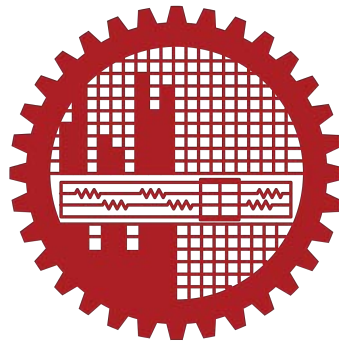


Efficient Deep Neural Network Architectures with Multi-receptive Feature Optimization for Multi-dimensional Data Processing

by
Tanvir Mahmud


Submitted to
Department of Electrical and Electronic Engineering
in fulfillment of the requirements for the degree of
Master of Science in Electrical and Electronic Engineering




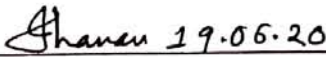
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Bangladesh
June 2021

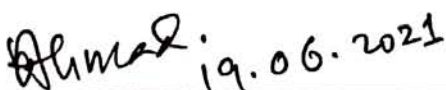
The thesis entitled "Efficient Deep Neural Network Architectures with Multi-receptive Feature Optimization for Multi-dimensional Data Processing", submitted by **Tanvir Mahmud**, Roll No. 1018062213, Session: October 2018, to the Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, has been accepted as satisfactory in partial fulfillment of the requirements for the degree of Master of Science in Electrical and Electronic Engineering and approved as to its style and contents. Examination held on **19 June, 2020**.

Board of Examiners

1. 

Dr. Shaikh Anowarul Fattah
Professor
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology, Dhaka.
Chairman
(Supervisor)
2. 

Dr. Md. Kamrul Hasan
Head and Professor
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology, Dhaka.
Member
(Ex-Officio)
3. 

Dr. Mohammed Imamul Hassan Bhuiyan
Professor
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology, Dhaka.
Member
4. 

Dr. Mohinuddin Ahmad
Professor
Department of Electrical and Electronic Engineering
Khulna University of Engineering & Technology, Khulna.
Member
(External)

Candidate's Declaration

This is declared that the work entitled "Efficient Deep Neural Network Architectures for Multi-receptive Feature Optimization for Multi-dimensional Data Processing" is the outcome of research carried out by me under the supervision of Dr. Shaikh Anowaul Fattah, in the Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, Dhaka 1000. It is also declared that this thesis or any part of it has not been submitted elsewhere for the award of any degree or diploma.

তানভীর মাহমুদ
২৭.০৬.২০২০

Tanvir Mahmud

Candidate

ID: 1018062213

Acknowledgement

I express my heartfelt gratitude to my supervisor, Dr. Shaikh Anowarul Fattah for his continuous support and guidelines throughout the journey. Starting from the very beginning to this far, I am indebted to him for his proper directions whenever I needed. To prepare myself for quality research works, I closely worked with him throughout all the phases including research idea evaluations, guided experimentations, manuscript preparation, and revision of the manuscripts that led to series of impactful contributions to the research community. I owe to him for helping me to bring out my inner potentials as a productive researcher that greatly influenced my perspectives towards any research problem. Working under his supervision at the beginning phase of my research career is a great privilege that will surely help me to establish myself as a prominent researcher in the coming days.

I would also want to thank the members of my thesis committee for their valuable suggestions. I thank Dr. Md. Kamrul Hasan, Dr. Mohammed Imamul Hassan Bhuiyan and specially the external member Dr. Mohiuddin Ahmad. I wish to give special thanks to Dr. Celia Shahnaz, for providing inspiration and guidance to walk in the right path.

I am very grateful to all my family members who consistently gave me support and inspiration to carry out the research works with proper enthusiasm.

Abstract

In this thesis, efficient deep neural network architectures are proposed for extracting features from diverse receptive fields by introducing numerous optimization, feature fusion, and data transformation schemes targeting numerous multi-dimensional applications. Firstly, to exploit effective features in data constrained environment from a single modality, feature learning from multiple perspectives is introduced through varying resolutions and novel data augmentation strategies. In addition, features from varying receptive fields have been extracted by introducing multi-kernel depthwise separable convolutions with varying dilation rates, and the performance is validated in the case of 1-D (electrocardiogram, ECG) and 2-D (chest X-ray image) data for disease classification. Afterward, feature spaces of multiple modalities have been explored by incorporating various transformed representations of multi-modal 1D time series sensor data. Moreover, a sequential training algorithm is proposed to gradually converge extracted features to the final objective and the overall scheme is deployed on human activity recognition application. For the purpose of multi-dimensional data segmentation (2D endoscopy images, and 3D CT volumes), instead of using conventional uni-scale feature propagation, multi-scale contextual feature aggregation and fusion-based building blocks are designed and incorporated in the DNN which offers improved feature sharing while minimizing the contextual information loss. Especially in the case of 3D data, a hybrid DNN architecture is proposed performing 2D slice-wise processing accompanied by lighter 3D-volumetric segmentation to reduce the complexity of the optimization process. Finally, a triple attention based learning scheme is proposed combining the channel, spatial, and pixel level attentions, which is incorporated in the DNN architectures to improve the feature sharing process targeting multiple objectives through hierarchical training and joint optimization. The proposed methods are validated in various multi-dimensional datasets targeting real-world applications such as disease classification, infection segmentation, disease severity prediction, and human activity recognition.

Contents

<i>Board of Examiners</i>	i
<i>Candidate's Declaration</i>	ii
<i>Acknowledgment</i>	iii
<i>Abstract</i>	iv
<i>Contents</i>	viii
<i>List of Figures</i>	xiii
<i>List of Tables</i>	xvi
1 Introduction	1
1.1 Multi-receptive Feature Optimization: Methods and Opportunities . . .	2
1.1.1 Architectural Modifications	3
1.1.2 Multi-transformed Data Representation	3
1.1.3 Multi-scale Contextual Feature Optimization for Segmenting Region-of-Interests	3
1.1.4 Multi-objective Learning with Multi-phase Optimization . . .	4
1.2 Multi-dimensional Data Processing in Real-world Applications	4
1.3 Literature Review	5
1.4 Objectives and Scopes	13
1.5 Organization of the Thesis	13
2 DeepArrnet: A Multi-receptive Neural Network for Arrhythmia Classification from ECG Beats	16
2.1 Preprocessing	17
2.1.1 Wavelet based denoising	17
2.1.2 R Peak Detection	18

2.1.3	ECG Beat Extraction	19
2.1.4	ECG Beat Augmentation	19
2.2	Proposed Deep Neural Network	21
2.2.1	Depthwise separable convolutions for 1D signal	21
2.2.2	Proposed Structural Unit	23
2.2.3	Proposed DeepArrNet Architecture	25
2.3	Results and Discussion	27
2.3.1	Dataset Description	27
2.3.2	Experimental Setup	27
2.3.3	Performance Evaluation	28
2.4	Conclusion	31
3	CovXNet: A Multi-dilation Neural Network for COVID-19 Diagnosis from Chest X-ray	32
3.1	Methodology	33
3.1.1	Preprocessing	34
3.1.2	Proposed Structural Units	35
3.1.3	Proposed CovXNet Architecture	37
3.1.4	Stacking of Multiple Networks	37
3.1.5	Proposed Transfer Learning Method on Novel Corona Virus Data Using CovXNet	38
3.2	Results and Discussions	39
3.2.1	Dataset Description	39
3.2.2	Experimental Setup	39
3.2.3	Performance Evaluation	40
3.3	Conclusion	46
4	Multi Stage Learning for Human Activity Recognition from Multimodal Wearable Sensors	48
4.1	Methodology	49
4.1.1	Transformations on Time Series Data	51
4.1.2	Proposed Deep Neural Network Architectures	53
4.1.3	Proposed Multi-Stage Training Scheme	54
4.1.4	Data Augmentation	58
4.2	Results and Discussions	58
4.2.1	Dataset Description	58
4.2.2	Experimental Setup	59
4.2.3	Performance Evaluation	59

4.3	Conclusion	63
5	PolypSegNet: A Modified Encoder Decoder Architecture for Polyp Segmentation from Endoscopy	64
5.1	Methodology	66
5.1.1	Proposed Depth Dilated Inception (DDI) Module	66
5.1.2	Proposed D-Unit Layer Structure	69
5.1.3	Proposed Deep Fusion Skip Module (DFSM)	70
5.1.4	Proposed Deep Reconstruction Module (DRM)	71
5.2	Results and Discussions	72
5.2.1	Database Description	73
5.2.2	Experimental Setup	74
5.2.3	Performance Evaluation	74
5.3	Conclusion	79
6	CovSegNet: A Multi Encoder Decoder Architecture for Improved Lesion Segmentation from COVID-19 Chest CT-scans	80
6.1	Methodology	82
6.1.1	Problem Formulation	83
6.1.2	Proposed CovSegNet architecture	83
6.1.3	Proposed Encoder/Decoder Structure	85
6.1.4	Proposed Multi-Scale Fusion (MSF) Module with Pyramid Fusion scheme	89
6.1.5	Proposed Pyramid Fusion (PF) Module	90
6.1.6	Structure of the Fusion Optimizer(\mathcal{O})	91
6.1.7	Loss Function	92
6.2	Results and Discussions	93
6.2.1	Dataset Description	93
6.2.2	Experimental Setup	93
6.2.3	Performance Analysis	94
6.3	Conclusion	101
7	CovTANet: A Multi Objective Learning Framework for Lesion Segmentation, Diagnosis, and Severity Prediction of COVID-19	103
7.1	Methodology	104
7.1.1	Proposed Tri-level Attention Scheme	105
7.1.2	Proposed Tri-level Attention-based Segmentation Network (TA-SegNet)	108

7.1.3	Proposed Regional Feature Extractor Module	110
7.1.4	Volumetric Feature Aggregation and Fusion Module	110
7.1.5	Loss Functions	111
7.2	Results and Discussions	112
7.2.1	Dataset Description	112
7.2.2	Experimental Setup	113
7.2.3	Performance Evaluation	113
7.3	Conclusion	117
8	Conclusion	119
8.1	Summary	119
8.2	Limitations and Future Works	121

List of Figures

2.1	A portion of ECG data collected from a patient is shown before and after denoising.	18
2.2	Augmentation on a right bundle branch block beat by applying 26 different combinations of operation on the original beat generated by the proposed augmentation scheme.	20
2.3	Proposed structural units utilizing (a) single temporal convolution, (b) multiple temporal convolution with various kernels in parallel with prior and post pointwise convolution.	22
2.4	DeepArrNet Unit Block utilizing multiple temporal convolutions with various kernels in parallel.	25
2.5	Architecture of the Proposed DeepArrNet. Here, ‘k’ stands for kernel size, ‘f’ for number of filters and ‘s’ for strides in the convolution.	26
2.6	Representing per class performance of the proposed network in sensitivity, positive predictive Value, and F1 score in each class.	29
3.1	The complete workflow is represented schematically.	33
3.2	Proposed structural units. Here, h, w, and c denote the height, width and no. of channels of the feature map, respectively, while ‘k’ stands for kernel size, ‘s’ for strides and ‘f’ for number of filters in the convolution. In depthwise convolution, dilation rate will be varied from 1 to ‘m’.	34
3.3	Dilated Convolution for different dilation rates with kernel size (3×3) are encompassing different receptive areas. With increased dilation rate, the receptive area also gets bigger, though kernel size is kept unchanged.	35

3.4	Schematic of the Proposed CovXNet architecture optimized for input shape (128, 128, 1). Each residual unit is replicated for ‘d’ times.	37
3.5	Individually optimized networks are stacked together by using the meta-learner for obtaining more-optimized predictions.	38
3.6	Proposed Transfer learning scheme on CovXNet for fine tuning with small number of images.	39
3.7	Sample X-ray images of normal, viral, bacterial and COVID-19 caused pneumonia patients are shown.	40
3.8	Multi-class validation accuracy in different training epochs is shown for different resolutions of inputs.	41
3.9	Effect of using proposed stacking algorithm in the initial training phase.	41
3.10	Effect of using proposed stacking algorithm in the transfer learning phase.	42
3.11	Effect of the choice of meta-learner in stacking.	42
3.12	Multi-class confusion matrices are shown before and after stacking.	43
3.13	Significant portions of the test X-rays that instigate the decision are localized by imposing the activation heatmap obtained from CovXNet.	47
4.1	Multiple training stages are utilized to incorporate features from numerous transformed representations of input sensor data. (a) Two stage training, and (b) Multiple sequential stages of training.	50
4.2	Proposed (a) 1D Convolutional Neural Network and (b) 2D Convolutional Neural Network.	54
4.3	Proposed (a) 1D unit residual block and (b) 2D unit residual block.	54
4.4	Schematic representation of the proposed multi-stage sequential training scheme. Here, (a) represents Individual training stage, (b) represents combined training stage, and (c), (d), (e) represent the sequential training stages.	57

4.5	Effect of various types of augmentation of the sample data. (a) Raw sample data collected from 3 axis accelerometer, with (b) scaling, (c) jittering, (d) permutation, (e) magni- tude warping, and (f) time warping applied on raw data.	58
4.6	Comparison of Average Cross-Validation IoU scores on var- ious activities of UCI HAR Database [190].	61
4.7	Average Cross-Validation IoU score in various training stages of multi-stage sequential training on different databases.	62
5.1	Some of the challenges presented by colonoscopy images are (a) blurred or low quality images, (b) varying shapes and tex- tures of polyps, (c) small visible differences among polyps, and d) background presence of extraneous matters.	65
5.2	Schematic diagram of the proposed PolypSegNet Architecture	67
5.3	Schematic diagram of the proposed Depth Dilation Inception (DDI) module.	68
5.4	Schematic diagram of the proposed D-Unit layer structure. Different number of DDI units (m) have been integrated in each layer depending on the feature map dimension.	70
5.5	Schematic representation of the Deep Fusion Skip Module (DFSM).	71
5.6	Schematic representation of the Deep reconstruction module (DRM).	73
5.7	Visual representation of the input colonoscopic images and the segmented polyp regions obtained using various archi- tectures on different databases. In segmented polyps, ‘blue’ denotes the false positive region and ‘green’ denotes the false negative region.	78
6.1	Workflow of the proposed scheme for segmenting lung lesions of COVID-19 in CT volume.	81
6.2	Schematic representation of the two-stage implementation of the proposed CovSegNet architecture where two sequen- tial encoder-decoder operational stages are employed with L subsequent levels.	84
6.3	Schematic representations of the proposed encoder and de- coder modules in five-level implementation.	86
6.4	Structure of the Encoder/Decoder Unit cells.	87

6.5	Schematic representations of the down transition unit (operating between level-3 and 4) and the up transition unit (operating between level- $(L - 2)$ and $(L - 3)$).	88
6.6	Schematic representation of the proposed Multi-Scale Fusion module.	89
6.7	Proposed pyramid fusion scheme for fusing multi-scale features.	90
6.8	Schematic of the fusion optimizer module optimizing the decoded feature maps generated from two decoding stages.	92
6.9	Visual representations of the segmentation performances of different state-of-the-art networks on the CT images from Database-1 and Database-2. Here, ‘yellow’, ‘red’, and ‘blue’ represent true positive (TP), false negative (FN), and false positive (FP) regions, respectively.	97
6.10	Visual representations of the segmentation performances obtained using single phase training (CovSegNet2D and CovSegNet3D) and multi-phase training (with hybrid 2D-3D networks) in Dataset-1.	98
6.11	Visual representations of the segmented multi-class lesions of the CT images from Database-2 obtained using different state-of-the-art networks. Here, ‘red’ represents the ‘Ground Glass Opacity (GGO)’ regions and ‘yellow’ represents the ‘Consolidation’ regions.	99
7.1	Graphical overview of the optimization scheme of CovTANet	104
7.2	Optimization flowchart of the proposed CovTANet network.	105
7.3	Schematic of the proposed channel, spatial, and pixel attention mechanisms.	106
7.4	Schematic of the proposed Tri-level Attention Unit (TAU) integrating channel attention (CA), spatial attention (SA), and pixel attention (PA) mechanisms.	108
7.5	Schematic representation of the proposed Tri-level Attention-based Segmentation Network (TA-SegNet).	109
7.6	Representation of the proposed regional feature extractor module	111
7.7	Proposed volumetric feature accumulation and fusion scheme used for severity and diagnostic feature extraction	113

7.8 Visualization of the lesion segmentation performance of some of the state-of-the-art networks in MosMedData [227]. Here, ‘green’ denotes the true positive (TP) region, ‘blue’ denotes the false positive region, and ‘red’ denotes the false negative regions. 116

List of Tables

2.1	Confusion Matrix for Proposed DeepArrNet	28
2.2	Effect of Proposed Data Augmentation and Denoising Methods on Sensitivity	28
2.3	Performance of the Proposed Scheme on Arrhythmia Detection Employing Total Arrhythmia and Normal Class	29
2.4	Performance of the Proposed Scheme on PTB Database [165]	29
2.5	Comparison of Proposed Scheme with Existing Methods on Average Values of Evaluation Metrics on MIT BIH Dataset [163]	30
3.1	Performance Comparison of the Proposed Method with Other State-of-the-Art Approaches in Non-COVID Pneumonia Detection	43
3.2	Performance Comparison of the Proposed Method with Other Traditional Networks on COVID-19 and Other Pneumonia Detection.	44
3.3	Performance Comparison of the Proposed Scheme with Other State-of-the-Art Approaches on COVID-19 and Other Pneumonia Detection	45
3.4	Performance on Additional 468 X-ray Images of COVID-19 patients [182]	45
4.1	Average Cross-Validation Performance Analysis on Various Activities of UCI HAR Dataset [190] for Proposed Two-Stage and Multi-Stage Training	60
4.2	Average Cross-Validation Performance Analysis on Various Activities of USC HAR Dataset [191] for Two-Stage and Multi-Stage Training	60
4.3	Average Cross-Validation Performance Analysis on Various Activities of SKODA Dataset [64] for Two-Stage and Multi-Stage Training	61

4.4	Comparison of the Proposed Schemes with Other Existing Approaches on Different Datasets	63
5.1	Structural Details of Different D-Unit layers in Encoder and Decoder Block	70
5.2	Effect of Different Network Configuration in the Performance on CVC-ClinicDB, Kvasir-SEG, and ETIS-Larib Databases	75
5.3	Comparison of Performance of the Proposed PolypSegNet with Other State-of-the-Art Approaches on CVC-ClinicDB, CVC-ColonDB, Kvasir-SEG, and ETIS-Larib Databases with Five-Fold Cross-validation Scheme	76
5.4	Comparative Analysis of Cross-Dataset Performances of Different State-of-the-Art Networks with Different Combinations of Training and Testing Dataset	76
5.5	Computational Performance Analysis of Different State-of-the-Art Networks	77
6.1	Ablation Study of the Effect of Different Modules in the Performance (Mean \pm Standard Deviation) of the Proposed CovSegNet2D Architecture	94
6.2	Performance Comparison (Mean \pm Standard Deviation) of the Proposed CovsegNet2D Architecture with Other State-of-the-Art Approaches on 2D-CT slices	95
6.3	Performance Comparison (Mean \pm Standard Deviation) of the CovSegNet3D Architecture with Other State-of-the-Art Networks on 3D-CT Volumes of Dataset-1	95
6.4	Effect of Vertical Expansions (Levels) and Horizontal Expansions (Stages) on the Dice Score (Mean \pm Standard Deviation) in Dataset-1	96
6.5	Ablation Study of the Effect of Different Modules in the Performance (Mean \pm Standard Deviation) of the Proposed CovSegNet3D Architecture in Dataset-1	96
6.6	Comparison of Performances (Mean \pm Standard Deviation) on Different Types of Infections (Ground Glass Opacity and Consolidation) in Different CT-slices of Dataset-2	97
6.7	Computational Efficiency Analysis of Numerous Architectures along with the Performances Obtained on Dataset-1	102

7.1	Performance (Mean \pm Standard Deviation) of the Ablation Study of the Proposed TA-SegNet on MosMedData	115
7.2	Comparison of Performances with Other the State-of-the-Art Networks on COVID Lesion Segmentation on MosMed-Data	115
7.3	Comparison of Performances in the Joint Diagnosis and Severity Prediction of COVID-19 with Different Networks on MosMed-Data	118

Chapter 1

Introduction

Machine learning approaches have been widely incorporated in wide range of applications including health science, computer vision, automation, and industrial applications [1], [2]. Particularly, these algorithms have brought about revolutionary improvements in diverse domains of health science including automated disease diagnosis, region-of-interest segmentation, severity prediction, and patient monitoring [3]. However, performances obtained with traditional hand-crafted feature based approaches with shallow neural networks and other machine learning algorithms demand domain expertise, and careful selection of features that limit their applicability in diverse applications [4]–[6]. Recently, deep learning approaches have been established as a paradigm shift for the automatic extraction of effective feature representation from the data with deep neural networks. With the widespread availability of the data in this era along with the massive computational advantages of modern hardware, such approaches have been incorporated in wide range of applications for achieving the optimum performance [7], [8].

Despite the widespread applicability with unprecedented performance of the deep learning algorithms, most of the successes came through huge amounts of labeled data that limits the applicability in data-constraint critical applications [9]. Hence, achieving considerable performance with a limited amount of labeled data is one of the major challenges of such schemes [10]. To exploit the available information embedded in the data, several methods have been studied in the literature [11]–[13]. Most of the traditional schemes primarily depend on the architectural variations of the deep network to incorporate effective features while keeping the data mostly unaltered [12], [13]. However, it is necessary to process the data further to incorporate multi-perspective features in operations along with the architectural variations. Yet, there exists a critical shortage of works to process data of a particular modality from multifarious observation windows to exploit the feature space. Moreover,

most of the traditional approaches deal with unimodal data that further limits the applicability [12]–[14]. When multimodal data are available, it is necessary to incorporate features from multiple modalities for making the feature extraction process more robust [15]. However, such objectives demand additional optimization strategies for effective feature selection, and fusion [15]. Transfer learning is one of the most widely used techniques to propagate knowledge from one domain to another. However, in practical conditions, it is difficult to get suitable applications with adequate labeled data to transfer knowledge. Without depending on few traditional architectures for all applications, it is necessary to design efficient deep neural network architectures targeting the final applications that demand reformulation of the objective function, domain-specific adaptation, effective training, optimization, and finer data processing schemes.

1.1 Multi-receptive Feature Optimization: Methods and Opportunities

For improving the feature quality, the training data should be processed from diverse operational perspectives to extract features from different receptive areas [16]. Several approaches can be incorporated in the data processing framework at different stages to represent distinctive features. Architectural modifications can be incorporated to improve the feature extraction process by exploring different receptive areas of the operational data [17]. Apart from solely depending on the architectural variations with raw data, numerous transformations can be introduced for facilitating the feature extraction process by sequentially integrating features from diverse perspectives through improved optimization strategies [18], [19]. Moreover, various scales of features generated at different subsequent stages of the deep neural network can be processed with group fusion schemes for extracting the multi-scale contextual features to improve the robustness of the features [20], [21]. Finally, multi-objective learning can be an effective strategy for integrating the effective feature representations from multiple objectives through joint multi-phase optimization schemes [22]. Such learning greatly improves the feature quality as different objectives extract features from various receptive areas and thus, provides the opportunity to learn the generic feature representation to reduce the bias in the optimization. These strategies for multi-receptive feature optimization are briefly discussed in the following discussions.

1.1.1 Architectural Modifications

Existing deep neural network architectures have incorporated numerous architectural renovations to make the network more robust for diverse applications [23], [24]. Nevertheless, different applications offer diverse challenges that makes it difficult to optimize the same network for all the applications despite the widespread generalization capability of the deep neural network [25], [26]. Hence, domain specific adaptation of the neural network architecture plays a significant role to achieve optimum performance in a particular application [27]. Moreover, the main limiting condition for achieving optimum performance with the deep neural network is their data hungry nature. To obtain optimum performance, it is of great importance to incorporate features from distinct receptive areas through architectural modification. Therefore, the topology of the network architecture is required to be adjusted for facilitating the feature extraction process.

1.1.2 Multi-transformed Data Representation

Along with the architectural modifications of the deep neural network, it is required to explore the available training data from diverse perspectives by introducing different representation of the data [18]. Due to environmental perturbations and many other noises, sometimes it becomes complicated to extract the optimum feature representation from one viewpoint with the unaltered raw data. Transformation on the raw data offers the opportunity to explore the data from different viewpoint that helps to improve the robustness of the system [28]. Nevertheless, to utilize the opportunities provided by these transformations, it is required to have effective feature extractors to extract the underlying feature representation. As different transformations provide varied challenges with their diversified representations, it is challenging to design the optimum feature extractors for all these transformed spaces [29]. However, deep learning algorithms have provided the great opportunity to utilize the data representation for obtaining the optimum feature extractors. Therefore, there exists significant research opportunity to explore diverse representations of the data with the end-to-end deep learning framework to generate the optimum feature representation for improving performance.

1.1.3 Multi-scale Contextual Feature Optimization for Segmenting Region-of-Interests

Segmentation of the region-of-interests from the images have great significance with numerous potential applications in healthcare, automation, and vision applications

[20], [21], [30]. It provides precise localization of the objects that greatly reduces the operational burden in diverse applications. It poses additional complexity to perform pixel-wise classification of the whole image that requires generalization of the different scales of contextual features. To improve the segmentation performance, it is of great importance to effectively extract the multi-scale contextual features that greatly helps to get the multi-perspective view of the target objects [31]. Moreover, volumetric segmentation puts forward further challenges to extract contextual features from the whole volume [32]. Since the network gets bigger, it demands larger amount of labelled 3D data for obtaining considerable performance as well as increases the computational cost. To improve the computational performance as well as to integrate the multi-perspective contextual features effectively from the whole 3D data frame, hybrid 2D-3D optimization technique can be explored in sequential optimization phases.

1.1.4 Multi-objective Learning with Multi-phase Optimization

Despite achieving satisfactory performances in different objectives separately, multi-objective learning provides further opportunity to improve performance, particularly in the case of applications with multiple objectives with shared feature spaces [33]. To improve feature sharing for facilitating the optimization, it is required to design the network in a customized manner targeting a particular application with multiple objectives [34]. Moreover, in many applications, some of the objectives provide additional challenges demanding larger amount of labelled data, whereas some other objectives comparatively offer lower burden on the feature extractors for having sufficient amount of labelled data [35]. In such cases, learning from a particular objective can be effectively transferred to learn other objectives with significantly smaller amount of labelled data [28]. Such multi-phase optimization provides the opportunity to efficiently extract the generic multi-perspective contextual features to achieve the optimum performance on all of the objectives.

1.2 Multi-dimensional Data Processing in Real-world Applications

In practice, most of the real world applications come with the challenges of multi-dimensional data processing [36]. With increasing dimensionality, the data becomes richer with contextual details that provide opportunities for improved feature ex-

traction. However, higher dimensionality increases computational complexity of the feature extraction schemes requiring advanced optimization strategies to exploit the finer details [37]. On the other hand, lower dimensional data provide lighter workload on the feature extraction scheme while being less robust with contextual information. Different applications demand the choice of dimensionality of the operating data targeting separate devices based on the computational overhead and intended users. Therefore, different dimensional data should be optimized with customized framework to get the optimum performance.

For mobile applications, lower dimensional data can be an effective choice for lower complexity with considerable performance [38]. Several real world applications have been incorporated in the mobile devices, such as wearable sensor based data processing, mobile healthcare, and utility applications. For the improved performance, it is required to process higher dimensional data with finer details. However, the increased complexity of the features demand robust feature extraction and optimization schemes for exploiting the higher dimensional features. Moreover, to optimize the deep neural network with higher dimensional data, considerably larger volume of data are required. The challenges offered with the smaller available data with higher dimensionality require customization of the deep neural network architectures and optimization schemes considering the requirement of the operating conditions.

In this thesis, 1-dimensional wearable sensor data from different modalities have been studied for human activity recognition [19]. Moreover, for representing mobile healthcare based applications, electrocardiogram data have been studied for cardiac arrhythmia diagnosis that have significant importance to prevent cardiovascular diseases [27]. Moreover, 2D-image data of chest X-ray have been studied for diagnosing pneumonia and COVID-19 under data constraint scenarios [20]. Moreover, endoscopy images have been studied for segmenting polyp lesions with diverse textures, types, and shapes [21]. Additionally, 3D chest CT volumes have been experimented for improving diagnosis, and chest lesion segmentation of COVID-19 [22], [30].

1.3 Literature Review

Efficient deep neural network architectures have been experimented for several real-world applications with multi-dimensional data representations. Firstly, multi-class arrhythmia detection and diagnosis application is experimented from electrocardiogram (ECG) data. On this task, a number of methods have already been presented

in the literature ranging from the traditional feature-based machine learning process to the end-to-end deep learning process in recent times. In feature-based arrhythmia detection techniques [39]–[48], various feature extraction approaches are employed, such as wavelet transform [41]–[45], principal component analysis [47], independent component analysis [48] and Hermite function [46]. For performing classification with the extracted features, support vector machine (SVM) [39], K-nearest neighbour [40], feed-forward neural network [43], [46], [48] and random forest [41] have been used. These approaches mostly depend on handcrafted feature extraction process that most often leads to loss of information required for the classification due to improperly chosen features or inadequate features. Automating the process of feature extraction and classification was the primary motivation behind the popularity of end-to-end deep learning-based frameworks.

A number of deep learning-based approaches have also been adopted recently for arrhythmia classification [25], [49]–[60]. In [49]–[54], 1D convolutional neural network (CNN) and in [58]–[60], recurrent neural network (RNN) and LSTM network are employed while in [55]–[57], 2D CNN is used by converting 1D ECG beats into 2D images. Most of the deep learning-based approaches are facing some common issues: (1) raw ECG data collected from patients are being directly fed to the deep neural network making the classification process complicated due to presence of various low and high-frequency noises, (2) for dealing with 1D ECG signal, data augmentation is not necessarily used and even if it is used, natural variational pattern of ECG isn't properly captured or preserved, and (3) most of the approaches use very deep CNNs with large number of parameters that not only increase the computational complexity but also lead to overfitting the model to training data. Hence, a deep CNN based arrhythmia classification scheme which can overcome the above problems and can provide very satisfactory classification performance with low computational burden is still in great demand.

Afterwards, wearable time-series data from multi-modal sensors have been incorporated for human activity recognition. Large varieties of approaches have been applied to make the correct recognition ranging from traditional feature-based approaches to the end-to-end deep neural network in recent times. Numerous handcrafted feature extraction process with shallow classifiers are explored in the literature for utilizing multimodal sensor data in activity recognition [61]–[68]. Though these types of handcrafted features perform well in limited training data scenario, the extraction of effective features gets very complicated with more number of sensors. Additionally, the process heavily demands domain expertise for proper selection of features which becomes harder with the presence of random noises that occurs very

often in practical conditions.

To automate the complicated feature extraction process, various types of deep neural networks have been studied in the literature to recognize human activity from wearable sensor data [69]–[79]. Most of these approaches directly employ the collected raw sensor data for automated feature extraction using the deep neural network, such as convolutional neural network (CNN) [69]–[71], recurrent neural network (RNN) [75], [76], long short term memory (LSTM) network [77], hybrid CNN-LSTM network [78], and a more complicated LSTM-CNN-LSTM based hierarchical network [79]. Most of these networks are very deep in structure and therefore, a large amount of data is required to train them properly. Moreover, due to random noises and perturbations in multi-modal sensor data from different sources, the process gets more intricate to operate with the raw data directly. Hence, with an increasing number of sensors, while having a small amount of data for some of the activity classes, this problem becomes critical for the automated extraction of features from raw sensor data using deep network that severely affects the performance.

In [80]–[83], various approaches have been introduced to represent the time series data in a modified space that makes the feature extraction process easier by reducing the effects of noise or random variations. These transformations on the time series sensor data provide more opportunities to explore the variations of features from different spaces. Though these transformations provide efficient representation of some of the features in a different space, some other features may become complicated to extract from that particular space. However, different transformations provide diverse viewpoints to explore the feature space of raw time series data. Hence, similar to these studies, depending solely on a single transformed space for feature exploration limits the scope of feature extraction that may result in smaller performance in many circumstances. If features extracted from different transformed spaces can be incorporated in the final decision-making process, it will provide a more robust opportunity to analyze the information on raw data. But, the challenging task of integrating effective features from diverse transformed spaces through joint optimization to reach the optimum performance in activity recognition is yet to be attempted.

Following that, automatic diagnosis of different types of pneumonia and Coronavirus Disease-2019 (COVID-19) from chest radiography (X-ray) have been experimented. different modalities of data have been experimented for the automatic diagnosis of the Coronavirus Disease-2019 (COVID-19). With a serious shortage of experts, while having large similarities of COVID-19 with traditional pneumonia,

an artificial intelligence (AI) assisted automated detection scheme can be a significant milestone towards a drastic reduction of testing time. The mortality rate is increasing alarmingly throughout the world demanding an early response to diagnose and prevent the rapid spread of this disease. Because of having no specific drugs and treatments, the situation has become frightening to billions of individuals [84]. Symptoms ranging from dry cough, sore throats, and fever to organ failure, septic shock, severe pneumonia, and Acute Respiratory Distress Syndrome (ARDS) are detected from COVID-19 patients [85]. Reverse transcription-polymerase chain reaction (RT-PCR), the most commonly used diagnostic test of COVID-19, suffers from low sensitivity in early stages with elongated test period assisting further transmission [86]. Furthermore, the extreme scarcity of this expensive test kit [87] exacerbating the situation. Hence, a chest scan such as X-rays and Computer tomography (CT) scans are prescribed to all individuals with potential pneumonia symptoms for faster diagnosis and isolation of the infected individuals. With a serious shortage of experts, while having large similarities of COVID-19 with traditional pneumonia, an artificial intelligence (AI) assisted automated detection scheme can be a significant milestone towards a drastic reduction of testing time.

In [88], [89], CT scans are used with deep learning-based systems for automated COVID-19 pneumonia detection. Though CT scans provide finer details, X-rays are quicker, easier to take, less injurious and more economical alternative. However, due to the scarcity of COVID-19 X-rays, it is extremely difficult to train a very deep network effectively. Hence, transfer-learning can be a viable solution in this circumstance that have been widely adopted in many recently proposed COVID-19 detection schemes [90]–[93]. Yet, the traditional scheme of transfer-learning that uses established deep networks pre-trained on the ImageNet database for transferring its initial learning can't be a good choice as the characteristics of COVID-19 X-rays are solely different from images intended for other applications. Therefore, an automated deep learning based approach for diagnosis COVID-19 and other traditional pneumonia from X-rays under such data constrained scenarios is of great demand.

Afterwards, efficient neural network architecture is experimented for automatic polyp segmentation from endoscopic images which has great clinical significance for preventing colorectal cancer. Colorectal cancer has become one of the major causes of death throughout the world. Early detection of Polyp, an early symptom of colorectal cancer, can increase the survival rate to 90%. Segmentation of Polyp regions from colonoscopy images can facilitate the faster diagnosis [94]. Due to varying sizes, shapes, and textures of polyps with subtle visible differences with

the background, automated segmentation of polyps still poses a major challenge towards traditional diagnostic methods [95], [96]. Though not all polyps lead to colorectal cancer, all colorectal cancer starts with polyps that become cancerous over time which makes the accurate detection, investigation, and analysis of the types, patterns, and structures of the polyps of primary importance to reduce spread of CRC. [97] Some of the rare types of polyps are visually difficult to distinguish due to flat natures that demands wide experiences and expertise of the endoscopists that may considerably increase the miss rate during colonoscopy. According to [98], most of the CRC events are found to occur (91% ~ 94%) in patients who aren't up-to-date with colonoscopic examinations, whereas 6% ~ 9% events still occur even after up-to-date colonoscopies. These discrepancies mainly arise from the higher miss-rate with the flat/sessile type of polyp which becomes critical for the proper diagnosis of polyps. Furthermore, risks associated with the adenomatous polyps turning to be malignant increases with the structural deformation of the polyp regions for abnormal growth. Also, increased number of polyps lead to higher risks of colorectal cancer demanding regular clinical investigation by expert endoscopist.

Numerous hand-crafted feature-based approaches have been explored for automatic polyp segmentation in the last two decades [99]. In [100], [101], a fuzzy c-mean clustering method is proposed followed by adaptive deformable models for separating polyp regions. In [102], the analysis of the color, shape, and curvatures of the contour regions is carried out for feature extraction. In [103], sparse autoencoder is incorporated for extracting super-pixel based features with different saliency methods to outline polyp regions. In [104], protrusion measurements using second principal curvature flow is introduced to extract structural features of polyps. However, due to the additional complexity in polyp region detection for its diversified textures, shapes, and colors, while having minute differences with the background, sub-optimal performance is achieved in most of these hand-crafted feature-based approaches.

Though not all polyps lead to colorectal cancer, all colorectal cancer starts with polyps that become cancerous over time which makes the accurate detection, investigation, and analysis of the types, patterns, and structures of the polyps of primary importance to reduce spread of CRC. Some of the rare types of polyps are visually difficult to distinguish due to flat natures that demands wide experiences and expertise of the endoscopists that may considerably increase the miss rate during colonoscopy. According to [98], most of the CRC events are found to occur (91% ~ 94%) in patients who aren't up-to-date with colonoscopic examinations, whereas 6% ~ 9% events still occur even after up-to-date colonoscopies. These

discrepancies mainly arise from the higher miss-rate with the flat/sessile type of polyp which becomes critical for the proper diagnosis of polyps. Furthermore, risks associated with the adenomatous polyps turning to be malignant increases with the structural deformation of the polyp regions for abnormal growth. Also, increased number of polyps lead to higher risks of colorectal cancer demanding regular clinical investigation by expert endoscopist. Hence, Computer Aided Detection (CAD) systems can contribute to this situation in a different avenue by potentially acting as a second observer that can complement the physician by pointing out overlooked polyps. Also, such systems would not require any notable alteration to the procedure of colonoscopy that makes the integration of this system more practical and easier. In addition, accurate segmentation of polyps may significantly reduce the miss-rate of polyps, and hence, can be an effective clinical tool for faster screening. However, such precise segmentation of the polyp regions is particularly complicated that requires extraction of effective features for precisely detecting the edges of diversified shapes of polyps. Therefore, an automated computer-aided scheme for properly segmenting regions of polyps from colonoscopy images/videos can be a significant contribution to expedite the process of early-stage polyp detection with more precision [26], [105].

Similar to other medical imaging applications, deep learning-based approaches have gained much attention in recent years for automating the feature extraction process to detect and segment polyp regions with unprecedented precision [106], [107]. In [108], mask-RCNN is incorporated with traditional CNN based feature extractors to provide bounding boxes in the polyp region. In [109], two mask-RCNN networks with different base CNN modules are ensembled for better prediction of bounding boxes. For obtaining pixel-level segmentation instead of such bounding boxes, a modified fully convolutional neural network (FCNN) is used with multiple decoders in [110]. To introduce more contextual information in polyp segmentation, a deep residual network with dilated kernels is incorporated in the FCNN module in [111]. Instead of a single encoder in traditional FCNN architecture, an encoder-decoder based structure is proposed, named as UNet, that increases the performance of FCNN considerably and has established as a popular choice in medical image segmentation [112]. However, there exist some architectural limitations in the traditional Unet architecture that opens the opportunity to improve the performance further.

Each level of both the encoder and decoder of Unet contains a series of traditional convolution operations that make it difficult to extract variational features from diverse receptive areas at different scales. To increase diversity, multi-dilated

residual units are introduced in [113], instead of traditional convolutions, followed by squeeze and excitation unit. In [114], a multires block is introduced that utilizes residual operation after separately integrating features learned from all sequential traditional convolutions. However, residual learning is more effective for the stack of deep convolutional layers at each level [115]. Incorporating residual learning with a single residual block at each level may hinder the learning process. In [116], unit block of densely connected convolutional layers are integrated at each level for better performance.

The skip inter-connection of Unet directly connects the output feature map generated from each level of the encoder to the corresponding level of decoder for propagating information that may be lost through subsequent pooling operation. In [114], a semantic gap between feature maps is noticed while merging two feature maps at each level of the decoder and a deep residual path is introduced in the shortcut skip connection. In [117], a dense residual operation is incorporated with convolutions of multiple dilation rates in the skip connection. However, most of these skip connections operate with the output feature maps generated from a particular level of the encoder to reduce the semantic gap with the corresponding decoder level. It is expected that more effective information flow between encoder and decoder can be achieved if all output feature maps from different encoder levels can be integrated for interconnecting with each decoder level. In [118], a nested convolutional stack is incorporated in between encoder and decoder modules to inter-connect diverse semantic levels. However, it increases the computational complexity considerably that hinders the optimization and propagation of the flow of information from the encoder module to the respective decoder level.

Moreover, the semantic information of the output mask is gradually aggregated in subsequent levels of the decoder and the reconstructed mask is generated considering only the final semantic level in traditional segmentation architectures. It makes the convergence of the network more complicated through such a deep stack of encoder-decoder layers mostly arising from the vanishing gradient problems. Since different scales of reconstructed feature maps are generated at various levels of the decoder, it is expected that more efficient reconstruction can be achieved during the final reconstruction phase through integrated and joint optimization of these multi-scale decoded feature maps. Hence, there exists several scopes for further improvement of the traditional Unet architecture for precise segmentation of polyp regions.

Afterwards, the CNN architecture for image segmentation have been extended further for volumetric segmentation of COVID-19 chest CT lesions. Several deep learning-based frameworks have been explored in recent times deploying automated

screening of chest radiography and computer tomography as one of the vital sources of information for COVID diagnosis [119]–[123]. However, owing to the relatively higher sensitivity and the provision of enhanced infection visualization in the three-dimensional representation, CT-based screening is a more viable alternative than the X-ray counterparts. Processing 3D CT volume at a whole increases computational complexity exponentially that makes the optimization and convergence more difficult limiting the architectural diversity of the network. The most widely used alternative of 3D-processing is to operate separately on 2D-slices extracted from the CT-volume [124]–[128]. However, such slice-based processing loses inter-slice contextual information that results in sub-optimal performance. In [129]–[132], smaller sub-volumes are extracted from the original 3D volumes to minimize the computational burden as well as to utilize 3D contextual information. However, such methods suffer from inter-volume contextual information loss by considering a smaller portion of the whole set at a time as well as increases complexity to process sub-volume level prediction into the final result.

Different architectural modifications have been explored in recent years to overcome some of these limitations. To increase the diversity of operations at each scale of feature maps, numerous established network building blocks are integrated in encoder/decoder module, e.g. residual block [133], dense block [134], inception block [135], dilated residual block [24], and multi-res block [136]. To reduce the semantic gap between a particular scale of encoder and decoder, a residual path is proposed in MultiResUnet architecture instead of a direct skip connection of Unet [136]. However, the semantic gap generated between multi-scale feature maps of encoder and decoder modules still persists. In Unet++ [137], a nested stack of convolutional layers is introduced to reduce the semantic gaps. But, it increases computational complexity considerably which makes convergence difficult. In [131], Vnet is proposed that utilizes residual building blocks in Unet architecture, while in [132], cascaded-Vnet is presented for performance improvement that utilizes a dual-stack of the cascaded encoder-decoder module. Nevertheless, with existing numerous architectural limitations of traditional U-shaped architecture in each stage, it increases semantic gaps with the additional encoding-decoding stage as well as increases vanishing gradient issues with contextual information loss that open up opportunities for further optimization.

However, most of the recent studies mostly opt for solving the daunting task of COVID-19 diagnosis partially where infection segmentation, diagnosis, or severity analysis have separately attempted [138]–[140]. However, there exists large degrees of correlation among all of these tasks. Under the data constrained scenarios, such

learning approaches may help to achieve better performances in all of the objectives compared to separate independent learning. For some of the challenging applications, such joint optimization can be designed in a sequential way to facilitate the information flow through all of the objectives. Hence, the joint multi-objective optimization a promising way of further research for better COVID-19 diagnosis, infected lesion segmentation, and severity prediction.

