. Pei, M. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, I. Ziyar, A. Shi, R. Zhang, L. Zheng, R. Hou, W. Shi,

X. Fu, Y. Duan, V. A. Huu, C. Wen, E. D. Zhang, C. L. Zhang, O. Li, X. Wang, M. A. Singer, X. Sun, J. Xu, A. Tafreshi, M. A. Lewis, H. Xia, and K. Zhang, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, 2018, doi:10.1016/j.cell.2018.02.010.

[114] H. H. Pham, N. H. Nguyen, T. T. Tran, T. N. Nguyen, and H. Q. Nguyen, "Pedicxr: An open, large-scale chest radiograph dataset for interpretation of common thoracic diseases in children," *Scientific Data*, vol. 10, 2023, doi:10.1038/s41597-023-02102-5.

[115] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, 2017, pp. 2097–2106, doi:10.1109/CVPR.2017.369.

[116] R. Yi, L. Tang, Y. Tian, J. Liu, and Z. Wu, "Identification and classification of pneumonia disease using a deep learning-based intelligent computational framework," *Neural Computing and Applications*, vol. 35, 2023, doi:10.1007/s00521-021-06102-7.

[117] E. Ayan, B. Karabulut, and H. M. Ünver, "Diagnosis of pediatric pneumonia with ensemble of deep convolutional neural networks in chest x-ray images," *Arabian Journal for Science and Engineering*, vol. 47, 2022, doi:10.1007/s13369-021-06127-z.

[118] S. Chattopadhyay, R. Kundu, P. K. Singh, S. Mirjalili, and R. Sarkar, "Pneumonia detection from lung x-ray images using local search aided sine cosine algorithm based deep feature selection method," *International Journal of Intelligent Systems*, vol. 37, 2022, doi:10.1002/int.22703.

[119] J. A. Prakash, C. R. Asswin, K. S. D. Kumar, A. Dora, V. Ravi, V. Sowmya, E. A. Gopalakrishnan, and K. P. Soman, "Transfer learning approach for pediatric pneumonia diagnosis using channel attention deep cnn architectures," *Engineering Applications of Artificial Intelligence*, vol. 123, 2023, doi:10.1016/j.engappai.2023.106416.

[120] M. Kaya, "Feature fusion-based ensemble cnn learning optimization for automated detection of pediatric pneumonia," *Biomedical Signal Processing and Control*, vol. 87, 2024, doi:10.1016/j.bspc.2023.105472.

[121] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, 2020, doi:10.1109/TMI.2019.2959609.

[122] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jul. 2017, p. 6517–6525, doi:10.1109/cvpr.2017.690.

[123] C. Lindner, S. Thiagarajah, J. Wilkinson, T. Consortium, G. Wallis, and T. Cootes, "Fully automatic segmentation of the proximal femur using random forest regression voting," *IEEE Transactions on Medical Imaging*, vol. 32, no. 8, p. 1462–1472, Aug. 2013, doi:10.1109/tmi.2013.2258030.

[124] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, Jun. 2020, pp. 702–703, doi:10.1109/cvprw50498.2020.00359.

[125] Z. Song, P. Du, J. Yan, K. Li, J. Shou, M. Lai, Y. Fan, and Y. Xu, "Nucleus-aware self-supervised pretraining using unpaired image-to-image translation for histopathology images," *IEEE Transactions on Medical Imaging*, vol. 43, no. 1, p. 459–472, Jan. 2024, doi:10.1109/tmi.2023.3309971.

[126] Z. Fei, M. Fan, L. Zhu, J. Huang, X. Wei, and X. Wei, "Masked auto-encoders meet generative adversarial networks and beyond," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2023, p. 24449–24459, doi:10.1109/cvpr52729.2023.02342.

[127] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick, "Masked autoencoders are scalable vision learners," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2022, pp. 16 000–16 009, doi:10.1109/cvpr52688.2022.01553.

[128] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, p. 139–144, Oct. 2020, doi:10.1145/3422622.

[129] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Procedings of the British Machine Vision Conference 2016*, ser. BMVC 2016. British Machine Vision Association, 2016, pp. 87.1–87.12, doi:10.5244/c.30.87.

[130] A. Mari, T. R. Bromley, J. Izaac, M. Schuld, and N. Killoran, "Transfer learning in hybrid classical-quantum neural networks," *Quantum*, vol. 4, p. 340, Oct. 2020, doi:10.22331/q-2020-10-09-340.