1.4 Objectives and Scopes

The objectives of this research with specific aims are as follows:

1. To develop efficient DNN architecture, building blocks, and data augmentation strategies for efficient feature learning from diverse receptive areas of unimodal data.
2. To incorporate effective features from multi-resolution images with multi-stage transfer learning strategies for sequential knowledge transfer and optimization.
3. To develop a multi-perspective feature integration scheme from numerous transformed spaces of multi-modal data through effective selection and fusion of features with deep neural networks.
4. To develop improved CNN architecture for multi-dimensional image segmentation applications by introducing multi-scale contextual feature aggregation and fusion based building blocks.
5. To investigate joint optimization strategies for developing a hybrid neural network architecture targeting multiple objectives through effective feature sharing and feature fusion schemes.

The possible outcome of this research is to develop efficient deep neural network architectures along with effective training, data processing, and optimization schemes to achieve optimum performance in disease classification, infection segmentation, and human action recognition.

1.5 Organization of the Thesis

In this thesis, several novel deep neural network architectures have been introduced along with diverse data representations, multi-phase optimization and customized

learning objectives for exploring the feature spaces from diverse receptive windows in several real-world applications with multi-dimensional data.

Firstly, a deep neural network architecture (DeepArrNet) is presented with multi-receptive building blocks for arrhythmia classification from denoised ECG beats. The presented scheme offers an end-to-end framework for precise diagnosis and classification of cardiac arrhythmia incorporating ECG beat extraction, denoising, beat augmentations for smaller classes, feature extractions with varying kernel windows, and regularized optimizations.

Secondly, the network architecture has been improvised with a novel multi-dilation building block for incorporating multi-receptive features from 2d chest X-ray images in order to diagnose and classify multi-class pneumonia including COVID-19. Additionally, to transform the observation perspectives, multi-resolution images have been introduced with customized neural networks along with a stacking algorithm for improved diagnosis. Moreover, for overcoming the scarcity of the available data for COVID-19, a multi-stage transfer learning approach has been presented for exploiting available chest X-ray images for learning generic feature representation.

Thirdly, different transformed representations of the multi-modal time series data from numerous wearable sensors are explored for improving the performance of human activity recognition objective by introducing diverse perspectives of the data. Moreover, efficient training and optimization algorithms have been proposed to gradually incorporate the effective feature representations from these representations. Furthermore, different augmentation techniques have been introduced for handling imbalance in the training data as well as to introduce different perspectives of the available data. Several real world datasets have been used for verifying the effectiveness of the proposed schemes under diverse environmental and operating conditions of different subjects.

Fourthly, a modified encoder-decoder based deep neural network architecture (PolypSegNet) is presented for segmenting polyp lesions from endoscopy images. This network generates various scales of feature representation from the endoscopy image through sequential encoding-decoding stage with inter-linked skip connection. Moreover, multi-scale feature maps generated from various levels of the network have been processed through group fusion and optimization schemes for extracting the general contextual information for precise segmentation. Such architectural building blocks facilitate the feature extraction process to cover diverse receptive areas for learning the generic representation of the region-of-interests. Several real-world datasets has been experimented to validate the effectiveness of the proposed multi-scale feature optimization framework.

Fifthly, a multi encoder-decoder architecture (CovSegNet) is introduced with horizontal and vertical expansion strategies for improved performance of COVID-19 lesion segmentation from CT volumes. Several encoding and decoding stages are stacked sequentially with multi-scale feature fusion schemes for gathering richer contextual details. The proposed architecture can be optimized for processing 2D images as well as can be designed for operating with 3D data. Though operating with 3D CT-volumes provides detailed contextual information, it increases the computational complexity of the optimization process demanding considerably larger amount of labelled data. To improve the computational performance as well as to integrate the multi-perspective contextual features effectively from the whole 3D data frame, hybrid 2D-3D optimization technique has been introduced in this chapter. Various architectural topology have been introduced with multi-phase training strategies to gradually optimize the system for achieving robust performance. The proposed approaches have been experimented on several publicly available datasets of COVID-19 CT volumes that represents the effectiveness of the proposed scheme outperforming traditional approaches.

Finally, a multi-objective learning framework is introduced for joint diagnosis, severity prediction, and infected lesion segmentation of COVID-19 chest CT volumes. Numerous architectural building blocks have been introduced including tri-level attention unit, a novel segmentation network, contextual feature aggregation along with an end-to-end framework for the joint optimization of the system of networks. Several objectives are integrated together in a hybrid-learning framework to optimize the system of networks with a multi-phase optimization scheme. Such optimization approaches have greatly reduced the burden of the feature extraction process that significantly improves the performance compared to the separate learning of these objectives.

Chapter 2

DeepArrnet: A Multi-receptive Neural Network for Arrhythmia Classification from ECG Beats

Cardiovascular diseases (CVDs) have become one of the most common causes of death throughout the world in recent times. Early recognition of cardiac abnormality is vital for proper treatment before occurring any major irreversible damages. Among various CVDs, the arrhythmia is one of the most common problems that describes irregularity and abnormality in heart beats [141]. There are various types of arrhythmia, such as ventricular fibrillation, premature atrial contraction and supraventricular arrhythmia [142], [143]. Electrocardiogram (ECG) signal, a recording of the heart's electrical potential to show the electrical activity of the heart, is most widely used by physicians to check the proper functionality of the heart. Arrhythmia detection based on manual inspection of ECG signals by experts is the commonly used approach which is often complicated, time-consuming, human error-prone and difficult due to lack of experts.

In this chapter, an efficient deep convolutional neural network (CNN) architecture is proposed based on depthwise temporal convolution along with a robust end-to-end scheme to automatically detect and classify arrhythmia from denoised electrocardiogram (ECG) signal, which is termed as 'DeepArrNet'. The major contributions of this chapter is summarized as follows:

1. A structural unit, namely PTP (Pointwise-Temporal-Pointwise Convolution) unit, is designed with its variants where depthwise temporal convolutions with varying kernel sizes are incorporated along with prior and post pointwise convolution.

2. A deep neural network architecture is constructed based on the proposed structural unit where series of such structural units are stacked together while increasing the kernel sizes for depthwise temporal convolutions in successive units along with the residual linkage between units through feature addition.
3. Considering the variational pattern of wavelet denoised ECG data, a realistic augmentation scheme is designed that offers a reduction in class imbalance as well as increased data variations.
4. Multiple depthwise temporal convolutions are introduced with varying kernel sizes in each structural unit to make the process more efficient while strided convolutions are utilized in the residual linkage between subsequent units to compensate the increased computational complexity.
5. Extensive experimentations are carried out on two publicly available datasets to validate the proposed scheme that results in outstanding performances in all traditional evaluation metrics outperforming other state-of-the-art approaches. The primary results of these experimentation are published in [27].

2.1 Preprocessing

The raw data collected from patients need to be pre-processed first to make it compatible with the deep neural network. This pre-processing stage consists of five different operations. Each of them is described in detail below.

2.1.1 Wavelet based denoising

The input raw ECG signal, $x[n]$ can be expressed as:

$$x[n] = \hat{x}[n] + v[n] \tag{2.1}$$

where $\hat{x}[n]$ is the original clean ECG signal that has a certain pattern, and $v[n]$ is the additive noise present in the raw data—generally random in nature—may significantly vary during train and test phase. If $x[n]$ is used for training, there is a chance of getting poor performance during the test even with the highly trained model due to the random nature of $v[n]$. Instead of using noisy raw data, if noise reduction is possible to achieve clean $\hat{x}[n]$, much better training and testing performance is achievable.

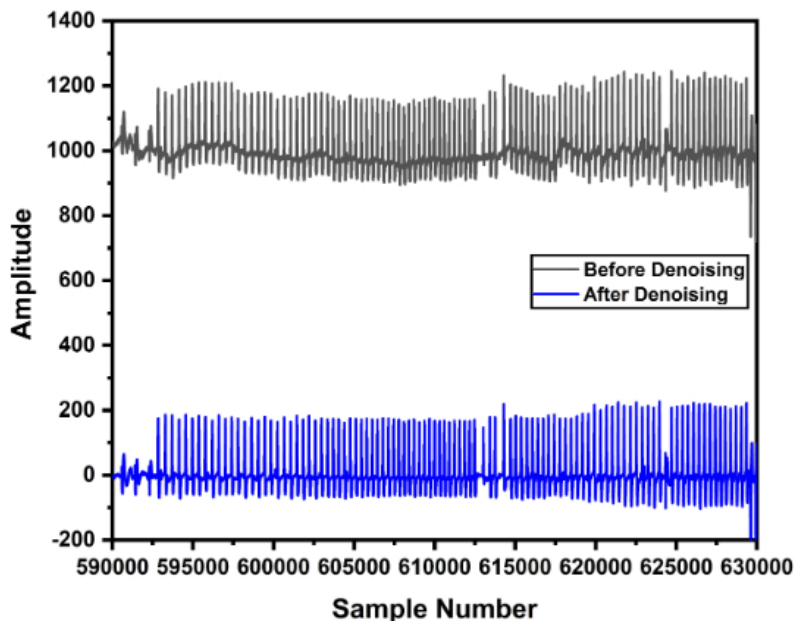


Figure 2.1: A portion of ECG data collected from a patient is shown before and after denoising.

ECG signals are corrupted by different types of noises, such as baseline wandering, power line interference, electromyogram (EMG) noise, electrode motion artifacts, and channel noise. Different noise reduction techniques are used in literature for removing noises from recorded ECG data [144]–[148]. Among them, wavelet transformation, a time-scale representation method, decomposes signals into basis functions of time and scale that makes it useful for data denoising. In this work, a wavelet transform based approach with soft thresholding scheme has been employed for the removal of the effects of noise. In Fig 2.1, the effect of denoising on a segment of raw data is shown.

2.1.2 R Peak Detection

Following the denoising, the R peak of each ECG beat is located from the continuous beat stream. Various methods are proposed for R peak detection in the literature [149]–[155]. As reported in [154], it provides very fast and precise detection of R-peak. Hence, this algorithm has been employed in this study. In this method, the denoised ECG signal is squared, firstly, to enhance large values and boost high-frequency components. Next, blocks of interest are generated using two event-related moving average to extract the QRS features and to extract the QRSs beat, respectively. Afterward, an even related threshold is applied to the generated blocks to separate the blocks that contain the R-peak from the blocks that include

noise. Finally, the maximum absolute value of each separated block is identified that provides the R-peak index.

2.1.3 ECG Beat Extraction

To process the ECG beats using deep neural networks, all the beat length should be uniform. In [56], the median of the R-R intervals is considered as the nominal period to segment each beat by maintaining equal length in both sides of the R peak. After segmenting each beat, zero padding is done to make the length uniform. In [52], [57], [58], equal length of portion is cropped centering the R-peak as an individual beat. In this work, each beat is segmented centering the R peak by cropping at the midpoint of the adjacent R-R intervals. Following that, further cropping or padding with the edge values is carried out centering the R-peak to make the length of each beat uniform as in some cases, the extracted beat length becomes larger or smaller than the predefined length. This process can be described as follows.

1. After R-peak detection, each beat is extracted centering the detected R peak and cropping at two adjacent edges depending on the position of adjacent R peaks. The cropping edges are decided to be the midpoint of adjacent R-R intervals. Hence, if the sample number of the extracted beat is denoted by \mathbf{n} and a, b, c representing the sample containing three adjacent R-peaks while b is representing the R-peak of the beat to be segmented, the range of the segmented beat can be written as,

$$n_{\min} = (a + b)/2 \quad (2.2)$$

$$n_{\max} = (b + c)/2 \quad (2.3)$$

2. As some of the segmented beats will be smaller/larger than the predefined length of beats, it is needed to be equalized for further processing on deep neural networks. Hence, by centering the R peak of the extracted beat, further cropping is done at the edges, if the beat is larger at the edges. Otherwise, padding with the edge value is carried out if the beat is shorter.

2.1.4 ECG Beat Augmentation

Due to the scarcity of data for rare diseases compared to normal cases, data imbalance is a common problem in almost every biomedical application. In ECG, the imbalance is more prominent due to the lack of arrhythmia beats as many of the

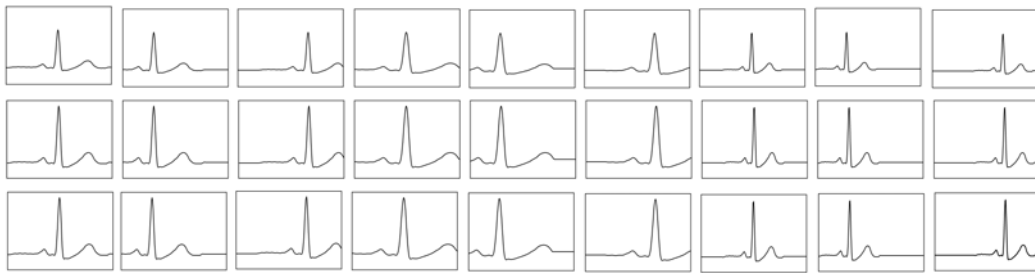


Figure 2.2: **Augmentation on a right bundle branch block beat by applying 26 different combinations of operation on the original beat generated by the proposed augmentation scheme.**

beats are normal even for a patient with cardiovascular disease. With such imbalance in the dataset, the trained model is prone to overfit with the large normal class, as it considers diseases with a smaller number of data as outliers. For 1D ECG signal, the problem still persists as the various augmentation techniques have not that much explored. In [57], 1D ECG data are converted to 2D images and some augmentations are done with different cropping techniques. Due to the 1D properties of ECG data, such techniques don't provide that many variations, especially in the case of smaller classes. In [58], the SMOTE augmentation technique is used which mainly operates through interpolation within various classes. As the position and shapes of the data still vary a significant amount within a class, such interpolations often lead to severe distortions in the generated synthetic data that result in a false representation of the original data class. In our previous work [53], we have provided different augmentation techniques for ECG signals. In this work, we have increased the augmentation techniques by modifying and combining these operations in an algorithmic way to incorporate more realistic variations in the dataset. The following steps are carried out sequentially while performing augmentations.

- **Step 1 (Amplitude Scaling):** The data, firstly, undergoes through amplification, attenuation or no operation. For amplification/attenuation, the scaling factor is chosen randomly from a range that is empirically determined. As in practical cases, there exist some variations in the relative value of amplitudes in ECG data, these introduced variations through amplitude scaling offer good augmented beats.
- **Step 2 (Time Scaling):** Next, the amplitude scaled beat undergoes through dilation, contraction or no operation on the time axis. For dilation, the beat is first over-sampled followed by cropping operation while for contraction, the beat is under-sampled followed by padding operation. The time scaling factor is randomly chosen from an empirically selected range. As in practical cases,

such dilations or contractions in beats are widely visible within a certain limit and the original morphology of the beat isn't changed, it offers a nice technique to introduce variations in the dataset.

- **Step 3 (Shifting):** Finally, the beat undergoes through left shifting, right shifting or no operation. After shifting, some samples of the particular beat are cropped at one edge while some samples are padded at the other edge. The number of samples shifted is chosen randomly from a predefined range. This operation leads to some variations in the information content that leads to making the classification action slightly challenging while demanding more priority to generalize the global features.

By iterating through various choices of these steps, 26 different combinations of operations are performed on the original beat that provides a more realistic representation of the synthesized beats. In Fig. 2.2, all such operations on a particular beat are shown.

2.2 Proposed Deep Neural Network

Once the pre-processed ECG beats are extracted, the next objective is to develop efficient deep convolutional neural network architecture for arrhythmia classification. In this case, instead of using traditional convolution, first, a structural unit is designed based on depthwise separable convolution in the 1D domain and then, a new deep CNN architecture is proposed utilizing the unit. Later, the topology and formation of the proposed deep neural network is presented in detail.

2.2.1 Depthwise separable convolutions for 1D signal

For conventional 2D convolution operation, both spatial convolution of each channel and the inter-channel convolution are performed jointly at the same time. This increases the number of arithmetic operations exponentially with the larger size of kernels. This overhead in computational cost also results in a larger network that becomes prone to overfitting the training data.

In depthwise separable 2D convolution, the spatial convolution and inter-channel convolution operations are performed separately. Generally, at first, a spatial convolution is done on each channel separately and then, pointwise convolution is carried out considering the inter-channel information together. This spatial convolution followed by pointwise convolution is jointly termed as separable convolution. This type

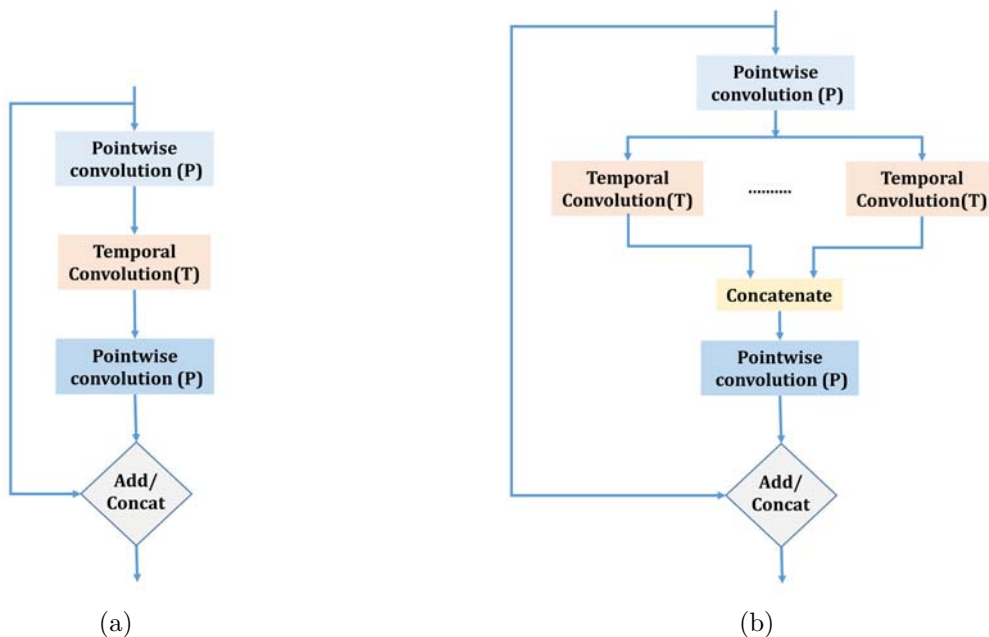


Figure 2.3: **Proposed structural units utilizing (a) single temporal convolution, (b) multiple temporal convolution with various kernels in parallel with prior and post pointwise convolution.**

of convolution offers a similar transformation with a small number of model parameters compared to traditional convolution. It was first proposed in [156] and is widely used for 2D image analysis [157]–[160]. Recently, depthwise separable convolution has also been incorporated for processing speech signals [161], [162].

The ECG signal is a 1D signal and in this case, the depthwise separable 2D convolution needs to be modified. Instead of block-wise spatial convolution on each 2D channel, simple 1D temporal convolution is required to capture temporal information individually. To classify among different classes of arrhythmia with a small size of available data for most of the abnormal classes, the network is highly prone to overfit with the large normal class. Hence, that opens the door of optimization between network overall capacity, i.e. the number of parameters and the network diversity to capture the minuscule difference in features of smaller classes. By utilizing the depthwise separable convolution operation in the 1D domain for ECG arrhythmia classification, both of these objectives can be achieved.

For better understanding, a comparison between traditional 1D convolution and depthwise separable 1D convolution in terms of computational complexity and required number of network parameters is presented here:

Let’s consider an input data of length l_i with C_i number of channels. Standard 1D convolution with kernel size k can be used to transform from input data (l_i, C_i)

to output data (l_i, C_o) , with C_o number of channels, which requires a computational cost of χ_s with ζ_s number of network parameters where:

$$\chi_s = l_i \times C_i \times C_o \times k \quad (2.4)$$

$$\zeta_s = k \times C_i \times C_o \quad (2.5)$$

In case of depthwise separable 1D convolution, the temporal 1D convolution requires a computational cost of $l_i \times C_i \times k$ with $k \times C_i$ number of network parameters and the pointwise convolution requires a computational cost of $C_i \times C_o \times l_i$ with $C_i \times C_o$ number of network parameters which results in total cost of χ_d with total network parameter of ζ_d where:

$$\chi_d = l_i \times C_i \times k + C_i \times C_o \times l_i \quad (2.6)$$

$$\zeta_d = k \times C_i + C_i \times C_o \quad (2.7)$$

Here, the depth of the output feature map is changed by a factor of C_o/C_i .

As a result, the reduction in computational cost is

$$\frac{\chi_d}{\chi_s} = \frac{1}{C_o} + \frac{1}{k} \quad (2.8)$$

and the reduction in the number of network parameter is

$$\frac{\zeta_d}{\zeta_s} = \frac{1}{C_o} + \frac{1}{k} \quad (2.9)$$

Therefore, depthwise separable convolution offers performance comparable to traditional convolution with a large reduction in computational cost with a smaller number of network parameters in case of 1D signals for the increased number of channels with the larger kernel.

2.2.2 Proposed Structural Unit

Based on depthwise separable 1D convolution, a structural unit is designed where before and after the depthwise temporal convolution, inter-channel pointwise convolutions are performed. A simplified schematic of the proposed structural unit is shown in Fig. 2.3a. The motivation and purpose of different convolution operations in this structural unit are described below.

- At first, a pointwise convolution is performed to combine the inter-channel input data information and project the information on a larger space with an

increased number of channels. The depth increase factor is chosen empirically for proper optimization.

- Next, to capture temporal information from each channel, a separate temporal convolution is carried out in the deeper feature map following the prior pointwise convolution. The kernel dimension for this temporal convolution can be varied.
- After that, another pointwise convolution is performed to merge the temporal information of different channels and to project the extracted feature information in a smaller space. The dimension reduction factor of the final space is empirically selected to provide new features from each structural unit after performing the temporal convolution from a larger window.
- Finally, these extracted new features will be combined with the input features through concatenation or addition before entering into the deeper structural unit. Such operations offer the opportunity to go deep with such units while reducing the vanishing/exploding gradient problems by establishing linkage between output and input feature map.

An alternate unit structure is shown in Fig. 2.3b where instead of doing single temporal convolution with larger kernels in deeper sequential structural units, multiple temporal convolutions with varying kernel dimensions are performed in parallel utilizing the broadened feature map from the first pointwise convolution. This will pave the way to combine various temporal correlation collected from smaller to broader time windows (by varying kernel dimension) at the same time with a small increase in computational cost as this temporal convolution will be performed separately on each channel. In this case, all the outputs from temporal convolutions are concatenated depthwise before the final pointwise convolution. Similar to Fig. 2.3a, the output of final pointwise convolution can be added/concatenated with the input feature map.

In the development of the entire architecture, this structural unit can be used (where temporal kernel dimension may be varied) repeatedly to incorporate features from a broader spectrum. Utilizing the variations in proposed structural units, different deep convolutional neural network architectures can be designed for arrhythmia detection and classification. In this study, the most effective and efficient form with multiple parallel temporal kernels (Fig. 2.3b) is used for the construction of ‘DeepArrNet’ architecture which is discussed with implementation details as below.

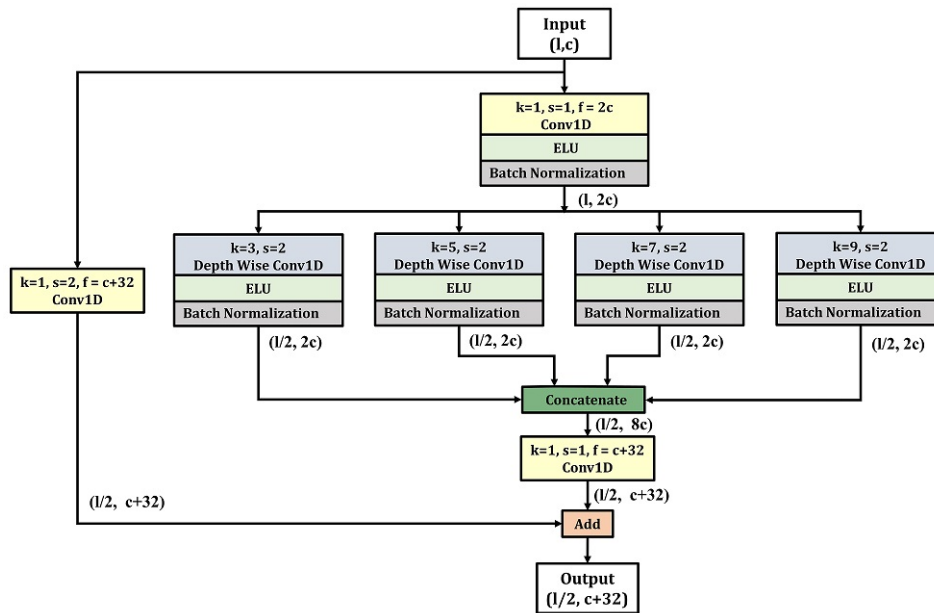


Figure 2.4: DeepArrNet Unit Block utilizing multiple temporal convolutions with various kernels in parallel.

2.2.3 Proposed DeepArrNet Architecture

In the proposed DeepArrNet architecture, to make more efficient use of temporal and pointwise convolution, based on the proposed structural unit shown in Fig. 2.3b, a DeepArrNet Unit Block is designed which is presented in Fig. 2.4. In this block, instead of using a single temporal convolution operation, multiple temporal convolution operations are performed in parallel using various kernel sizes at the same time. In order to limit the computational complexity in parallel temporal convolution operations, a strided temporal convolution is performed which also reduces the length of the output feature map. As a result, this will reduce the computational complexity in the subsequent stages of the proposed DeepArrNet architecture as well, while extracting more generalized features combining various temporal windows. The detail description of the operations performed in ‘DeepArrNet Unit Block’ is summarized below.

- At first, the input data undergo through pointwise convolution with depth increase factor of 2 and nonlinear activation function followed by normalization.
- Following that, the data are passed through four parallel paths to perform separate temporal convolution with strides of 2 on each of them considering varying temporal kernel dimensions (e.g. kernel sizes of 3,5,7 and 9 are chosen here). This strided multi-kernel temporal convolution operations not only reduce the computational complexity but also provide adequate temporal

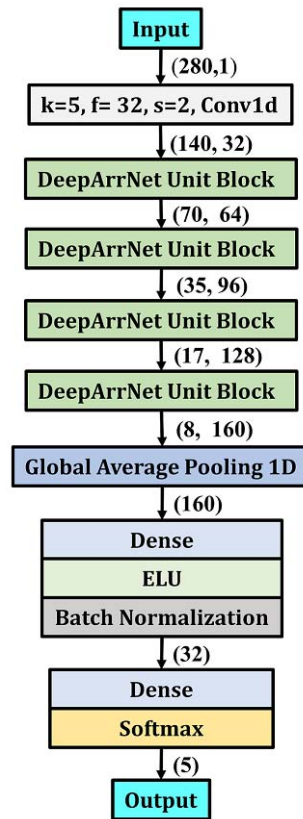


Figure 2.5: **Architecture of the Proposed DeepArrNet.** Here, ‘k’ stands for kernel size, ‘f’ for number of filters and ‘s’ for strides in the convolution.

information extracted from varying observation perspective.

- Next, after undergoing through nonlinear activation and normalization, the output feature maps from all four paths are concatenated vertically which causes an increase in the number of channels.
- Thereafter, pointwise convolution is performed to project the resultant concatenated features on a smaller space. Here, the number of channels in the output feature map is increased by 32 from the input feature map.
- Finally, the output of this pointwise convolution is added with the input feature map after being passed through a strided pointwise convolution operation to generate the final output feature map.

The complete architecture of the proposed DeepArrNet is presented in Fig. 2.5 where the input data are first passed through a standard convolution block. Next, the output is fed to consecutive four ‘DeepArrNet Unit Block’s, In each block, the length of the transformed feature map is reduced while increasing the depth. Hence, after passing through these blocks, final features are obtained with a reduced

length of 8 with a depth of 160 channels. After that, the global average 1D pooling operation followed by traditional classification operations are performed.

Therefore, the proposed DeepArrNet architecture is capable of performing the depthwise temporal and pointwise convolutions efficiently to merge features from different observation windows. Consequently, this results in a very light deep neural network, which can also capture the complex functionality of the data to distinguish among minuscule variations in features of different classes effectively.

2.3 Results and Discussion

In this section, for the purpose of demonstrating the performance of the proposed method experimental results are presented along with performance comparison and detail analysis on the effects of various parameters on the performance. Two publicly available datasets are used to carry out the experimentation. Description of the datasets and method of training/testing are first discussed. Next, the results and analysis are presented focusing on major findings.

2.3.1 Dataset Description

In this work, a very widely used publicly available MIT BIH arrhythmia dataset is used for analyzing the ECG beats [163]. The dataset contains 48 half-hours of two-channel (MLII and V1) ECG recordings collected from 47 patients, which are digitized at 360 samples per second per channel with 11-bit resolution. There are approximately 110,000 ECG beats in this dataset and the beats are classified into five broad categories by the Association for the Advancement of Medical Instrumentation (AAMI) [143]. It is to be noted that out of two channels, similar to most of the research works, only MLII channel is considered as it provides better information regarding the condition of the heart [164].

PTB database [165], another publicly available dataset, is also used for this study. It contains 290 records of which 148 are diagnosed as MI, 52 are healthy and rest ones are with 7 different diseases. Each record contains ECG signals sampled to the sampling frequency of 1000 samples/sec.

2.3.2 Experimental Setup

For performance evaluation of multi-class arrhythmia classification methods, commonly used four performance criteria are used, namely accuracy, sensitivity, positive predictive value (PPV) and F1 score [166].

Table 2.1: Confusion Matrix for Proposed DeepArrNet

Actual	Predicted				
	<i>N</i>	<i>S</i>	<i>V</i>	<i>F</i>	<i>Q</i>
<i>N</i>	89955	118	272	109	135
<i>S</i>	11	2749	8	4	7
<i>V</i>	19	25	7170	7	15
<i>F</i>	6	2	3	789	3
<i>Q</i>	16	7	12	5	7999

* N: Normal, S: Supraventricular, V: Ventricular, F: Fusion, Q: Unknown Beat.

Table 2.2: Effect of Proposed Data Augmentation and Denoising Methods on Sensitivity

Class	Without Proposed Augmentation	Without denoising	Proposed Method
N	98.5	98.8	99.3
S	96.4	98.5	98.9
V	98.2	98.7	99.1
F	95.9	97.8	98.3
Q	98.1	98.6	99.5

* N: Normal, S: Supraventricular, V: Ventricular, F: Fusion, Q: Unknown Beat.

2.3.3 Performance Evaluation

At first, the performance of the proposed architecture is analyzed from different perspectives. Later, the performances of some existing approaches are compared with that of the proposed method.

The proposed architecture is trained on the dataset after completing the pre-processing stage. In Tables 2.1, the confusion matrix obtained by evaluating the proposed architecture is provided. This matrix represents the overall performance of the proposed architecture during testing. It is observed that diagonal values of this matrix, representing the number of correctly predicted beats, are much higher compared to others. Though the number of incorrect predictions seems to be large for the normal class, with a large number of tested beats, these belong to a very small percentage of total normal beats. Moreover, the proposed network consists of 238,629 number of total parameters and maximum accuracy is reached in 99 epochs. Hence, this network is very lightweight that converges to the optimum performance in considerably smaller number of iterations.

In Fig. 2.6, the performance of the proposed network is presented in terms of sensitivity, positive predictive value and F1 score for each class (N, S, V, F, Q). Generally, the presence of a large number of training beats in a particular class can create a bias towards predicting that class in most of the cases which may result in

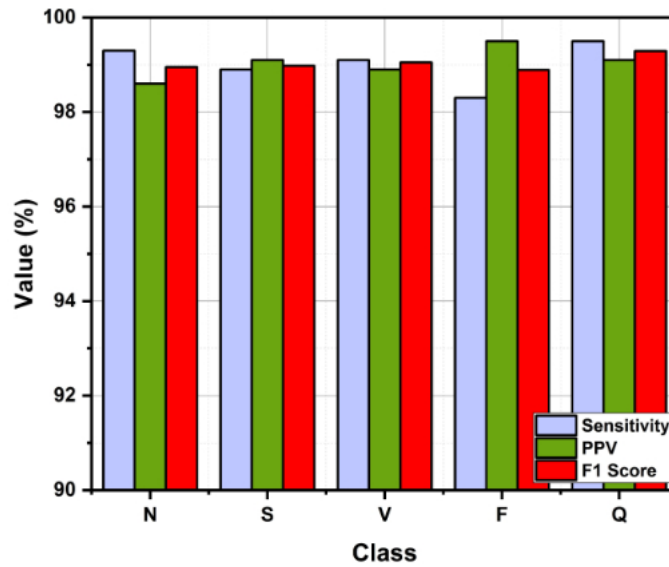


Figure 2.6: Representing per class performance of the proposed network in sensitivity, positive predictive Value, and F1 score in each class.

Table 2.3: Performance of the Proposed Scheme on Arrhythmia Detection Employing Total Arrhythmia and Normal Class

Metric	Value(%)
Positive Predictive Value	99.59
Sensitivity	99.71
Specificity	99.91
Accuracy	99.87

Table 2.4: Performance of the Proposed Scheme on PTB Database [165]

Metric	Value(%)
Accuracy	99.21
Average Precision	99.03
Average Recall	99.12
Average F1 Score	99.08

a higher value of sensitivity in that class. However, this problem is almost overcome with the proposed scheme providing quite satisfactory sensitivity values in all classes. On the other hand, due to the tendency of biasing towards a larger sized class, the number of false-positive predictions from other smaller sized classes likely to be higher that results in lower value of positive predictive value in the larger sized class. It is noticeable that the DeepArrNet architecture provides noteworthy predictive value in all classes consistently. Moreover, the F1 score combines the results of positive predictive value with sensitivity to provide a robust way of comparison. It can be inferred that proposed scheme provides significant performance with high

Table 2.5: Comparison of Proposed Scheme with Existing Methods on Average Values of Evaluation Metrics on MIT BIH Dataset [163]

Work	Method	No. of Class	Accuracy (%)	Average Sensitivity(%)	Average PPV(%)	Average F1 Score(%)
Matris <i>et al.</i> [44]	DWT + SVM	5	93.8	91.5	87.9	89.06
Li <i>et al.</i> [41]	DWT + Random Forest	5	94.6	92.4	91.9	92.15
Derya <i>et al.</i> [59]	RNN	4	97.07	97.15	97.03	97.09
Acharya <i>et al.</i> [52]	1D CNN	5	93.5	93.35	93.47	93.41
Kiranyaz <i>et al.</i> [50]	1D CNN	5	96.4	79.2	68.8	74.1
Proposed DeepArrNet	1D CNN	5	99.28	99.13	99.08	99.11

consistency in the F1 metric also for all five classes.

Due to the scarcity of data for classes with a very small number of members, the trained network is prone to struggle to extract proper features for classification. A proper augmentation method is thus necessary to increase the overall sensitivity of the network. From Table 2.2, significant improvement in the sensitivity of minority ‘S’ and ‘F’ beat classes are noticeable in the proposed method comparing with the one unaccompanied by the proposed realistic augmentation scheme. However, to make fair comparison, in the scheme of without proposed augmentation, the minority classes are oversampled by taking aliases of the existing beats in each training fold to reduce the class imbalance. Hence, this comparison represents the significance of the varieties of realistic augmentation techniques employed during training. Moreover, the denoising operation applied to raw data reduces the complexity of processing and offers more opportunities to generalize various classes. The effect of this denoising operation is clearly visible in Table 2.2 with a noticeable increase in the sensitivity.

In many cases, it is required to detect arrhythmia rather than detecting all detailed arrhythmia classes. In order to demonstrate the performance of the proposed method in such arrhythmia detection problems, all the classes of arrhythmia are considered to be the diseased class as a whole and the proposed methods are applied to identify them with the normal beats. In Table 2.3, the performance of the proposed network in arrhythmia detection is presented. As it becomes a binary classification problem, the proposed network is performing even better in this case. In Tab. 2.4, performance of the proposed architecture is presented on the secondary PTB database [165]. It is clear that the proposed architecture performs consistently on this database also.

A comparative analysis of various approaches with the proposed one is presented in Table 2.5 in terms of the evaluated metrics. Our 1D CNN based proposed architecture with the applied techniques of augmentations and data denoising provide outstanding results that outperform most other approaches. Moreover, the accuracy metric that mainly represents the total number of correct predictions as a whole, provides a significant improvement compared to others. Derya *et al.* [59] provided

an RNN based approach with comparable results in all metrics. However, as RNN is difficult to train compared to CNN while suffering from vanishing and exploding gradient problems, proposed 1D CNN based methods provide better performance. Li *et al.* [41] and Matris *et al.* [44] used traditional handcrafted feature-based approaches using discrete wavelet transform with traditional classifiers commonly used for shallow networks, which offers unsatisfactory performance as expected.

2.4 Conclusion

In this chapter, a deep CNN architecture is proposed for arrhythmia detection and classification from ECG data. It is observed that the proposed architecture provides better generalization among various smaller numbered classes. Moreover, the proposed architecture utilizes various temporal windows in parallel while reducing the feature map using strided convolution. This offers a very lightweight architecture with great generalization capability that becomes the best fit for arrhythmia classification and provides state of the art result in all the evaluation metrics. It is expected that the proposed architecture can also be used in other applications similar to arrhythmia classification employing various other 1D bio-signals.

Chapter 3

CovXNet: A Multi-dilation Neural Network for COVID-19 Diagnosis from Chest X-ray

Coronavirus disease (COVID-19), caused by SARS-CoV-2, has been declared as a global pandemic by WHO that almost collapsed the healthcare systems in many of the countries [85], [167]. Reverse transcription-polymerase chain reaction (RT-PCR), the most commonly used diagnostic test of COVID-19, suffers from low sensitivity in early stages with elongated test period assisting further transmission [86]. Furthermore, the extreme scarcity of this expensive test kit [87] exacerbating the situation. Hence, a chest scan such as X-rays and Computer tomography (CT) scans are prescribed to all individuals with potential pneumonia symptoms for faster diagnosis and isolation of the infected individuals. Though CT scans provide finer details, X-rays are quicker, easier to take, less injurious and more economical alternative. However, due to the scarcity of COVID-19 X-rays, it is extremely difficult to train a very deep network effectively.

In this chapter, an efficient scheme is proposed utilizing relevant available X-ray images for training an efficient deep neural network so that the trained parameters can be effectively utilized for detecting COVID-19 cases even with very smaller size of available COVID-19 X-rays. The major contributions of this chapter is summarized as follows:

1. Instead of using other traditional databases used for disparate applications, a larger database containing X-rays from normal and other non-COVID pneumonia patients are used for transfer learning.
2. A deep neural network is proposed, named as CovXNet, to detect COVID-

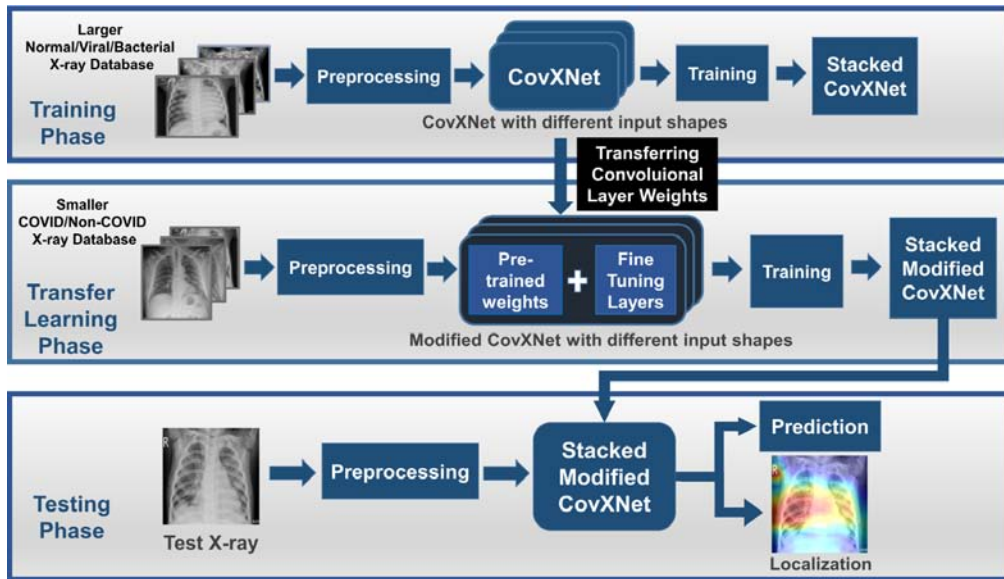


Figure 3.1: **The complete workflow is represented schematically.**

19 from X-rays, which is built from a basic structural unit utilizing depthwise convolutions with varying dilation rates to incorporate local and global features extracted from diversified receptive fields.

3. A stacking algorithm is developed that utilizes a meta-learner to optimize the predictions of different forms of CovXNet operating with different resolutions of X-rays and thus covering diverse receptive fields.
4. The initially trained convolutional layers are transferred directly with some additional fine-tuning layers to train on the smaller COVID-19 X-rays along with other X-rays. This modified network incorporates all its initial learning on X-rays into further exploration of the COVID-19 X-rays for proper diagnosis.
5. A gradient-based localization is integrated for further investigation by circumscribing the significant portions of X-rays that instigated the prediction.
6. Intense experimentations of the proposed methods exhibit significant performance in all traditional evaluation metrics. The primary results obtained from these experimentation are published in [20].

3.1 Methodology

The workflow of the proposed method is schematically shown in Fig. 3.1. As pneumonia caused by COVID-19 contains a high degree of similarity with traditional

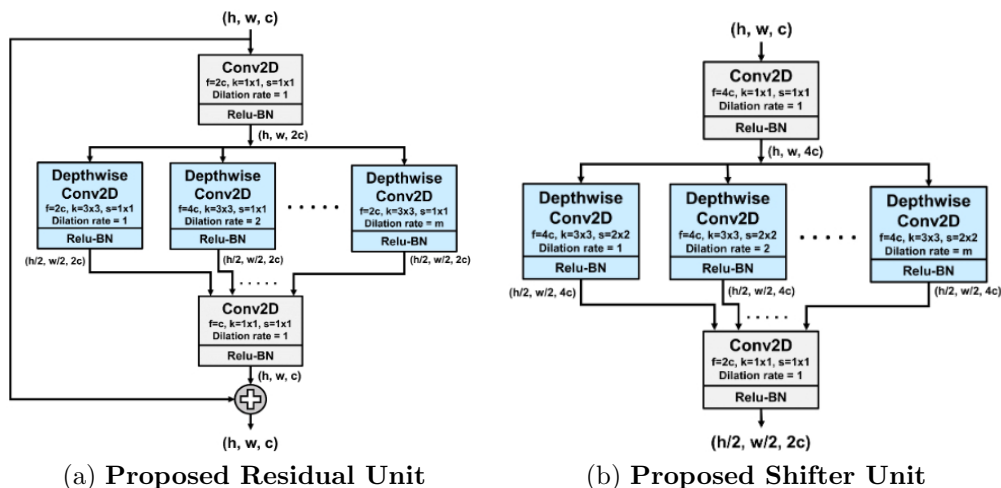


Figure 3.2: **Proposed structural units.** Here, h , w , and c denote the height, width and no. of channels of the feature map, respectively, while ‘ k ’ stands for kernel size, ‘ s ’ for strides and ‘ f ’ for number of filters in the convolution. In depthwise convolution, dilation rate will be varied from 1 to ‘ m ’.

pneumonia from both clinical and physiological perspectives [168], [169], transferring knowledge gained from a large number of chest X-rays collected from normal and other traditional pneumonia patients can be an effective way to utilize smaller COVID-19 X-rays for extracting additional features. Therefore, in the initial training phase, a larger database containing X-rays collected from normal and other non-COVID viral/bacterial pneumonia patients are used for training the proposed CovXNet. Here, after pre-processing, different resolutions of input X-rays are deployed to separately train different CovXNet architectures. Afterward, a stacking algorithm is employed to optimize the predictions of all these networks through a meta-learner. As the convolutional layers are optimized to extract significant spatial features from X-rays, weights of these layers are directly transferred in the transfer learning phase. Next, a smaller database containing COVID-19 and other pneumonia patients are used to train the additional fine-tuning layers integrated with the CovXNet. Finally, in the testing phase, this trained, fine-tuned, stacked modified CovXNet is employed to efficiently predict the test X-ray image class. Moreover, a gradient-based localization algorithm is used to visually localize the significant portion of X-ray that mainly contribute to the decision.

3.1.1 Preprocessing

The collected X-rays pass through minimal preprocessing to make the testing process faster and easier to implement. Images are reshaped to uniform sizes followed by

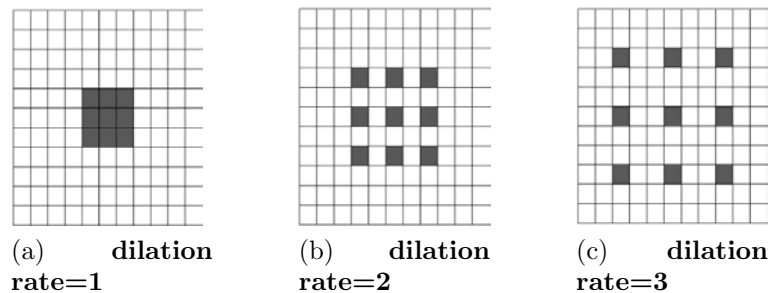


Figure 3.3: **Dilated Convolution for different dilation rates with kernel size (3×3) are encompassing different receptive areas. With increased dilation rate, the receptive area also gets bigger, though kernel size is kept unchanged.**

min-max normalization for further processing with the proposed CovXNet.

3.1.2 Proposed Structural Units

Two structural units are proposed, as shown in Fig. 3.2, which are the main building blocks of the proposed CovXNet architecture. Depthwise dilated convolutions are efficiently introduced in these units to effectively extract distinctive features from X-rays to identify pneumonia.

As the features of pneumonia can be very localized (consolidated) or diffusely distributed over a larger area of the X-rays, it is necessary to incorporate features from different levels of observations [168], [170], [171]. In [172], dilated convolution is introduced to broaden the receptive field of the convolution without increasing the total number of parameters of kernels by increasing dilation rates. This process is presented visually in Fig. 3.3. Various features extracted from different convolutions with varying dilation rates will integrate more diversity in the feature extraction process.

Moreover, traditional convolution can be divided into depthwise convolution followed by a pointwise convolution that makes the process extremely computationally efficient [173]. In depthwise convolution, i.e. a spatial convolution, each input channel is individually filtered by separate filters without combining them. Afterward, a pointwise convolution, i.e. traditional convolution with 1×1 windows, is performed

for projecting the inter-channel features into a new space.

$$\text{DepthwiseConv}(W, y)_{(i,j)} = \sum_{k,l}^{K,L} W_{(k,l)} \odot y_{(i+k,j+l)} \quad (3.1)$$

$$\text{PointwiseConv}(W, y)_{(i,j)} = \sum_m^M W_m \odot y_{(i,j,m)} \quad (3.2)$$

In the proposed structural units, depthwise dilated convolutions along with pointwise convolutions are introduced efficiently. Firstly, the input feature map undergoes through a pointwise convolution to project the inter-channel information into a broader space. Following that, numerous depthwise convolutions are performed with different spatial kernels with varying dilation rates starting from dilation rate of 1 to a max-dilation rate of m . The value of m is adjusted according to the shape of the input feature map for covering the necessary receptive area. Hence, these depthwise convolutions are extracting spatial features from various receptive fields ranging from very localized features to broader perspective generalized features. Thereafter, all these variegated features go through another pointwise convolution to merge these inter-channel features into a constricted space.

In the proposed residual unit, as shown in Fig. 3.2a, this pointwise-depthwise-pointwise convolutional mapping is set to fit a residual mapping by adding the output with the input feature map. This type of residual learning, introduced in [174], is used to capture the identity mapping that helps to produce a very deep network without overfitting. If the proposed residual mapping is denoted by R with input tensor X such that $X \mapsto R(X)$, the final output mapping F can be represented as $F : X \rightarrow [X + R(X)]$. These residual units can be stacked in more numbers to produce a deeper network.

In the proposed shifter unit, as presented in Fig. 3.2b, the input feature map undergoes through some dimensional transformations. Firstly, the depth of the input feature map is increased by 4 times to introduce more processing for spatial reduction. Later, the spatial dimensions are halved through strided depthwise convolution instead of traditional pooling operation as it loses positional information [5]. Such spatial reduction helps to broaden the receptive field for further processing to introduce more generalization. Finally, the depth of the output feature map is doubled in the final pointwise convolution to increase the filtering operations in later stages.

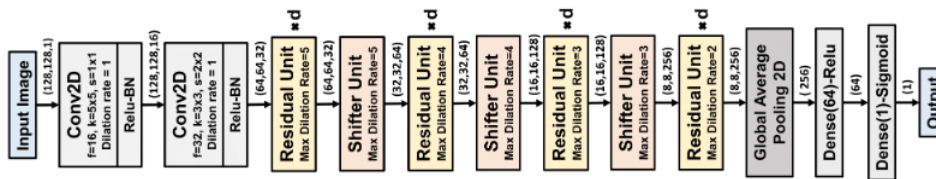


Figure 3.4: Schematic of the Proposed CovXNet architecture optimized for input shape (128, 128, 1). Each residual unit is replicated for ‘d’ times.

3.1.3 Proposed CovXNet Architecture

The residual and shifter units are the main building blocks of the proposed CovXNet architecture, as shown in Fig. 3.4. Firstly, the input image undergoes convolutions with broader kernels to process the information with the larger receptive area. The following convolution introduces some dimensional transformation. Afterward, it passes through a series of residual units. Depth of this stack of residual learning (d) can be increased to produce a deeper network. Shifter units are incorporated in between such stacks to introduce dimensional transformation to generalize the extracted the information further. However, the maximum dilation rate (m) of each residual unit is determined based on the dimension of the input feature map. For processing larger features, m is set to be higher to increase the maximum receptive area of the residual unit accordingly to encompass more variations in the extracted features. Finally, the processed feature map passes through global average pooling followed by some densely connected layers before providing final prediction. Moreover, the rectified linear unit (Relu) is instigated after each convolution for non-linear activation with batch normalization to make the convergence faster.

3.1.4 Stacking of Multiple Networks

The proposed CovXNet architecture can be optimized for input images with different resolutions by adjusting the number and maximum dilation rates of the structural residual and shifter units. Such introduced architectural variations with changing resolutions of X-rays will force these networks to explore the information content from different levels of observations. Though with the reduction of the resolution, information content of an image decreases, it insists the network on focusing the generalized features by broadening receptive area. In the proposed scheme, a stacking algorithm is incorporated to learn the generalizability of these networks by optimizing their predictions to produce a more accurate final prediction. This step can be considered as a meta-learning process and it is schematically presented in Fig. 3.5.

Firstly, total training data is divided into two portions: one for training all

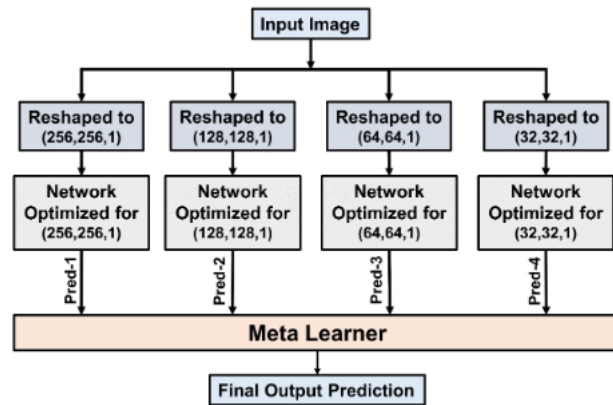


Figure 3.5: Individually optimized networks are stacked together by using the meta-learner for obtaining more-optimized predictions.

the individual networks, while other for training the meta-learner. Next, all the individual networks are trained separately with the resized representations of input images. These networks analyze the data from different perspectives for proper prediction. After being properly optimized, these networks are used to generate a prediction on the other portion of data kept for meta learner training. Finally, the meta learner is being optimized by exploring the predictions of all the individual networks to generate the final output. This approach offers the meta learner to optimize the analysis by inspecting diversified viewpoints. As the meta learner deals with the predictions of individually optimized networks, a very small portion of training data is used to train the meta-learner. Hence, shallow neural networks along with other traditional machine learning techniques can be utilized to build the meta-learner.

3.1.5 Proposed Transfer Learning Method on Novel Corona Virus Data Using CovXNet

As the CovXNet is optimized for analyzing X-rays using very deep architectures with a large number of convolutional layers, this knowledge can be effectively transferred to learn the representation of novel COVID-19 X-rays. This scheme is presented in Fig. 3.6. All the convolutional layers including all residual and shifter units that were initially trained on non-COVID X-rays are directly transferred with their pre-trained weights. Additionally, two more convolutional layers are integrated at the bottom for fine-tuning. Afterward, a traditional global pooling layer with a series of densely connected layers are also incorporated for training. As very few images of COVID-19 X-rays are available, it is difficult to train very deep architecture using them. Nevertheless, as most of the pre-trained convolutional layers are directly

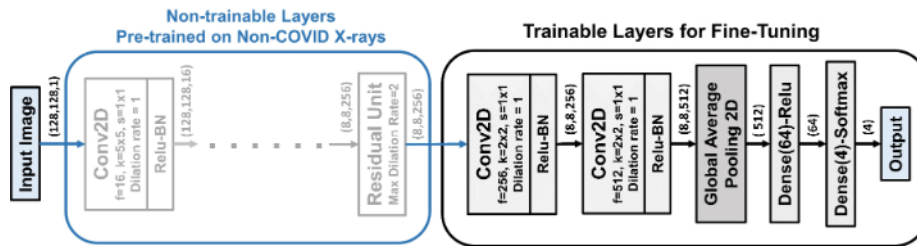


Figure 3.6: **Proposed Transfer learning scheme on CovXNet for fine tuning with small number of images.**

utilized without further training, very few parameters need to be fine-tuned for the newly integrated layers.

3.2 Results and Discussions

In this section, the performances of the proposed schemes are presented with the visual interpretations of the spatial localization from clinical perspectives. Different cases are analyzed with COVID-19 X-rays to explore the robustness of the method. Finally, some state-of-the-art methods for pneumonia detection along with some traditional networks are also compared.

3.2.1 Dataset Description

One of the datasets used in this study is a collection of total 5856 images consisting 1583 normal X-rays, 1493 non-COVID viral pneumonia X-rays and 2780 bacterial pneumonia X-rays collected in Guangzhou Medical Center, China [175]. Another database containing 305 X-rays of different COVID-19 patients is collected from Sylhet Medical College, Bangladesh which is also verified by expert radiologist panel. Finally, a smaller balanced database is created combining all the COVID-19 X-rays with equal number of normal, viral, bacterial pneumonia X-rays (305 X-rays in each class) that are employed for the transfer learning phase (sample images are shown in Fig. 3.7). The rest of the X-rays (Normal, viral, bacterial pneumonia) are utilized for the initial training phase. In both these phases, five fold cross validation scheme is employed for the evaluation of the proposed method.

3.2.2 Experimental Setup

Different hyper-parameters of the network are chosen through experimentation for better performance. Numerous traditional metrics of classification tasks are used

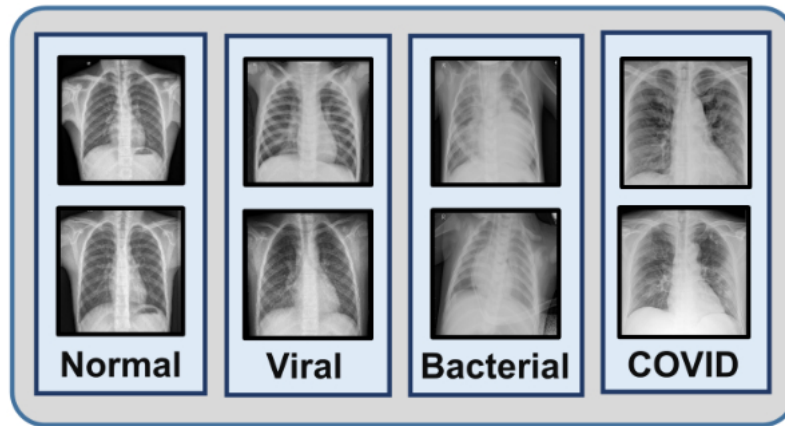


Figure 3.7: Sample X-ray images of normal, viral, bacterial and COVID-19 caused pneumonia patients are shown.

for evaluating the performance of the proposed architectures, such as accuracy, sensitivity, specificity, area under curve (AUC) score, precision, recall, and F1 score.

3.2.3 Performance Evaluation

At the initial training phase, the network is optimized for the normal and other non-COVID viral/bacterial pneumonia X-rays. Different combinations of output classes are experimented for analyzing the inter-class relationships. As the CovXNet architecture is highly scalable to adjust the receptive area depending on the input data, performance with different resolutions of images are experimented with targeting different classes of pneumonia. From the multi-class validation accuracy plot for different resolutions over the training epochs, as shown in Fig. 3.8, it can be observed that the networks with a higher resolution of X-rays lead over smaller ones throughout all the epochs. Nevertheless, the smallest representation still provides comparable performance that indicates the higher generalizability of the proposed CovXNet which can still perform well with very small-scale of information. As a result, utilizing images of different resolutions in the proposed meta-learner, the prediction accuracy is further improved, as shown in Fig. 3.9. It is clearly observed that the meta-learner optimizes the predictions generated from a different level of data representation and provides a significant rise in accuracy for all types of classifications. As different optimized networks are analyzing the data from diversified perspectives, optimizing all of these predictions through additional meta-learner provides a more generalized decision.

After completing the initial training on non-COVID X-rays, these highly optimized convolutional layers are transferred to train with a smaller database containing COVID-19 X-rays. In this transfer learning phase, COVID-19 X-rays are

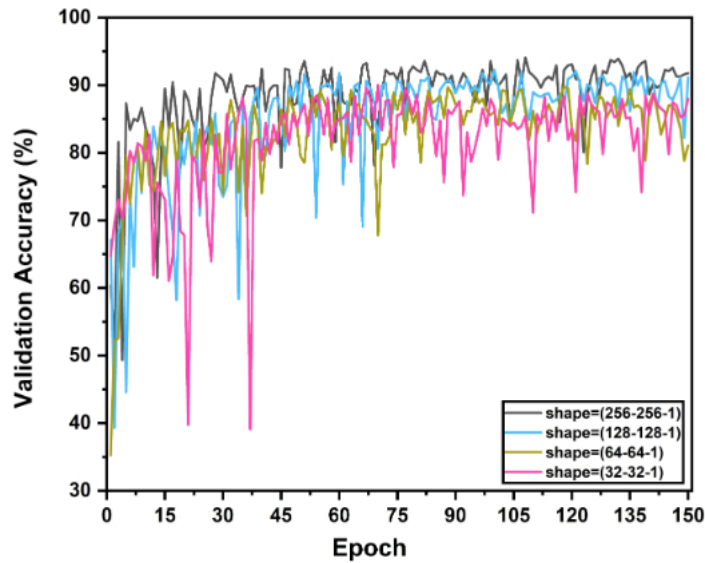


Figure 3.8: Multi-class validation accuracy in different training epochs is shown for different resolutions of inputs.

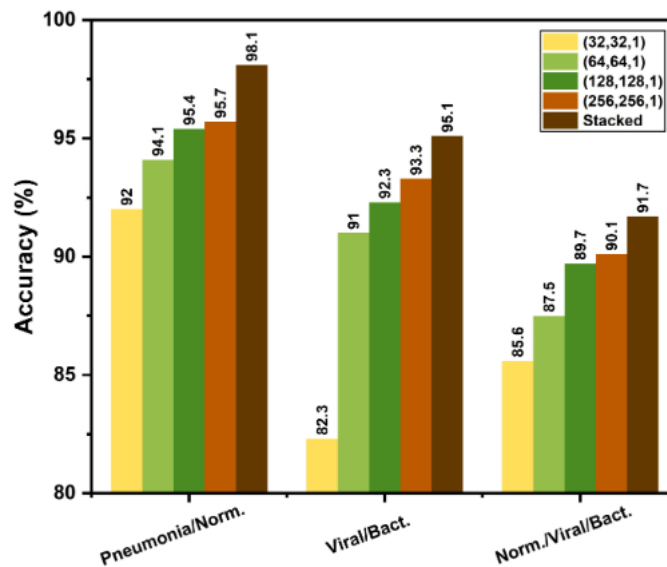


Figure 3.9: Effect of using proposed stacking algorithm in the initial training phase.

experimented with different output classes of normal/traditional pneumonia through fine-tuning of the additionally added layers. Similar to the initial training phase, an additional meta-learner is trained to optimize the predictions obtained from different variants of modified CovXNet that are optimized for different resolutions of input X-rays. The performance of these individually trained networks along with the performance obtained after stacking with meta-learner is shown in Fig. 3.10. As COVID-19 caused pneumonia contains a significant overlap of features with other viral pneumonia [168], [169], it is difficult to isolate these two categories. Hence,

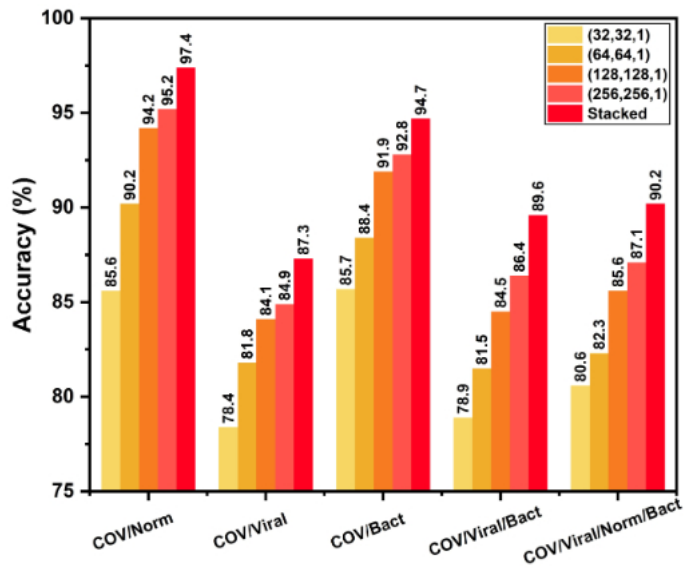


Figure 3.10: Effect of using proposed stacking algorithm in the transfer learning phase.

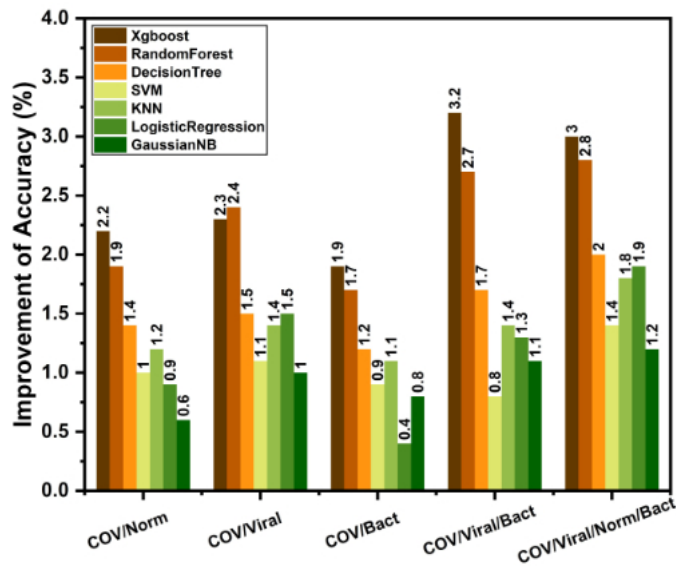


Figure 3.11: Effect of the choice of meta-learner in stacking.

comparably smaller accuracy is noticeable for separating COVID-19 and other viral pneumonia X-rays. However, due to significant variations of features between COVID-19 and normal/bacterial pneumonia X-rays [176], [177], higher accuracy is obtained in such cases. Moreover, stacking with meta-learner provides improved performance in all the classification tasks relating to COVID-19. For example, stacking provides 2.2% improvement of accuracy with respect to the best performing individual network in COVID-19/Normal classification. However, this improvement of accuracy may vary depending on the type of supervised classifier to be used in the

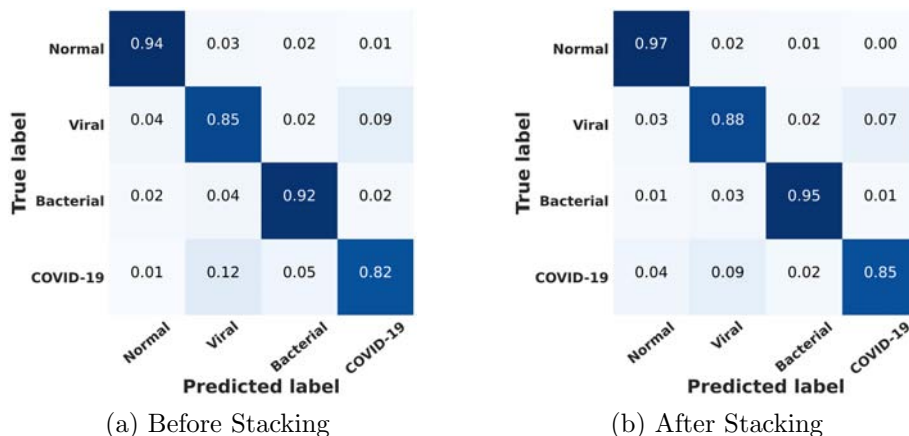


Figure 3.12: Multi-class confusion matrices are shown before and after stacking.

Table 3.1: Performance Comparison of the Proposed Method with Other State-of-the-Art Approaches in Non-COVID Pneumonia Detection

Task	Methods	Accuracy(%)	AUC Score(%)	Precision(%)	Recall(%)	Specificity(%)	F1 Score(%)
Normal/ Pneumonia	Proposed	98.1	99.4	98.0	98.5	97.9	98.3
	Residual	91.2	96.4	90.7	95.9	84.1	93.4
	Inception	88.7	92.6	88.9	94.1	80.2	91.1
	VGG-19	87.2	90.7	85.6	91.1	77.9	89.3
	[178]	95.7	99.0	95.1	98.3	91.5	96.7
	[179]	92.8	96.8	-	93.2	90.1	-
	[180]	96.4	99.3	93.3	99.6	-	-
Viral/ Bacterial Pneumonia	Proposed	95.1	97.6	94.9	96.1	94.3	95.5
	Residual	89.5	92.4	88.3	96.9	78.1	92.4
	Inception	85.8	90.6	84.5	93.8	72.1	88.9
	VGG-19	83.2	88.5	81.1	91.3	71.7	86.6
	[178]	93.6	96.2	92.0	98.4	86.0	95.1
[179]	90.7	94.0	-	88.6	90.9	-	
Normal/ Viral/ Bacterial/ Pneumonia	Proposed	91.7	94.1	92.9	92.1	93.6	92.6
	Residual	86.3	88.5	86.3	88.5	93.5	87.4
	Inception	81.1	84.6	75.4	84.9	86.2	78.9
	VGG-19	79.8	83.1	74.5	82.9	83.4	77.9
	[178]	91.7	93.8	91.7	90.5	95.8	91.1

meta learner phase. For experimentation, different classifiers are tested, such as Xgboost, random forest, decision tree, SVM, KNN, logistic regression and Gaussian naive bias algorithm. Improvement of performance with different meta-learners are shown in Fig. 3.11 for different classification tasks. Xgboost and RandomForest algorithm provide the best performance as these learners provide prediction after further ensembling of several boosting and bagging algorithms, respectively.

The multi-class confusion matrix is provided in Fig. 3.12. As expected, due to a high degree of overlapping features, a few COVID-19/viral cases exhibit misclassification. However, very satisfactory performance is obtained for other classification cases. However, recall of all of the classes can be improved further by incorporating the meta-learner through the stacking of different networks.

The performance of the proposed schemes in the initial training phase on non-

Table 3.2: Performance Comparison of the Proposed Method with Other Traditional Networks on COVID-19 and Other Pneumonia Detection.

Task	Methods	Accuracy(%)	AUC score(%)	Precision(%)	Recall(%)	Specificity(%)	F1 Score(%)
COVID/ Normal	Proposed	97.4	96.9	96.3	97.8	94.7	97.1
	Residual	92.1	91.2	90.4	93.4	89.2	91.9
	Inception	89.5	84.3	89.1	87.7	83.2	88.4
	VGG-19	85.3	82.7	86.3	83.9	79.9	85.1
COVID/ Viral Pneumonia	Proposed	87.3	92.1	88.1	87.4	85.5	87.8
	Residual	80.4	78.9	81.1	79.3	77.1	80.2
	Inception	78.2	75.5	76.8	79	75.4	77.9
	VGG-19	72.1	67.7	70.9	74.7	69.3	72.8
COVID/ Bacterial Pneumonia	Proposed	94.7	95.1	93.5	94.4	93.3	93.9
	Residual	84.2	80.3	86.7	83.5	82.4	85.1
	Inception	83.1	79.9	82.2	85.2	83.6	83.7
	VGG-19	77.2	75.5	73.3	80.3	71.4	76.8
COVID/ Viral/ Bacterial Pneumonia	Proposed	89.6	90.7	88.5	90.3	87.6	89.4
	Residual	82.1	79.8	81.5	80.3	78.5	80.9
	Inception	84.3	83.1	81.4	85.9	80.8	83.7
	VGG-19	79.1	77.5	76.5	80.7	77.2	78.6
COVID/ Normal/ Viral/ Bacterial	Proposed	90.2	91.1	90.8	89.9	89.1	90.4
	Residual	82.3	80.7	82.7	79.5	80.7	81.1
	Inception	82.9	79.8	80.6	84.3	82.4	82.5
	VGG-19	80.8	78.5	77.4	81.6	78.1	79.5

COVID X-rays is compared with other existing approaches in Table 3.1. Here, the performance of different traditional architectures [174], [183], [184], developed for other computer vision applications, are compared with our proposed CovXNet. Additionally, performance of some state-of-the-art AI-based pneumonia detection schemes [178]–[180] are also compared. Rajraman *et al.* [178], Kermayn *et al.* [179], and Chouhan *et al.* [180] utilized conventional transfer learning schemes using pre-trained networks on ImageNet database for traditional pneumonia detection. The proposed schemes outperform most other approaches by a considerable margin.

In Table 3.2, the performance of the proposed CovXNet is compared with other traditional networks on COVID-19 and other types of pneumonia detection. It can be observed that the proposed CovXNet architecture provides significantly better performance in different classification tasks handling with COVID-19 X-rays compared to other traditional architectures. Moreover, in Table 3.3, the proposed method is compared with other existing state-of-the-art approaches for COVID-19 detection from X-rays. As the proposed schemes utilized all the non-COVID X-rays in the initial learning phase, final training and evaluation is carried out on the separated balanced database containing X-rays of COVID patients. Ozturk *et al.* [181] proposed a deep neural network based approach without applying transfer learning strategies. Whereas, Wang. *et al.* [90], Ioannis *et al.* [91], Sethy *et al.* [92], and Narin *et al.* [93] used traditional networks with conventional transfer learning scheme from ImageNet database. In most of these cases, the obtained result is biased due to the small amount of COVID-19 X-rays. It should be noticed that the proposed schemes provide consistent performance in different combinations of classification with balanced set of data. Moreover, the larger number of non-COVID

Table 3.3: **Performance Comparison of the Proposed Scheme with Other State-of-the-Art Approaches on COVID-19 and Other Pneumonia Detection**

Work	Amount of Chest X-rays	Architecture	Accuracy(%)
Ozturk <i>et al.</i> [181]	125 COVID-19 + 500 No finding	DarkCovidNet	98.08
	125 COVID-19 + 500 Pneumonia + 500 No finding		87.02
Wang <i>et al.</i> [90]	53 COVID-19 + 5526 Non-COVID	COVID-Net	92.4
Ioannis <i>et al.</i> [91]	224 COVID-19 + 700 Pneumonia + 504 Normal	VGG-19	93.48
Sethy <i>et al.</i> [92]	25 COVID-19+ 25 Non-COVID	ReNet-50/SVM	95.38
Narin <i>et al.</i> [93]	50 COVID-19 + 50 Non-COVID	ResNet-50	98
Proposed	305 COVID-19+ 305 Normal	Stacked Multi-resolution CovXNet	97.4
	305 COVID-19 + 305 Viral Pneumonia		87.3
	305 COVID-19+ 305 Bacterial pneumonia		94.7
	305 COVID-19 + 305 Viral Pneumonia + 305 Bacterial pneumonia		89.6
	305 COVID-19+ 305 Normal+ 305 Viral Pneumonia+ 305 Bacterial Pneumonia		90.3

Table 3.4: **Performance on Additional 468 X-ray Images of COVID-19 patients [182]**

Metrics	Values
True Positive	403 (86.1%)
False Negative	65 (13.9%)
Total	468

X-rays are properly utilized for initial training phase that is effectively transferred for diagnosing COVID-19 and other pneumonias in the final transfer learning phase. Additionally, performance of the proposed scheme has been tested on additional 468 chest X-ray images of COVID-19 patients collected from [182]. Here, the proposed scheme provides consistence performance that validates its robustness on real-time test scenario.

Gradient-based class activation mapping (Grad-CAM) algorithm [185] is integrated with the proposed CovXNet to generate the class activation mapping for localizing the particular portion of the X-rays that mainly instigated the decision. By superimposing the heatmap with the input X-rays, such localizations are studied further to interpret the learning of the network from the clinical perspective. In Fig. 3.13, some of the X-rays with imposed localization are shown. Following findings are summarized:

- In normal X-rays, no kind of opacity is present that isolates the normal patients from all kinds of pneumonia patients having some form of opacities [170], [171],

[186]. In Fig. 3.13, it is observed that no significant region is localized for normal X-rays. As it is more distinguishable, it is easier to isolate from other patients.

- By carefully examining the heatmaps generated for traditional viral pneumonia, it can be observed that our model has localized regions with bilateral multifocal ground-glass opacities (GGO) along with patchy consolidations in some of the cases. Additionally, some localized regions contain diffused GGOs and multilobar infiltrations. These localized features are also commonly approved radiological features of traditional viral pneumonia [86], [168], [170], [186].
- In the case of bacterial pneumonia, the localized activation heatmaps are mainly involving opacities with consolidation on lower and upper lobes. Additionally, there is also the involvement of both unilateral and bilateral along with peripheral. According to [170], [171], these features mainly represent bacterial pneumonia.
- According to [168], [169], there are lots of similarities between COVID-19 and traditional viral pneumonia both demonstrating bilateral GGOs along with some patchy consolidations. Some more likely features of COVID-19 caused pneumonia are reported in [168], [169], [176], [177], such as peripheral and diffuse distribution, vascular thickening, fine reticular opacity along with the conventional viral-like ground-glass opacities. By carefully examining the generated heatmap from some of the COVID-19 infected X-rays (Fig. 3.13), it is distinguishable that peripheral and diffuse distribution of such opacities is diagnosed. Moreover, vascular thickening is also localized for some of the cases along with other traditional viral features.

Therefore, the radiological features extracted and localized by the proposed CovXNet provide substantial information about the underlying reasons for pneumonia. This type of localization can assist the clinicians to analyze the prediction obtained from the proposed scheme.

3.3 Conclusion

Due to significant overlapping characteristics between COVID-19 and other pneumonia, by transferring the initially trained convolutional layers with some additional fine-tuning layers, a very satisfactory result is obtained with a smaller database

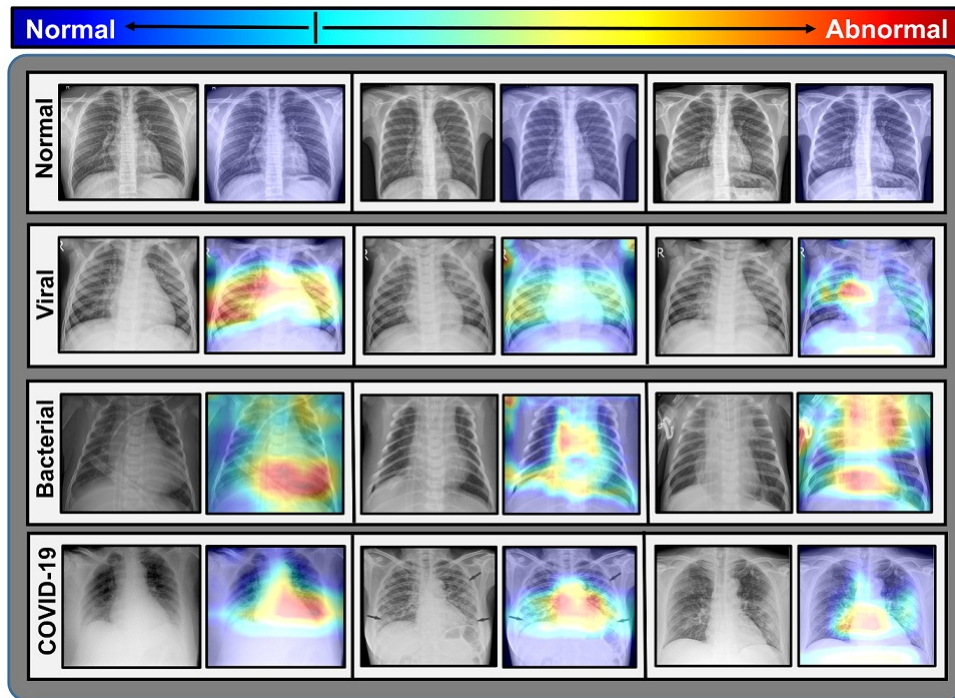


Figure 3.13: Significant portions of the test X-rays that instigate the decision are localized by imposing the activation heatmap obtained from CovXNet.

containing COVID-19 X-rays. Moreover, it is observed that a stacking algorithm provides additional performance improvement by further optimizing predictions obtained from different variations of CovXNet that are primarily optimized with various resolutions of input X-rays. Furthermore, a generated class activation map provides discriminative localization of the abnormal zones that can assist to diagnose the variations of clinical features of pneumonia on X-rays. Experimental results obtained from extensive simulations suggest that it can be a very effective choice for faster diagnosis of COVID-19 and other pneumonia patients.

Chapter 4

Multi Stage Learning for Human Activity Recognition from Multi-modal Wearable Sensors

Activity recognition using wearable sensors has been a trending topic of research for its widespread applicability on diverse domains ranging from health care services to military applications [187]. With the ubiquitous availability of modern mobile devices such as smartphones, tablets, and smartwatches, various types of sensor data are available that can be utilized effectively in numerous applications like activity recognition. Various types of sensor data along with image and video data have been employed for recognizing human activity [188]. In particular, time series wearable sensor data, e.g. accelerometer, gyroscope, and magnetometer are easy to obtain even with our smart devices and can be used to recognize human activity from distant position on real-time basis as these sensors' data are very small in volume and easy to share through internet. In the previous chapters, architectural modifications are explored for incorporating features from diverse receptive areas. Along with the architectural modifications, numerous transformations on the raw data can be introduced with novel optimization strategies for introducing newer perspectives on the feature extraction process.

In this chapter, we have proposed a novel multi-stage training methodology to make accurate recognition of human activity from multi-modal time-series sensor wearable sensor data by efficiently employing a multitude of time-series transformations that facilitates the exploration of diversified feature spaces. The major contributions of this chapter is summarized as follows:

1. Instead of relying on a single transformed space, features from numerous transformed spaces are integrated together to make the process more resilient from

noise and other random perturbations. The proposed approach opens scopes for optimization of diversified features extracted from numerous representations of the raw data.

2. An efficient deep convolutional neural network architecture is proposed that can be separately tuned and optimized as efficient feature extractors from different transformed spaces.
3. A two-stage training algorithm is proposed to combine the separately optimized networks operating on different transformed spaces into an unified architecture utilizing an additional combined training stage.
4. A multi-stage sequential training algorithm is proposed for sequentially converging the optimum feature representations obtained from numerous transformed spaces through sequential weighting, optimization and integration of multi-transformed features. This scheme makes it possible to optimize the unified architecture with a smaller amount of available training data in several stages.
5. Different types of realistic data augmentation techniques have been introduced to increase the variations of the available data.
6. Results of intense experimentations have been presented using three publicly available datasets that provide very satisfactory performance compared to other state-of-the-art approaches. The primary results of the experimentation are published in [19].

4.1 Methodology

The proposed multi-stage training approach is represented in Fig. 4.1. In the first stage of training, individual feature extractors operating on different transformed spaces are trained in parallel with separate classifiers. In the literature, varieties of feature extractors from time-series data have been explored ranging from PCA, ICA, wavelet-based methods to modern CNN, DNN, LSTM, and numerous deep learning methods [69], [73]–[75]. To overcome additional complexities mainly arising from the difficulty of feature selection and optimization from different diversified transformed representations of time series data, we have proposed deep CNN architectures as feature extractors from different transformed domains. As it is completely data-driven, deep CNN architecture can be trained as an efficient feature extractor from any representation of data. For joint optimization of multiple transformed feature

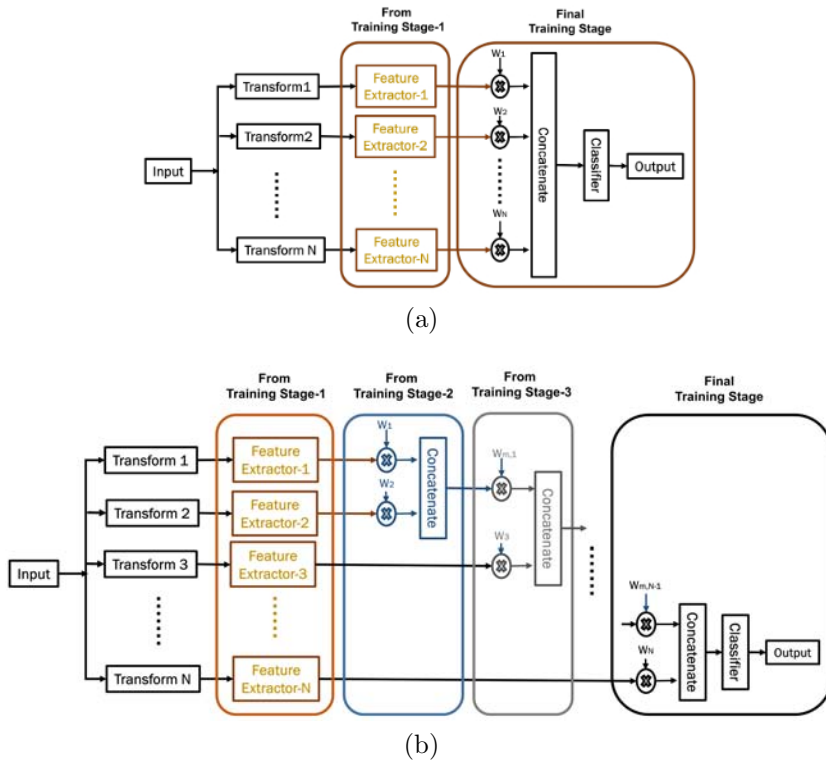


Figure 4.1: Multiple training stages are utilized to incorporate features from numerous transformed representations of input sensor data. (a) Two stage training, and (b) Multiple sequential stages of training.

spaces, learning of this first training stage is transferred into a unified structure utilizing another combined training stage (Fig. 4.1a) or utilizing a number of sequential training stages (Fig. 4.1b).

After completing the first stage of training, all the separate classifiers of individual networks are removed. As a result, when input data is fed to the network, representational features extracted from different transformed domains utilizing the trained feature extractors are available which were fed into separate classifier units in the first training stage. However, the feature quality can be varied with the transformation of the raw data which can be visible by evaluating the performance of the separate feature extractors in the first stage. Hence, in the second and final stage of training (Fig. 4.1a), these feature vectors are multiplied by separate weighting vectors to increase the selectivity of the system. Later, all these weighted feature vectors are concatenated together and a common dense classifier unit is trained to provide the exact prediction from these concatenated features. Therefore, these weighting vectors, along with the combined dense classifier unit, are supposed to learn in this stage of training utilizing the data again.

In Fig. 4.1b, the proposed multi-stage sequential training is shown. In the two-

stage training, as described, the final second-stage training learns the weighting vector for each feature map at the same time with the combined classifier. However, in the multi-stage sequential training, weighting vectors for only two feature vectors, extracted by the feature extractors trained in the previous stage, are learned along with the combined classifier at a time. In the following stage, the classifier is removed and the merged weighted feature vectors of these two transformations undergo through similar next stage of training with one of the remaining feature vectors representing different transformation. Thus, in each stage of sequential training, weighted feature vector from an additional transformed space is accumulated with the combined feature extractors trained in the previous stage. This method of sequential training offers additional opportunity to converge individual feature representations corresponding to variegated transformed spaces to the final decision-making process by optimizing two feature vectors sequentially. Moreover, in deep learning-based approaches, these weighting vectors applied on separate feature vectors can be easily integrated by introducing a separate densely connected layer operating on each feature vectors accompanied by different weighting vectors.

4.1.1 Transformations on Time Series Data

Different types of transformations on time series data have been utilized in the proposed approach. These are described briefly as below.

Gramian Angular Field Transformation (GAF)

Gramian angular field transformation maps the elements of a $1D$ time-series data into a $2D$ matrix representation. This encoding scheme preserves the temporal dependency of the original time series data along the diagonal of the encoded matrix while the non-diagonal entries essentially represent the correlation between samples [80]. In this transformation, $G : R^N \rightarrow R^{N \times N}$, the input time series, X , is transformed into polar coordinate (r, ϕ) after normalization.

$$\phi_i = \cos^{-1}(x_i), \quad -1 \leq x_i \leq 1, \quad x_i \in X \quad (4.1)$$

$$r_i = \frac{t_i}{N}, \quad t_i \in \mathbb{N} \quad (4.2)$$

Here, t_i the time stamp and N is a constant factor to regularize the span of the polar coordinate system. These polar angles are utilized to get the final transformed matrix G , which is,

$$G_{i,j} = \cos(\phi_i + \phi_j), \quad i, j = 1, 2, \dots, n \quad (4.3)$$

Recurrence Plotting

The recurrence plot portrays the inherent recurrent behavior of time-series, e.g. irregular cyclicality and periodicity, into a $2D$ matrix [81]. This method provides a way to explore the m -dimensional phase space trajectory of time series data for generating a $2D$ representation by searching points of some trajectories that have returned to the previous state and is represented by,

$$R_{i,j} = \theta(\epsilon - \|\mathbf{s}_i - \mathbf{s}_j\|), \mathbf{s}(\cdot) \in R^m, i, j = 1, 2, \dots, K \quad (4.4)$$

where K is the number of considered states \mathbf{s} , ϵ is a threshold distance, $\|\cdot\|$ a norm and $\theta(\cdot)$ is the Heaviside function.

Scattering Wavelet Transformation

Scattering wavelet transform offers representational features of the time-series data those are rotation/translation-invariant while remaining stable to deformations. This technique provides the opportunity to extract features from a very small number of data [82]. A mortlet wavelet function, defined as mother wavelet, undergoes through convolution operation with the raw time series data while being scaled and rotated, and thus creates different levels of representational features.