[131] M. Benedetti, E. Lloyd, S. Sack, and M. Fiorentini, "Parameterized quantum circuits as machine learning models," *Quantum Science and Technology*, vol. 4, no. 4, p. 043001, Nov. 2019, doi:10.1088/2058-9565/ab4eb5.

[132] E. Ovalle-Magallanes, J. G. Avina-Cervantes, I. Cruz-Aceves, and J. Ruiz-Pinales, "Hybrid classical–quantum convolutional neural network for stenosis detection in x-ray coronary angiography," *Expert Systems with Applications*, vol. 189, p. 116112, Mar. 2022, doi:10.1016/j.eswa.2021.116112.

[133] Z. Ning, Z. Jiang, and D. Zhang, "To combat multiclass imbalanced problems by aggregating evolutionary hierarchical classifiers," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 3, p. 5258–5272, Mar. 2025, doi:10.1109/tnnls.2024.3383672.

[134] P. Chen, "Knee osteoarthritis severity grading dataset," Mendeley Data, V1, 2018. [Online]. Available: https://data.mendeley.com/datasets/56rmx5bjcr/1

[135] S. Gornale and P. Patravali, "Digital knee x-ray images," Mendeley Data, V1, 2020. [Online]. Available: https://data.mendeley.com/datasets/t9ndx37v5h/1

[136] M. Tao, "Osteoarthritis prediction," IEEE DataPort, 2023, doi:10.21227/ewx7-b315.

[137] D. Shaw, "Osteoarthritis," IEEE DataPort, 2024, doi:10.21227/mszn-gr21.

[138] H. H. Nguyen, M. B. Blaschko, S. Saarakkala, and A. Tiulpin, "Clinically-inspired multi-agent transformers for disease trajectory forecasting from multimodal data," *IEEE Transactions on Medical Imaging*, vol. 43, no. 1, p. 529–541, Jan. 2024, doi:10.1109/tmi.2023.3312524.

[139] C.-M. Lo and K.-L. Lai, "Deep learning-based assessment of knee septic arthritis using transformer features in sonographic modalities," *Computer Methods and Programs in Biomedicine*, 2023, doi:10.1016/j.cmpb.2023.107575.

[140] Y. X. Teoh, A. Othmani, K. W. Lai, S. L. Goh, and J. Usman, "Stratifying knee osteoarthritis features through multitask deep hybrid learning: Data from the osteoarthritis initiative," *Computer Methods and Programs in Biomedicine*, vol. 242, p. 107807, Dec. 2023, doi:10.1016/j.cmpb.2023.107807.

[141] K. Hu, W. Wu, W. Li, M. Simic, A. Zomaya, and Z. Wang, "Adversarial evolving neural network for longitudinal knee osteoarthritis prediction," *IEEE Transactions on Medical Imaging*, vol. 41, no. 11, p. 3207–3217, Nov. 2022, doi:10.1109/tmi.2022.3181060.

[142] S. Azizi, L. Culp, J. Freyberg, B. Mustafa, S. Baur, S. Kornblith, T. Chen, N. Tomasev, J. Mitrović, P. Strachan, S. S. Mahdavi, E. Wulczyn, B. Babenko, M. Walker, A. Loh, P.-H. C. Chen, Y. Liu, P. Bavishi, S. M. McKinney, J. Winkens, A. G. Roy, Z. Beaver, F. Ryan, J. Krogue, M. Etemadi, U. Telang, Y. Liu, L. Peng, G. S. Corrado, D. R. Webster, D. Fleet, G. Hinton, N. Houlsby, A. Karthikesalingam, M. Norouzi, and V. Natarajan, "Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging," *Nature Biomedical Engineering*, vol. 7, no. 6, p. 756–779, Jun. 2023, doi:10.1038/s41551-023-01049-7.

[143] Y. Cai, H. Chen, X. Yang, Y. Zhou, and K.-T. Cheng, "Dual-distribution discrepancy with self-supervised refinement for anomaly detection in medical images," *Medical Image Analysis*, vol. 86, p. 102794, May 2023, doi:10.1016/j.media.2023.102794.

[144] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, V. Natarajan, and M. Norouzi, "Big self-supervised models advance medical image classification," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2021, p. 3458–3468, doi:10.1109/iccv48922.2021.00346.

[145] J. Huo, X. Ouyang, L. Si, K. Xuan, S. F. Wang, W. Yao, Y. Liu, J. Xu, D. Qian, Z. Xue, Q. Wang, D. Shen, and L. Zhang, "Automatic grading assessments for knee mri cartilage defects via self-ensembling semi-supervised learning with dual-consistency," *Medical Image Analysis*, vol. 80, p. 102508, Aug. 2022, doi:10.1016/j.media.2022.102508.