Let's consider, W_j and U_j to be the averaging operation and complex modulus of the averaged signal, respectively, for order j ($0, 1, \dots, L$) of the scattering coefficients, and these coefficients can be described as

$$S_j = W_j U_j S_{j-1} * |\psi_j| * \phi_j, \quad (4.5)$$

where ϕ_j represents the Gaussian low pass filter and ψ_j represents the mortlet wavelet function of order j . Therefore, a scattering representation, S_X of time series data, X , is obtained by concatenating the scattering coefficients of a different order,

$$S_X = [S_0 X, S_1 X, \dots, S_L X] \quad (4.6)$$

As multi-channel sensor data collected from numerous sensors have been used in this work, each channel of such time-series data is transformed individually using any of these transformations, and all such transformed data are stacked together maintaining a similar time information in all the channels. Later, they undergo through the feature extraction process utilizing deep neural networks.

4.1.2 Proposed Deep Neural Network Architectures

For feature extraction and classification, two deep CNN architectures are proposed, as shown in Fig. 4.2a and Fig. 4.2b, optimized to operate in 1D and 2D domain, respectively. Both of them are very similar to each other, as the objective of them is to extract features for activity recognition, with some modifications to operate in different domains for handling different dimensions of data. In general, the proposed CNN architecture mainly consists of a CNN base part followed by a top classifier layer. The CNN base part involves a number of convolution and pooling operations while the top classifier layer consists of a series of densely connected layers followed by the final activation layer to generate activity prediction. The operations performed here are discussed below.

- i. The input 1D time-series data undergo an initial transformation operation as discussed above before starting the convolutional filtering in the deep network.
- ii. Next, the tensor enters the convolutional base part where it passes through a series of unit residual block operations to extract deep features from a broad spectrum. Different representations of these unit residual blocks are shown in Fig. 4.3 with some variations in operations for handling 1D (Fig. 4.3a) and 2D (Fig. 4.3b) data. In these blocks, the input tensor passes through two different operations in parallel and the transformed tensors get added later to produce the final output tensor. Subsequently, a global average pooling operation is performed to extract the global features from each channel of the transformed tensor. This CNN base part extracts effective temporal/spatial features through convolutional filtering and pooling operations required for the final decision.
- iii. After that, the tensor propagates through the top classifier block where series of densely connected layers explore the extracted features of the CNN base part to get higher level of representation with the softmax activation layer at the end to merge these representations into a specified class of action.

The values of different convolutional kernel sizes, number of convolutional layers in each unit block, and number of unit residual blocks are established through experimentation to reach the maximum performance. Shallower networks are prone to underfit with the training data while deeper networks are prone to overfit. However, the proposed network effectively utilizes efficient separable convolutions along with residual operations to reduce vanishing gradient and overfitting issues for achieving optimum performance.

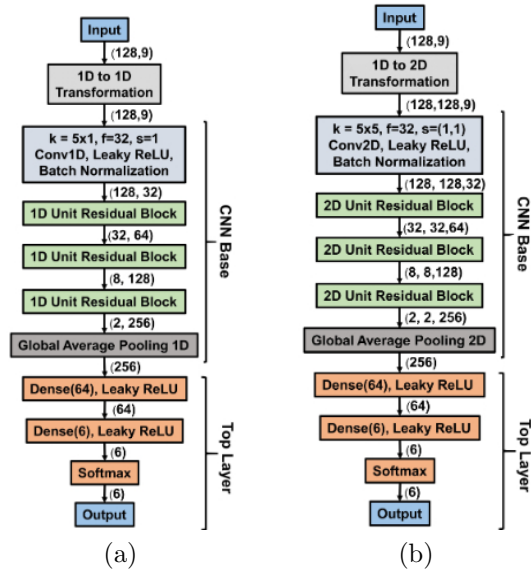


Figure 4.2: Proposed (a) 1D Convolutional Neural Network and (b) 2D Convolutional Neural Network.

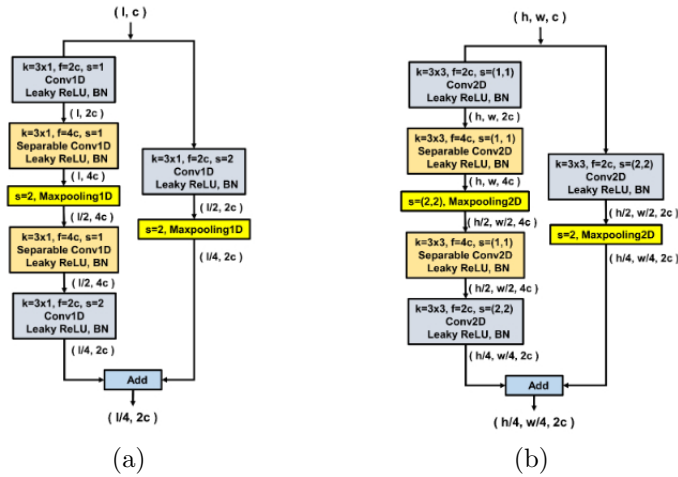


Figure 4.3: Proposed (a) 1D unit residual block and (b) 2D unit residual block.

4.1.3 Proposed Multi-Stage Training Scheme

In the proposed training method, a number of training stages have been utilized to combine features from different transformed spaces. In Fig. 4.4, this scheme is represented schematically. These optimizations of individually trained feature extractors can be done in two stages or number of sequential stages. Algorithm 1 and 2 are executed for implementing two-stage training scheme, and multi-stage sequential training scheme, respectively. Operations performed in different stages are described below.

Algorithm 1: Proposed Two-Stage Training Method

```

Data: training sample  $\mathbf{X}$ ; training label  $y_{actual}$ 
Result: weight matrices  $\mathbf{D}, \mathbf{F}$ 
/* Individual training begins */
1 for  $i \leftarrow 1$  to  $N$  do
2   Calculate  $\hat{X}_i = T_i(X)$ ;
3   Randomize  $D_{1,i}^l$  and  $F_i$ , for  $l = [1, \dots, L]$ ;
4   while The training error threshold is unsatisfied do
5     Calculate  $f_i = F_i(\hat{X}_i)$ ;
6     Find  $y_{pred,i}^1 = D_{1,i}^L(D_{1,i}^{L-1}(\dots(D_{1,i}^1(f_i))))$ ;
7     Find loss  $L_{1,i} = \mathcal{L}(y_{pred,i}^1, y_{actual})$ ;
8     Update  $D_{1,i}^l$  and  $F_i$ , for  $l = [1, \dots, L]$ ;
9   end
10  Calculate  $d_i = F_i(\hat{X}_i)$ ;
11 end
/* Combined training stage begins */
12 Randomize  $D_2^m$ , for  $m = 1, \dots, L'$ ;
13 while The training error threshold is unsatisfied do
14   for  $i \leftarrow 1$  to  $N$  do
15     Set,  $f_i = D_{2,i}^1(d_i)$ ;
16   end
17   Set feature mapping group,  $f = [f_1, f_2, \dots, f_N]$ ;
18   Find  $y_{pred}^2 = D_2^{L'}(D_2^{L'-1}(\dots(D_2^2(f))))$ ;
19   Find loss  $L_2 = \mathcal{L}(y_{pred}^2, y_{actual})$ ;
20   Update  $D_2^m$ , for  $m = 1, \dots, L'$ ;
21 end
    
```

F_i denotes the CNN base part of i^{th} transform.

D_n^l denotes the l^{th} densely connected layer of n_{th} training stage.

T_i denotes the i_{th} transformation on raw data.

- i. **Individual training stage:** This stage is common for both two-stage and multi-stage training schemes. In this stage, separate CNN base parts with associate dense classifiers are trained individually to prepare the CNN base part as an efficient feature extractor for the respective transformed domain, as shown in Fig. 4.4a. Here, the identity transform is also used to incorporate features from unaltered raw data along with other transformations. However, some of these transformations contain more distinctive features related to the final activity recognition compared to others that lead to variations of performance after being trained.
- ii. **Combined training stage:** After the first training stage, an additional combined training stage is employed to combine all these individually trained

Algorithm 2: Proposed Sequential Training Method

```

Data: training sample  $\mathbf{X}$ ; training label  $y_{actual}$ 
Result: weight matrices  $\mathbf{D}, \mathbf{F}$ 
/* Individual training begins */
1 for  $i \leftarrow 1$  to  $N$  do
2   Calculate  $\hat{X}_i = T_i(X)$ ;
3   Randomize  $D_{1,i}^l$  and  $F_i$ , for  $l = [1, \dots, L]$ ;
4   while The training error threshold is unsatisfied do
5     Calculate  $f_i = F_i(\hat{X}_i)$ ;
6     Find  $y_{pred,i}^1 = D_{1,i}^L(D_{1,i}^{L-1}(\dots(D_{1,i}^1(f_i))))$ ;
7     Find loss  $L_{1,i} = \mathcal{L}(y_{pred,i}^1, y_{actual})$ ;
8     Update  $D_{1,i}^l$  and  $F_i$ , for  $l = [1, \dots, L]$ ;
9   end
10 end
/* Sequential training begins */
11 Initialize  $F_{merged,1} = F_1$ ;
12 for  $n \leftarrow 2$  to  $N$  do
13   Set  $\hat{X}_{merged,n} = [\hat{X}_1, \dots, \hat{X}_{n-1}]$ ;
14   Randomize  $D_n^m$ , for  $m = 1, \dots, L'$ ;
15   while The training error threshold is unsatisfied do
16     Set  $f_{1,n} = D_n^1(F_{merged,n-1}(\hat{X}_{merged,n}))$ ;
17     Set  $f_{2,n} = D_n^2(F_n(\hat{X}_n))$ ;
18     Set feature mapping group,  $f_n = [f_{1,n}, f_{2,n}]$ ;
19     Find  $y_{pred}^n = D_n^{L'}(D_n^{L'-1}(\dots(D_n^3(f_n))))$ ;
20     Find loss  $L_n = \mathcal{L}(y_{pred}^n, y_{actual})$ ;
21     Update weights of  $D_n^m$ , for  $m = 1, \dots, L'$ ;
22   end
23   Calculate  $\hat{F}_{n-1} = D_n^1 \circ F_{merged,n-1}$ ;
24   Calculate  $\hat{F}_n = D_n^2 \circ F_n$ ;
25   Set  $F_{merged,n} = [\hat{F}_{n-1}, \hat{F}_n]$ ;
26 end
    
```

F_i denotes the CNN base part for i^{th} transform.

D_n^l denotes the l^{th} densely connected layer of n_{th} training stage.

T_i denotes the i_{th} transformation on raw data.

feature extractors for the proposed two-stage training scheme, as shown in Fig. 4.4b. In this stage, all individual top dense classifier blocks trained in the first stage are removed while all CNN base parts are used unaltered as they are finely tuned as efficient feature extractors. Next, a separate densely connected layer is introduced on top of each CNN base part to reduce the extracted spatial/temporal features into more general representation. These separate densely connected layers act as the weighting vectors for feature se-

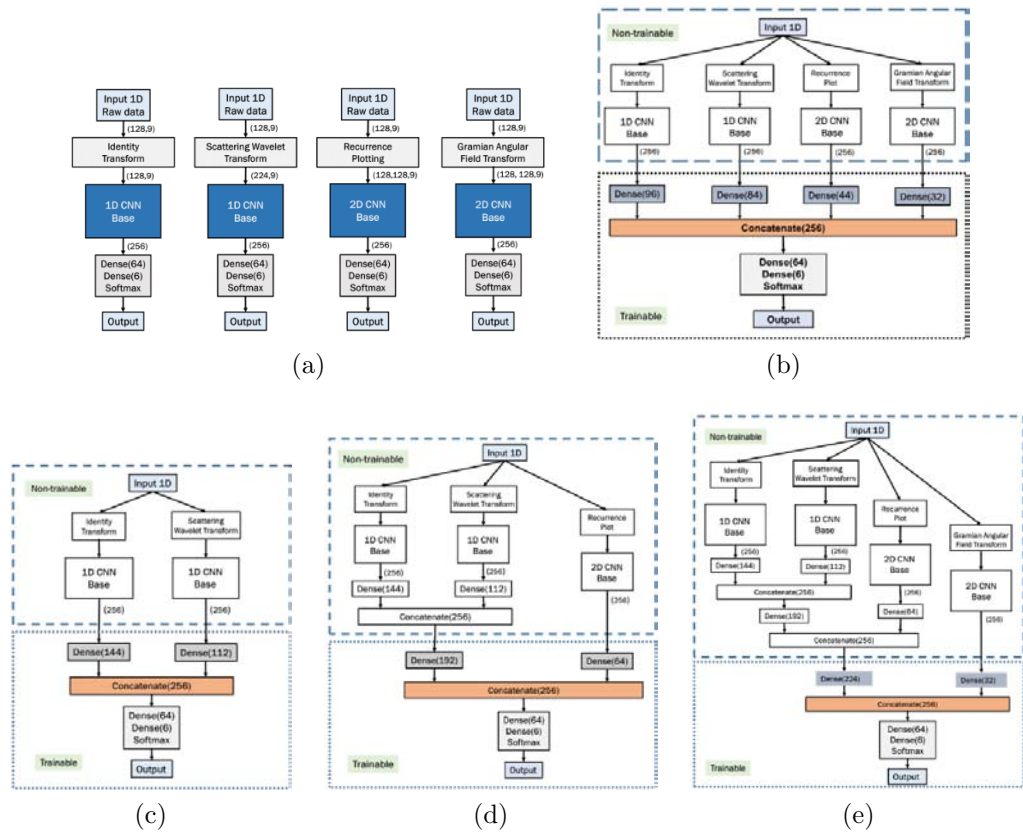


Figure 4.4: Schematic representation of the proposed multi-stage sequential training scheme. Here, (a) represents Individual training stage, (b) represents combined training stage, and (c), (d), (e) represent the sequential training stages.

lection from different transformed domains as introduced in Fig. 4.1. Here, the number of nodes in these densely connected layers are varied for incorporating more features from the feature extractors that contain more information for final classification.

- iii. **Sequential training stages:** In the proposed multi-stage sequential training scheme, individually trained feature extractors are optimized and converged in a unified architecture through series of sequential training stages, as shown in Fig. 4.4c, 4.4d and 4.4e. In this approach, two of the CNN base units operating on different transformed spaces are optimized together at a time by training an individual densely connected layer for each of the base units followed by feature concatenation and combined dense classifier unit, as shown in Fig. 4.4c. Later, these combined two feature extractors are considered as an individual unit and further merged with the next CNN base part. Similarly, in the next stage, another separate densely connected layers with a combined

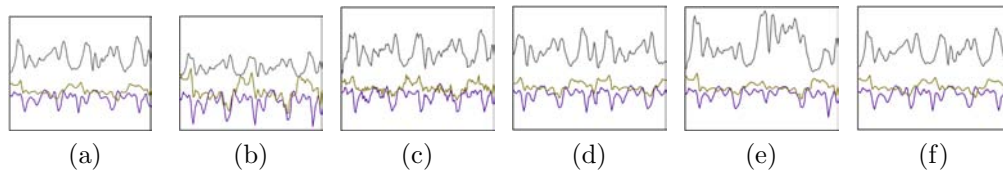


Figure 4.5: **Effect of various types of augmentation of the sample data.** (a) **Raw sample data collected from 3 axis accelerometer, with (b) scaling, (c) jittering, (d) permutation, (e) magnitude warping, and (f) time warping applied on raw data.**

dense classifier unit are trained, as shown in Fig. 4.4d and 4.4e. Therefore, through each training stage, a new CNN base part corresponding to another transformation is combined with the merged feature extractor. Moreover, each such stage merges these base feature extractor units by introducing a newly trained densely connected layers for providing the most optimized features at a whole utilizing all the existing features.

4.1.4 Data Augmentation

As imbalance in the dataset makes the training process complicated for learning the distribution of minority class, data augmentation is a viable approach to mitigate such problems. We have utilized the combination of five techniques that incorporate realistic variations in the data and make the process more robust [189]. However, all such augmentations are applied to the training data leaving the testing data unaltered for proper evaluation of the proposed methods. In Fig. 4.5, the individual effect of these augmentations are shown on raw sample data.

4.2 Results and Discussions

Three publicly available datasets used for this study are described below. Detailed comparative analysis of the results obtained is discussed later.

4.2.1 Dataset Description

UCI HAR database [190] contains 6 activities collected from 30 subjects with 50 Hz sampling rate using 3 axis accelerometer, gyroscope, and magnetometer embedded on a smartphone placed on the waist. USC HAR database [191] contains 12 activities collected from 14 subjects with 100 Hz sampling rate using 3 axis accelerometer and gyroscope. SKODA database [64] contains 11 activities collected from a single

subject in a car maintenance scenario using only a 3 axis accelerometer sampled at 98 Hz.

4.2.2 Experimental Setup

A five-fold cross-validation scheme is carried out for evaluation of the proposed scheme on each database separately. The performances of the evaluation metrics obtained from each test fold are averaged to get the final values. The Wilcoxon rank-sum test is used for statistical analysis of the average accuracy improvement obtained from the proposed scheme. The accuracies of the proposed schemes are statistically analyzed and the statistical significance level is set to $\alpha = 0.01$. The null hypothesis is that no significant improvement of average accuracy is achieved using the proposed scheme over the other existing best performing approaches.

4.2.3 Performance Evaluation

The performance of the optimized networks is evaluated using the test data of various datasets. Traditional evaluation metrics for the multi-class classification task, i.e accuracy, precision, recall, and intersection-over-union (IoU) score, are employed for analyzing the performance. In Tab. 4.1, the score of averaged cross-validation evaluation metrics are provided for both these training approaches. It is clear that both these approaches provide a considerable performance of over 98% in most of these classes that are separated almost perfectly. However, the two-stage method slightly struggles to separate features between walking and ascending upstairs activities as these activities contain close inter-relation in the feature space. But, in the case of multi stage-training, this problem is reduced which signifies the robust optimization capability of this method as it can separate features with proximity.

In Fig. 4.6, the average cross-validation IoU score of the optimized networks on different transformed spaces along with the final converged networks using both two-stage and multi-stage training are compared for all the activities. It is visible that identity transform representing the unaltered raw data provides better performance with more than 2% improvement in most classes compared to other transformed spaces in case of individual training. However, irrespective of the performance, all the networks operating on separate transformed spaces extract features that are significantly different as they work with diversified representations of the data. Through optimization of these features, as visible in Fig. 4.6, the proposed two-stage, and multi-stage training approach provide a sharp increase in IoU scores in all the activity classes compared to the individual training stage. However,

Table 4.1: Average Cross-Validation Performance Analysis on Various Activities of UCI HAR Dataset [190] for Proposed Two-Stage and Multi-Stage Training

Met- rics	Prop. Meth.	Class					
		Walk	Up Stairs	Down Stairs	Sit	Stand	Lay
Prec. (%)	2-Stg.	98.53	96.34	99.14	98.75	99.58	100
	M-Stg.	99.27	98.36	99.32	99.46	99.81	100
Rec. (%)	2-Stg.	94.93	98.72	100	99.61	98.64	100
	M-Stg.	97.44	99.26	100	99.83	99.35	100
IoU Sc. (%)	2-Stg.	96.31	97.41	99.49	99.07	99.02	100
	M-Stg.	98.24	98.68	99.52	99.59	99.47	100

Table 4.2: Average Cross-Validation Performance Analysis on Various Activities of USC HAR Dataset [191] for Two-Stage and Multi-Stage Training

Class	Two stage Training			Multi Stage Training		
	Prec. (%)	Rec. (%)	IoU (%)	Prec. (%)	Rec. (%)	IoU (%)
Walking Forward	99.2	98.3	98.5	99.7	99.4	99.4
Walking Left	99.1	98.5	98.7	99.6	99.3	99.3
Walking Right	99.2	99.4	99.1	99.5	99.6	99.5
Walking Upstairs	99.3	98.6	98.8	99.5	99.1	99.2
Walking Down	98.2	98.4	98.1	99.1	98.8	98.7
Running	99.0	98.2	98.4	99.3	98.7	98.9
Jumping	97.2	97.4	97.1	97.9	98.6	98.1
Sitting	99.1	99.2	99.0	99.4	99.5	99.2
Standing	97.5	98.1	97.8	98.5	98.8	98.4
Sleeping	100	99.5	99.6	100	99.7	99.7
In Elevator	98.1	98.3	98.1	98.4	98.6	98.4

lower performing transformed spaces are de-emphasized through a smaller number of densely connected nodes and with smaller weights generated in the later training stages while merging, as shown in Fig. 4.4. All of the transformed spaces contribute some new and valuable information that may be indistinguishable even on other space that provides significantly better performance. Moreover, in multi-stage sequential training, two of the feature spaces are optimized at a time by integrating an additional feature space to the resultant feature space (Fig. 4(c)-4(e)). It should be noticed that more number of nodes are provided in the densely connected layer following the features space of respective transformation to emphasize the features from those space that provided higher performance during the individual training stage.

In Tab. 4.2, the average cross-validation performance of the proposed schemes on different activity classes of this dataset is provided. It is clear that both these approaches provide consistent performance over 99% for most of the classes. However, multi-stage training provides a slight increase in incorrect predictions for some

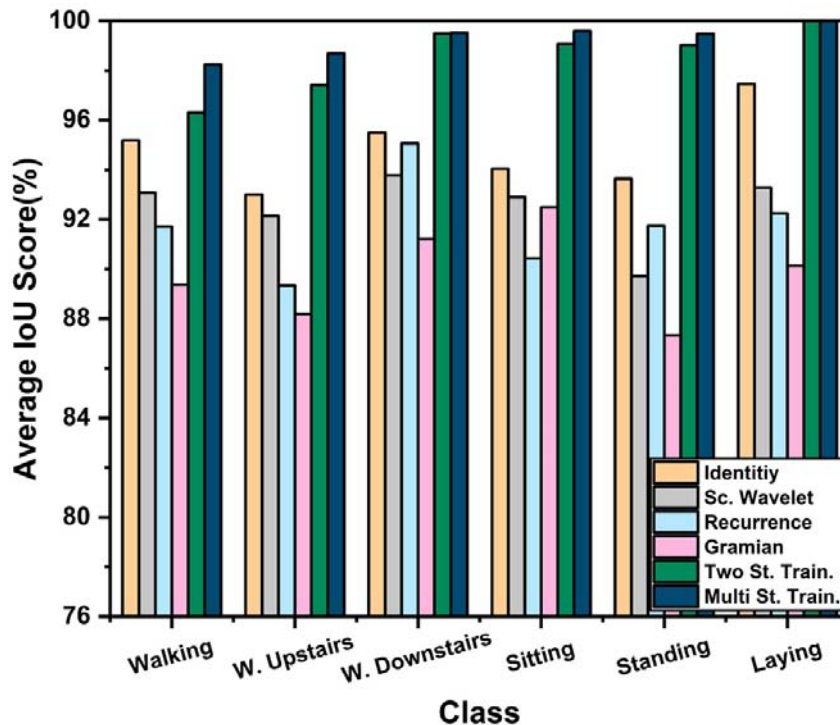


Figure 4.6: Comparison of Average Cross-Validation IoU scores on various activities of UCI HAR Database [190].

Table 4.3: Average Cross-Validation Performance Analysis on Various Activities of SKODA Dataset [64] for Two-Stage and Multi-Stage Training

Class	Two stage Training			Multi Stage Training		
	Prec. (%)	Rec. (%)	IoU (%)	Prec. (%)	Rec. (%)	IoU (%)
Null	95.3	96.4	95.8	96.2	96.5	96.1
Write on notepad	98.5	97.1	97.6	99.1	98.6	98.6
Open hood	95.2	94.5	94.7	97.3	95.4	96.3
Close hood	95.7	96.1	95.7	96.5	96.9	96.5
Check gaps on front door	96.3	97.5	96.5	97.8	99.1	98.3
Open left front door	96.6	95.8	96.1	97.5	95.7	96.4
Close left front door	96.7	95.6	95.9	97.2	95.9	96.3
Check trunk gaps	98.1	98.4	98.1	99.2	98.7	98.8
Open and close trunks	97.2	98.1	97.4	97.7	99.2	98.3
Check steering wheel	97.4	98.5	97.8	97.9	99.4	98.4

closely related activities like among various walking actions, between standing and sitting actions. In Tab. 4.3, the average cross-validation performance of both the training approaches is presented on the SKODA dataset. Though most of the activities contain close inter-relation in this dataset, our proposed training methods provide consistent performance over 95% for most of the classes. However, some activities like opening and closing hood, opening, and closing doors, are difficult to separate as expected. Despite that, comparable performances have been achieved

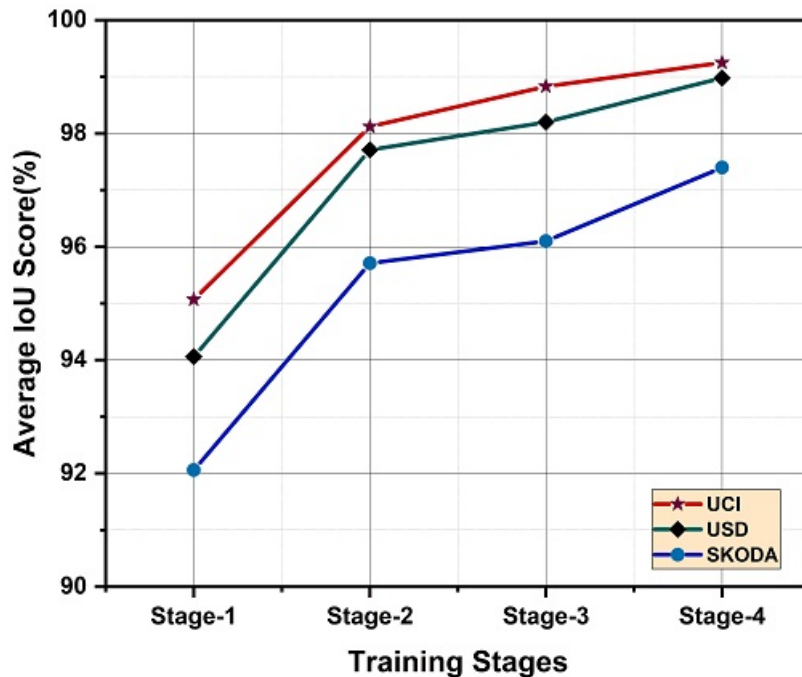


Figure 4.7: Average Cross-Validation IoU score in various training stages of multi-stage sequential training on different databases.

in these classes utilizing the proposed scheme.

In Fig. 4.7, the average IoU score in different stages are shown for multi-stage training. It is clear that each stage provides some improvement in performance by incorporating new features. However, in the first two stages, the trained network has achieved significant performance improvement with more than 3% improvement in the average IoU score mostly achieved utilizing the features from identity transformation and scattering wavelet transformation with the 1D deep CNN feature extractor. Nevertheless, features from other transformations exploited at the later stages still provide a considerable contribution with around 1 ~ 2% improvement in total to make the final network more optimized to separate challenging classes and thus to attain a higher average IoU score. Hence, integration of features from four transformed spaces in the proposed sequential training approach, 4 ~ 6% improvement of average IoU score is achieved in total compared to operating with raw sensor data alone.

Various existing approaches are compared with the proposed ones in Tab. 4.4 on different datasets. Average accuracies obtained from the proposed two-stage and multi-stage training methods are compared with the reported accuracy of varieties of state-of-the-art approaches. It can be noted that the proposed multi-stage scheme has improved average accuracy from 86.1% to 99.29% (13.19% improvement) in UCI

Table 4.4: Comparison of the Proposed Schemes with Other Existing Approaches on Different Datasets

UCI HAR Database [190]				USC HAR Database [191]				SKODA Database [64]			
Work	Method	Acc.(%)	P-value	Work	Method	Acc.(%)	P-value	Work	Method	Acc.(%)	P-value
[61]	MLP	86.1	NA	[192]	MLP, J48	89.2	NA	[64]	HMM	86	NA
[70]	CNN	94.2	NA	[66]	Random Forest	90.7	NA	[67]	DBN	89.4	NA
[65]	DTW	95.3	NA	[74]	CNN	93.2	NA	[73]	Deep Conv LSTM	91.2	NA
[190]	SVM	96	NA	[68]	LS-SVM	95.6	NA	[72]	CNN	91.7	NA
[75]	Deep RNN	96.7	NA	[69]	CNN	97	NA	[193]	Ensemble LSTM	92.4	NA
[62]	SVM	97.1	NA	[75]	Deep RNN	97.8	NA	[75]	DeepRNN	92.6	NA
Prop. 2-Stage	CNN	98.63	3.4e-5			98.57	2.5e-6			96.51	4.2e-4
Prop. M-Stage	CNN	99.29	5.1e-5			99.02	1.3e-6			97.21	2.8e-4

database, from 89.2% to 99.02% (9.82% improvement) in USC database, and from 86% to 97.21% (11.21% improvement) in SKODA database. The improvement in the multi-stage approach is around 1% higher over the two-stage training approach for its increased opportunity of optimization through multiple stages. However, the training complexity also increases as more number of training stages need to be adjusted. As the p -values obtained from the statistical significance test on different databases are considerably smaller from the predefined threshold of 0.01, we have to reject the null hypothesis and it suggests that considerable improvement of average accuracy is achieved using the proposed schemes over other existing approaches.

4.3 Conclusion

In this chapter, it is shown that instead of utilizing trained CNN as a feature extractor from a single space if multiple trained CNNs dealing with numerous transformed spaces can be utilized together, much better representation of features can be obtained. Such an idea of multiple training stages utilizing the initially trained CNN models from the preceding stages operating on different transformed spaces can offer a significant increase in performance with 4 ~ 6% improvement in average IoU scores. This method outperforms other state-of-the-art approaches in different datasets by a considerable margin with an average accuracy of 98.51% (11.49% average improvement) over three databases.

Chapter 5

PolypSegNet: A Modified Encoder Decoder Architecture for Polyp Segmentation from Endoscopy

Colorectal cancer (CRC) is the second most prevalent cause of cancer-related death in the United States with a death toll of around 53,200 in 2020 [94]. Most colorectal cancers start as adenomatous polyps (adenoma), initially benign growth on the inner lining of the colon and rectum, which can become malignant over time and spread to nearby organs. Early-stage diagnosis of polyps can increase the survival rate of CRC to 90%, whereas the 5-year relative survival rate of distant-stage patients can be as low as 14% [95] that makes early detection and removal of polyps vital for survival. Though colonoscopy is the gold standard tool for polyp detection, accurate detection is still a major challenge due to the varying size, position, and textures of polyps (shown in Fig. 5.1) along with differing colonoscopic withdrawal techniques, bowel preparation quality, and skills of the colonoscopist [96], [97], [194].

Numerous hand-crafted feature-based approaches have been explored for automatic polyp segmentation in the last two decades [99]–[102]. With the advent of deep learning, Unet architecture has become widely popular for image segmentation applications. However, there exist some architectural limitations in the traditional Unet architecture that opens the opportunity to improve the performance further, such as semantic gap between corresponding encoder-decoder level, simpler building blocks at each level, and sequential reconstruction of the output feature map. In this chapter, an improved encoder-decoder architecture, named as PolypSegNet, is proposed for efficient polyp segmentation by resolving all these architectural limitations in traditional segmentation networks. The major contributions introduced in this chapter can be summarized as follows:



Figure 5.1: Some of the challenges presented by colonoscopy images are (a) blurred or low quality images, (b) varying shapes and textures of polyps, (c) small visible differences among polyps, and d) background presence of extraneous matters.

- i. For encompassing diverse receptive areas in the feature extraction process, a depth dilated inception (DDI) module is introduced.
- ii. An efficient and generalized D-Unit layer structure of the encoder/decoder module is proposed that incorporates several sequential DDI modules for deep feature extraction.
- iii. To reduce the semantic gap in traditional skip connections of Unet, a deep fusion skip module (DFSM) is proposed that aggregates various scales of feature representations from different levels of the encoder.
- iv. For introducing more efficient reconstruction, a deep reconstruction module (DRM) is proposed for aggregation and joint optimization of multi-scale decoded feature maps generated at various levels of the decoder module.
- v. Extensive experimentations on four publicly available datasets provide significant improvement of performance in all evaluation metrics compared to other existing state-of-the-art approaches. The primary results of these experimentations are published in [21].

5.1 Methodology

Colonoscopic images collected from patients undergo through minimal pre-processing before extracting the segmented polyp regions using the proposed PolypSegNet. All the images are reshaped to uniform sizes followed by the amplitude normalization operation before feeding into the PolypSegNet architecture. The workflow of the proposed network is shown schematically in Fig. 5.2. Operations in this network can be divided into separate encoder and decoder modules in general. The encoder module generates different scales of feature maps in subsequent layers. In each of the proposed D-Unit layer, a feature map is processed through deep convolutional filtering utilizing several Depth Dilation Inception (DDI) units for generating a particular scale of feature representation that is down-sampled through strided convolution to extract more general feature representation in the following unit layer. Afterwards, different scales of feature maps generated from each D-Unit layer of the encoder are processed together in the proposed deep fusion skip module (DFSM). In stead of separately connecting different encoder and decoder layers, this module generates different scales of representations through a deep fusion of multi-scale encoded feature maps and these representations are passed to the decoder module to be employed in the reconstruction process. Similar to the encoder module, each D-Unit layer in the decoder module operates with a respective scale of the feature map. The output of each D-Unit layer is upsampled through a deconvolution operation. Hence, each D-Unit layer in the decoder module takes two input feature maps: one from the DFSM module and other from the deconvolution operation, which are concatenated and processed together in the respective unit layer. Hence, the decoded feature maps gradually gather finer and finer details of the segmentation mask in the subsequent layers. Similar to the DFSM module, different scales of decoded feature maps are aggregated and processed together for joint optimization in the proposed deep reconstruction module (DRM). Instead of only considering the decoded map from final layers, this DRM module generates the final segmentation mask through a deep fusion of different scales of decoded feature maps in the reconstruction process. The operations of different modules of the proposed PolypSegNet architecture are discussed in detail in the following subsections.

5.1.1 Proposed Depth Dilated Inception (DDI) Module

One of the main building blocks of the proposed PolypSegNet architecture is the depth dilated Inception (DDI) module, as shown in Fig. 5.3. In this block, advantages of the inception module, dilated convolution, depthwise separable convolution,

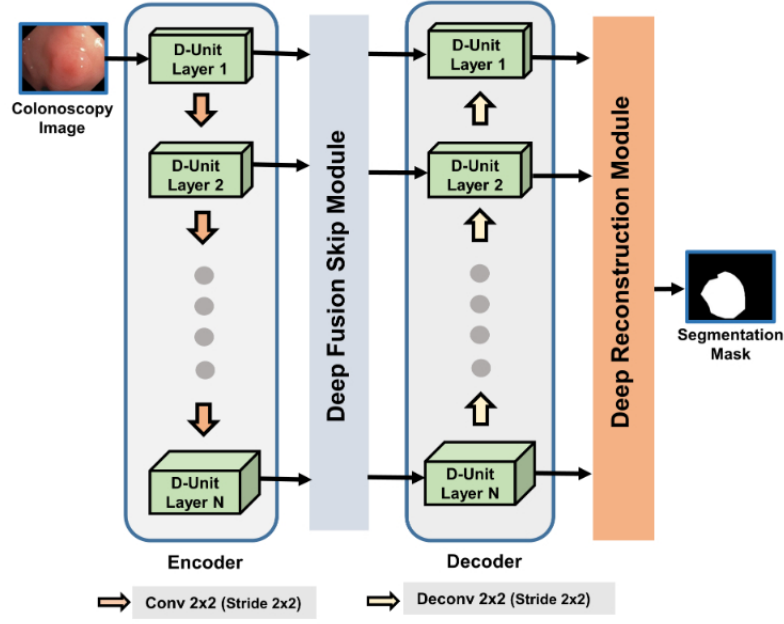


Figure 5.2: Schematic diagram of the proposed PolypSegNet Architecture

and asymmetric filtering are exploited for efficient feature extraction. The inception block was introduced in [195] that utilized convolutions with different sized kernels to extract features from diverse receptive areas. However, dilated convolutions [196] offers the advantages of covering diverse receptive area without increasing computational complexity by utilizing varying dilation rates, which can be represented by

$$y(i, j) = \sum_p \sum_q x(i + rp, j + rq) w(p, q) \quad (5.1)$$

where y is the output of convolution with dilation rate of r , (i, j) represents the center of the convolution, x represents the input feature map and w is the convolution filter/kernel.

Moreover, a traditional convolution operation operates both the spatial and inter-channel filtering simultaneously. For more efficient operation, it can be divided into separate depthwise spatial convolution followed by inter-channel point-wise convolution (kernel 1×1) [197], that can be represented as

$$\text{DWConv}(W, y)_{(i,j)} = \sum_p \sum_q W_{(p,q)} \odot y_{(i+p,j+q)} \quad (5.2)$$

$$\text{PWConv}(W, y)_{(i,j)} = \sum_m W_m \odot y_{(i,j,m)} \quad (5.3)$$

where $W \in \mathbb{R}^{P \times Q}$ is convolutional kernel, y is the feature map with M number of

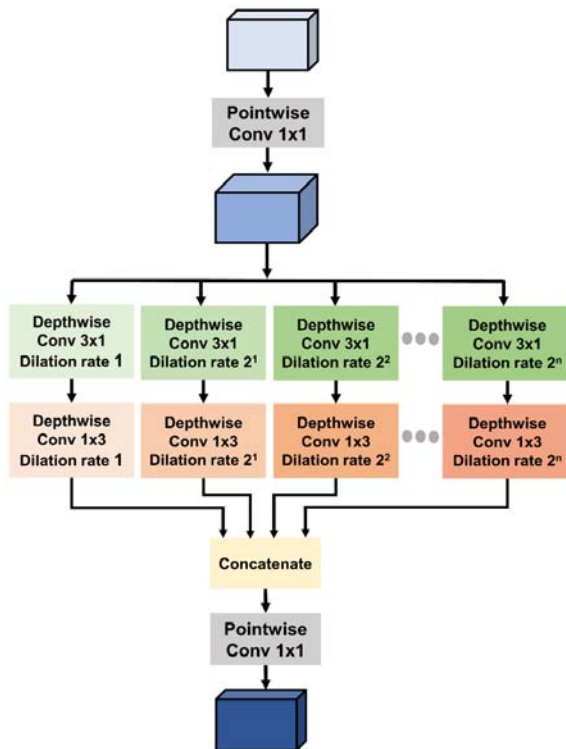


Figure 5.3: Schematic diagram of the proposed Depth Dilation Inception (DDI) module.

channels, (i, j) is the centre of the convolution operation, and \odot denotes element-wise multiplication.

Furthermore, asymmetric spatial convolutional filtering is found to be more computationally efficient compared to traditional spatial filtering [198]. Therefore, a $(n \times n)$ dimensional kernel can be efficiently divided into two sequential convolutional filtering operations with $(n \times 1)$ and $(1 \times n)$ kernels.

In the proposed DDI block, a pointwise-depthwise-pointwise operation is effectively incorporated for feature aggregation from diverse receptive areas. The initial pointwise convolution deepens the input feature map for introducing more convolutional filtering in the subsequent stage. Following that, depthwise convolutions are carried out on the deepened feature map with increasing dilation rates in parallel which accumulate features from diverse receptive areas. In each parallel path, two sequential depthwise convolutions are carried out with asymmetric kernels of (3×1) and (1×3) while maintaining the similar dilation rates. These two asymmetric convolutions provide an efficient alternative of symmetric (3×3) kernel for depthwise convolution by covering the similar spatial receptive area while incorporating less number of parameters. The depth growth ratio (d) of the initial pointwise convolution along with the total number (n) of parallel depthwise convolution with

increasing dilation rates are adjusted considering the dimension of the input feature map. For a deeper input feature map, the depth growth ratio (d) of initial convolution is reduced accordingly to limit the computational complexity. Moreover, when the spatial resolution of the input feature map gets smaller, the maximum dilation factor, n , is also reduced to limit the observation window accordingly. Afterwards, parallel outputs generated from multiple depth dilated convolutions are aggregated through concatenation and final pointwise convolution is carried out to reduce the depth through inter-channel filtering. Hence, the proposed transformation in the DDI module provides effective feature extraction with considerable diversity utilizing depthwise dilated convolutions with varying dilation rates.

5.1.2 Proposed D-Unit Layer Structure

In the proposed D-Unit layer structure (shown in Fig. 5.4), several depth dilated Inception (DDI) blocks are placed in series to extract deep features from the input feature map. After that, different levels of feature maps with similar dimensions generated by each DDI module are accumulated to gather the information of each transformation. Generally, deeper feature maps are generated at the later unit layers of the encoder module which contain more generalized representation with reduced spatial dimension and increased number of channels. As a result, the computational complexity also increases accordingly in the DDI unit for filtering more number of channels at a time. Hence, fewer DDI units are used to limit the computation for the D-Unit layers with deeper input feature maps. In this way, the total number of DDI units (m) incorporated in each layer is adjusted according to the depth of the input feature map. Finally, different stages of transformed feature maps are aggregated from individual DDI module and an inter-channel pointwise convolution is carried out to reduce the depth of the accumulated feature map by extracting the more-generalized feature representation utilizing contributions of all DDI modules. Therefore, the stack of DDI modules is effectively integrated into the proposed D-Unit layer structure to generate various scales of feature maps. Both in the encoder and decoder module of PolypSegNet architecture, different scales of feature maps are generated utilizing the D-Unit layer structure in subsequent layers. Structural details of all D-Unit layers operating with various scales of feature maps in the encoder/decoder module are provided in Table 5.1.

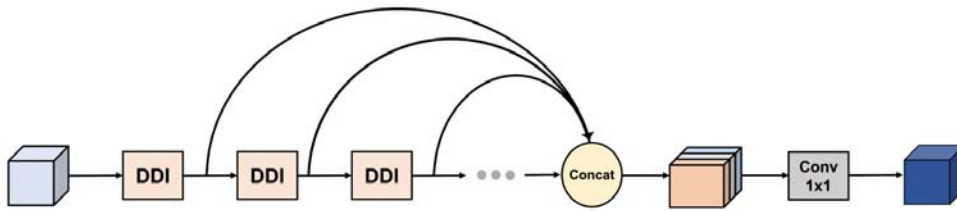


Figure 5.4: Schematic diagram of the proposed D-Unit layer structure. Different number of DDI units (m) have been integrated in each layer depending on the feature map dimension.

Table 5.1: Structural Details of Different D-Unit layers in Encoder and Decoder Block

Layer	Output Dimension	Depth Growth Ratio (d)	Max dilation factor (n)	No. of DDI Units (m)
D-Unit Layer 1	256x256x16	4	5	6
D-Unit Layer 2	128x128x32	3	4	5
D-Unit Layer 3	64x64x64	2	4	4
D-Unit Layer 4	32x32x128	2	3	4
D-Unit Layer 5	16x16x256	1	2	3

5.1.3 Proposed Deep Fusion Skip Module (DFSM)

In traditional Unet architecture, different scales of feature maps are generated from various layers of the encoder that are directly passed to the decoder to establish skip interconnection. In the decoder module, these different scales of feature maps are used in the subsequent layers for reconstruction. However, these types of connections only contain feature representation from a particular level of encoder that limits the information flow between encoder and decoder. In the proposed deep fusion skip module (DFSM), skip inter-connections are generated utilizing different scales of feature representations from all individual layers of the encoder (shown in Fig. 5.5). Operations in this DFSM module is designed to be divided into two general stages:

- i. A fusion feature vector creation for combining effects of different levels of the encoder.
- ii. Multi-scale feature maps creation for establishing skip interconnections between encoder and decoder.

Thus, a fusion feature vector is created utilizing the outputs from all encoder layers which is repeatedly used later to produce different scales of representation for skip interconnections. Firstly, different scales of encoded feature maps, generated from various D-Unit layers, are upsampled to make them uniform in spatial resolution. After that, these upsampled variants of feature maps are aggregated through concatenation to produce a fusion feature vector.

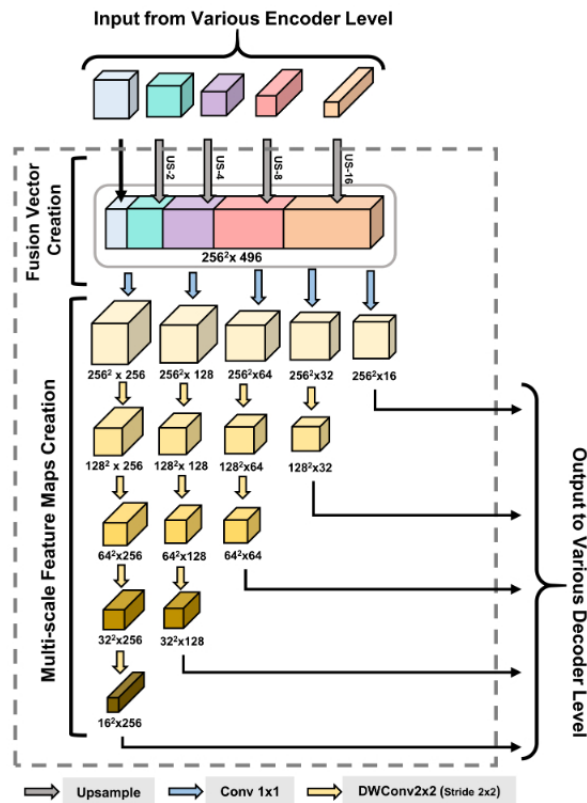


Figure 5.5: Schematic representation of the Deep Fusion Skip Module (DFSMS).

For generating multiple scales of feature representations, this fusion vector is passed through the separate depth and spatial scaling operations using pointwise convolutions followed by series of depthwise separable convolutions, respectively. Firstly, a pointwise convolution (kernel, 1×1) is used to scale the depth of the fusion vector according to the depth of the respective decoder layer without changing the spatial resolution (256×256). Following that, series of depthwise separable convolutions with kernel (2×2) is carried out with stride of (2×2) to reduce the spatial resolution with sequential spatial filtering to match the resolution with respective decoder layer.

Hence, the proposed DFSMS module generates feature maps for skip interconnection through the deep fusion of multi-scale features from all respective layers of the encoder.

5.1.4 Proposed Deep Reconstruction Module (DRM)

Different scales of encoded feature maps generated from the encoder module are passed through the decoder module that gradually decodes the segmentation mask.

The initial layers of the decoder module contain more generalized and global representations of the mask that gradually incorporate finer details in the subsequent layers before final reconstruction. In traditional Unet and other architectures, the final representation obtained from the top of the decoder block is used for the final reconstruction. However, considering the only single scale of decoded feature map for final reconstruction limits the gradient flow in the network as well as cannot fully utilize different scales of contextual information generated from various levels of the decoder module.

In the proposed deep reconstruction module (DRM), different scales of contextual information aggregated from various decoder layers are used for final reconstruction that not only increases the gradient flow through joint optimization but also incorporates more information for effective reconstruction (shown in Fig. 5.6). Similar to the DFSM block, operations in the DRM block can be divided into two stages:

- i. A fusion feature vector generation for accumulating effect of various decoder levels.
- ii. A segmentation mask creation utilizing the fusion feature vector.

Hence, all the decoded feature maps generated from various levels of the decoder are made spatially uniform through upsampling for feature aggregation. Afterwards, parallel convolution operations with different kernels are carried out to generate diversified representations through diverse convolutional filtering operations for fusing multi-scale contextual decoded features. Next, all these transformed representations are added together to generate the equivalent effect of fusion through various kernels. Finally, a pointwise convolution is carried out with sigmoid activation to produce the final binary segmentation map. Therefore, the proposed deep reconstruction module (DRM) generates the final segmentation map by integrating a deep fusion of different scales of decoded feature maps.

5.2 Results and Discussions

Experimentations are carried out on several colonoscopy image datasets to validate the robustness and applicability of the proposed scheme for segmenting regions of polyps. Finally, performances achieved from extensive experimentations are discussed and analyzed from different perspectives.

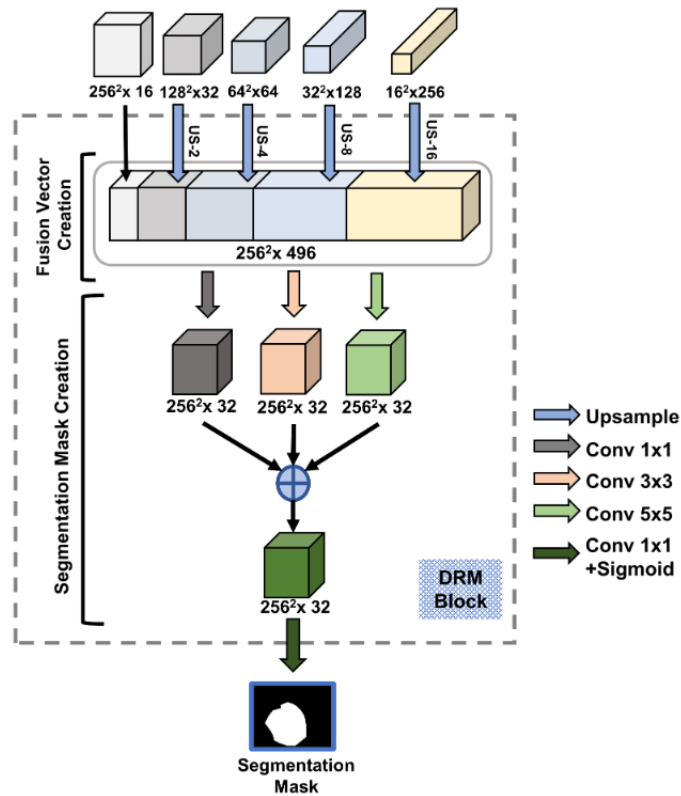


Figure 5.6: Schematic representation of the Deep reconstruction module (DRM).

5.2.1 Database Description

Four publicly available databases are used for training and evaluation of the proposed PolypSegNet. Details of these databases are summarized as below:

- CVC-ClinicDB [199] dataset consists of 612 images from 31 different types of polyps along with the corresponding ground truth of defined polyp regions which are manually annotated by experts. All the images originally have a resolution of 384×288 .
- Kvasir-SEG [200] dataset contains 1000 polyp images and their corresponding ground truth masks manually annotated by expert endoscopists from Oslo University Hospital (Norway). The resolutions of the images varied from 332×482 to 1920×1072 .
- ETIS-Larib [201] dataset contains 36 different types of polyps in 196 images with resolution of 1225×966 . These images were extracted from colonoscopy videos and the ground truths for the mask were annotated.

- CVC-ColonDB [202] dataset contains 300 polyp images and their corresponding pixel level annotated polyp masks which are extracted from 15 video sequences. The images had a resolution of 574×500 .

All the images are resized to uniform resolutions of (256×256) for operating with the proposed PolypSegNet.

5.2.2 Experimental Setup

Different hyper-parameters of the network are chosen through experimentation for better performance. A number of traditional evaluation metrics are used for evaluation of performance. Five-fold cross-validation scheme is carried out separately on these databases for evaluation of the proposed scheme. The Wilcoxon rank-sum test is used for statistical analysis of the performance improvement obtained from the proposed scheme. The performances of the proposed schemes are statistically analyzed and the statistical significance level is set to $\alpha = 0.01$. The null hypothesis is that no significant improvement of performance is achieved using the proposed scheme over the other best performing approaches.

5.2.3 Performance Evaluation

In the proposed PolypSegNet, depth dilation block (DDB) based D-Unit layer, deep fusion skip module (DFSM), and deep reconstruction module (DRM) are introduced. The effects of all these blocks are separately studied to validate the effectiveness of these blocks. For the baseline model, traditional Unet architecture is considered. And these modules are gradually integrated into the Unet architecture and the performance improvement is compared in different combinations of the proposed modifications. In Table 5.2, the performances of the network in different evaluation metrics obtained from these modifications are summarized. It should be noticed that the integration of the proposed modules provides consistent improvement of performance in all evaluation metrics compared to the baseline Unet architecture.

To evaluate the improvement in more detail, let's consider the ETIS-Larib database as it is more challenging due to the smaller amount of available data. Integration of the DDI module into the baseline Unet provides 3.34% improvement in dice coefficient and 1.75% improvement in the mean IoU. Whereas, integration of the DFSM module into the Unet provides 5.53% improvement in the dice score and 3.34% improvement in IoU. And, incorporation of the DRM module provides a 4.42% improvement in dice coefficient and 2.92% improvement in IoU. Hence, it can be observed that the DFSM module and the DRM module have higher effects on

Table 5.2: Effect of Different Network Configuration in the Performance on CVC-ClinicDB, Kvasir-SEG, and ETIS-Larib Databases

Model	ETIS-Larib				CVC-ClinicDB				Kvasir-SEG			
	Prec.(%)	Rec.(%)	Dice(%)	IoU(%)	Prec.(%)	Rec.(%)	Dice(%)	IoU(%)	Prec.(%)	Rec.(%)	Dice(%)	IoU(%)
Unet	82.12	71.26	73.31	66.92	95.21	89.35	89.13	83.52	75.21	89.33	82.14	74.52
Unet + DDI	85.71	73.82	76.65	68.67	95.64	89.96	90.38	84.61	79.61	90.66	84.43	76.46
Unet + DFSM	87.52	75.56	78.84	70.26	95.79	90.34	90.61	84.94	81.36	90.84	85.65	77.17
Unet + DRM	86.83	74.52	77.73	69.84	95.72	90.13	90.47	84.65	80.85	90.77	84.26	76.53
Unet + DFSM+DRM	92.62	81.58	81.99	75.33	96.08	90.97	91.12	85.76	87.51	92.23	86.82	80.42
Unet + DDI+DFSM	90.58	79.87	80.22	72.92	95.87	90.65	91.01	85.58	85.67	91.96	86.17	79.54
Unet + DDI+DRM	89.95	79.41	81.56	73.78	95.93	90.54	91.09	85.47	84.58	91.34	85.39	78.73
PolypSegNet	95.71	84.34	84.79	78.32	96.21	91.13	91.52	86.22	91.68	92.54	88.72	82.56

performance improvement. This occurs due to their increased opportunity in the information flow and gradient propagation throughout the network as both these modules integrate multi-scale feature maps that are generated at different levels of encoder and decoder. After observing the individual effect of these three modules, different combinations of two of these modules are further studied. It can be noticed that highest improvement is achieved when DFSM and DRM modules are integrated (8.68% improvement in dice score and 8.61% improvement in IoU) while DDI module combining with DFSM module provides 6.91% improvement in dice score, and 8.25% improvement in dice score is achieved when combined with DRM module. Finally, all three modules are integrated into the baseline Unet that results in the proposed PolypSegNet architecture and it provides the highest performance in all the metrics compared to other combinations. It provides 13.59% improvement in precision, 13.08% improvement in recall, 11.48% improvement in dice score, and 11.4% improvement in mean IoU. Hence, it justifies that all three modules contribute to the performance improvement of the proposed PolypSegNet architecture compared to the baseline Unet.

The performance of the proposed PolypSegNet is compared with other state-of-the-art segmentation networks. To make a fair comparison, most of these networks are re-produced using their open-source implementations while maintaining a similar training condition. Moreover, similar optimizer and loss functions are used for the performance evaluation of all the networks that prioritize the architectural contributions in performance improvements.

In Table 5.3, the performances of these networks are summarized. It should be noted that the proposed network consistently outperforms other networks improving dice score from 80.92% to 91.52% (10.6% improvement) in CVC-ClinicDB database, from 71.1% to 92.8% (21.7% improvement) in CVC-ColonDB database, from 64.25% to 88.72% (24.47% improvement) in Kvasir-SEG database, and from 59.71% to 84.79% (25.08% improvement) in ETIS-Larib database. Moreover, considerable improvements are also achieved in mean IoU, precision, and recall metrics. All these improvements are found to be statistically significant ($p < 0.01$). As the pro-

Table 5.3: Comparison of Performance of the Proposed PolypSegNet with Other State-of-the-Art Approaches on CVC-ClinicDB, CVC-ColonDB, Kvasir-SEG, and ETIS-Larib Databases with Five-Fold Cross-validation Scheme

Model	CVC-ClinicDB [199]					CVC-ColonDB [202]					Kvasir-Seg [201]					ETIS-Larib [200]				
	Pre	Rec	Dice	IoU	p-Val	Pre	Rec	Dice	IoU	p-Val	Pre	Rec	Dice	IoU	p-Val	Pre	Rec	Dice	IoU	p-Val
FCN [203]	81.4	79.4	80.9	75.3	-	95.5	73.3	71.1	67.9	-	61.9	63.8	64.2	60.8	-	68.2	61.3	59.7	53.9	-
Unet [112]	95.2	89.3	88.1	83.5	-	99.0	79.7	79.5	74.2	-	75.2	89.3	82.1	74.5	-	82.1	71.2	73.3	66.9	-
Unet++ [118]	92.1	82.2	87.4	77.3	-	97.1	71.1	75.8	70.8	-	82.4	70.1	76.7	63.6	-	79.6	56.6	65.3	54.8	-
MultiResUnet [114]	96.3	85.1	88.5	82.7	-	98.8	77.2	77.8	72.8	-	89.7	71.7	80.1	73.1	-	95.7	62.5	69.1	63.6	-
ResUnet++ [204]	87.3	70.6	79.4	79.8	-	94.9	72.3	73.2	69.9	-	79.2	81.7	71.3	66.9	-	69.3	65.1	60.9	56.9	-
LinkNet [205]	92.2	85.8	86.2	79.5	-	91.8	64.1	66.9	63.4	-	79.2	68.6	72.2	63.4	-	77.8	61.4	58.2	48.6	-
Double-Unet [206]	95.4	82.9	87.7	84.0	-	98.4	85.5	85.8	81.1	-	87.8	76.5	82.6	78.9	-	83.9	73.3	76.2	72.1	-
BA-Net [207]	93.8	88.2	88.2	84.3	-	99.1	91.6	90.2	86.9	-	90.9	82.1	87.8	82.0	-	89.8	76.1	78.3	72.5	-
Dil. ResFCN [208]	94.5	86.9	87.1	83.8	-	97.6	83.8	84.7	79.4	-	86.6	79.6	82.3	77.3	-	87.7	71.9	75.5	71.1	-
PolypSegNet(Ours)	96.2	91.1	91.5	86.2	9e-4	99.2	93.1	92.8	88.2	2e-3	91.7	84.5	88.7	82.5	6e-3	95.7	84.3	84.8	78.3	4e-5

Table 5.4: Comparative Analysis of Cross-Dataset Performances of Different State-of-the-Art Networks with Different Combinations of Training and Testing Dataset

Training Dataset	Testing Dataset	Dice Score(%)						
		Unet [112]	Unet++ [118]	MultiResUnet [114]	Double-Unet [206]	BA-Net [207]	ResUnet++ [204]	PolypSegNet(Ours)
CVC-ClinicDB	CVC-ColonDB	65.5	62.9	64.9	71.1	72.3	60.4	74.7
CVC-ColonDB	CVC-ClinicDB	72.5	69.8	73.6	74.2	74.9	63.8	76.1
CVC-ColonDB+ETIS-Larib	CVC-ClinicDB	75.7	71.2	74.8	75.7	77.1	66.1	80.4
CVC-ClinicDB	ETIS-Larib	57.5	54.9	58.6	61.2	63.7	41.8	68.6
CVC-ColonDB	ETIS-Larib	52.7	50.1	51.3	56.8	60.1	36.9	63.7
Kvasir-Seg	ETIS-Larib	60.2	58.3	60.5	64.4	67.1	44.7	71.8
Kvasir-Seg	CVC-ClinicDB	75.0	71.1	73.8	75.3	76.6	65.4	78.1
CVC-ClinicDB	Kvasir-Seg	66.8	62.3	65.1	67.6	68.4	52.7	70.7
CVC-ClinicDB+CVC-ColonDB	Kvasir-Seg	70.1	67.2	68.9	71.4	72.5	56.2	75.3

posed PolypSegNet incorporates three major modifications in the traditional Unet architecture, these modifications improved the extracted feature quality by introducing more opportunities to optimize at a whole for better gradient propagation throughout the network. Moreover, the proposed PolypSegNet effectively utilizes multi-scale feature representations generated at different levels of the encoder to reduce the semantic gap with the decoder as well as incorporates better reconstruction strategy through combining multi-scale decoded representations from different levels of the decoder, which offer the considerable performance improvements over other segmentation networks.

In Table 5.4, cross-dataset evaluation has been carried out for measuring the generalization capability of different networks where trained models are tested on different datasets. As the trained networks are tested on data collected from other sources, it is expected to be very challenging for source variations. Different combinations of train-test datasets are formed by utilizing the four available datasets. It should be noticed that the proposed network consistently provides better performance compared to other existing methods, though the achievable performances are lower than those of the similar train-test data experimentation. For instance, while the networks are trained on CVC-ClinicDB and tested on ETIS-Larib, our proposed PolypSegNet provides 4.9% higher performance than next closer performing BA-

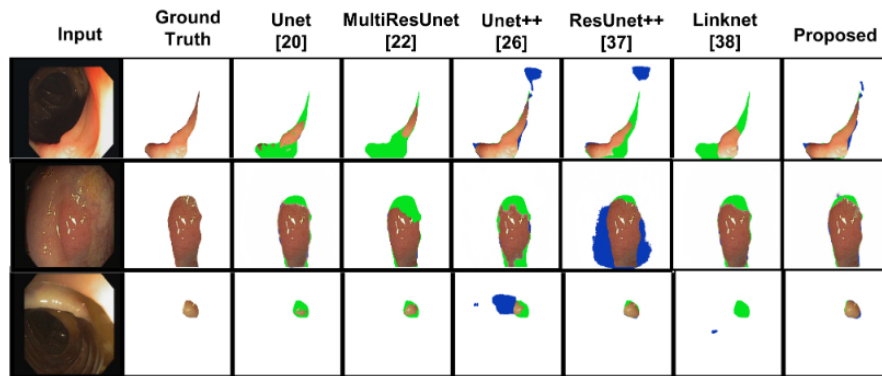
Table 5.5: **Computational Performance Analysis of Different State-of-the-Art Networks**

Networks	Total Parameters (M)	Inference Time(ms)	Speed (FPS)	Mean Dice Score (%)
Unet [112]	34.5	42	24	80.7
Unet++ [118]	8.8	35	28	76.3
MultiResUnet [114]	7.2	33	30	78.9
Double-Unet [206]	29.3	47	22	83.1
ResUnet++ [204]	16.2	48	21	71.2
LinkNet [205]	20.3	50	20	70.9
PolypSegNet(Ours)	5.5	39	25	89.5

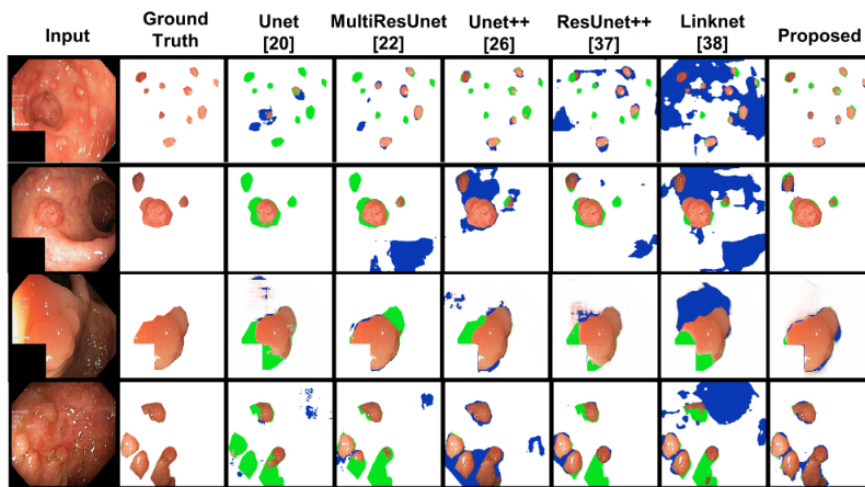
Net, and provides 11.1% higher dice score than Unet. These improvements indicate the better generalizability of the proposed PolypSegNet that can be an effective choice for practical applications with considerable data-variations.

In Table 5.5, computational performances of numerous networks are summarized including number of parameters used and operational speed. It is to be noted that the proposed PolypSegNet has lowest number of parameters requiring 5.5M parameters which is considerably smaller compared to other existing networks. For efficiently exploiting all of the network parameters while increasing feature diversity, the proposed network integrates some novel architectural building blocks that causes slight increase in processing time. However, the PolypSegNet operates with comparable speed of other state-of-the-art networks with inference time of 39ms which leads to 25 FPS processing speed. Hence, the proposed network is capable of performing at near real time speed with minimal memory requirement that can be vital for mobile devices. Though MultiResUnet and Unet++ provide slightly higher FPS compared to the proposed PolypSegNet, these come with comparatively lower mean dice score over four datasets. Therefore, this network provides an efficient means for precisely segmenting polyp regions with considerably lightweight architectures operating with near-real time speed that can be deployed to real time video colonoscopy data processing.

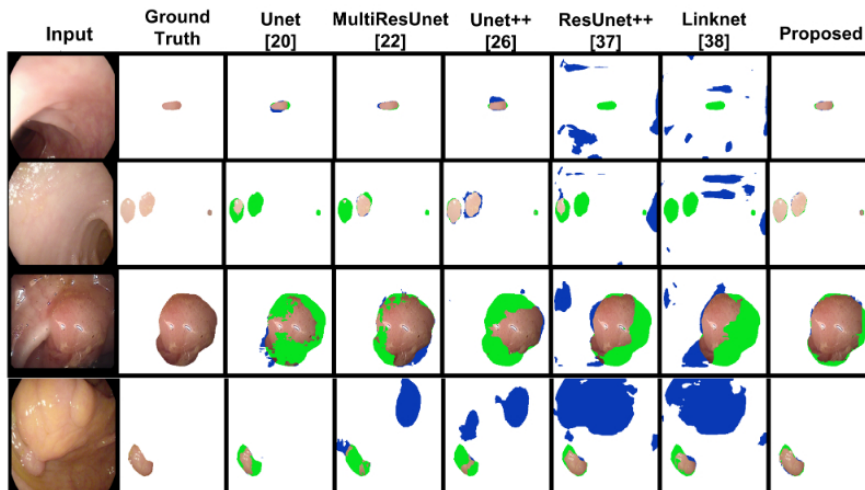
In Fig. 5.7a, 5.7b, and 5.7c some of the challenging cases collected from four databases are used for visual discrimination of the performance of different networks. The predicted segmentation masks generated by the networks are used for segmenting the region of polyps and corresponding false positives and false negative regions are identified comparing with the provided ground truth. It is noticeable that the proposed PolypSegNet extracts the polyp regions effectively that provides consistent performance in most challenging cases with minimum false positive and false negative regions compared to other state-of-the-art networks. Moreover, the performance improvement is prominently noticeable when the boundary of the polyps are



(a) Kvasir-SEG [200]



(b) CVC-ClinicDB [199]



(c) ETIS-Larib [201]

Figure 5.7: Visual representation of the input colonoscopic images and the segmented polyp regions obtained using various architectures on different databases. In segmented polyps, ‘blue’ denotes the false positive region and ‘green’ denotes the false negative region.

difficult to visually discriminate from the background, when the polyp regions are much smaller, and when there exist considerable differences in the contrast over the image. Hence, it can be said that the proposed PolypSegNet provides considerably better segmentation of the polyp regions in many challenging cases while most-other methods provide sub-optimal performance.

5.3 Conclusion

In this chapter, it is shown that combining all three modules (DDI, DRM, DFSM) in the proposed PolypSegNet, significant improvements in evaluation metrics are achieved consistently on all four databases used for extensive experimentation. Moreover, considerable qualitative visible improvements are obtained mostly in challenging conditions where traditional networks are supposed to under-perform. Furthermore, in the cross dataset performance analysis, it is found that the proposed network also provides higher generalizability over other state-of-the-art approaches that makes it more suitable for practical applications where wide-variations on data occur frequently.

Chapter 6

CovSegNet: A Multi Encoder Decoder Architecture for Improved Lesion Segmentation from COVID-19 Chest CT-scans

Chest radiography has already been proven to be an effective source for COVID diagnostics due to its major implications relating to various levels of lung infections [209]. Computer tomography (CT) scan and chest X-ray have been extensively explored in the literature to establish an automated AI-based COVID diagnostic scheme [210], [211]. Despite the easier access to chest X-ray, CT scans are more widely accepted due to its finer details leveraging the accurate diagnosis of COVID infections. Precise segmentation of lung lesions in chest CT scans is one of the most demanding and challenging aspects for faster diagnosis of COVID-19 due to the shortage of annotated data, diverse levels of infections, and novel types and characteristics of the infections [124].

A wide variety of approaches have been introduced in recent years for segmenting the region-of-interest in diverse applications. In [212], a fully connected network (FCN) is introduced that produces multiple scales of encoded feature maps and reconstructs the segmentation mask utilizing these encoded representations. In [213], Unet architecture is introduced by integrating an inverted decoder module following the encoder module to gradually reconstruct the mask that gains much popularity over the years. However, several architectural limitations of Unet are identified that provides suboptimal performance.

- The skip connection introduced in Unet generates semantic gap between corresponding feature scale of encoder-decoder modules, which mainly arises from

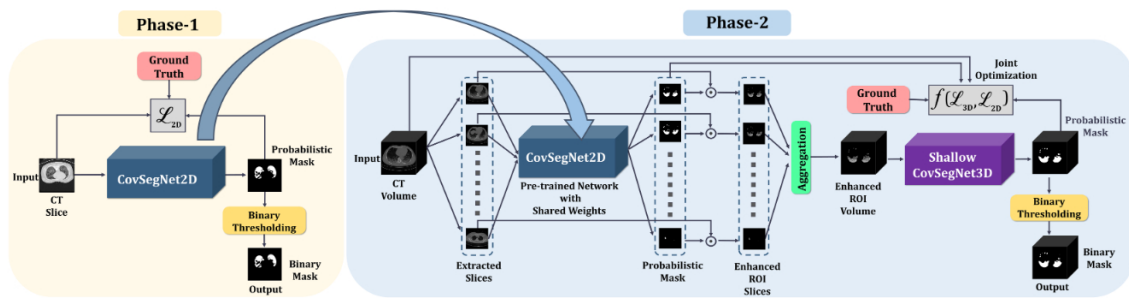


Figure 6.1: Workflow of the proposed scheme for segmenting lung lesions of COVID-19 in CT volume.

the direct concatenation of two semantically dissimilar feature maps.

- Contextual information loss occurs in traditional pooling/strided convolution-based downsampling operations that become more eminent with deeper architecture.
- The vanishing gradient problem rises in a deeper structure for sequential optimization of multi-scale features.
- Simplistic sequential convolutional layers are integrated into each level of encoder/decoder modules that lack enough architectural diversity to extract features from a broader spectrum.

In this chapter, an improved, automated scheme is proposed for precise lesion segmentation of COVID-19 chest CT volumes by overcoming the limitations of traditional approaches with a novel deep neural network architecture, named as CovSegNet. The major contributions of this work are summarized below:

- i. Along with the opportunity of vertical expansion, a horizontal expansion strategy is introduced in the CovSegNet architecture.
- ii. For further replenishing the loss of contextual information in traditional pooling/upsampling operations, a scale transition scheme is introduced in the encoder/decoder module by incorporating multi-scale feature maps from preceding levels.
- iii. For reducing semantic gaps among corresponding feature scales of the encoder-decoder modules, a multi-scale fusion module is introduced in between successive encoder-decoder modules.
- iv. A multi-phase training approach is introduced for integrating the advantages of both the 2D and 3D data processing scheme to reach the optimum performance.

- v. The proposed CovSegNet architecture is designed in a modular and structured way that can be adapted to its lightweight, shallow form to reduce complicity with considerable performance.
- vi. Extensive experimentations have been carried out to validate the effectiveness of the proposed scheme on two publicly available datasets containing chest CT scans from COVID-19 patients. The primary results of these experimentation are published in [30].

6.1 Methodology

The proposed scheme splits the segmentation of CT volumes into two subsequent phases to reduce the computational complexity of 3D convolution as well as to take the advantages of multi-scale 2D convolutions (Fig. 6.1). In the first phase of training, 2D slices are extracted from the 3D CT-volumes and these are used for the optimization of CovSegNet2D (i.e. 2D variant of the proposed CovSegNet architecture) from randomized initial state. After the optimization, the trained CovSegNet2D is capable of extracting lesions from 2D slices. However, slice-based processing of input CT volumes will lead to loss of inter-slice contextual information resulting in sub-optimal performance. Nevertheless, 2D-processing are computationally efficient and easy to optimize compared to the complete 3D processing. To introduce further optimization for integrating the inter-slice contextual information of particular CT volume, phase-2 of the training stage is incorporated. Here, a hybrid volumetric processing scheme is introduced where the CovSegNet2D is initialized with the pre-trained weights obtained from the phase-1 of the training. Thus, the complete 3D-CT volume is split into several 2D-slices that are processed through the CovSegNet2D to extract the region-of-interest in the 2D CT-slices. Afterwards, these enhanced 2D CT-slices are aggregated to generate the ROI-enhanced CT-volume where most of the redundant parts are suppressed. Nevertheless, to extract the inter-slice contextual information for further optimization, a lighter variant of CovSegNet3D is incorporated to operate on the ROI-enhanced CT-volume. In the second phase of the training, CovSegNet3D will be optimized from scratch to extract the inter-slice contextual information, while CovSegNet2D will be fine-tuned for better extraction of the intra-slice features. Hence, this joint optimization operation in phase-2 is supposed to optimize a very lighter variant of CovSegNet3D (as it operates on the ROI-enhanced Volume), which reduces the computational burden of complete 3D-processing with very deep network. Moreover, as the CovSegNet2D is initially pretrained in the phase-1 for efficient 2D-slicewise processing, it greatly

reduces the optimization complexity in phase-2 through generating ROI-enhanced CT-volume. Hence, this hybrid networking scheme is capable of utilizing both the inter-slice and intra-slice contextual information while greatly reducing the computational complexity of complete 3D-processing.

6.1.1 Problem Formulation

Let consider the set of CT volumes as \mathbf{X} , and their corresponding ground truths as \mathbf{Y} , such that $X_i \in \mathbb{R}^{h \times w \times s \times c}$, $Y_i \in \mathbb{R}^{h \times w \times s \times c}$, and $i = \{1, 2, 3, \dots, N\}$, where (h, w, s, c) denote height, width, number of slices, and channels per slice, respectively, of a particular CT volume from total N number of CT volumes. Moreover, let consider $\mathbf{x}_{i,j} \in \mathbb{R}^{h \times w \times c}$ as the i_{th} slice from total S slices of j_{th} CT volume and $\mathbf{y}_{i,j} \in \mathbb{R}^{h \times w \times c}$ as its corresponding mask, such that $i = \{1, 2, \dots, S\}$, and $j = \{1, 2, \dots, N\}$. In the first phase of training, the objective function for slice-based optimization of CovSegNet2D is

$$\text{Phase 1: } \operatorname{argmin}_{\theta} \mathcal{L}_{2D}(\theta, \mathbf{y}^P, \mathbf{y}) \quad (6.1)$$

where, θ denotes the network parameter of CovSegNet2D, $\mathbf{x}, \mathbf{y}^P, \mathbf{y}$ denote the input 2D-slice, predicted probability mask, and corresponding ground truth mask.

In the phase-2 of training, the pre-trained CovSegNet2D network obtained from phase-1 is employed to generate ROI enhanced CT volume \mathbf{X}' , and thus

$$\mathbf{x}' = \mathbf{x} \odot \mathbf{y}^P ; \quad \forall \mathbf{x}' \in \mathbf{X}', \mathbf{x} \in \mathbf{X}, \mathbf{y}^P \in \mathbf{Y}^P \quad (6.2)$$

where \odot denotes element-wise multiplication and \mathbf{x} denotes 2D-CT slice, \mathbf{x}' denotes ROI-enhanced CT-slice, and \mathbf{y}^P denotes the predicted probability mask.

Afterwards, optimization of the CovSegNet3D is carried out utilizing ROI-enhanced CT-volume, while CovSegNet2D is fine-tuned to generate more accurate probability masks from 2D-slices, and the joint optimization objective function \mathcal{F} can be formulated as

6.1.2 Proposed CovSegNet architecture

The proposed CovSegNet architecture is a generic representation of a network with a wide range of flexibility for increasing its applicability in different challenging conditions. This architecture can be designed for efficient operations in both 2D and 3D domains. Moreover, it can be made deeper/lighter according to the requirement of the applications.

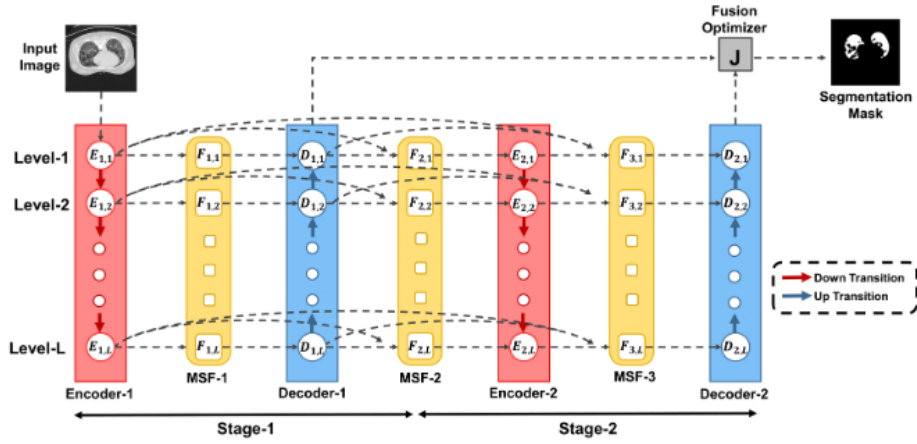


Figure 6.2: Schematic representation of the two-stage implementation of the proposed CovSegNet architecture where two sequential encoder-decoder operational stages are employed with L subsequent levels.

In CovSegNet architecture, multiple stages of sequential encoding and decoding operations are carried out along with a fusion scheme of multi-scale features in between subsequent encoder/decoder module. Each stage of the network consists of an encoder module and a corresponding decoder module. Hence, the network, \mathcal{N} , can be represented as

$$\mathcal{N} = \mathbf{D}_m(\mathbf{E}_m \dots (\mathbf{D}_1(\mathbf{E}_1(\theta_{\mathbf{E}_1}), \theta_{\mathbf{D}_1}), \dots, \theta_{\mathbf{E}_m}), \theta_{\mathbf{D}_m}) \quad (6.3)$$

where $\mathbf{E}_i, \mathbf{D}_i$ represents the encoder and decoder modules, respectively, of i_{th} stage from total m stages, and $\theta_{\mathbf{E}_i}, \theta_{\mathbf{D}_i}$ represents their respective parameters. Two-stage implementation of this architecture is schematically presented in Fig. 6.2.

This network can be extended from level-1 to level-L to produce a deeper variant. The encoder/decoder module constitutes of several unit cells operating at each level of the network. To generate a deeper network, additional unit cells are integrated in each of the encoder/decoder module to increase number of levels. Here, $E_{i,j}, D_{i,j}$ represent the i_{th} unit cell of j_{th} stage of encoder and decoder, respectively, where $i = \{1, 2, \dots, L\}$, and $j = \{1, 2, \dots, m\}$. Hence, L number of different scales of representative feature maps are obtained from each encoder/decoder module. Moreover, scale transition of feature maps is carried out in between succeeding encoder/decoder unit cells, and effective transformation on each scale of feature maps are integrated utilizing the generalized unit cell structure in encoder/decoder module.

In between successive encoder/decoder modules, a multi-scale fusion (MSF) module is introduced to reduce the semantic gap with preceding stages as well as to im-

prove the gradient propagation through parallel linkage of multi-scale features. Similar to encoder/decoder module, each MSF module consists of several operational unit cells operating at different levels. Let consider, \mathbf{F}_i represents the i_{th} MSF module, $F_{i,j}$ represents the i_{th} unit cell of j_{th} MSF module, such that $i = \{1, 2, \dots, L\}$, $j = \{1, 2, \dots, 2m - 1\}$, and $F_{i,j} \in \mathbf{F}_i$.

Each MSF module takes all scales of feature representations as input from all preceding encoder/decoder stages, and generates L number of different feature maps for the following encoder/decoder stage through deep fusion of multi-scale features obtained from preceding stages. In each unit cell of MSF module, multi-scale feature aggregation and pyramid fusion scheme is employed, which can be represented as

$$F_{i,j} = \mathcal{F}(\mathbf{E}_1, \mathbf{E}_2, \dots, \mathbf{E}_{\frac{j}{2}}, \mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_{\frac{j}{2}}) \quad (6.4)$$

$$\forall i = \{1, 2, \dots, L\}, j = \{1, 2, \dots, 2m - 1\}$$

where $\mathcal{F}(\cdot)$ represents the functional operations in the MSF unit cell considering L scale of representations from each of the preceding encoder/decoder module.

From final level of the sequential decoder modules, several decoded feature representations are obtained which are processed together in the fusion optimizer unit (\mathcal{O}) to produce the final segmentation mask, and it can be given by,

$$\mathcal{O} = \mathcal{F}(D_{1,1}, D_{1,2}, \dots, D_{1,m}) \quad (6.5)$$

where $\mathcal{O}(\cdot)$ represents the fusion optimizer function.

All the basic building blocks of the CovSegNet architecture are generic and can be designed and optimized for both 2D and 3D operations. In the following discussions, different building blocks of the CovSegNet architecture are presented in detail.

$$\text{Phase 2: } \underset{\Theta_1, \Theta_2}{\operatorname{argmin}} \mathcal{F}\{\mathcal{L}_{2D}(\Theta_1, \mathbf{y}^P, \mathbf{y}), \mathcal{L}_{3D}(\Theta_2, \mathbf{Y}^P, \mathbf{Y})\} \quad (6.6)$$

where Θ_1 denotes the network parameters of CovSegNet2D, Θ_2 denotes the network parameters of CovSegNet3D, \mathbf{X}' , \mathbf{Y}^P , \mathbf{Y} denote the ROI enhanced CT volume, predicted 3D mask, and corresponding 3D ground truth.

6.1.3 Proposed Encoder/Decoder Structure

The encoder and decoder modules are structurally similar that are successively used in the sequential stages of CovSegNet. Encoder/decoder modules are schematically presented in Fig. 6.3. These encoder/decoder modules are composed of several operational unit cells with transitional dense interconnections. The operations of

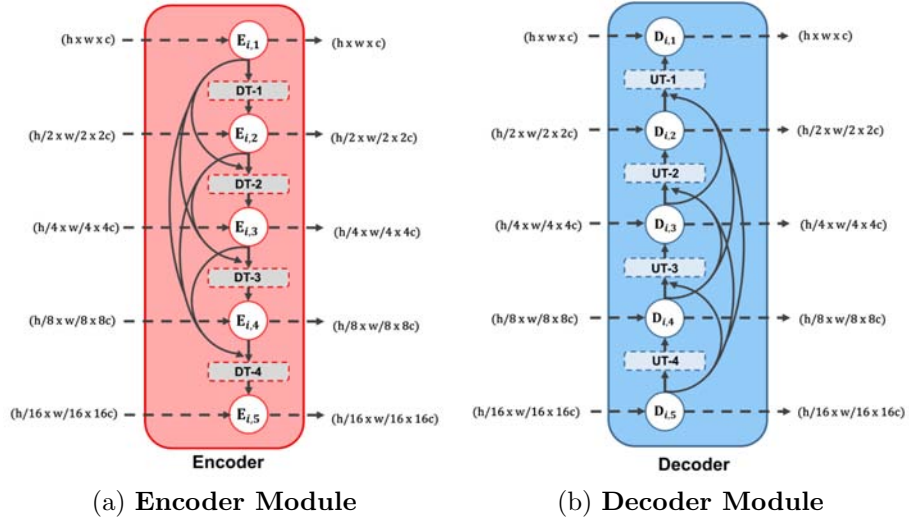


Figure 6.3: Schematic representations of the proposed encoder and decoder modules in five-level implementation.

encoder/decoder modules can be divided into two categories: unit cell operations and transitional operation.

Encoder/Decoder Unit Cell operation

In Fig. 6.4, the unit cell structure of the encoder/decoder module is presented. In each unit cell, two input feature map is entered, one from the transitional unit and the other from the preceding MSF unit while the output feature map is passed through following transitional and multi-scale fusion operations. Moreover, each unit cell consists of four densely interconnected convolutional layers, where each convolutional layer provides two sequential convolutional filtering with (1×1) and (3×3) kernels. Such dense interconnection between convolutional operations has been proven to be effective in numerous applications. No dimensional scaling has been carried out in each of this unit cell as it is employed for introducing adequate transformation in the feature space to encode/decode effective representation. Hence, this unit cell operations can be functionally represented as, $E, D : \mathbb{R}^{h \times w \times c} \rightarrow \mathbb{R}^{h \times w \times c}$, where (h, w, c) represents the height, width and channel of the feature map.

Encoder Down-transitional Operation

During down transitional operations between subsequent unit cells of the encoder module, the spatial dimension of the feature map is reduced for generalizing the feature map, whereas the channel depth is increased to incorporate more filtering operations in subsequent levels for generating more sparser features. It can be

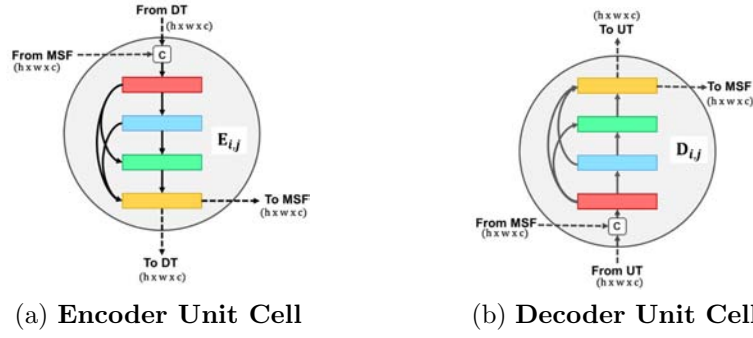


Figure 6.4: Structure of the Encoder/Decoder Unit cells.

functionally presented as, $f : \mathbb{R}^{h \times w \times c} \rightarrow \mathbb{R}^{h/2 \times w/2 \times 2c}$, where spatial resolution is downscaled by 2 and channel depth is increased by 2 from the input feature map obtained from the previous level. However, traditional downsampling operations using pooling/strided convolutions results in loss of contextual information. Moreover, it can be more prominent while incorporating a deep stack of unit cells in the encoder module. To mitigate the loss of contextual information in down transitional operation, a higher level of dense interconnection is proposed among multi-scale feature maps generated from different unit cells. In Fig. 6.5a, the structure of such a down transition unit is schematically presented. In each of such down transition unit, encoded feature representations generated from all higher levels of unit cells are considered for generating the down-scaled feature map. Hence, contextual information lost in each transitional operation can be recovered from very deep stack of unit cells as feature representations from all preceding cells are considered during transition. To converge multi-scale feature maps from preceding levels, firstly, pooling operations with different kernels are carried out to make their spatial dimension uniform and subsequently, channelwise feature aggregation is carried out. The aggregated feature map, $F_{agg,DT}$, generated at i_{th} level can be represented as

$$F_{agg,DT}^i = E^i \oplus P^{(2 \times 2)}(E^{i-1}) \dots \oplus P^{(2^{i-1} \times 2^{i-1})}(E^1) \quad (6.7)$$

where \oplus indicates the feature concatenation, $P^{(2 \times 2)}$ represents pooling operation with (2×2) window, E^i represents the output of i_{th} unit cell of the encoder.

Finally, a convolutional operation with (2×2) kernel is carried out with a stride of (2×2) for generating the downscaled feature map by filtering the aggregated feature vector.