[146] W. Burton, C. Myers, and P. Rullkoetter, "Semi-supervised learning for automatic segmentation of the knee from mri with convolutional neural networks," *Computer Methods and Programs in Biomedicine*, vol. 189, p. 105328, Jun. 2020, doi:10.1016/j.cmpb.2020.105328.

[147] H. H. Nguyen, S. Saarakkala, M. B. Blaschko, and A. Tiulpin, "Semixup: In- and out-of-manifold regularization for deep semi-supervised knee osteoarthritis severity grading from plain radiographs," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, p. 4346–4356, Dec. 2020, doi:10.1109/tmi.2020.3017007.

[148] S. Poppi, M. Cornia, L. Baraldi, and R. Cucchiara, "Revisiting the evaluation of class activation mapping for explainability: A novel metric and experimental analysis," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2021, pp. 2299–2304, doi:10.1109/CVPRW53098.2021.00260.

[149] Z. Huang, X. Jin, C. Lu, Q. Hou, M.-M. Cheng, D. Fu, X. Shen, and J. Feng, "Contrastive masked autoencoders are stronger vision learners," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, p. 2506–2517, Apr. 2024, doi:10.1109/tpami.2023.3336525.

[150] C. Prabhakar, H. Li, J. Yang, S. Shit, B. Wiestler, and B. Menze, "Vit-ae++: improving vision transformer autoencoder for self-supervised medical image representations," in *Medical Imaging with Deep Learning*. PMLR, 2024, pp. 666–679. [Online]. Available: https://proceedings.mlr.press/v227/prabhakar24b/prabhakar24b.pdf

[151] M. Yeung, E. Sala, C. B. Schönlieb, and L. Rundo, "Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation," *Computerized Medical Imaging and Graphics*, vol. 95, 2022, doi:10.1016/j.compmedimag.2021.102026.

[152] P. Bilic, P. Christ, H. B. Li *et al.*, "The liver tumor segmentation benchmark (lits)," *Medical Image Analysis*, vol. 84, p. 102680, 2 2023, doi:10.1016/J.media.2022.102680.

[153] L. Soler, A. Hostettler, V. Agnus, A. Charnoz, J.-B. Fasquel, J. Moreau, A.-B. Osswald, M. Bouhadjar, and J. Marescaux, "3d image reconstruction for comparison of algorithm database," 2010. [Online]. Available: https://www.ircad.fr/research/data-sets/liver-segmentation-3d-ircadb-01/

[154] F. Quinton, R. Popoff, B. Presles, S. Leclerc, F. Meriaudeau, G. Nodari, O. Lopez, J. Pellegrinelli, O. Chevallier, D. Ginhac, J. M. Vrigneaud, and J. L. Alberini, "A tumour and liver automatic segmentation (atlas) dataset on contrast-enhanced magnetic resonance imaging for hepatocellular carcinoma," *Data*, vol. 8, 2023, doi:10.3390/data8050079.

[155] M. Liao, H. Tang, X. Li, P. Vijayakumar, V. Arya, and B. B. Gupta, "A lightweight network for abdominal multi-organ segmentation based on multi-scale context fusion and dual self-attention," *Information Fusion*, vol. 108, p. 102401, 2024, doi:10.1016/j.inffus.2024.102401.

[156] Y. Li, B. Zou, P. Dai, M. Liao, H. X. Bai, and Z. Jiao, "Ac-e network: Attentive context-enhanced network for liver segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, 2023, doi:10.1109/JBHI.2023.3278079.

[157] D. T. Kushnure and S. N. Talbar, "Hfru-net: High-level feature fusion and recalibration unet for automatic liver and tumor segmentation in ct images," *Computer Methods and Programs in Biomedicine*, vol. 213, 2022, doi:10.1016/j.cmpb.2021.106501.

[158] J. Li, G. Huang, J. He, Z. Chen, C. M. Pun, Z. Yu, W. K. Ling, L. Liu, J. Zhou, and J. Huang, "Shift-channel attention and weighted-region loss function for liver and dense tumor segmentation," *Medical Physics*, vol. 49, 2022, doi:10.1002/mp.15816.

[159] Z. Qian, Y. Lv, D. Lv, H. Gu, K. Wang, W. Zhang, and M. M. Gupta, "A new approach to polyp detection by pre-processing of images and enhanced faster r-cnn," *IEEE Sensors Journal*, vol. 21, 2021, doi:10.1109/JSEN.2020.3036005.