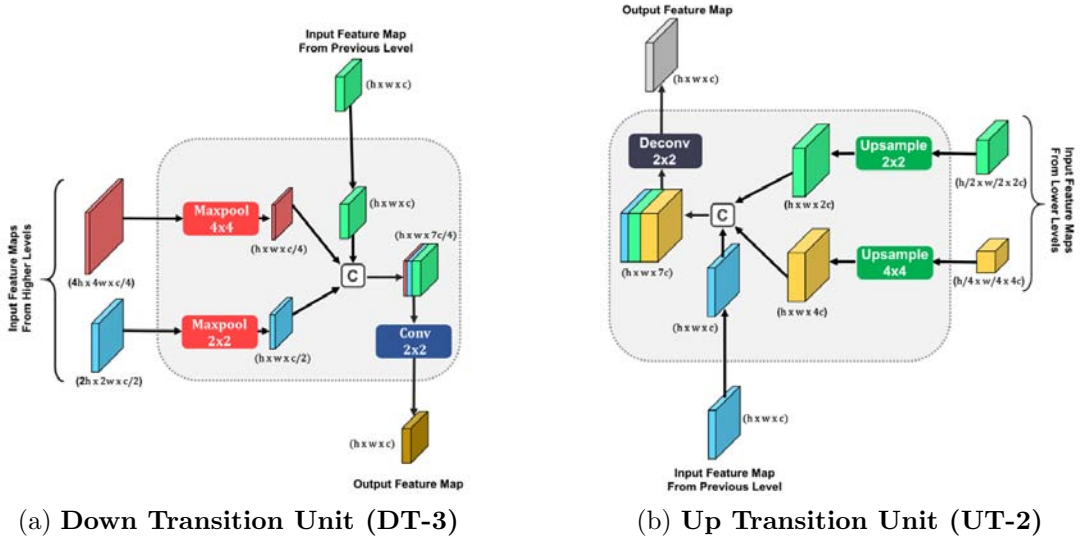


Figure 6.5: Schematic representations of the down transition unit (operating between level-3 and 4) and the up transition unit (operating between level- $(L - 2)$ and $(L - 3)$).

Decoder Up-transitional Operation

On the contrary, up transitional operations are carried out in between successive decoder unit cells to provide the dimensional shifting towards the reconstruction of the final segmentation mask. In each of such up-transition operations, spatial resolution is upscaled by 2 while channel depth is reduced by 2 to get closer to the final reconstruction mask and it can be represented as, $f' : \mathbb{R}^{h \times w \times c} \rightarrow \mathbb{R}^{2h \times 2w \times c/2}$. Similar to the down-transitional operation in Encoder, all the preceding representations of multi-scale decoded feature maps generated from different unit cells are taken into consideration in the up-transition operation to gather more contextual information (Fig. 6.5b). Firstly, spatially uniform feature maps are created through bi-linear interpolation upsampling with different windows, and feature aggregation is carried out to generate aggregated feature vector, $F_{agg,UT}$, which is given by

$$F_{agg,UT}^i = D^i \oplus U^{(2 \times 2)}(D^{i+1}) \dots \oplus U^{(2^{i-1} \times 2^{i-1})}(D^L) \quad (6.8)$$

where $U^{(2 \times 2)}$ represents bilinear upsampling operation with (2×2) window, D^i represents the output of i_{th} unit cell of the decoder.

Finally, the aggregated feature map is processed using a deconvolution operation with (2×2) kernel to incorporate the necessary dimensional up-scaling for further processing in the following unit cell.

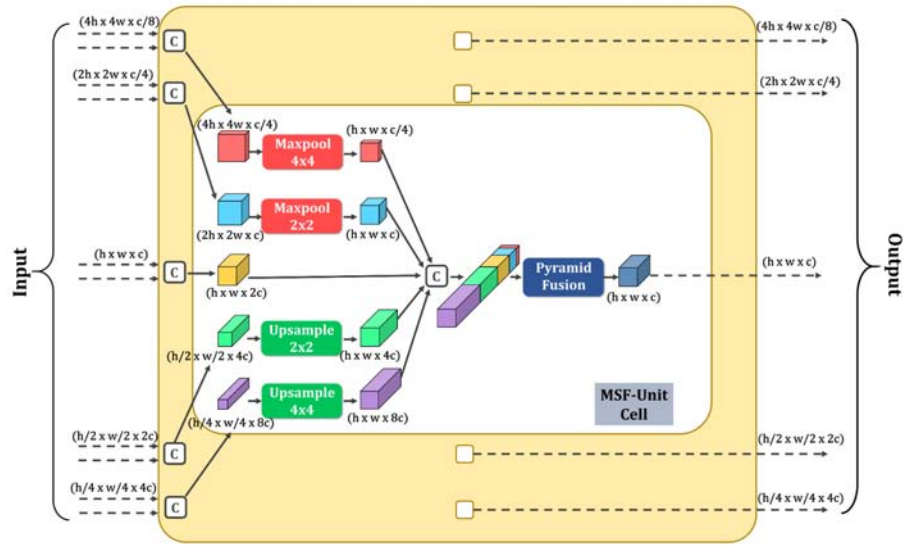


Figure 6.6: Schematic representation of the proposed Multi-Scale Fusion module.

6.1.4 Proposed Multi-Scale Fusion (MSF) Module with Pyramid Fusion scheme

During sequential encoding-decoding operations, a semantic gap is generated between a similar scale of encoded and decoded feature maps. Moreover, in traditional architecture, the gradient has to propagate sequentially that sometimes gives rise to vanishing gradient problems for deeper encoder/decoder module particularly. As multiple stages of encoding and decoding operations are integrated into the CovSegNet, this problem is supposed to be more prominent if all the encoder and decoder modules are sequentially connected. To overcome these limitations, a multi-scale fusion module is proposed that develops parallel interconnection among different scales of feature maps of the encoder/decoder modules utilizing a pyramid fusion scheme.

As shown in Fig. 6.6, each MSF module consists of several MSF-unit cells where each cell considers multi-scale feature maps generated from different levels of preceding encoder/decoder modules and generates feature map for the unit cell of the following encoder/decoder module. Here, similar scale of feature representations generated from different levels of the preceding encoder/decoder modules are concatenated, firstly, to produce L number of multi-scale feature maps. Afterward, all the L scales of feature maps are made spatially equivalent in dimension through pooling and bi-linear upsampling with different windows, and channelwise feature concatenation is carried out to generate the aggregated feature vector.

Afterward, the aggregated feature vector is passed through a pyramid fusion

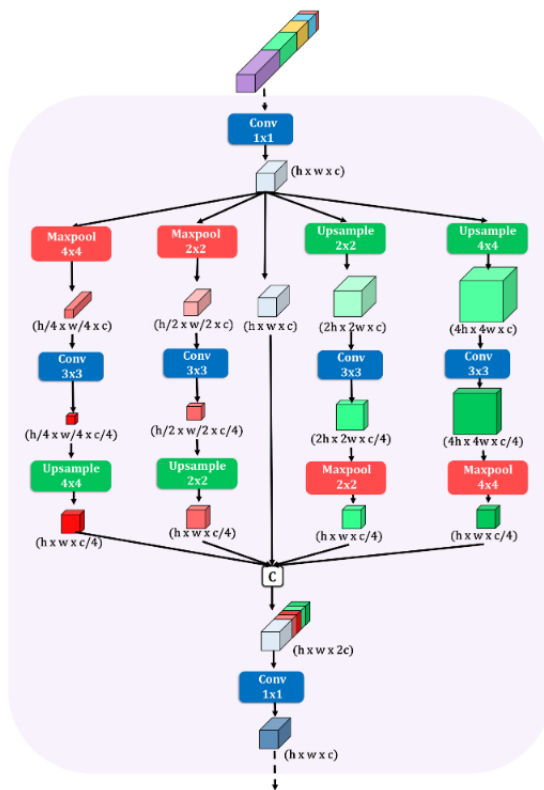


Figure 6.7: **Proposed pyramid fusion scheme for fusing multi-scale features.**

scheme to generate the output feature vector that will be fed to the corresponding encoder/decoder unit cell of the following module. Hence, the generated output feature map from each MSF unit cell contains information from all preceding modules and thus, establishes a parallel flow of optimization for efficient gradient propagation.

6.1.5 Proposed Pyramid Fusion (PF) Module

The pyramid fusion (PF) module incorporates pyramid fusion scheme into the aggregated feature map of MSF unit cell ($F_{agg,MSF}$) utilizing the combinations of sequential multi-window pooling and upsampling operations (shown in Fig. 6.7). Firstly, the depth of the aggregated vector, $F_{agg,MSF}$, is reduced through a pointwise convolution (kernel, 1×1) to generate feature vector f_a , and thus, $F_{agg,MSF} \mapsto f_a$, where $f_a \in \mathbb{R}^{h \times w \times c}$.

Afterwards, the generated vector, f_a , passes through multiple spatial scaling-vertical scaling-inverse spatial scaling operations in parallel with different scaling factors. Spatial scaling operation is carried out utilizing pair of pooling and upsampling operations with different kernel windows, while vertical scaling is employed utilizing convolutional filtering (kernel, 3×3) to reduce the channel depth by one-

fourth of the initial depth. Initial reduction followed by expansion of the feature map assists in gathering the more general feature representation, while initial expansion followed by reduction of the feature map gathers the more detailed information from a sparser domain. These operations pave the way to extract the most generalized representations through analyzing from diverse feature domains, which can be represented by

$$P_r : \mathbb{R}^{h \times w \times c} \rightarrow \mathbb{R}^{h*r \times w*r \times c} \rightarrow \mathbb{R}^{h*r \times w*r \times c/4} \rightarrow \mathbb{R}^{h \times w \times c/4}$$

$$\forall r = \{0.25, 0.5, 2, 4\} \quad (6.9)$$

where P_r denotes one of the parallel operational paths in the PF module with a spatial scaling factor of r .

Afterwards, feature aggregation operation is carried out utilizing different representations generated at multiple paths along with the input representation to generate the aggregated vector $F_{agg,PF}$, where $F_{agg,PF} \in \mathbb{R}^{h \times w \times 2c}$. Finally, a final pointwise convolution (kernel, 1×1) is carried out to generate the output feature map $f_{out,PF}$, where $f_{out,PF} \in \mathbb{R}^{h \times w \times c}$.

6.1.6 Structure of the Fusion Optimizer(\mathcal{O})

The decoded feature maps generated from the top of decoder modules are considered for final reconstruction through a fusion optimization process. This process is schematically shown in Fig. 6.8. Initially, an aggregated feature vector $F_{agg,\mathcal{O}}$, is created considering all the output feature maps from different decoder modules which can be given by

$$F_{agg,\mathcal{O}} = D_{1,1} \oplus D_{1,2} \oplus \dots \oplus D_{1,S} \quad (6.10)$$

where S denotes total number of stages.

Afterward, pyramid fusion scheme is employed on aggregated vector to obtain the more generalized representation utilizing multi-scale decoded representations. Finally, another convolutional filtering (kernel, 3×3) is carried out to generate the final segmentation mask f_{mask} , utilizing binary activation function, and these can be represented as

$$f_{mask} = \sigma(Conv(PF(F_{agg,\mathcal{O}}))) \quad (6.11)$$

where $\sigma(\cdot)$ denotes the non-linear activation.

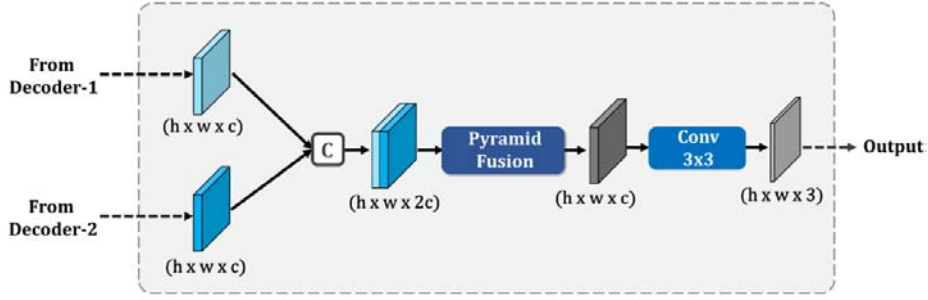


Figure 6.8: Schematic of the fusion optimizer module optimizing the decoded feature maps generated from two decoding stages.

6.1.7 Loss Function

Tversky Index is introduced in [214] for better generalization of the the dice index by balancing out false positives and false negatives, which is given by

$$TI = \frac{\sum_{i=1}^P p_{1i}g_{1i} + \epsilon}{\sum_{i=1}^P p_{1i}g_{1i} + \alpha \sum_{i=1}^P p_{0i}g_{1i} + \beta \sum_{i=1}^P p_{1i}g_{0i} + \epsilon} \quad (6.12)$$

where g_{0i} , p_{0i} indicate the ground truth and prediction probability of pixel i being in a normal region, while g_{1i} , p_{1i} indicate the ground truth and prediction probability of pixel i being in an abnormal region, P is the total number of pixels on a certain image, α , β are used to shift emphasize for balancing class imbalance such that $\alpha + \beta = 1$, and $\epsilon(10^{-8})$ is used to avoid division-by-zero as safety factor.

To put more emphasis on hard training examples, a Focal Tversky loss function is introduced in [215] utilizing the Tversky Index, which is given by

$$\mathcal{L} = \sum_c (1 - TI_c)^{\frac{1}{\gamma}} \quad (6.13)$$

where γ is used to emphasize the challenging less accurate predictions. Due to the better generalization over a large number of datasets according to [215], $\alpha = 0.7$, $\beta = 0.3$, $\gamma = \frac{4}{3}$ are used for all experimentations in this study.

If \mathbf{y} , \mathbf{y}^P denote slice-wise mask ground truth and corresponding probability prediction, respectively, while \mathbf{Y} , \mathbf{Y}^P denote volumetric mask ground truth and corresponding probability prediction, respectively, the objective loss functions for separately optimizing CovSegNet2D and CovSegNet3D can be represented as

$$\mathcal{L}_{2D} = \mathcal{L}(\mathbf{y}, \mathbf{y}^P); \mathbf{y}, \mathbf{y}^P \in \mathbb{R}^{h \times w \times c} \quad (6.14)$$

$$\mathcal{L}_{3D} = \mathcal{L}(\mathbf{Y}, \mathbf{Y}^P); \mathbf{Y}, \mathbf{Y}^P \in \mathbb{R}^{h \times w \times s \times c} \quad (6.15)$$

The joint optimization objective function used in phase-2 combining slice-wise and volumetric operations is given by

$$\mathcal{F} = \lambda \left(\frac{1}{S} \sum_{i=1}^S \mathcal{L}_{2D}^i \right) + \mathcal{L}_{3D} \quad (6.16)$$

where λ denotes the scaling factor of 2D-loss term, and s denotes total number of 2D-slices per volume. Here, $\lambda = 0.2$ is used for optimization to provide more emphasis on CovSegNet3D in phase-2 as CovSegNet2D is pre-trained in phase-1 and is supposed to be fine-tuned in phase-2.

6.2 Results and Discussions

Experimentations have been carried out on three publicly available datasets to validate the effectiveness of the proposed scheme on numerous segmentation tasks. Performances of CovSegNet2D and CovSegNet3D have been separately studied along with the proposed hybrid scheme of joint optimization combining CovSegNet2D and CovSegNet3D.

6.2.1 Dataset Description

Dataset-1 contains 20 CT volumes with 1800+ slices annotated by expert radiologist panel [216]. All the slices have annotations for both lung and infection regions. Each slices are of resolution (630×630) which are resized to (512×512) . Dataset-2 is the ‘‘COVID-19 CT Segmentation dataset’’ that contains 110 axial CT images collected by the Italian Society of Medical and Interventional Radiology from 40 different COVID-patients [217]. All the images are of resolution (512×512) . Each slice contains multi-class annotations of infections.

6.2.2 Experimental Setup

Different hyper-parameters of the network are chosen through experimentation for better performance. A number of traditional evaluation metrics are used for the evaluation of performance, such as IoU, dice score, sensitivity, and specificity. A five-fold cross-validation scheme is carried out separately on these databases for evaluation of the proposed scheme. Mean and standard deviations of the evaluation metrics obtained from different test folds are reported. The Wilcoxon rank-sum test ($\alpha = 0.01$) is used for statistical analysis of the performance improvement obtained from the proposed scheme.

Table 6.1: Ablation Study of the Effect of Different Modules in the Performance (Mean±Standard Deviation) of the Proposed CovSegNet2D Architecture

Network	Dataset-1					Dataset-2				
	Sen.(%)	Spec.(%)	Dice(%)	IoU(%)	p-Value	Sen.(%)	Spec.(%)	Dice(%)	IoU(%)	p-Value
Baseline (V1)	82.7±0.49	97.4±0.09	84.1±0.29	79.8±0.21	-	71.7±0.12	95.8±0.18	71.9±0.33	65.8±0.27	-
Baseline+ DT (V2)	83.8±0.29	97.8±0.12	85.8±0.36	81.1±0.08	0.0033	73.6±0.31	96.5±0.15	73.4±0.14	67.6±0.21	0.0023
Baseline+ UT (V3)	83.1±0.25	97.7±0.08	85.4±0.16	80.9±0.13	0.0017	73.1±0.55	96.3±0.18	73.1±0.19	67.2±0.35	0.0044
Baseline+ DT+UT (V4)	84.9±0.41	98.1±0.11	86.7±0.27	82.3±0.32	0.0021	74.6±0.17	97.1±0.12	74.8±0.34	69.4±0.18	0.0012
Baseline+(MSF-w/o PF) (V5)	86.9±0.15	98.3±0.07	87.3±0.28	82.9±0.26	0.0019	76.2±0.27	97.9±0.16	77.2±0.29	72.8±0.24	0.0034
Baseline+ MSF (V6)	88.4±0.28	98.7±0.08	89.2±0.32	84.1±0.21	0.0041	78.8±0.25	98.4±0.11	79.5±0.21	74.1±0.25	0.0048
CovSegNet2D (V7)	90.8±0.32	99.1±0.13	91.1±0.25	86.9±0.09	0.0011	81.5±0.22	98.9±0.13	82.7±0.08	77.5±0.14	0.0009

6.2.3 Performance Analysis

To analyze the effectiveness of different modules of the proposed CovSegNet architecture, an ablation study is carried out. The baseline model is defined as the two-stage implementations with encoder and decoder modules only excluding the down-transition (DT) units, up-transition units (UT), and multi-scale fusion modules. The statistical significance test is carried out to validate the improvement of dice-scores over the baseline model.

- i. **Effects of the transition unit:** Instead of proposed down-transition units and up-transition units, traditional max-pooling and upsampling operations are used, respectively, in the baseline model according to the conventions of traditional Unet architecture. Performances with different combinations of transition units are provided in (V2-V4) of Table 6.1 for 2D analysis. The inclusion of down-transition unit (V2) in encoder modules provides 1.7% improvement and 1.5% improvement of dice scores in Database-1 and 2, respectively, over the baseline. Moreover, the inclusion of up-transition unit (V3) in decoder modules provides 1.3% and 1.2% improvements of dice scores, while the inclusion of both of the transition units (V4) provide 2.6% and 2.9% improvements of dice scores in Database-1 and 2, respectively. Hence, both of the up-transition units and down-transition units are contributing considerable improvements over the baseline performance. Similar improvements can be noticeable for 3D variants of the transition units also (from $V2_{3D}$ to $V4_{3D}$) that are summarized in Table 6.5. All the improvements are found to be statistically significant ($p < 0.01$).
- ii. **Effects of the multi-scale fusion (MSF) module:** The MSF modules are proposed in place of the traditional direct skip connection scheme of Unet architecture to reduce the semantic gaps between subsequent encoder and decoder modules. In the baseline model, direct skip connections are used between succeeding modules instead of the MSF module. In Table 6.1, the change of performance with the inclusion of the MSF module in the 2D-baseline model

Table 6.2: Performance Comparison (Mean±Standard Deviation) of the Proposed CovsegNet2D Architecture with Other State-of-the-Art Approaches on 2D-CT slices

Network	Dataset-1					Dataset-2				
	Sen.(%)	Spec.(%)	Dice(%)	IoU(%)	p-Value	Sen.(%)	Spec.(%)	Dice(%)	IoU(%)	p-Value
Unet [213]	75.9±0.34	88.9±0.12	79.3±0.26	74.9±0.18	-	52.9±0.29	86.2±0.09	43.3±0.34	38.8±0.32	-
Unet++ [137]	78.6±0.17	91.1±0.18	81.1±0.23	76.2±0.21	-	57.7±0.32	89.2±0.11	52.3±0.31	48.1±0.37	-
MultiResUnet [136]	77.2±0.33	90.3±0.24	82.7±0.28	77.4±0.15	-	56.9±0.27	86.9±0.15	50.8±0.28	45.2±0.22	-
Attention-Unet-2D [218]	81.1±0.29	92.2±0.11	85.1±0.14	79.6±0.28	-	60.8±0.25	88.4±0.12	57.7±0.36	51.9±0.26	-
CPF-Net [219]	78.9±0.27	91.7±0.14	84.4±0.25	79.3±0.25	-	62.2±0.14	91.1±0.14	60.4±0.25	56.1±0.21	-
Semi-Inf-Net [124]	82.7±0.26	94.8±0.21	86.9±0.34	81.1±0.18	-	72.9±0.44	95.8±0.19	74.1±0.24	68.1±0.32	-
CovSegNet2D(Ours)	90.8±0.32	99.1±0.13	91.1±0.25	86.9±0.09	0.0008	81.5±0.22	98.9±0.13	82.7±0.08	77.5±0.14	0.0013

Table 6.3: Performance Comparison (Mean±Standard Deviation) of the CovSegNet3D Architecture with Other State-of-the-Art Networks on 3D-CT Volumes of Dataset-1

Network	Sen.(%)	Spec.(%)	Dice(%)	IoU(%)	p-Value
Unet-3D [213]	77.1±0.22	89.8±0.18	84.2±0.27	79.4±0.24	-
Unet++-3D [137]	79.2±0.17	91.7±0.25	85.1±0.29	80.2±0.26	-
MultiResUnet-3D [136]	78.7±0.27	90.9±0.16	84.5±0.31	78.9±0.18	-
Attention-Unet-3D [218]	82.5±0.26	93.1±0.31	85.9±0.24	81.4±0.29	-
CPF-Net-3D [219]	80.1±0.23	92.6±0.23	85.2±0.18	80.8±0.34	-
VNet-3D [131]	84.3±0.29	93.9±0.17	85.7±0.31	81.3±0.19	-
CovSegNet3D(Ours)	91.1±0.26	99.3±0.09	92.3±0.15	87.7±0.23	0.0024
CovSegNet-Hybrid(Ours)	92.6±0.25	99.5±0.07	94.1±0.19	90.2±0.27	0.0011

is provided in V6. It should be noticed that 5.1% improvement of dice-score, 4.3% improvement of IoU score have been achieved in Database-1, while 7.6% improvement of dice-score, 8.3% improvement of IoU score have been achieved in Database-2. Similar performance improvements can be noticed for the incorporation of MSF module in the 3D-baseline model ($V6_{3D}$ in Table 6.5). These improvements are found to be statistically significant ($p < 0.01$).

- iii. **Effects of the pyramid fusion scheme in MSF module:** Pyramid fusion (PF) modules are integrated into the MSF modules to operate on the aggregated multi-scale feature vector in the MSF module. Instead of the PF module, a point-wise convolution with (1×1) kernel can be performed to reduce and transform the aggregated vector into the output vector. The performance of the 2D-baseline model including this simplified version of the MSF module is reported in V5 of Table 6.1. It is to be noted that 2.3% improvement of dice score is achieved in Database-1 and 3.4% improvement is achieved in Database-2 over the baseline model using these simplified MSF modules, and these improvements are statistically significant ($p < 0.01$). However, 3.2% and 5.3% reduction of dice scores can be noticed in Database-1 and 2, respectively, from the baseline model with original MSF modules ($V6$) incorporating PF scheme. Similarly, considerable improvement is also achieved for the incorporation of 3D-pyramid fusion scheme in the 3D variants of MSF module which

Table 6.4: Effect of Vertical Expansions (Levels) and Horizontal Expansions (Stages) on the Dice Score (Mean±Standard Deviation) in Dataset-1

Level	CovSegNet2D			CovSegNet3D		
	1-stage	2-stage	3-stage	1-stage	2-stage	3-stage
2	49.9±0.37	75.3±0.13	78.12±0.21	57.3±0.18	79.8±0.18	82.1±0.19
3	64.8±0.23	85.8±0.32	88.5±0.15	69.3±0.35	89.2±0.26	90.2±0.25
4	75.2±0.32	89.6±0.27	90.8±0.22	79.8±0.29	92.3±0.15	91.8±0.17
5	83.5±0.19	91.1±0.25	89.9±0.12	84.5±0.43	90.2±0.34	89.7±0.28
6	86.7±0.27	90.9±0.21	89.1±0.11	89.3±0.21	89.8±0.41	87.9±0.36

Table 6.5: Ablation Study of the Effect of Different Modules in the Performance (Mean±Standard Deviation) of the Proposed CovSegNet3D Architecture in Dataset-1

Network	Dataset-1				
	Sensitivity(%)	Specificity(%)	Dice Score(%)	IoU(%)	p-Value
Baseline3D (V1 _{3D})	84.5±0.21	97.9±0.12	85.2±0.23	80.8±0.32	-
Baseline3D + DT (V2 _{3D})	85.7±0.31	98.2±0.19	86.1±0.25	82.3±0.29	0.0011
Baseline3D + UT (V3 _{3D})	85.2±0.18	98.1±0.08	85.9±0.18	82.0±0.21	0.0008
Baseline3D + DT+UT (V4 _{3D})	86.7±0.22	98.7±0.14	88.3±0.28	83.5±0.27	0.0017
Baseline3D+(MSF-w/o PF) (V5 _{3D})	87.4±0.25	97.9±0.11	88.2±0.21	83.8±0.31	0.0032
Baseline3D+ MSF (V6 _{3D})	89.6±0.19	98.4±0.15	89.9±0.17	85.1±0.19	0.0021
CovSegNet3D	91.1±0.26	99.3±0.09	92.3±0.15	87.7±0.23	0.0025

can be noticed from $V5_{3D}$ and $V6_{3D}$ in Table 6.5. It justifies the effectiveness of the pyramid fusion scheme in the MSF module.

iv. **Effects of vertical and horizontal scaling** The proposed CovSegNet architecture is designed in a modular way with the opportunity for both vertical and horizontal expansions for integrating more number of levels and stages, respectively. In Table 6.4, the performances of the CovSegNet architecture with different numbers of levels and stages are provided. It should be noticed that the optimum dice score of 91.1% is obtained for CovSegNet2D with 5-levels and 2-stages. The best performance on single stage implementation is found to be 86.7%, which is 4.4% lower than the best of the 2-stage implementation. Similar analyses have been carried out on CovSegNet3D using volumetric data where the highest dice score of 92.3% is achieved with 3-levels and 2-stages implementation. Moreover, when more stages are included, comparably higher performances are obtained in a lower number of levels, e.g. best dice score of 90.8% in the 3-stage setup of CovSegNet2D has been achieved with 4-levels. With the horizontal expansion, the model gathers more amount of contextual information in a lower number of stages that result in higher performances. However, more expansion in both directions starts to increase the complexity that causes a decrease in performance due to overfitting issues.

v. **Effects of the hybrid 2D-3D joint optimization scheme with two-**

Table 6.6: Comparison of Performances (Mean±Standard Deviation) on Different Types of Infections (Ground Glass Opacity and Consolidation) in Different CT-slices of Dataset-2

Network	Consolidation					Ground-Glass Opacity				
	Sen.(%)	Spec.(%)	Dice(%)	IoU(%)	p-Value	Sen.(%)	Spec.(%)	Dice(%)	IoU(%)	p-Value
Unet [213]	41.1±0.26	96.2±0.12	40.3±0.28	35.5±0.28	-	35.1±0.27	98.2±0.09	44.1±0.27	39.8±0.25	-
Unet++ [137]	48.8±0.23	97.8±0.16	42.6±0.26	38.2±0.19	-	41.2±0.32	96.6±0.14	49.9±0.22	45.7±0.27	-
MultiResUnet [136]	46.6±0.28	97.1±0.14	42.1±0.19	37.6±0.27	-	44.5±0.28	97.3±0.11	47.7±0.18	43.1±0.28	-
Attention-Unet-2D [218]	44.8±0.19	96.8±0.08	44.5±0.25	40.1±0.33	-	55.3±0.31	95.4±0.08	52.9±0.17	47.6±0.35	-
CPF-Net [219]	49.9±0.18	97.4±0.15	44.1±0.23	39.9±0.29	-	53.5±0.22	96.9±0.13	56.9±0.26	51.1±0.34	-
Semi-Inf-Net [124]	50.9±0.22	96.7±0.11	45.8±0.31	41.4±0.18	-	62.2±0.34	96.1±0.18	62.7±0.22	58.4±0.23	-
CovSegNet2D(Ours)	63.8±0.17	98.4±0.09	56.8±0.24	51.9±0.25	0.0017	73.3±0.25	98.9±0.12	70.9±0.31	66.1±0.19	0.0028

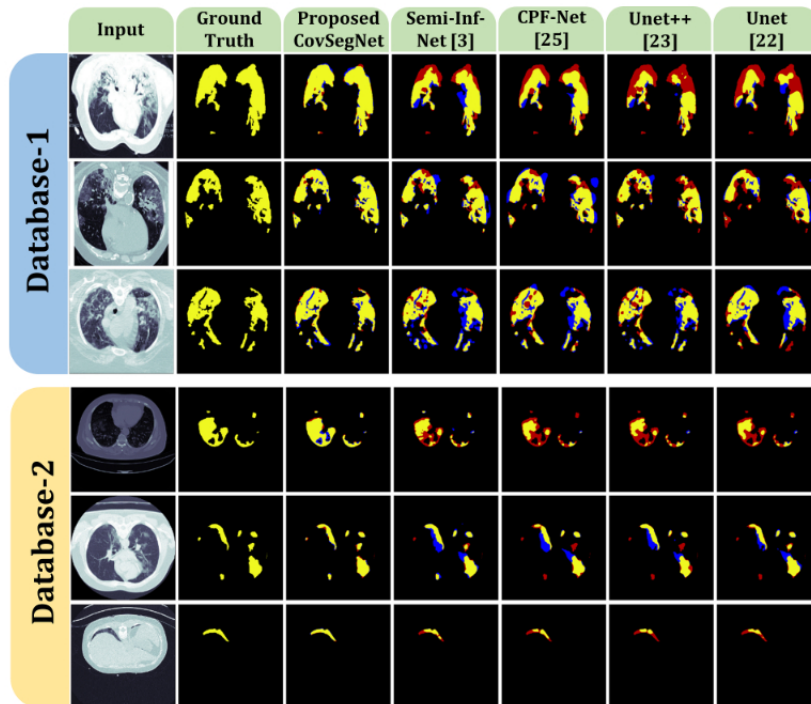


Figure 6.9: Visual representations of the segmentation performances of different state-of-the-art networks on the CT images from Database-1 and Database-2. Here, ‘yellow’, ‘red’, and ‘blue’ represent true positive (TP), false negative (FN), and false positive (FP) regions, respectively.

phase training: The proposed 2-phase training scheme exploits the advantages of both the slice-based optimization and volumetric optimization. Quantitative performances obtained using CovSegNet2D, CovSegNet3D, and the hybrid scheme are provided in Table 6.2 and 6.3. Slice based processing provides the advantages of employing deeper networks for lighter 2D-convolutions, while loses the inter-slice contextual information that results in sub-optimal performance. On the other hand, 3D-volumetric analysis incorporates more contextual information while increasing the computational burden of optimization for the expensive 3D-kernels processing. The best variant of CovSegNet3D provides 1.2% higher dice score, and 0.8% higher IoU score over the best variant of CovSegNet2D. Thus, the performances of the proposed

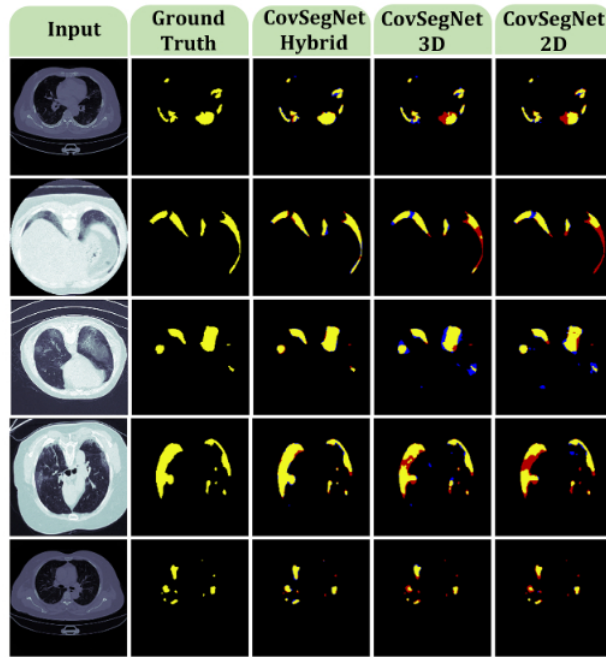


Figure 6.10: **Visual representations of the segmentation performances obtained using single phase training (CovSegNet2D and CovSegNet3D) and multi-phase training (with hybrid 2D-3D networks) in Dataset-1.**

CovSegNet architectures are quite comparable in both 2D and 3D processing with minor variations. It is to be noted that, more improvements can be achieved with the expensive 3D-processing if the number of training CT-volumes can be increased substantially for exploiting the advantages of the complete 3D-processing. However, by combining the advantages of both these schemes in the proposed multi-phase hybrid training approach, 3% and 1.8% higher dice scores are achieved compared to the best performing CovSegNet2D and CovSegNet3D architectures, respectively. In the hybrid scheme, to reduce the computational burden of 3D-data processing, only 2-level and dual-stage implementation of the CovSegNet3D is employed accompanied by the 4-level and dual-stage implementation of the CovSegNet2D that provides the optimal performance with minimal complexity. Since a very shallower variant of CovSegNet3D is employed in the hybrid network compared to the best performing variant of CovSegNet3D, the operational complexity is greatly reduced in the hybrid network that led to the optimum performance with the available CT-volumes. This improvement signifies the effectiveness of the hybrid networking scheme in multi-phase training ($p < 0.01$). Moreover, qualitative analysis of the performances of the individual networks and hybrid networks are presented in Fig. 6.10 with different levels of infection. It should be noticed that both

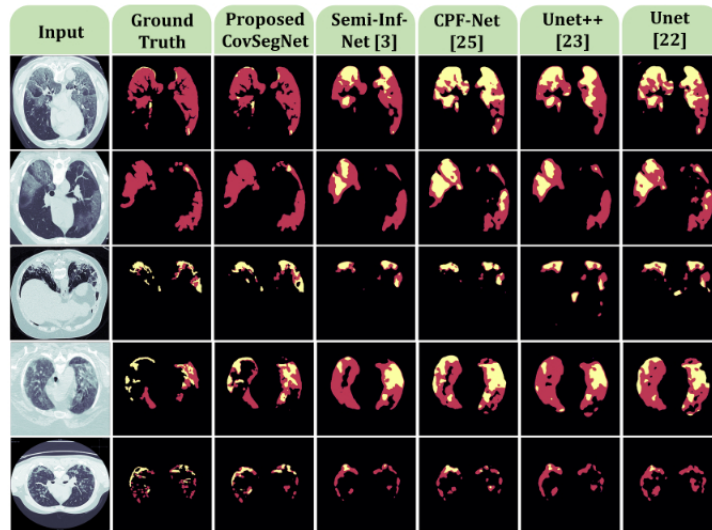


Figure 6.11: Visual representations of the segmented multi-class lesions of the CT images from Database-2 obtained using different state-of-the-art networks. Here, ‘red’ represents the ‘Ground Glass Opacity (GGO)’ regions and ‘yellow’ represents the ‘Consolidation’ regions.

of the false positive and false negative regions are reduced in the segmented mask for the hybrid scheme compared to the individual networks. Therefore, for the proper optimization with the hybrid networking scheme through multi-phase training, optimum performance is achieved compared to the independent 2D/3D data processing.

- vi. **Comparison with Other Existing Approaches:** To compare the performances of the proposed CovSegNet architecture, several state-of-the-art networks are considered. To compare on a fair platform, most of these networks are implemented using their open-source implementation, and same train-test folds are used for performance evaluation. Infection segmentation performances using slice-based 2D-operations and volumetric 3D-operations are summarized in Table 6.2 and 6.3, respectively. CovSegNet2D provides a 4.2% higher dice score in Database-1, and an 8.6% improvement in dice score in Database-2 compared to the second-highest score (Semi-Inf-Net). Hence, consistent improvements in performances have been achieved in 2D-slice based analysis using CovSegNet2D. Moreover, in the volumetric analysis approach, CovSegNet3D provides an 8.4% higher dice score and 9.4% higher IoU score compared to the next-best performing model (VNet). Thus, the 3D-variant of CovSegNet provides consistent improvements over other 3D-counterparts of existing networks. It should be noticed that the proposed hybrid scheme combining CovSegNet2D and CovSegNet3D provides the most optimum per-

formance with a dice score of 94.1% and IoU score of 90.2%. Some of the qualitative visualizations of performances obtained in different challenging conditions are shown in Fig. 6.9. For having the volumetric information of the Database-1, the proposed hybrid scheme is employed here, while only 2D-slice based analysis is carried out in Database-2 using CovSegNet2D. It should be noted that the proposed scheme performs consistently better compared to other networks in segmenting most of the challenging diffused, blurred, and varying shaped edges of COVID lesions. Moreover, quantitative performances on challenging multi-class lesion segmentation, including separate ground-glass opacity (GGO) and consolidation regions, are summarized in Table 6.6, where 8.2% improvement in dice score is obtained in GGO segmentation and 11% improvement in consolidation segmentation using CovSegNet architecture over the other best-performing approaches. Additionally, from the visual analysis of the performances shown in Fig. 6.11, it can be easily noted that the proposed network considerably reduces the false predictions even in these challenging conditions compared to other state-of-the-art approaches.

- vii. **Computational Efficiency Analysis of Numerous Approaches:** The proposed CovSegNet architecture ensures the proper optimization of all the network parameters through improved parallelization that enhances efficient gradient propagation in the whole network. However, this improved parallelism also poses some computational burden for the effective exploitation of the network parameters. Nevertheless, the CovSegNet architecture provides additional opportunity for horizontal scaling as well as vertical scaling that facilitates the performance improvement with much shallower variant. On the contrary, other traditional networks solely depend on vertical scaling that exponentially increases the computational burden with exponential increase of the number of convolutional filters in the deeper layers. In Table 6.7, the computational efficiency of different networks are summarized, where performances of different variants of CovSegNet is summarized based on the number of levels (L) and stages (S). For 2D-processing, it is to be noted that. the CovSegNet2D-v2 achieves 3.5% higher dice score compared to the Unet while incorporating only 3-levels (L-3), and two-horizontal stages (S-2). Due to lower number of filtering operations in the upper vertical levels, significantly lower number of parameters (reduced 94.8%) are incorporated. However, for proper optimization of these parameter with improved parallelism in the network, comparatively lower gain is achieved in terms of the GPU consumption (reduced 14.2%) and inference time (reduced 30%) with respect to the Unet. A

similar observation can be carried out for 3D analysis with CovSegNet3D. It is clear that 3D processing increases computational complexity greatly compared to the 2D networks. However, it should be noticed that CovSegNet-Hybrid provides the best achievable dice score (94.1%) while consisting of 0.09x parameters of Unet3D with 0.08s reduction of inference time. This significant reduction in parameter counts is mainly achieved by integrating a shallower variant of CovSegNet3D with the CovSegNet2D. Moreover, this hybrid processing effectively extracts both the inter-slice and intra-slice contextual information that are responsible for the highest dice score. Therefore, this hybrid scheme provides considerable advantages over other existing 3D variants in terms of parameters, and dice scores with comparable processing speed.

6.3 Conclusion

In this chapter, an automated scheme is proposed with an efficient neural network architecture (CovSegNet) for very precise lung lesion segmentation of COVID CT scans that provides outstanding performances with 8.4% average improvement of dice score over two datasets. It is found that horizontal expansion mechanism with multi-stage encoder-decoder modules assists in further improvements for gathering more multi-scale contextual information when coupled with the traditional vertical expansion mechanism. Furthermore, the two-phase optimization scheme with hybrid 2D-3D processing provides considerable improvement over traditional single domain approaches for introducing more contextual information to gather finer details.

Table 6.7: Computational Efficiency Analysis of Numerous Architectures along with the Performances Obtained on Dataset-1

2-Dimensional Analysis							3-Dimensional Analysis						
Architecture	Details	Total Parameters(M)	GPU Usage(GB)	Inference Time(s)	Mean Dice(%)		Architecture	Details	Total Parameters(M)	GPU Usage(GB)	Inference Time(s)	Mean Dice(%)	
Unet [213]	-	31.0	2.1	0.10	82.3		Unet3D [213]	-	90.3	13.2	1.22	84.2	
Semi-Inf-Net [124]	-	33.3	6.8	0.18	86.9		Vnet3D [131]	-	45.1	15.1	1.16	85.7	
Unet++ [137]	-	27.0	6.5	0.17	84.1		MultiResUnet3D [136]	-	18.1	12.9	1.15	84.5	
CPF-Net [219]	-	32.4	2.3	0.12	84.4		Attention Unet3D [218]	-	103.5	20.9	1.13	85.9	
CovSegNet2D-v1 (ours)	L-2, S-2	0.37	1.1	0.05	75.3		CovSegNet3D-v1 (ours)	L-2, S-2	1.1	7.0	1.02	79.8	
CovSegNet2D-v2 (ours)	L-3, S-2	1.60	1.8	0.07	85.8		CovSegNet3D-v2 (ours)	L-3, S-2	4.6	13.7	1.21	89.2	
CovSegNet2D-v3 (ours)	L-4, S-2	6.70	3.3	0.11	89.6		CovSegNet3D-v3 (ours)	L-4, S-2	19.0	22.2	1.85	92.3	
CovSegNet2D-v4 (ours)	L-5, S-2	27.0	7.0	0.20	91.1		CovSegNet Hybrid (ours)	2D(L-4, S-2) 3D(L-2,S-2)	7.8	10.5	1.14	94.1	

Chapter 7

CovTANet: A Multi Objective Learning Framework for Lesion Segmentation, Diagnosis, and Severity Prediction of COVID-19

With a large number of asymptomatic patients, early detection of COVID-19 through CT imaging is still a stupendously challenging task due to significantly smaller, scattered, and obscure regions of infections that are difficult to distinguish [220]. These diverse heterogeneous characteristics of infections among different subjects also make the severity prediction to be an extremely difficult objective to achieve [221]. The scarcity of considerably large reliable datasets further increases the complexity of the endeavor. Most of the recent studies mostly opt for solving this daunting task partially where infection segmentation, diagnosis, or severity analysis have separately attempted [138]–[140]. Such methods lack the complete integration of the objectives for providing a robust clinical tool.