[160] T. J. Jaspers, T. G. Boers, C. H. Kusters, M. R. Jong, J. B. Jukema, A. J. de Groof, J. J. Bergman, P. H. de With, and F. van der Sommen, "Robustness evaluation of deep neural networks for endoscopic image analysis: Insights and strategies," *Medical Image Analysis*, vol. 94, p. 103157, 2024, doi:10.1016/j.media.2024.103157.

[161] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, vol. 11962 LNCS, 2020, pp. 451–462, doi:10.1007/978-3-030-37734-2_37.

[162] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015, doi:10.1016/j.compmedimag.2015.02.007.

[163] E. Smistad, T. L. Falch, M. Bozorgi, A. C. Elster, and F. Lindseth, "Medical image segmentation on gpus - a comprehensive review," *Medical Image Analysis*, vol. 20, 2015, doi:10.1016/j.media.2014.10.012.

[164] Z. Liu, S. Zheng, X. S. Brandon, Z. Zhu, Y. Zhao, X. Yang, and Y. Zhao, "The devil is in the boundary: Boundary-enhanced polyp segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, 2024, doi:10.1109/TCSVT.2023.3348598.

[165] B. L. Chen, J. J. Wan, T. Y. Chen, Y. T. Yu, and M. Ji, "A self-attention based faster r-cnn for polyp detection from colonoscopy images," *Biomedical Signal Processing and Control*, vol. 70, 2021, doi:10.1016/j.bspc.2021.103019.

[166] X. Jia, Y. Shen, J. Yang, R. Song, W. Zhang, M. Q. Meng, J. C. Liao, and L. Xing, "Polypmixnet: Enhancing semi-supervised polyp segmentation with polyp-aware augmentation," *Computers in Biology and Medicine*, vol. 170, 2024, doi:10.1016/j.compbiomed.2024.108006.

[167] J. Zheng, Y. Yan, L. Zhao, and X. Pan, "Cgma-net: Cross-level guidance and multi-scale aggregation network for polyp segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, pp. 1424–1435, 2024, doi:10.1109/JBHI.2023.3345479.

[168] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 834–848, 2018, doi:10.1109/TPAMI.2017.2699184.

[169] Y. Wu and K. He, "Group normalization," *International Journal of Computer Vision*, vol. 128, pp. 3–19, 2020, doi:10.1007/s11263-019-01198-w.

[170] S. Feng, H. Zhao, F. Shi, X. Cheng, M. Wang, Y. Ma, D. Xiang, W. Zhu, and X. Chen, "Cpfnet: Context pyramid fusion network for medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, 2020, doi:10.1109/TMI.2020.2983721.

[171] S. Pan, X. Liu, N. Xie, and Y. Chong, "Eg-transunet: a transformer-based u-net with enhanced and guided models for biomedical image segmentation," *BMC Bioinformatics*, vol. 24, p. 85, 2023, doi:10.1186/s12859-023-05196-1.

[172] A. Haider, M. Arsalan, S. H. Nam, J. S. Hong, H. Sultan, and K. R. Park, "Multi-scale feature retention and aggregation for colorectal cancer diagnosis using gastrointestinal images," *Engineering Applications of Artificial Intelligence*, vol. 125, 2023, doi:10.1016/j.engappai.2023.106749.

[173] R. G. Dumitru, D. Peteleaza, and C. Craciun, "Using duck-net for polyp image segmentation," *Scientific Reports*, vol. 13, p. 9803, 2023, doi:10.1038/s41598-023-36940-5.

[174] P. Sharma, A. Gautam, P. Maji, R. B. Pachori, and B. K. Balabantaray, "Li-segpnet: Encoder-decoder mode lightweight segmentation network for colorectal polyps analysis," *IEEE Transactions on Biomedical Engineering*, vol. 70, pp. 1330–1339, 2023, doi:10.1109/TBME.2022.3216269.

[175] J. Zhu, M. Ge, Z. Chang, and W. Dong, "Crcnet: Global-local context and multi-modality cross attention for polyp segmentation," *Biomedical Signal Processing and Control*, vol. 83, 2023, doi:10.1016/j.bspc.2023.104593.

[176] N. K. Tomar, D. Jha, U. Bagci, and S. Ali, "Tganet: Text-guided attention for improved polyp segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 13433 LNCS, 2022, pp. 151–160, doi:10.1007/978-3-031-16437-8_15.