In this chapter, CovTANet, an end-to-end hybrid neural network is proposed, that is capable of performing precise segmentation of COVID lesions along with accurate diagnosis and severity predictions. The intricate network of the proposed scheme emerges as an effective solution by overcoming the limitations of the traditional approaches. The major contributions of this work can be summarized as follows:

- i. A novel tri-level attention guiding mechanism is proposed combining channel, spatial and pixel domains for feature recalibration and better generalization.
- ii. A tri-level attention based segmentation network (TA-SegNet) network is pro-

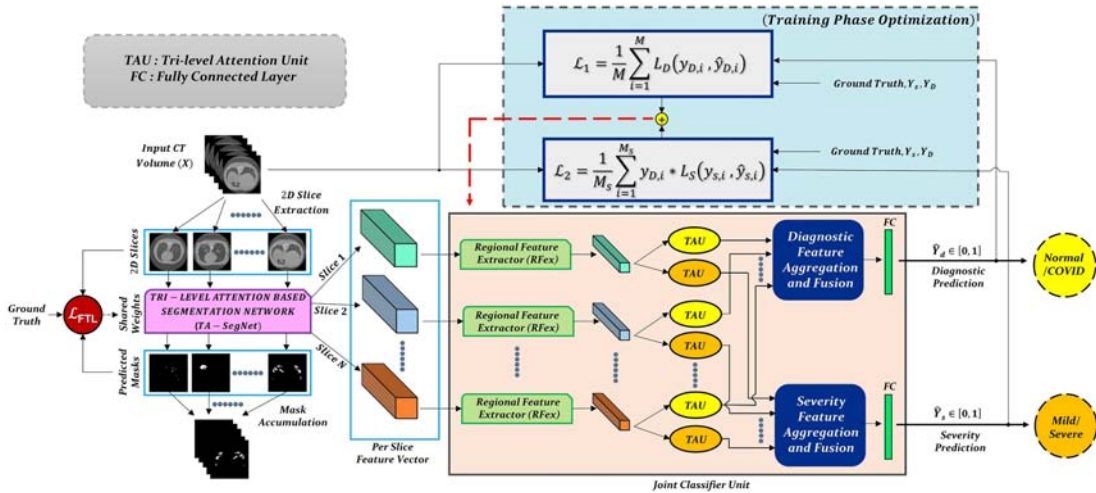


Figure 7.1: Graphical overview of the optimization scheme of CovTANet

posed for precise segmentation of COVID lesions integrating the triple attention mechanisms with parallel multi-scale feature optimization and fusion.

- iii. A multi-phase optimization scheme is introduced by effectively integrating the initially optimized TA-SegNet with the joint diagnosis and severity prediction framework.
- iv. A system of networks is proposed for efficient processing of CT-volumes to integrate all three objectives for improving performance in challenging conditions.
- v. Extensive experimentations have been carried out over a large number of subjects with diverse levels and characteristics of infections. The primary results of these experimentation are published in [22].

7.1 Methodology

The proposed CovTANet network is developed in a modular way focusing on diverse clinical perspectives including precise COVID diagnosis, automated lesion segmentation, and effective severity prediction. The whole scheme is represented in Fig. 7.1. Here, a hybrid neural network (CovTANet) is introduced for segmenting COVID lesions from CT-slices as well as for providing effective features of the region-of-lesions which are later integrated for the precise diagnosis and severity prediction tasks.

The complete optimization process is divided into two sequential stages for efficient processing. Firstly, a neural network, named as Tri-level Attention-based Segmentation Network (TA-SegNet), is designed and optimized for slicewise lesion

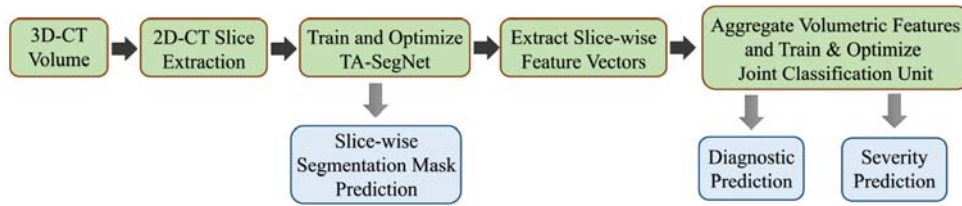


Figure 7.2: **Optimization flowchart of the proposed CovTANet network.**

segmentation from a particular CT-volume. A tri-level attention gating mechanism is introduced in this network with multifarious architectural renovations to overcome the limitations of the traditional Unet network (Section 7.1.2), which gradually accumulates effective features for precise segmentation of COVID lesions. Because of the pertaining complicacy with blurred, diffused, and scattered patterns of COVID lesions, it is quite obvious that direct utilization of the final segmented portions for diagnosis may result in loss of information due to some false positive estimations. The proposed CovTANet aims to resolve this issue by extracting effective features regarding the regions-of-infection utilizing the initially optimized TA-SegNet as it is optimized for precisely segmenting COVID lesions with diverse levels, types, and characteristics.

Additionally, separate regional feature extractors are employed for generating more generalized forms of the slicewise feature vectors from different lung regions. Subsequently, these generalized feature representations of CT-slices are guided into separate volumetric feature aggregation and fusion schemes through the proposed tri-level attention mechanism for extracting the significant diagnostic features as well as severity based features. The diagnostic path is supposed to extract the more generalized representation of infections while the severity path is more concerned with the levels of infections. Both the diagnostic and severity predictions are optimized through a joint optimization strategy with an amalgamated loss function. The optimization flow of the complete CovTANet network is shown in Fig. 7.2. Several architectural submodules of the CovTANet are discussed in detail in the following sections.

7.1.1 Proposed Tri-level Attention Scheme

Attention mechanism, first proposed in [222] for enhanced contextual information extraction in natural language processing, has been adopted in numerous fields including medical image processing [223], [224]. This mechanism assists faster convergence with considerable performance improvement by eliminating the redundant parts while putting more attention on the region-of-interests through the general-

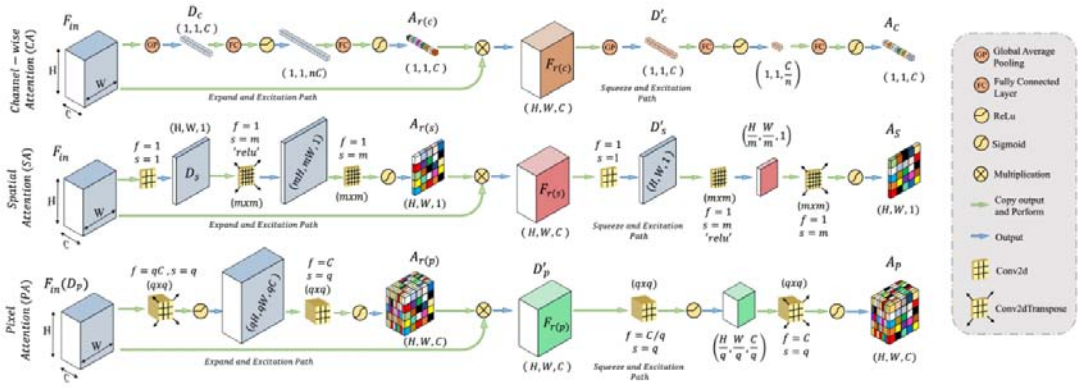


Figure 7.3: Schematic of the proposed channel, spatial, and pixel attention mechanisms.

ization of the predominant contextual information. Here, we have proposed a novel self-supervised attention mechanism combining three levels of abstraction for improved generalization of the relevant contextual features, i.e. channel-level, spatial-level, and pixel-level. The channel attention (CA) mechanism operates on a broader perspective to emphasize the corresponding channels containing more information, while the spatial attention (SA) mechanism concentrates more on the local spatial regions containing region of interests, and finally, the pixel attention (PA) mechanism operates on the lowest level to analyze the feature relevance of each pixel. However, relying only on the higher level of attention causes loss of information while relying on lower/local levels may weaken the effect of generalization. Hence, to reach the optimum point of generalization and re-calibration of feature space, we have introduced a tri-level attention unit (TAU) mechanism that integrates the advantages of all three levels of attention. This TAU unit module is repeatedly used all over the CovTANet network (Fig. 7.1) to improve the feature relevance through feature recalibration.

In general, the proposed attention mechanisms operating at different levels of abstraction (shown in Fig. 7.3) can be divided into two phases: a feature re-calibration phase followed by a feature generalization phase. In each phase, a statistical description of the intended level of generalization is extracted, which is processed later for generating the corresponding attention map. Let, $F_{in} \in \mathbb{R}^{H \times W \times C}$ be the input feature map where (H, W, C) represent the height, width, and channels of the feature map, respectively. Here, channel description, $D_c \in \mathbb{R}^{1 \times 1 \times C}$ is generated by taking the global averages of the pixels of particular channels, while the spatial description, $D_s \in \mathbb{R}^{H \times W \times 1}$ is created by convolutional filtering, and the input feature map, F_{in} represents the pixel description, $D_p \in \mathbb{R}^{H \times W \times C}$ itself.

Afterwards, the feature re-calibration phase is carried out by projecting the de-

scriptor vector D to a higher dimensional space followed by the restoration process of the original dimension to generate the re-calibration attention map A_r , which is utilized to obtain the re-calibrated feature map F_r . This process assists in the redistribution of the feature space in the subsequent feature generalization phase for better generalization of features through sharpening the effective representative features. It can be represented as:

$$F_r = F_{in} \otimes A_r = F_{in} \otimes \sigma(W_R(W_E(D))) \quad (7.1)$$

$$= F_{in} \otimes \sigma(W_R(W_E(W_D(F_{in})))) \quad (7.2)$$

where \otimes represents the element-wise multiplication with the required dimensional broadcasting operation, W_D denotes the statistical descriptor extractor, W_E represents the dimension expansion filtering, W_R represents the dimension restoration filtering, and $\sigma(\cdot)$ represents the sigmoid activation. For the channel-attention mechanism, W_E and W_R are realized by fully connected layers, while for spatial and pixel attention, convolutional filters are employed.

Subsequently, the feature generalization operation is carried out through the squeeze and excitation operation on the re-calibrated feature space, F_r to generate the effective attention map A . In this phase, the extracted feature descriptor, D' is projected into a lower-dimensional space to extract the most effective representational features and thereafter, reconstructed back to the original dimension. Such sequential dimension reduction and reconstruction operations provide an opportunity to emphasize the generalized features while reducing the redundant features. Hence, the generated attention map A provides the opportunity to reduce the effect of redundant features by providing more attention to the effective features, and it can be represented as:

$$A = \sigma(W_{R'}(W_S(D'))) = \sigma(W_{R'}(W_S(W_{D'}(F_r)))) \quad (7.3)$$

where $W_S, W_{R'}$ represents the corresponding squeeze and restoration filtering, respectively, while $W_{D'}$ represents the statistical descriptor extractor. Therefore, three levels of attention maps are generated, i.e. a channel attention map $A_C \in \mathbb{R}^{1 \times 1 \times C}$, a spatial attention map $A_S \in \mathbb{R}^{H \times W \times 1}$, and a pixel attention map $A_P \in \mathbb{R}^{H \times W \times C}$. The tri-level attention unit (TAU), represented in Fig. 7.4, generates the effective volumetric, triple attention mask A_T integrating all three maps, which is given by:

$$A_T = A_P \otimes (A_S \otimes A_C) \quad (7.4)$$

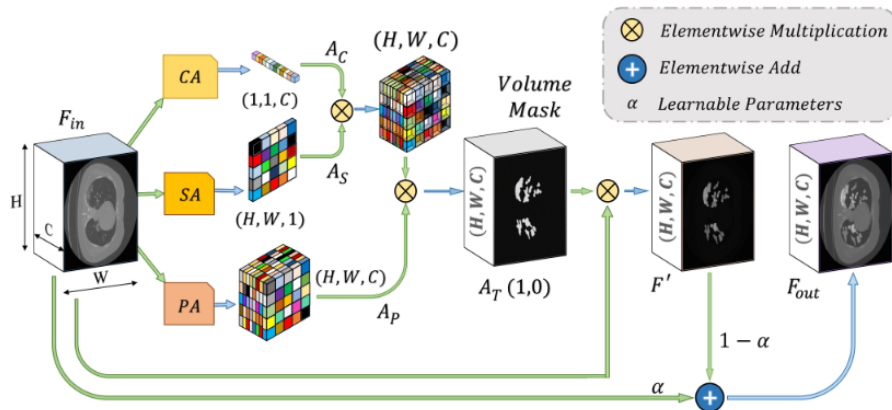


Figure 7.4: Schematic of the proposed Tri-level Attention Unit (TAU) integrating channel attention (CA), spatial attention (SA), and pixel attention (PA) mechanisms.

Later, this accumulated attention mask A_T is used to transform the input feature map F_{in} to F' for enhancing the region-of-interest, and finally the output feature map, F_{out} is generated through the weighted addition of the input and transformed feature maps, and these can be summarized as:

$$F' = F_{in} \otimes A_T \quad (7.5)$$

$$F_{out} = T(F_{in}) = \alpha F_{in} + (1 - \alpha)F' \quad (7.6)$$

where $T(\cdot)$ represents the proposed Tri-level attention mechanism, α is a learnable parameter that is optimized through the back-propagation algorithm along with other parameters.

7.1.2 Proposed Tri-level Attention-based Segmentation Network (TA-SegNet)

The proposed TA-SegNet network is deployed for segmenting the infected lesions as well as for extracting features for the following joint diagnosis and segmentation tasks (as shown in Fig. 7.1). For better segmentation, this network introduces several modifications over traditional networks which are mostly based on Fully convolutional networks (FCN) and Unet networks generally.

The proposed TA-SegNet network (shown in Fig. 7.5) integrates the advantages of both Unet and FCN by introducing an encoder-decoder based network with reduced semantic gaps along with the opportunity of parallel optimization of multi-scale features. Firstly, the input images pass through sequential encoding stages with convolutional filtering followed by sequential decoding operations similar to the

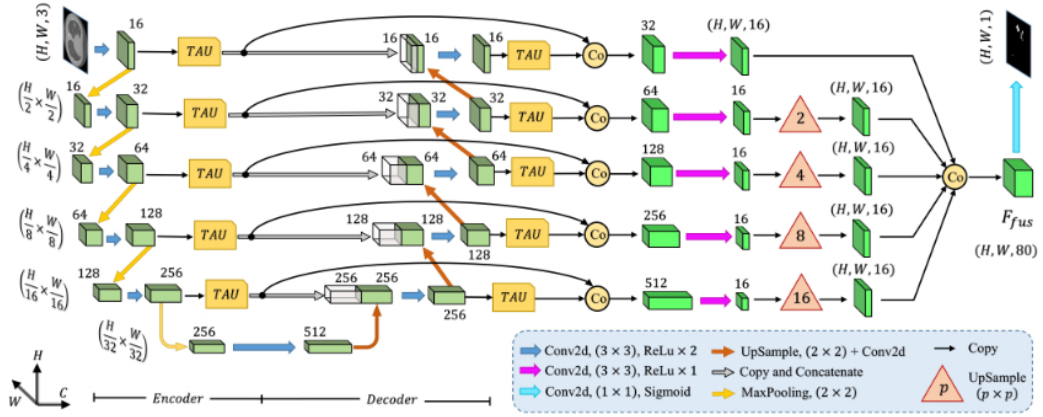


Figure 7.5: **Schematic representation of the proposed Tri-level Attention-based Segmentation Network (TA-SegNet).**

Unet. Moreover, the output feature map generated from each layer of the encoder unit is connected to the corresponding decoder layer through a Tri-level Attention Unit (TAU) mechanism for better reconstruction in the decoder unit. For further generalization and refinement of contextual features, all scales of decoded feature representations also pass through another stage of the attention mechanism. Afterwards, for introducing joint optimization of multi-scale features, the attention gated, refined feature maps generated at different stages of encoder and decoder modules are accumulated through a series of operational stages. Initially, sequential concatenation of corresponding encoder-decoder layer outputs (after attention-gating) are carried out. Following that, channel downscaling operations through convolutional filtering and bi-linear spatial upsampling operations are employed to produce feature vectors with uniform dimensions. Afterwards, these uniform feature vectors are accumulated through channel-wise concatenation to generate the fusion vector F_{fus} , and it can be represented as:

$$F_{fus} = \mathcal{F}_{i=1}^N (T(E_i) \oplus T(D_i)) \quad (7.7)$$

where \oplus represents feature concatenation, E_i, D_i stand for i_{th} level of feature representations from total N levels of the encoder, and decoder modules, respectively, $T(\cdot)$ represents the tri-level attention unit operation, and $\mathcal{F}(\cdot)$ represents the multi-scale feature fusion operation.

Afterwards, the final convolutional filtering is operated on the fusion feature map (F_{fus}) to produce the output segmentation mask. Moreover, to introduce transfer-learning in this TA-SegNet similar to other networks, the encoder module can be replaced by different pre-trained backbone networks for better optimization. Hence,

the proposed TA-SegNet facilitates faster convergence through parallel optimization of the multi-scale features while effectively extracting the region-of-interest from each scale of representation with the novel tri-level attention gating mechanism for providing the optimum performance even in the most challenging conditions.

7.1.3 Proposed Regional Feature Extractor Module

To overcome the loss of information especially for the early stage of infection, the final fusion vector F_{fus} generated at TA-SegNet is incorporated into further processing, instead of the segmented lesion, as it contains the effective feature representations of the region-of-infections. For further emphasizing the COVID lesion features, a regional feature extractor module (RF_{ex}) is also proposed that separately operates on each of the slice-wise fusion vector F_{fus} and thus generates the effective regional feature representation F_{reg} . From Fig. 7.1, it is to be noted that such regional feature extractor module separately operates on the extracted feature vectors of each CT-slice and hence, enhance the effective regional features regarding the infection. The architectural details of this module are presented in Fig. 7.6. It consists of several stages of convolutional filtering while incorporating the Tri-level Attention Unit at each stage. These attention units operated at different stages are supposed to execute different roles. As we go deeper into this RF_{ex} module, more generalized feature representations are created through subsequent pooling operations where the information is made more sparsely distributed among increased channels. Therefore, the regional feature extractor module effectively incorporates the proposed tri-level attention mechanism to extract the most generalized representative features of infections from different regions of the respective CT volume.

7.1.4 Volumetric Feature Aggregation and Fusion Module

This module accumulates the volumetric features from the generalized feature representation of each slice as well as introduces an effective fusion of features to generate the corresponding representative feature vector of the CT-volume. Moreover, this module plays an influential role in the proper selection of features especially in the early stage of infection when few of the slices contain infected lesions. To facilitate the feature selection process, the processing of severity based features and diagnostic features are isolated. In Fig. 7.1, separate volumetric feature aggregation and fusion modules are integrated to separately optimize the diagnostic and severity features. Though similar operational modules are employed in both of these cases, another stage of attention-gating operations is employed to guide the effec-

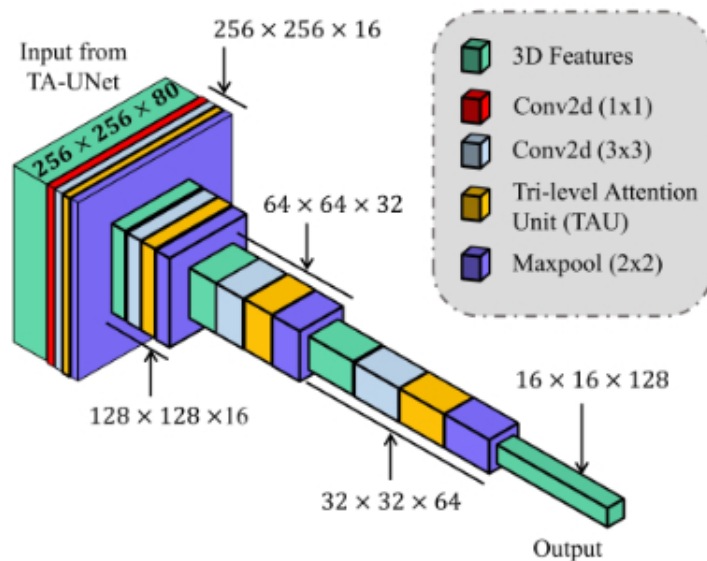


Figure 7.6: **Representation of the proposed regional feature extractor module**

tive slice-wise features in these operational modules with different objectives (shown in Fig. 7.1). This module is schematically presented in Fig. 7.7. Firstly, the volumetric feature accumulation is carried out to produce the aggregated feature vector F_{agg} from the regional features (F_{reg}) of all slices. Thereupon, the fusion scheme is employed utilizing dilated convolutions [225] which provides the opportunity to explore features from diverse receptive areas. Firstly, a pointwise convolution (1×1) is carried out for depth reduction of the aggregated vector F_{agg} . Subsequently, several dilated convolutions are operated with varying dilation rates for the effective fusion of features, and outputs of these convolutions are processed through another stage of aggregation, convolutional filtering, and global pooling operations to generate a 1D-representational feature vector. Finally, several fully connected layer operations are employed for generating the final prediction for a specific CT-volume.

7.1.5 Loss Functions

The optimization of the whole process is divided into two phases where the TA-SegNet is optimized in the first phase and joint optimization of the diagnostic and severity prediction tasks are carried out in the second phase utilizing the optimized TA-SegNet from phase-1. A focal Tversky loss function (\mathcal{L}_{FTL}) is proposed in [226] utilizing the Tversky index that performs well over a large range of applications which is used as the objective function to optimize TA-SegNet.

In general, both the COVID diagnosis and severity predictions are defined as binary-classification tasks, where normal/disease classes are considered for diagno-

sis while mild/severe classes are considered for severity predictions. For joint optimization of the diagnosis and severity prediction, an objective loss function (L_{obj}) is defined by combining the objective loss functions for diagnosis (L_d) and severity prediction (L_s). The severity prediction task will only be initiated for the infected volumes where $y_d = 1$, while for the normal cases ($y_d = 0$), this task is ignored. However, the diagnosis task is carried out for all normal/infectious volumes. Hence, the objective loss function (L_{obj}) can be expressed as:

$$\begin{aligned} L_{obj} &= L_d(\mathbf{Y}_d, \mathbf{Y}_d^p) + L_s(\mathbf{Y}_d, \mathbf{Y}_s, \mathbf{Y}_s^p) \\ &= \frac{1}{M} \sum_{i=1}^M \mathcal{L}_B(y_{d,i}, y_{d,i}^p) + \frac{1}{M_I} \sum_{i=1}^{M_I} y_{d,i} \mathcal{L}_B(y_{s,i}, y_{s,i}^p) \end{aligned} \quad (7.8)$$

where \mathbf{Y}_d and \mathbf{Y}_s represent the set of diagnosis and severity ground truths while $\mathbf{Y}_d^p, \mathbf{Y}_s^p$ represent the corresponding set of predictions, \mathcal{L}_B denotes binary cross-entropy loss, M denotes the total number of CT-volumes, and M_I represents the total number of infected volumes. Hence, the proposed CovTANet network can be effectively optimized for joint segmentation, diagnosis, and severity predictions of COVID-19 utilizing this two phase optimization scheme.

7.2 Results and Discussions

In this section, results obtained from extensive experimentation on a publicly available dataset are presented and discussed from diverse perspectives to validate the effectiveness of the proposed scheme.

7.2.1 Dataset Description

This study is conducted using ‘‘MosMedData: Chest CT Scans with COVID-19 Related Findings’’ [227], one of the largest publicly available datasets in this domain. The dataset, being collected from the hospitals in Moscow, Russia, contains 1110 anonymized CT-volumes with severity annotated COVID-19 related findings, as well as without such findings. Each one of the 1110 CT-volumes is acquired from different persons and 30-46 slices per patient are available. Pixel annotations of the COVID lesions are provided for 50 CT-volumes which are used for training and evaluation of the proposed TA-SegNet. For carrying out the diagnosis and severity prediction tasks, all the CT-volumes are divided into normal, mild (<25% lung parenchyma) and severe (>25% lung parenchyma) lung infection categories.

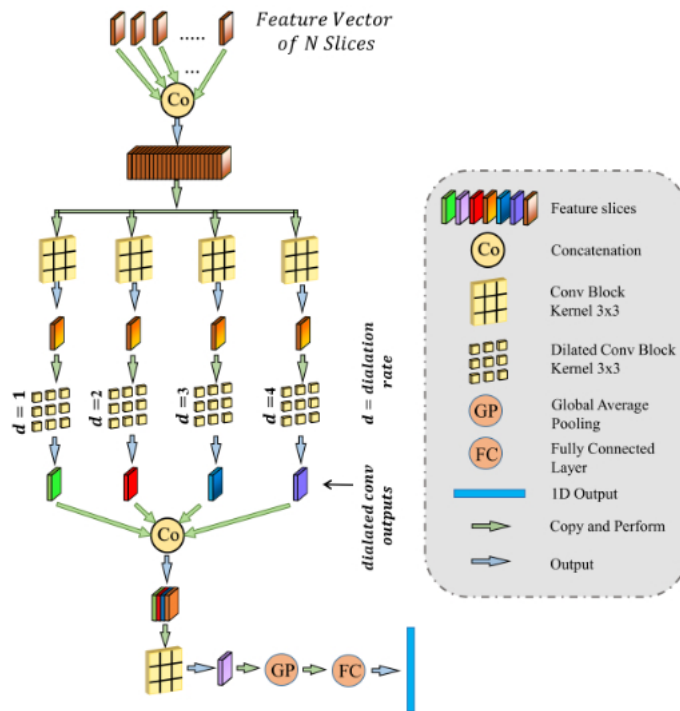


Figure 7.7: Proposed volumetric feature accumulation and fusion scheme used for severity and diagnostic feature extraction

7.2.2 Experimental Setup

With a five-fold cross-validation scheme over the MosMed dataset, all the experimentations have been implemented on the google cloud platform with NVIDIA P-100 GPU as the hardware accelerator. For evaluation of the segmentation performance, some of the traditional metrics are used, such as accuracy, precision, dice score, and intersection-over-union (IoU) score, while for assessing the severity classification performance, accuracy, sensitivity, specificity, and F1-score are used.

7.2.3 Performance Evaluation

Similar to chapter 6, an ablation study is carried out to analyze the effect of different building blocks in the TA-SegNet for segmentation objective. Afterwards, the performance of the best performing variant is compared with other networks from qualitative and quantitative perspectives.

Traditional Unet network has been used as a baseline model (V1) and five other schemes/modules have been incorporated in the baseline model to analyze the contribution of different modules in the performance improvement of the proposed TA-SegNet (V8). For ease of comparison, only Dice score is used as it is the most widely used metric for segmentation tasks. From Table 7.1, it can be noted that the encoder

TAUs (V4) provide 4.1% improvement of the Dice score from the baseline, while the decoder TAUs (V5) provide a 2.9% improvement and when both these are combined (V6), 6.6% improvement is achieved. As the encoder TAUs contribute significantly to the reduction of semantic gaps with the corresponding decoder feature maps, while the decoder TAU units guide the decoded feature maps with finer details for better generalization of multi-scale features, considerable performance improvement is achieved when employed in combination. Moreover, all the multi-scale feature maps generated from various encoder levels are guided to the reconstruction process through a deep fusion scheme along with the multi-scale decoded feature maps. The integration of these multi-scale features from the encoder-decoder modules in the fusion process (V3) contributes to the efficient reconstruction, and 4.4% improvement of Dice score is achieved over the baseline. Moreover, 9.7% improvement of Dice score is achieved when the fusion scheme is combined with two-stage TAU-units (V7). Additionally, for introducing transfer learning, pre-trained models on the ImageNet database can be used as the backbone of the encoder module of the TA-SegNet similar to most other segmentation networks. It should be noted that with the pre-trained EfficientNet network as the backbone of the encoder module (V8), the performance gets improved by 2.1% compared to the TA-SegNet framework without such backbone (V7).

In Table 7.2, performances of some of the state-of-the-art networks are summarized. It should be noticed that the proposed TA-SegNet outperforms all the methods compared by a considerable margin in all the metrics. Using the proposed framework, 11.8% improvement of Dice score over Unet, and 26.7% improvement of Dice score over the FCN have been achieved. Furthermore, our network improves the dice score of the second-best method (Inf-Net) by about 10.5%, which intuitively indicates its excellent capabilities over the rest of the models. The robustness of the proposed scheme and the enhanced capability of our model in terms of infected region identification is further demonstrated by the high sensitivity score (99.6%) reported. This signifies the fact that the model integrates the symmetric encoding-decoding strategy of Unet as well as exploits the parallel optimization advantages of FCN that provides this large improvement. Most other state-of-the-art variants of the Unet provide sub-optimal performances for increasing complexity considerably that makes the optimization difficult in most of the challenging cases. However, due to the smaller amount of infections in the annotated CT-volumes used for training and optimization of the segmentation networks, a higher amount of false positives have been generated in most of the networks which reduced the precision. The proposed TA-SegNet has considerably reduced the false positives along with false

Table 7.1: Performance (Mean \pm Standard Deviation) of the Ablation Study of the Proposed TA-SegNet on MosMedData

Version	EfficientNet Backbone	Encoder TAU Unit	Decoder TAU Unit	Encoder in Fusion	Decoder in Fusion	Dice(%)
V1	✗	✗	✗	✗	✗	50.5 \pm 0.26
V2	✗	✗	✗	✗	✓	52.4 \pm 0.17
V3	✗	✓	✗	✗	✗	54.9 \pm 0.14
V4	✗	✓	✗	✗	✗	54.6 \pm 0.14
V5	✗	✗	✓	✗	✗	53.4 \pm 0.19
V6	✗	✓	✓	✗	✗	57.1 \pm 0.33
V7	✗	✓	✓	✓	✓	60.2 \pm 0.26
V8	✓	✓	✓	✓	✓	62.3 \pm 0.18

Table 7.2: Comparison of Performances with Other the State-of-the-Art Networks on COVID Lesion Segmentation on MosMedData

Networks	Sensitivity(%)	Precision(%)	Dice(%)	IoU(%)
FCN [228]	78.8 \pm 0.23	58.9 \pm 0.16	35.6 \pm 0.36	29.3 \pm 0.45
Unet [229]	94.3 \pm 0.34	74.4 \pm 0.32	50.5 \pm 0.26	40.3 \pm 0.23
Vnet [230]	84.5 \pm 0.42	64.6 \pm 0.54	40.2 \pm 0.33	36.4 \pm 0.26
Unet++ [231]	78.1 \pm 0.15	65.1 \pm 0.25	37.2 \pm 0.27	33.3 \pm 0.32
CPF-Net [232]	82.4 \pm 0.25	71.3 \pm 0.29	48.9 \pm 0.21	37.6 \pm 0.38
COPE-Net [140]	85.5 \pm 0.18	73.1 \pm 0.20	51.1 \pm 0.21	41.2 \pm 0.38
Mini-SegNet [139]	81.5 \pm 0.25	69.1 \pm 0.19	43.7 \pm 0.23	35.2 \pm 0.38
Inf-Net [138]	92.8 \pm 0.27	76.9 \pm 0.34	51.8 \pm 0.31	41.6 \pm 0.27
TA-SegNet (Prop.)	99.6\pm0.09	84.8\pm0.26	62.3\pm0.18	51.7\pm0.29

negatives and has improved both precision and sensitivity.

In Fig. 7.8, qualitative representations of the segmentation performances of different networks are shown in some challenging conditions. The comparable dimensions of the small infected regions and the arteries, veins embedded in the thorax cavity with varying anatomical appearance might be attributed to the large occurrences of the false positives. It is evident that most other networks struggle to extract the complicated, scattered, and diffused COVID-19 lesions, while the proposed TA-SegNet considerably improves the segmentation performance in such challenging conditions. This depiction conforms to the fact that our network can correctly segment both of the large and small infected regions. Furthermore, our framework consistently demonstrates almost non-existent false negatives compared to the other models while considerably reducing the false positive predictions as it can distinguish sharper details of the lesions and effectively perform for precise diagnosis of the infection.

In Table 7.3, the performances obtained from the joint diagnosis and severity prediction tasks are summarized. To analyze the effectiveness of the proposed multi-phase optimization scheme, some of the state-of-the-art networks are also evaluated for the slice-wise processing of the CT-volumes in the joint-classification scheme discarding the TA-SegNet.

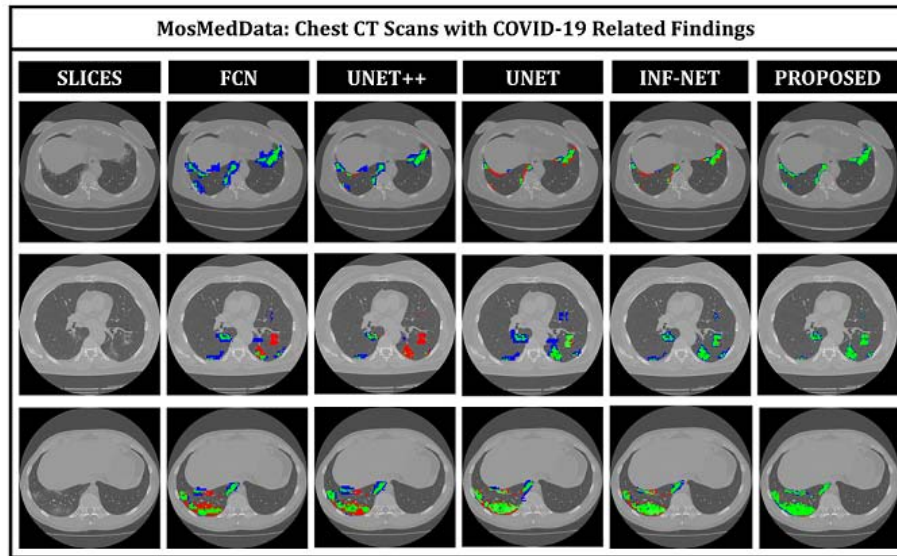


Figure 7.8: Visualization of the lesion segmentation performance of some of the state-of-the-art networks in MosMedData [227]. Here, ‘green’ denotes the true positive (TP) region, ‘blue’ denotes the false positive region, and ‘red’ denotes the false negative regions.

The diagnosis performances with mild and severe cases of COVID-19 are separately reported to distinguish the severity prediction performance. The proposed CovTANet provides 85.2% accuracy in isolating the COVID patients even with mild symptoms, while the accuracy is as high as 95.8% when the CT volumes contain severe infections. However, the other networks operating without the TA-SegNet noticeably suffer especially in the mild lung infection phase, as it is difficult to isolate the small infection patches from the CT-volume.

In the joint optimization process based on the amount of infected lung parenchyma, mild and severe patients are also categorized. Despite the additional challenges regarding the isolation and quantification of the abnormal tissues, the proposed scheme generalizes the problem quite well which provides 91.7% accuracy in categorizing mild and severe patients. It should be noted that the highest achievable severity prediction accuracy with a traditional network is 64.8% (using ResNet50) with considerably smaller results in most other metrics. Traditional network directly operates on the whole CT-volume to extract effective features for severity prediction which makes the task more complicated. Whereas, the proposed hybrid CovTANet with multiphase optimization effectively integrates features regarding infections from the TA-SegNet for considerably simplifying the feature extraction process in the joint-classification process that results in higher accuracy.

7.3 Conclusion

In this chapter, it is shown that the proposed joint classification scheme not only provides better diagnosis at severe infection stages but also is capable of categorizing mild and severe lung infections with outstanding precision. It can be interpreted that this high early diagnostic accuracy of CovTANet is significantly contributed by the multi-phase optimization scheme that incorporates the highly optimized TA-SegNet for extracting the most effective lesion features to mitigate the effect of redundant healthy parts. Furthermore, considerable performances have been achieved in severity screening that would facilitate a faster clinical response to substantially reduce the probable damages. The proposed scheme can be a valuable tool for the clinicians to combat this pernicious disease through faster-automated mass-screening.

Table 7.3: Comparison of Performances in the Joint Diagnosis and Severity Prediction of COVID-19 with Different Networks on MosMedData

Network	Diagnostic Prediction												Severity Prediction		
	Normal vs. Mild Lung Infection				Normal vs. Severe Lung Infection				Mild vs. Severe Lung Infection				Spec.(%)	Acc.(%)	F1(%)
	Sen.(%)	Spec.(%)	Acc.(%)	F1(%)	Sen.(%)	Spec.(%)	Acc.(%)	F1(%)	Sen.(%)	Spec.(%)	Acc.(%)				
VGG-19	54.4±0.37	63.4±0.28	58.4±0.32	58.6±0.33	63.4±0.32	70.8±0.37	65.9±0.28	66.9±0.39	62.7±0.25	65.5±0.23	61.9±0.25	64.1±0.28			
ResNet-50	61.1±0.29	65.7±0.33	62.5±0.44	63.3±0.37	66.5±0.26	69.3±0.31	69.1±0.17	67.8±0.33	61.1±0.36	63.8±0.29	64.8±0.34	62.4±0.37			
Xception	56.8±0.38	57.9±0.41	59.9±0.27	57.3±0.43	64.8±0.19	67.2±0.26	66.7±0.21	65.9±0.25	62.9±0.33	65.1±0.32	63.1±0.22	63.9±0.34			
DenseNet121	59.7±0.27	64.6±0.21	61.1±0.39	62.1±0.28	65.1±0.23	70.1±0.19	67.8±0.29	67.5±0.25	60.2±0.28	64.4±0.21	60.6±0.26	62.2±0.27			
InceptionV3	60.4±0.31	62.1±0.38	59.3±0.23	61.2±0.35	66.6±0.28	69.8±0.24	66.2±0.32	68.2±0.31	62.8±0.29	67.9±0.35	61.4±0.19	65.3±0.33			
CovTANet (ours)	83.8±0.25	90.3±0.27	85.2±0.19	86.9±0.22	93.9±0.13	96.6±0.11	95.8±0.17	94.2±0.21	90.9±0.16	93.4±0.23	91.7±0.19	92.1±0.11			

Chapter 8

Conclusion

In this chapter, we will conclude the discussions on efficient deep neural network architectures for multi-dimensional data processing. A brief summary of the thesis is presented where the key contributions of this thesis are highlighted. Afterwards, some of the future research directions have been presented for improving the performance of the multi-receptive feature extraction and optimization process.

8.1 Summary

Despite the widespread applicability of the deep neural network in diverse applications, limitations of the available data with enormous computational burden open up the opportunity for further improvements of the feature extraction process. To make the process more resilient for extracting the optimum feature representation, multi-receptive feature spaces are intensely explored through various data representation approaches and architectural modifications of the deep neural network. A multi-receptive neural network (DeepArrNet) is introduced for exploring diverse kernel windows of time-series data with pointwise-depthwise-pointwise (PTP) convolutional building blocks. This architecture is optimized for arrhythmia diagnosis from the time series ECG data along with numerous data processing strategies including denoising, beat segmentation, and various time-series data augmentations for handling class-imbalance.

Afterwards, a deep neural network architecture namely CovXNet is proposed to efficiently detect COVID-19 and other types of pneumonia with distinctive localization from chest X-rays. Instead of using traditional convolution, efficient depthwise convolution is used with varying dilation rates that integrates features from diversified receptive fields to analyze the abnormalities in X-rays from different perspectives. To utilize the small number of COVID-19 X-rays, a larger database is

utilized containing X-rays from normal and other traditional pneumonia patients for initially training the deep network. Moreover, the proposed CovXNet is highly scalable with enormous receptive capacity that can also be employed in varieties of other computer vision applications.

Furthermore, various types of human activities are recognized utilizing the proposed multi-stage training method. Firstly, the raw data undergo through numerous transformations to interpret the information encoded in raw data in different spaces and thus to obtain a diversified representation of the features. Afterwards, separate deep CNN architectures are trained on each space to be an optimized feature extractor from that particular space for the final prediction of activity. Later, these tuned feature extractors are merged into a final form of deep network effectively through a combined training stage or through sequential stages of training by exploring the extracted feature spaces exhaustively to attain the most robust and accurate feature representation. Therefore, the proposed scheme opens up a new approach of employing multiple training stages for deep CNNs deploying various transformed representations of data which can also be utilized in very diversified applications by increasing the diversity of the extracted features.

Next, an improved architecture is proposed, namely PolypSegNet, for proper segmentation of the polyp regions from colonoscopy images. To reduce some major architectural limitations of the traditional Unet architecture, three major building blocks are incorporated in the baseline Unet architecture, i.e. DDI module based D-Unit layer, DFSM block, and DRM block. For efficient feature extraction from diverse receptive fields, a DDI module is introduced and repeated use of DDI modules in the D-Unit layer of encoder/decoder improved the performance over the traditional baseline Unet model. Higher improvement is achieved using the proposed DFSM block as it reduces the semantic gap between encoder and decoder utilizing a deep fusion of multi-scale features generated at various encoder levels. Moreover, the proposed DRM block also provides comparable performance improvement for its efficient reconstruction through the incorporation of multiscale decoded feature maps in the final reconstruction. Though the proposed PolypSegNet is extensively studied for polyp segmentation in this work, it can be easily extended for any medical image segmentation related applications that can be a better alternative to other traditional networks.

Additionally, a multi encoder-decoder based architecture (CovSegNet) is introduced with numerous architectural renovations that assist in achieving state-of-the-art performance on COVID lesion segmentation. The horizontal and vertical expansion mechanisms provide the opportunity to incorporate more detailed features

as well as more generalized features, which improved the feature quality considerably that is particularly effective in distinguishing multi-class, scattered COVID lesions with widely varied shapes. Moreover, the improved gradient flow throughout the network, achieved with the introduction of multi-scale fusion module and scale transition modules, have greatly reduced the contextual information loss in the generalization process and have also ensured the best optimization of all network parameters that particularly contribute to recover and distinguish the blurry, diffused edges of COVID lesions as well as the very minute instances of abnormalities. Furthermore, the integration of a hybrid 2D-3D networking scheme exploits both the intra-slice and inter-slice contextual information without increasing computational burden that results in more precise, finer segmentation performance mostly in challenging conditions.

Finally, a multi-phase optimization scheme is proposed with a hybrid neural network (CovTANet) where an efficient lesion segmentation network is integrated into a complete optimization framework for joint diagnosis and severity prediction of COVID-19 from CT-volume. The tri-level attention mechanism and parallel optimization of multi-scale encoded-decoded feature maps which are introduced in the segmentation network (TA-SegNet) have improved the lesion segmentation performance substantially. Moreover, the effective integration of features from the optimized TA-SegNet is found to be extremely beneficial in diagnosis and severity prediction by de-emphasizing the effects of redundant features from the whole CT-volumes.

8.2 Limitations and Future Works

Despite significant performance is achieved using the proposed schemes presented in this thesis, some other perspectives can be explored in future studies. Firstly, similar to the most other established studies in arrhythmia classification [41], [44], [50], [51], five arrhythmia classes are considered for experimentation and very satisfactory performance is achieved. However, the proposed scheme with DeepArrNet can be extended considering more number of arrhythmia classes as considered in [54], [57]. Secondly, to incorporate effective features from new representational space with the sequential learning algorithm, separate CNN-based feature extractors need to be incorporated which will increase the total size of the network accordingly. Nevertheless, the proposed training scheme separately optimizes individual deep feature extractors and integrates the extracted feature spaces in a sequential manner providing a significant advantage over the traditional training approaches. Thirdly, the

CovXNet model can be made more accurate and robust through the incorporation of more data. However, the proposed scheme is highly adaptive and the CovXNet can be more finely tuned in the transfer learning phase with additional COVID-19 X-rays. Hence, further research should be carried out with more diversified data for a thorough investigation of the clinical features of COVID-19. Lastly, although consistent performances have been achieved on COVID-19 lesion segmentation, the proposed segmentation approaches will be extended with the incorporation of diversified datasets including patient-based study considering age, sex, health conditions, and geographical locations of the patients. An in-depth, closer, patient-specific study should be carried out for better understandings of the nature of the infection. Moreover, to understand the mutation and evolution of this deadly virus, the proposed hybrid multi-task learning should be extended considering patients from diverse geographic locations.

Bibliography

- [1] M. W. Libbrecht and W. S. Noble, “Machine learning applications in genetics and genomics,” *Nature Reviews Genetics*, vol. 16, no. 6, pp. 321–332, 2015.
- [2] Q. Yao, M. Wang, Y. Chen, W. Dai, Y.-F. Li, W.-W. Tu, Q. Yang, and Y. Yu, “Taking human out of learning applications: A survey on automated machine learning,” *arXiv preprint arXiv:1810.13306*, 2018.
- [3] J. Wiens, S. Saria, M. Sendak, M. Ghassemi, V. X. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, M. Saeed, *et al.*, “Do no harm: A roadmap for responsible machine learning for health care,” *Nature medicine*, vol. 25, no. 9, pp. 1337–1340, 2019.
- [4] L. Zhang, J. Tan, D. Han, and H. Zhu, “From machine learning to deep learning: Progress in machine intelligence for rational drug discovery,” *Drug discovery today*, vol. 22, no. 11, pp. 1680–1685, 2017.
- [5] R. C. Guido, “A tutorial review on entropy-based handcrafted feature extraction for information fusion,” *Information Fusion*, vol. 41, pp. 161–175, 2018.
- [6] H. Wang, J. Hu, and W. Deng, “Face feature extraction: A complete review,” *IEEE Access*, vol. 6, pp. 6001–6039, 2017.
- [7] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, “Object detection with deep learning: A review,” *IEEE transactions on neural networks and learning systems*, vol. 30, no. 11, pp. 3212–3232, 2019.
- [8] O. Faust, Y. Hagiwara, T. J. Hong, O. S. Lih, and U. R. Acharya, “Deep learning for healthcare applications based on physiological signals: A review,” *Computer methods and programs in biomedicine*, vol. 161, pp. 1–13, 2018.
- [9] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, “Deep learning for healthcare: Review, opportunities and challenges,” *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

- [10] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.
- [11] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, “A survey on deep transfer learning,” in *International conference on artificial neural networks*, Springer, 2018, pp. 270–279.
- [12] C. Dai, X. Liu, and J. Lai, “Human action recognition using two-stream attention based lstm networks,” *Applied soft computing*, vol. 86, p. 105 820, 2020.
- [13] B. Bhandari, G. Lee, and J. Cho, “Body-part-aware and multitask-aware single-image-based action recognition,” *Applied Sciences*, vol. 10, no. 4, p. 1531, 2020.
- [14] G. Lorre, J. Rabarisoa, A. Orcesi, S. Ainouz, and S. Canu, “Temporal contrastive pretraining for video action recognition,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 662–670.
- [15] J. Gao, P. Li, Z. Chen, and J. Zhang, “A survey on deep learning for multi-modal data fusion,” *Neural Computation*, vol. 32, no. 5, pp. 829–864, 2020.
- [16] Z. He, Y. Cao, L. Du, B. Xu, J. Yang, Y. Cao, S. Tang, and Y. Zhuang, “Mrfn: Multi-receptive-field network for fast and accurate single image super-resolution,” *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 1042–1054, 2019.
- [17] M. Zhong, B. Verma, and J. Affum, “Multi-receptive atrous convolutional network for semantic segmentation,” in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020, pp. 1–8.
- [18] A. Bagnall, J. Lines, J. Hills, and A. Bostrom, “Time-series classification with cote: The collective of transformation-based ensembles,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 9, pp. 2522–2535, 2015.
- [19] T. Mahmud, A. S. Sayyed, S. A. Fattah, and S.-Y. Kung, “A novel multi-stage training approach for human activity recognition from multimodal wearable sensor data using deep neural network,” *IEEE Sensors Journal*, vol. 21, no. 2, pp. 1715–1726, 2020.
- [20] T. Mahmud, M. A. Rahman, and S. A. Fattah, “Covxnet: A multi-dilation convolutional neural network for automatic covid-19 and other pneumonia detection from chest x-ray images with transferable multi-receptive feature optimization,” *Computers in biology and medicine*, vol. 122, p. 103 869, 2020.

- [21] T. Mahmud, B. Paul, and S. A. Fattah, “Polypsegnet: A modified encoder-decoder architecture for automated polyp segmentation from colonoscopy images,” *Computers in Biology and Medicine*, vol. 128, p. 104119, 2021.
- [22] T. Mahmud, M. J. Alam, S. Chowdhury, S. N. Ali, M. M. Rahman, S. A. Fattah, and M. Saquib, “Covtanet: A hybrid tri-level attention based network for lesion segmentation, diagnosis, and severity prediction of covid-19 chest ct scans,” *IEEE Transactions on Industrial Informatics*, 2020.
- [23] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [24] L. Mou, L. Chen, J. Cheng, Z. Gu, Y. Zhao, and J. Liu, “Dense dilated network with probability regularized walk for vessel detection,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 5, pp. 1392–1403, 2019.
- [25] E. J. da S Luz, W. R. Schwartz, G. Cámara-Chávez, and D. Menotti, “ECG-based heartbeat classification for arrhythmia detection: A survey,” *Computer Methods and Programs in Biomedicine*, vol. 127, pp. 144–164, 2016.
- [26] P. Wang, X. Xiao, J. R. G. Brown, T. M. Berzin, M. Tu, F. Xiong, X. Hu, P. Liu, Y. Song, D. Zhang, *et al.*, “Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy,” *Nature biomedical engineering*, vol. 2, no. 10, pp. 741–748, 2018.
- [27] T. Mahmud, S. A. Fattah, and M. Saquib, “Deeparnet: An efficient deep cnn architecture for automatic arrhythmia detection and classification from denoised ecg beats,” *IEEE Access*, vol. 8, pp. 104788–104800, 2020.
- [28] A. A. Ewees, M. Abd Elaziz, and D. Oliva, “A new multi-objective optimization algorithm combined with opposition-based learning,” *Expert Systems with Applications*, vol. 165, p. 113844, 2021.
- [29] R. Salles, K. Belloze, F. Porto, P. H. Gonzalez, and E. Ogasawara, “Non-stationary time series transformation methods: An experimental review,” *Knowledge-Based Systems*, vol. 164, pp. 274–291, 2019.
- [30] T. Mahmud, M. A. Rahman, S. A. A. Fattah, and S.-Y. Kung, “Covsegnet: A multi encoder-decoder architecture for improved lesion segmentation of covid-19 chest ct scans,” *IEEE Transactions on Artificial Intelligence*, 2021.
- [31] D. Lin, Y. Ji, D. Lischinski, D. Cohen-Or, and H. Huang, “Multi-scale context intertwining for semantic segmentation,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 603–619.

- [32] T. D. Bui, J. Shin, and T. Moon, “3d densely convolutional networks for volumetric segmentation,” *arXiv preprint arXiv:1709.03199*, 2017.
- [33] F. Wang, Y. Li, F. Liao, and H. Yan, “An ensemble learning based prediction strategy for dynamic multi-objective optimization,” *Applied Soft Computing*, vol. 96, p. 106592, 2020.
- [34] O. Sener and V. Koltun, “Multi-task learning as multi-objective optimization,” *arXiv preprint arXiv:1810.04650*, 2018.
- [35] P. Sen and D. Ganguly, “Towards socially responsible ai: Cognitive bias-aware multi-objective learning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 2685–2692.
- [36] L. Peng, M. Peng, B. Liao, G. Huang, W. Li, and D. Xie, “The advances and challenges of deep learning application in biological big data processing,” *Current Bioinformatics*, vol. 13, no. 4, pp. 352–359, 2018.
- [37] P. Chu and H. Ling, “Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6172–6181.
- [38] J. Wang, B. Cao, P. Yu, L. Sun, W. Bao, and X. Zhu, “Deep learning towards mobile applications,” in *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*, IEEE, 2018, pp. 1385–1393.
- [39] F. Melgani and Y. Bazi, “Classification of electrocardiogram signals with support vector machines and particle swarm optimization,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 5, pp. 667–677, 2008.
- [40] J. Park, K. Lee, and K. Kang, “Arrhythmia detection from heartbeat using k-nearest neighbor classifier,” in *2013 IEEE International Conference on Bioinformatics and Biomedicine*, IEEE, 2013, pp. 15–22.
- [41] T. Li and M. Zhou, “ECG classification using wavelet packet entropy and random forests,” *Entropy*, vol. 18, no. 8, 2016. [Online]. Available: <https://doi.org/10.3390/e18080285>.
- [42] S.-N. Yu and Y.-H. Chen, “Electrocardiogram beat classification based on wavelet transformation and probabilistic neural network,” *Pattern Recognition Letters*, vol. 28, no. 10, pp. 1142–1150, 2007.
- [43] İ. Güler and E. D. Übeyli, “ECG beat classifier designed by combined neural network model,” *Pattern Recognition*, vol. 38, no. 2, pp. 199–208, 2005.

- [44] R. J. Martis, U. R. Acharya, and L. C. Min, "ECG beat classification using PCA, LDA, ICA and discrete wavelet transform," *Biomedical Signal Processing and Control*, vol. 8, no. 5, pp. 437–448, 2013.
- [45] S. Banerjee and M. Mitra, "Application of cross wavelet transform for ECG pattern analysis and classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 2, pp. 326–333, 2013.
- [46] T. H. Linh, S. Osowski, and M. Stodolski, "On-line heart beat recognition using hermite polynomials and neuro-fuzzy network," *IEEE Transactions on Instrumentation and Measurement*, vol. 52, no. 4, pp. 1224–1231, 2003.
- [47] R. J. Martis, U. R. Acharya, C. M. Lim, and J. S. Suri, "Characterization of ECG beats from cardiac arrhythmia using discrete cosine transform in PCA framework," *Knowledge-Based Systems*, vol. 45, pp. 76–82, 2013.
- [48] S.-N. Yu and K.-T. Chou, "Integration of independent component analysis and neural networks for ECG beat classification," *Expert Systems with Applications*, vol. 34, no. 4, pp. 2841–2846, 2008.
- [49] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Medicine*, vol. 25, no. 1, pp. 65–69, 2019.
- [50] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific ECG classification by 1-d convolutional neural networks," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 3, pp. 664–675, 2015.
- [51] N. Strodthoff and C. Strodthoff, "Detecting and interpreting myocardial infarction using fully convolutional neural networks," *arXiv:1806.07385*, 2018. [Online]. Available: <https://arxiv.org/abs/1806.07385>.
- [52] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, M. Adam, A. Gertych, and R. San Tan, "A deep convolutional neural network model to classify heartbeats," *Computers in Biology and Medicine*, vol. 89, pp. 389–396, 2017.
- [53] T. Mahmud, A. R. Hossain, and S. A. Fattah, "ECGDeepNET: A deep learning approach for classifying ECG beats," in *Proc. 7th Int. Conf. Robot Intelligence Technology and Applications (RiTA)*, IEEE, 2019, pp. 32–37.
- [54] E. Ihsanto, K. Ramli, D. Sudiana, and T. S. Gunawan, "An efficient algorithm for cardiac arrhythmia classification using ensemble of depthwise separable convolutional neural networks," *Applied Sciences*, vol. 10, no. 2, p. 483, 2020. [Online]. Available: <https://doi.org/10.3390/app10020483>.

- [55] M. Salem, S. Taheri, and J.-S. Yuan, “ECG arrhythmia classification using transfer learning from 2-dimensional deep cnn features,” in *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, IEEE, 2018, pp. 1–4.
- [56] M. Kachuee, S. Fazeli, and M. Sarrafzadeh, “ECG heartbeat classification: A deep transferable representation,” in *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, IEEE, 2018, pp. 443–444.
- [57] T. J. Jun, H. M. Nguyen, D. Kang, D. Kim, D. Kim, and Y.-H. Kim, “ECG arrhythmia classification using a 2-d convolutional neural network,” *arXiv:1804.06812*, 2018. [Online]. Available: <http://arxiv.org/abs/1804.06812>.
- [58] S. Mousavi and F. Afghah, “Inter-and intra-patient ECG heartbeat classification for arrhythmia detection: A sequence to sequence deep learning approach,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 1308–1312.
- [59] E. D. Übeyli, “Combining recurrent neural networks with eigenvector methods for classification of ECG beats,” *Digital Signal Processing*, vol. 19, no. 2, pp. 320–329, 2009.
- [60] Ö. Yildirim, “A novel wavelet sequence based on deep bidirectional lstm network model for ECG signal classification,” *Computers in Biology and Medicine*, vol. 96, pp. 189–202, 2018.
- [61] R.-A. Voicu, C. Dobre, L. Bajenaru, and R.-I. Ciobanu, “Human physical activity recognition using smartphone sensors,” *Sensors*, vol. 19, no. 3, p. 458, 2019.
- [62] A. Jain and V. Kanhangad, “Human activity classification in smartphones using accelerometer and gyroscope sensors,” *IEEE Sensors Journal*, vol. 18, no. 3, pp. 1169–1177, 2017.
- [63] R. C. Kumar, S. S. Bharadwaj, B. Sumukha, and K. George, “Human activity recognition in cognitive environments using sequential elm,” in *2016 Second International Conference on Cognitive Computing and Information Processing (CCIP)*, IEEE, 2016, pp. 1–6.
- [64] P. Zappi, C. Lombriser, T. Stiefmeier, E. Farella, D. Roggen, L. Benini, and G. Tröster, “Activity recognition from on-body sensors: Accuracy-power trade-off by dynamic sensor selection,” in *European Conference on Wireless Sensor Networks*, Springer, 2008, pp. 17–33.

- [65] S. Seto, W. Zhang, and Y. Zhou, “Multivariate time series classification using dynamic time warping template selection for human activity recognition,” in *2015 IEEE Symposium Series on Computational Intelligence*, IEEE, 2015, pp. 1399–1406.
- [66] P. Vaka, F. Shen, M. Chandrashekar, and Y. Lee, “Pemar: A pervasive middleware for activity recognition with smart phones,” in *2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, IEEE, 2015, pp. 409–414.
- [67] M. A. Alsheikh, A. Selim, D. Niyato, L. Doyle, S. Lin, and H.-P. Tan, “Deep activity recognition models with triaxial accelerometers,” in *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
- [68] Y. Zheng, “Human activity recognition based on the hierarchical feature selection and classification framework,” *Journal of Electrical and Computer Engineering*, vol. 2015, 2015.
- [69] W. Jiang and Z. Yin, “Human activity recognition using wearable sensors by deep convolutional neural networks,” in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1307–1310.
- [70] V. Bianchi, M. Bassoli, G. Lombardo, P. Fornacciari, M. Mordonini, and I. De Munari, “Iot wearable sensor and deep learning: An integrated approach for personalized human activity recognition in a smart home environment,” *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8553–8562, 2019.
- [71] B. Zhou, J. Yang, and Q. Li, “Smartphone-based activity recognition for indoor localization using a convolutional neural network,” *Sensors*, vol. 19, no. 3, p. 621, 2019.
- [72] D. Ravi, C. Wong, B. Lo, and G.-Z. Yang, “Deep learning for human activity recognition: A resource efficient implementation on low-power devices,” in *2016 IEEE 13th international conference on wearable and implantable body sensor networks (BSN)*, IEEE, 2016, pp. 71–76.
- [73] F. J. Ordóñez and D. Roggen, “Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition,” *Sensors*, vol. 16, no. 1, p. 115, 2016.
- [74] Y. Chen and Y. Xue, “A deep learning approach to human activity recognition based on single accelerometer,” in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, IEEE, 2015, pp. 1488–1492.

- [75] A. Murad and J.-Y. Pyun, “Deep recurrent neural networks for human activity recognition,” *Sensors*, vol. 17, no. 11, p. 2556, 2017.
- [76] A. Gumaiei, M. M. Hassan, A. Alelaiwi, and H. Alsalman, “A hybrid deep learning model for human activity recognition using multimodal body sensing data,” *IEEE Access*, vol. 7, pp. 99 152–99 160, 2019.
- [77] S. Chung, J. Lim, K. J. Noh, G. Kim, and H. Jeong, “Sensor data acquisition and multimodal sensor fusion for human activity recognition using deep learning,” *Sensors*, vol. 19, no. 7, p. 1716, 2019.
- [78] M. Lv, W. Xu, and T. Chen, “A hybrid deep convolutional and recurrent neural network for complex activity recognition using multimodal sensors,” *Neurocomputing*, vol. 362, pp. 33–40, 2019.
- [79] H. Yu, G. Pan, M. Pan, C. Li, W. Jia, L. Zhang, and M. Sun, “A hierarchical deep fusion framework for egocentric activity recognition using a wearable hybrid sensor system,” *Sensors*, vol. 19, no. 3, p. 546, 2019.
- [80] Z. Wang and T. Oates, “Encoding time series as images for visual inspection and classification using tiled convolutional neural networks,” in *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
- [81] N. Hatami, Y. Gavet, and J. Debayle, “Classification of time-series images using deep convolutional neural networks,” in *Tenth International Conference on Machine Vision (ICMV 2017)*, International Society for Optics and Photonics, vol. 10696, 2018, 106960Y.
- [82] S. Mallat, “Group invariant scattering,” *Communications on Pure and Applied Mathematics*, vol. 65, no. 10, pp. 1331–1398, 2012.
- [83] W. Lu, F. Fan, J. Chu, P. Jing, and S. Yuting, “Wearable computing for internet of things: A discriminant approach for human activity recognition,” *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2749–2759, 2018.
- [84] H. A. Rothan and S. N. Byrareddy, “The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak,” *Journal of Autoimmunity*, p. 102 433, 2020.
- [85] C.-C. Lai, T.-P. Shih, W.-C. Ko, H.-J. Tang, and P.-R. Hsueh, “Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and corona virus disease-2019 (COVID-19): The epidemic and the challenges,” *International journal of antimicrobial agents*, p. 105 924, 2020.
- [86] T. Franquet, “Imaging of pulmonary viral pneumonia,” *Radiology*, vol. 260, no. 1, pp. 18–39, 2011.

- [87] X. Xie, Z. Zhong, W. Zhao, C. Zheng, F. Wang, and J. Liu, "Chest CT for typical 2019-nCoV pneumonia: Relationship to negative RT-PCR testing," *Radiology*, p. 200343, 2020.
- [88] O. Gozes, M. Frid-Adar, H. Greenspan, P. D. Browning, H. Zhang, W. Ji, A. Bernheim, and E. Siegel, "Rapid AI development cycle for the coronavirus (COVID-19) pandemic: Initial results for automated detection & patient monitoring using deep learning CT image analysis," *arXiv preprint arXiv:2003.05037*, 2020.
- [89] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, *et al.*, "Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT," *Radiology*, p. 200905, 2020.
- [90] L. Wang and A. Wong, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest radiography images," *arXiv preprint arXiv:2003.09871*, 2020.
- [91] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: Automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," *Physical and Engineering Sciences in Medicine*, p. 1, 2020.
- [92] P. K. Sethy and S. K. Behera, "Detection of coronavirus disease (covid-19) based on deep features," *Preprints*, vol. 2020030300, p. 2020, 2020.
- [93] A. Narin, C. Kaya, and Z. Pamuk, "Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks," *arXiv preprint arXiv:2003.10849*, 2020.
- [94] R. L. Siegel, K. D. Miller, A. Goding Sauer, S. A. Fedewa, L. F. Butterly, J. C. Anderson, A. Cercek, R. A. Smith, and A. Jemal, "Colorectal cancer statistics, 2020," *CA: a cancer journal for clinicians*, 2020.
- [95] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA: a cancer journal for clinicians*, vol. 69, no. 1, pp. 7–34, 2019.
- [96] G.-C. Lou, J.-M. Yang, Q.-S. Xu, W. Huang, and S.-G. Shi, "A retrospective study on endoscopic missing diagnosis of colorectal polyp and its related factors," *Turk J Gastroenterol*, vol. 25, no. Suppl 1, pp. 182–6, 2014.
- [97] D. A. Corley, C. D. Jensen, A. R. Marks, W. K. Zhao, J. K. Lee, C. A. Doubeni, A. G. Zauber, J. de Boer, B. H. Fireman, J. E. Schottinger, *et al.*, "Adenoma detection rate and risk of colorectal cancer and death," *New england journal of medicine*, vol. 370, no. 14, pp. 1298–1306, 2014.

-
- [98] C. M. Rutter, E. Johnson, D. L. Miglioretti, M. T. Mandelson, J. Inadomi, and D. S. Buist, “Adverse events after screening and follow-up colonoscopy,” *Cancer Causes & Control*, vol. 23, no. 2, pp. 289–296, 2012.
- [99] V. Prasath, “Polyp detection and segmentation from video capsule endoscopy: A review,” *Journal of Imaging*, vol. 3, no. 1, p. 1, 2017.
- [100] J. Yao, M. Miller, M. Franaszek, and R. M. Summers, “Colonic polyp segmentation in ct colonography-based on fuzzy clustering and deformable models,” *IEEE Transactions on Medical Imaging*, vol. 23, no. 11, pp. 1344–1352, 2004.
- [101] J. Yao and R. M. Summers, “Adaptive deformable model for colonic polyp segmentation and measurement on ct colonography,” *Medical physics*, vol. 34, no. 5, pp. 1655–1664, 2007.
- [102] A. Sanchez-Gonzalez, B. Garcia-Zapirain, D. Sierra-Sosa, and A. Elmaghraby, “Automatized colon polyp segmentation via contour region analysis,” *Computers in biology and medicine*, vol. 100, pp. 152–164, 2018.
- [103] Y. Yuan, D. Li, and M. Q.-H. Meng, “Automatic polyp detection via a novel unified bottom-up and top-down saliency approach,” *IEEE journal of biomedical and health informatics*, vol. 22, no. 4, pp. 1250–1260, 2017.
- [104] C. Van Wijk, V. F. Van Ravesteijn, F. M. Vos, and L. J. Van Vliet, “Detection and segmentation of colonic polyps on implicit isosurfaces by second principal curvature flow,” *IEEE Transactions on Medical Imaging*, vol. 29, no. 3, pp. 688–698, 2010.
- [105] C. Hassan, M. B. Wallace, P. Sharma, R. Maselli, V. Craviotto, M. Spadacini, and A. Repici, “New artificial intelligence system: First validation study versus experienced endoscopists for colorectal polyp detection,” *Gut*, vol. 69, no. 5, pp. 799–800, 2020.
- [106] A. Nogueira-Rodríguez, R. Dominguez-Carbajales, H. López-Fernández, Á. Iglesias, J. Cubiella, F. Fdez-Riverola, M. Reboiro-Jato, and D. Glez-Peña, “Deep neural networks approaches for detecting and classifying colorectal polyps,” *Neurocomputing*, 2020.
- [107] P. Wang, X. Liu, T. M. Berzin, J. R. G. Brown, P. Liu, C. Zhou, L. Lei, L. Li, Z. Guo, S. Lei, *et al.*, “Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (cade-db trial): A double-blind randomised study,” *The Lancet Gastroenterology & Hepatology*, vol. 5, no. 4, pp. 343–351, 2020.

- [108] H. A. Qadir, Y. Shin, J. Solhusvik, J. Bergsland, L. Aabakken, and I. Balasingham, “Polyp detection and segmentation using mask r-cnn: Does a deeper feature extractor cnn always perform better?” In *2019 13th International Symposium on Medical Information and Communication Technology (ISMICT)*, IEEE, 2019, pp. 1–6.
- [109] J. Kang and J. Gwak, “Ensemble of instance segmentation models for polyp segmentation in colonoscopy images,” *IEEE Access*, vol. 7, pp. 26 440–26 447, 2019.
- [110] H. A. Qadir, J. Solhusvik, J. Bergsland, L. Aabakken, and I. Balasingham, “A framework with a fully convolutional neural network for semi-automatic colon polyp annotation,” *IEEE Access*, vol. 7, pp. 169 537–169 547, 2019.
- [111] Y. B. Guo and B. Matuszewski, “Giana polyp segmentation with fully convolutional dilation neural networks,” in *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, SCITEPRESS-Science and Technology Publications, 2019, pp. 632–641.
- [112] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, pp. 234–241.
- [113] L. Wang, R. Chen, S. Wang, N. Zeng, X. Huang, and C. Liu, “Nested dilation network (ndn) for multi-task medical image segmentation,” *IEEE Access*, vol. 7, pp. 44 676–44 685, 2019.
- [114] N. Ibtehaz and M. S. Rahman, “Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation,” *Neural Networks*, vol. 121, pp. 74–87, 2020.
- [115] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [116] M. Al Ghamdi, M. Abdel-Mottaleb, and F. Collado-Mesa, “Du-net: Convolutional network for the detection of arterial calcifications in mammograms,” *IEEE Transactions on Medical Imaging*, 2020.
- [117] L. Mou, L. Chen, J. Cheng, Z. Gu, Y. Zhao, and J. Liu, “Dense dilated network with probability regularized walk for vessel detection,” *IEEE transactions on medical imaging*, vol. 39, no. 5, pp. 1392–1403, 2019.

- [118] 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, no. 6, pp. 1856–1867, 2019.
- [119] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, *et al.*, “Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT,” *Radiology*, vol. 296, no. 2, E65–E71, 2020.
- [120] L. Huang, R. Han, T. Ai, P. Yu, H. Kang, Q. Tao, and L. Xia, “Serial quantitative chest CT assessment of COVID-19: Deep-learning approach,” *Radiology: Cardiothoracic Imaging*, vol. 2, no. 2, e200075, 2020.
- [121] M. Abdel-Basset, V. Chang, H. Hawash, R. K. Chakraborty, and M. Ryan, “FSS-2019-nCov: A deep learning architecture for semi-supervised few-shot segmentation of COVID-19 infection,” *Knowledge-Based Systems*, p. 106 647, 2020.
- [122] T. Mahmud, M. A. Rahman, and S. A. Fattah, “CovXNet: A multi-dilation convolutional neural network for automatic COVID-19 and other pneumonia detection from chest X-ray images with transferable multi-receptive feature optimization,” *Computers in Biology and Medicine*, p. 103 869, 2020.
- [123] M. Abdel-Basset, V. Chang, and R. Mohamed, “HSMA_WOA: A hybrid novel slime mould algorithm with whale optimization algorithm for tackling the image segmentation problem of chest X-ray images,” *Applied Soft Computing*, vol. 95, p. 106 642, 2020.
- [124] D.-P. Fan, T. Zhou, G.-P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, “Inf-net: Automatic COVID-19 lung infection segmentation from CT images,” *IEEE Transactions on Medical Imaging*, 2020.
- [125] N. Saeedizadeh, S. Minaee, R. Kafieh, S. Yazdani, and M. Sonka, “COVID TV-UNet: Segmenting COVID-19 chest CT images using connectivity imposed u-net,” *arXiv preprint arXiv:2007.12303*, 2020.
- [126] T. Zhou, S. Canu, and S. Ruan, “An automatic COVID-19 CT segmentation network using spatial and channel attention mechanism,” *arXiv preprint arXiv:2004.06673*, 2020.
- [127] Q. Yan, B. Wang, D. Gong, C. Luo, W. Zhao, J. Shen, Q. Shi, S. Jin, L. Zhang, and Z. You, “COVID-19 chest CT image segmentation—a deep convolutional neural network solution,” *arXiv preprint arXiv:2004.10987*, 2020.

- [128] Y. Qiu, Y. Liu, and J. Xu, “Miniseg: An extremely minimum network for efficient COVID-19 segmentation,” *arXiv preprint arXiv:2004.09750*, 2020.
- [129] J. Ma, Y. Wang, X. An, C. Ge, Z. Yu, J. Chen, Q. Zhu, G. Dong, J. He, Z. He, *et al.*, “Towards efficient COVID-19 CT annotation: A benchmark for lung and infection segmentation,” *arXiv preprint arXiv:2004.12537*, 2020.
- [130] D. Müller, I. S. Rey, and F. Kramer, “Automated chest CT image segmentation of COVID-19 lung infection based on 3d U-Net,” *arXiv preprint arXiv:2007.04774*, 2020.
- [131] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in *2016 fourth international conference on 3D vision (3DV)*, IEEE, 2016, pp. 565–571.
- [132] L. Zhang, J. Zhang, P. Shen, G. Zhu, P. Li, X. Lu, H. Zhang, S. A. Shah, and M. Bennamoun, “Block level skip connections across cascaded V-Net for multi-organ segmentation,” *IEEE Transactions on Medical Imaging*, 2020.
- [133] 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.
- [134] M. Al Ghamdi, M. Abdel-Mottaleb, and F. Collado-Mesa, “DU-Net: Convolutional network for the detection of arterial calcifications in mammograms,” *IEEE Transactions on Medical Imaging*, 2020.
- [135] Z. Zeng, W. Xie, Y. Zhang, and Y. Lu, “RIC-Unet: An improved neural network based on Unet for nuclei segmentation in histology images,” *IEEE Access*, vol. 7, pp. 21 420–21 428, 2019.
- [136] N. Ibtehaz and M. S. Rahman, “MultiResUNet: Rethinking the u-net architecture for multimodal biomedical image segmentation,” *Neural Networks*, vol. 121, pp. 74–87, 2020.
- [137] 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, no. 6, pp. 1856–1867, 2019.
- [138] D. -P. Fan, T. Zhou, G. -P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, “Inf-Net: Automatic COVID-19 lung infection segmentation from CT images,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2626–2637, 2020.

- [139] Y. Qiu, Y. Liu, and J. Xu, “Miniseg: An extremely minimum network for efficient COVID-19 segmentation,” *arXiv preprint arXiv:2004.09750*, 2020.
- [140] G. Wang, X. Liu, C. Li, Z. Xu, J. Ruan, H. Zhu, T. Meng, K. Li, N. Huang, and S. Zhang, “A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2653–2663, 2020.
- [141] S. C. Wang, D. Meng, H. Yang, X. Wang, S. Jia, P. Wang, Y.-F. Wang, *et al.*, “Pathological basis of cardiac arrhythmias: Vicious cycle of immune-metabolic dysregulation,” *Cardiovascular Disorders and Medicine*, vol. 3, pp. 1–7, 2018.
- [142] H. J. Wellens, “Cardiac arrhythmias: The quest for a cure: A historical perspective,” *Journal of the American College of Cardiology*, vol. 44, no. 6, pp. 1155–1163, 2004.
- [143] A. for the Advancement of Medical Instrumentation *et al.*, “Testing and reporting performance results of cardiac rhythm and st segment measurement algorithms,” *ANSI/AAMI EC38:1998*, 1998.
- [144] M. A. Kabir and C. Shahnaz, “Denoising of ECG signals based on noise reduction algorithms in EMD and wavelet domains,” *Biomedical Signal Processing and Control*, vol. 7, no. 5, pp. 481–489, 2012.
- [145] B. M. M. A. Tinati *et al.*, “ECG baseline wander elimination using wavelet packets,” *World Scademy of Science, Engineering and Technology*, vol. 3, no. 2005, pp. 14–16, 2005.
- [146] R. Sameni, M. B. Shamsollahi, C. Jutten, and G. D. Clifford, “A nonlinear bayesian filtering framework for ECG denoising,” *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 12, pp. 2172–2185, 2007.
- [147] P. Karthikeyan, M. Murugappan, and S. Yaacob, “ECG signal denoising using wavelet thresholding techniques in human stress assessment,” *International Journal on Electrical Engineering and Informatics*, vol. 4, no. 2, pp. 306–319, 2012.
- [148] M. Blanco-Velasco, B. Weng, and K. E. Barner, “ECG signal denoising and baseline wander correction based on the empirical mode decomposition,” *Computers in Biology and Medicine*, vol. 38, no. 1, pp. 1–13, 2008.
- [149] Q. Qin, J. Li, Y. Yue, and C. Liu, “An adaptive and time-efficient ECG R-peak detection algorithm,” *Journal of Healthcare Engineering*, vol. 2017, 2017. [Online]. Available: <https://doi.org/10.1155/2017/5980541>.

- [150] S. Chouakri, F. Berekxi-Reguig, and A. Taleb-Ahmed, “QRS complex detection based on multi wavelet packet decomposition,” *Applied Mathematics and Computation*, vol. 217, no. 23, pp. 9508–9525, 2011.
- [151] P. J. M. Fard, M. Moradi, and M. Tajvidi, “A novel approach in R peak detection using hybrid complex wavelet (HCW),” *International Journal of Cardiology*, vol. 124, no. 2, pp. 250–253, 2008.
- [152] D. Sadhukhan and M. Mitra, “R-peak detection algorithm for ECG using double difference and RR interval processing,” *Procedia Technology*, vol. 4, pp. 873–877, 2012.
- [153] M. S. Manikandan and K. Soman, “A novel method for detecting R-peaks in electrocardiogram (ECG) signal,” *Biomedical Signal Processing and Control*, vol. 7, no. 2, pp. 118–128, 2012.
- [154] M. Elgendi, “Fast QRS detection with an optimized knowledge-based method: Evaluation on 11 standard ECG databases,” *PLoS One*, vol. 8, no. 9, 2013. [Online]. Available: <https://doi.org/10.1371/journal.pone.0073557>.
- [155] M. Elgendi, A. Mohamed, and R. Ward, “Efficient ECG compression and QRS detection for e-health applications,” *Scientific Reports*, vol. 7, no. 1, pp. 1–16, 2017.
- [156] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” *arXiv:1704.04861*, 2017. [Online]. Available: <https://arxiv.org/abs/1704.04861>.
- [157] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.
- [158] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.
- [159] G. Li, I. Yun, J. Kim, and J. Kim, “Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation,” *arXiv:1907.11357*, 2019. [Online]. Available: <https://arxiv.org/abs/1907.11357>.
- [160] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, “Shufflenet v2: Practical guidelines for efficient cnn architecture design,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 116–131.

- [161] L. Kaiser, A. N. Gomez, and F. Chollet, “Depthwise separable convolutions for neural machine translation,” *arXiv:1706.03059*, 2017. [Online]. Available: <https://arxiv.org/abs/1706.03059>.
- [162] S. Krıman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang, “Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions,” *arXiv:1910.10261*, 2019. [Online]. Available: <https://arxiv.org/abs/1910.10261>.
- [163] G. B. Moody and R. G. Mark, “The impact of the MIT-BIH arrhythmia database,” *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, 2001. [Online]. Available: <http://physionet.org/physiobank/database/mitdb/>.
- [164] L. Sörnmo and P. Laguna, “Electrocardiogram (ECG) signal processing,” *Wiley Encyclopedia of Biomedical Engineering*, 2006.
- [165] R. Boussejot, D. Kreiseler, and A. Schnabel, “Use of the ptb’s ECG signal database CARDIODAT via the internet,” *Biomedizinische Technik/Biomedical Engineering*, vol. 40, no. s1, pp. 317–318, 1995, (in German). [Online]. Available: <https://www.physionet.org/physiobank/database/ptbdb/>.
- [166] M. Hossin and M. Sulaiman, “A review on evaluation metrics for data classification evaluations,” *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, p. 1, 2015.
- [167] C. Sohrabi, Z. Alsafi, N. O’Neill, M. Khan, A. Kerwan, A. Al-Jabir, C. Iosifidis, and R. Agha, “World health organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19),” *International Journal of Surgery*, 2020.
- [168] H. X. Bai, B. Hsieh, Z. Xiong, K. Halsey, J. W. Choi, T. M. L. Tran, I. Pan, L.-B. Shi, D.-C. Wang, J. Mei, *et al.*, “Performance of radiologists in differentiating COVID-19 from viral pneumonia on chest CT,” *Radiology*, p. 200 823, 2020.
- [169] M. Chung, A. Bernheim, X. Mei, N. Zhang, M. Huang, X. Zeng, J. Cui, W. Xu, Y. Yang, Z. A. Fayad, *et al.*, “CT imaging features of 2019 novel coronavirus (2019-nCoV),” *Radiology*, p. 200 230, 2020.
- [170] T. Franquet, “Imaging of pneumonia: Trends and algorithms,” *European Respiratory Journal*, vol. 18, no. 1, pp. 196–208, 2001.
- [171] J. Vilar, M. L. Domingo, C. Soto, and J. Cogollo, “Radiology of bacterial pneumonia,” *European journal of radiology*, vol. 51, no. 2, pp. 102–113, 2004.

- [172] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” *arXiv preprint arXiv:1511.07122*, 2015.
- [173] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.
- [174] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [175] “Labeled optical coherence tomography (OCT),”
- [176] H. Shi, X. Han, N. Jiang, Y. Cao, O. Alwalid, J. Gu, Y. Fan, and C. Zheng, “Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: A descriptive study,” *The Lancet Infectious Diseases*, vol. 20, no. 4, pp. 425–434, 2020.
- [177] M.-Y. Ng, E. Y. Lee, J. Yang, F. Yang, X. Li, H. Wang, M. M.-s. Lui, C. S.-Y. Lo, B. Leung, P.-L. Khong, *et al.*, “Imaging profile of the COVID-19 infection: Radiologic findings and literature review,” *Radiology: Cardiothoracic Imaging*, vol. 2, no. 1, 2020.
- [178] S. Rajaraman, S. Candemir, I. Kim, G. Thoma, and S. Antani, “Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs,” *Applied Sciences*, vol. 8, no. 10, p. 1715, 2018.
- [179] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, *et al.*, “Identifying medical diagnoses and treatable diseases by image-based deep learning,” *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.
- [180] V. Chouhan, S. K. Singh, A. Khamparia, D. Gupta, P. Tiwari, C. Moreira, R. Damaševičius, and V. H. C. de Albuquerque, “A novel transfer learning based approach for pneumonia detection in chest X-ray images,” *Applied Sciences*, vol. 10, no. 2, p. 559, 2020.
- [181] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. R. Acharya, “Automated detection of covid-19 cases using deep neural networks with x-ray images,” *Computers in Biology and Medicine*, p. 103792, 2020.
- [182] J. P. Cohen, P. Morrison, and L. Dao, “Covid-19 image data collection,” *arXiv 2003.11597*, 2020. [Online]. Available: <https://github.com/ieee8023/covid-chestxray-dataset>.

- [183] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.
- [184] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [185] 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,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
- [186] H. J. Koo, S. Lim, J. Choe, S.-H. Choi, H. Sung, and K.-H. Do, “Radiographic and CT features of viral pneumonia,” *Radiographics*, vol. 38, no. 3, pp. 719–739, 2018.
- [187] N. Islam, Y. Faheem, I. U. Din, M. Talha, M. Guizani, and M. Khalil, “A blockchain-based fog computing framework for activity recognition as an application to e-healthcare services,” *Future Generation Computer Systems*, vol. 100, pp. 569–578, 2019.
- [188] A. Jalal, Y.-H. Kim, Y.-J. Kim, S. Kamal, and D. Kim, “Robust human activity recognition from depth video using spatiotemporal multi-fused features,” *Pattern recognition*, vol. 61, pp. 295–308, 2017.
- [189] T. T. Um, F. M. Pfister, D. Pichler, S. Endo, M. Lang, S. Hirche, U. Fietzek, and D. Kulić, “Data augmentation of wearable sensor data for parkinson’s disease monitoring using convolutional neural networks,” in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 216–220.
- [190] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, “A public domain dataset for human activity recognition using smartphones,” in *Esann*, 2013.
- [191] M. Zhang and A. A. Sawchuk, “Usc-had: A daily activity dataset for ubiquitous activity recognition using wearable sensors,” in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, 2012, pp. 1036–1043.
- [192] C. Catal, S. Tufekci, E. Pirmit, and G. Kocabag, “On the use of ensemble of classifiers for accelerometer-based activity recognition,” *Applied Soft Computing*, vol. 37, pp. 1018–1022, 2015.

- [193] Y. Guan and T. Plötz, “Ensembles of deep lstm learners for activity recognition using wearables,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 2, pp. 1–28, 2017.
- [194] D. K. Rex, “Colonoscopic withdrawal technique is associated with adenoma miss rates,” *Gastrointestinal endoscopy*, vol. 51, no. 1, pp. 33–36, 2000.
- [195] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
- [196] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” *arXiv preprint arXiv:1511.07122*, 2015.
- [197] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.
- [198] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.
- [199] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilarino, “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.
- [200] 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 *International Conference on Multimedia Modeling*, Springer, 2020, pp. 451–462.
- [201] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, “Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer,” *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, no. 2, pp. 283–293, 2014.
- [202] J. Bernal, J. Sánchez, and F. Vilarino, “Towards automatic polyp detection with a polyp appearance model,” *Pattern Recognition*, vol. 45, no. 9, pp. 3166–3182, 2012.
- [203] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

- [204] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, and H. D. Johansen, “Resunet++: An advanced architecture for medical image segmentation,” in *2019 IEEE International Symposium on Multimedia (ISM)*, IEEE, 2019, pp. 225–2255.
- [205] A. Chaurasia and E. Culurciello, “Linknet: Exploiting encoder representations for efficient semantic segmentation,” in *2017 IEEE Visual Communications and Image Processing (VCIP)*, IEEE, 2017, pp. 1–4.
- [206] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, “Doubleu-net: A deep convolutional neural network for medical image segmentation,” *arXiv preprint arXiv:2006.04868*, 2020.
- [207] R. Wang, S. Chen, C. Ji, J. Fan, and Y. Li, “Boundary-aware context neural network for medical image segmentation,” *arXiv preprint arXiv:2005.00966*, 2020.
- [208] Y. Guo, J. Bernal, and B. J Matuszewski, “Polyp segmentation with fully convolutional deep neural networks—extended evaluation study,” *Journal of Imaging*, vol. 6, no. 7, p. 69, 2020.
- [209] P. Huang, T. Liu, L. Huang, H. Liu, M. Lei, W. Xu, X. Hu, J. Chen, and B. Liu, “Use of chest CT in combination with negative RT-PCR assay for the 2019 novel coronavirus but high clinical suspicion,” *Radiology*, vol. 295, no. 1, pp. 22–23, 2020.
- [210] T. Mahmud, M. A. Rahman, and S. A. Fattah, “CovXNet: A multi-dilation convolutional neural network for automatic COVID-19 and other pneumonia detection from chest x-ray images with transferable multi-receptive feature optimization,” *Computers in Biology and Medicine*, p. 103 869, 2020.
- [211] H. Kang, L. Xia, F. Yan, Z. Wan, F. Shi, H. Yuan, H. Jiang, D. Wu, H. Sui, C. Zhang, *et al.*, “Diagnosis of coronavirus disease 2019 (COVID-19) with structured latent multi-view representation learning,” *IEEE Transactions on Medical Imaging*, 2020.
- [212] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
- [213] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, pp. 234–241.

- [214] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, “Tversky loss function for image segmentation using 3d fully convolutional deep networks,” in *International Workshop on Machine Learning in Medical Imaging*, Springer, 2017, pp. 379–387.
- [215] N. Abraham and N. M. Khan, “A novel focal tversky loss function with improved attention u-net for lesion segmentation,” in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, IEEE, 2019, pp. 683–687.
- [216] *COVID-19 CT lung and infection segmentation dataset*, [online]. Available: <https://doi.org/10.5281/zenodo.3757476>, 2020.
- [217] *COVID-19 CT segmentation dataset*, [online]. Available: <https://medicalsegmentation.com/covid19/>, 2020.
- [218] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, *et al.*, “Attention U-net: Learning where to look for the pancreas,” *arXiv preprint arXiv:1804.03999*, 2018.
- [219] 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*, 2020.
- [220] J. P. Kanne, B. P. Little, J. H. Chung, B. M. Elicker, and L. H. Ketaj, “Essentials for radiologists on COVID-19: An update—radiology scientific expert panel,” *Radiology*, vol. 296, no. 2, E113–E114, 2020.
- [221] J. T. Wu, K. Leung, M. Bushman, N. Kishore, R. Niehus, P. M. de Salazar, B. J. Cowling, M. Lipsitch, and G. M. Leung, “Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China,” *Nature Medicine*, vol. 26, no. 4, pp. 506–510, 2020.
- [222] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.
- [223] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *2018 IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
- [224] Z. Tan, Y. Yang, J. Wan, H. Hang, G. Guo, and S. Z. Li, “Attention-based pedestrian attribute analysis,” *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 6126–6140, 2019.

- [225] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” *arXiv preprint arXiv:1511.07122*, 2015.
- [226] N. Abraham and N. M. Khan, “A novel focal tversky loss function with improved attention U-net for lesion segmentation,” in *2019 IEEE 16th International Symposium on Biomedical Imaging*, 2019, pp. 683–687.
- [227] *MosMedData: Chest CT Scans with COVID-19 Related Findings*, Accessed: 28 April, 2020. [online]. Available: https://mosmed.ai/datasets/covid19_1110, 2020.
- [228] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *2015 IEEE 28th conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
- [229] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *2015 18th International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241.
- [230] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in *2016 4th International Conference on 3D Vision*, 2016, pp. 565–571.
- [231] 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, no. 6, pp. 1856–1867, 2019.
- [232] 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, no. 10, pp. 3008–3018, 2020.