Northumbria University
NEWCASTLE

DEVELOPMENT OF MEDICAL IMAGE/VIDEO SEGMENTATION VIA DEEP LEARNING MODELS

MANSOUREH PEZHMAN POUR

PhD

2021

DEVELOPMENT OF MEDICAL IMAGE/VIDEO SEGMENTATION VIA DEEP LEARNING MODELS

MANSOUREH PEZHMAN POUR

A thesis submitted in partial fulfilment of the requirements of the University of Northumbria at Newcastle for the degree of Doctor of Philosophy

Faculty of Computer and Information Sciences

July 2021

Abstract

Image segmentation has a critical role in medical diagnosis systems as it is mostly the initial stage, and any error would be propagated in the subsequent analysis. Certain challenges, including Irregular border, low quality of images, small Region of Interest (RoI) and complex structures such as overlapping cells in images impede the improvement of medical image analysis. Deep learning-based algorithms have recently brought superior achievements in computer vision. However, there are limitations to their application in the medical domain including data scarcity, and lack of pre-trained models on medical data. This research addresses the issues that hinder the progress of deep learning methods on medical data. Firstly, the effectiveness of transfer learning from a pre-trained model with dissimilar data is investigated. The model is improved by integrating feature maps from the frequency domain into the spatial feature maps of Convolutional Neural Network (CNN). Training from scratch and the challenges ahead were explored as well. The proposed model shows higher performance compared to state-of-the-art methods by $\%2.2$ and $\%17$ in Jaccard index for tasks of lesion segmentation and dermoscopic feature segmentation respectively. Furthermore, the proposed model benefits from significant improvement for noisy images without preprocessing stage. Early stopping and drop out layers were considered to tackle the overfitting and network hyper-parameters such as different learning rate, weight initialization, kernel size, stride and normalization techniques were investigated to enhance learning performance. In order to expand the research into video segmentation, specifically left ventricular segmentation, U-net deep architecture was modified. The small RoI and confusion between overlapped organs are big challenges in MRI segmentation. The consistent motion of LV and the continuity of neighbor frames are important features that were used in the proposed architecture. High level features including optical flow and contourlet were used to add temporal information and the RoI module to the Unet model. The proposed model surpassed the results of original Unet model for LV segmentation by a $\%7$ increment in Jaccard index.

Contents

Abstract iii

Acknowledgements xiii

Declaration xv

Published Contributions xvii

1 Introduction 1

1.1 Background 1

1.2 Motivation 5

1.3 Aims and Objectives 7

1.4 Thesis Structure 8

2 Review of Segmentation Methods 11

2.1 Introduction 11

2.2 Traditional Methods 11

2.2.1 Methods based on Thresholding 12

2.2.2 Edge based Segmentation 13

2.2.3 Region Based Segmentation 14

2.2.4 Neural Network Based Methods 14

2.2.5 Hybrid Methods 15

2.2.6 Why Deep Learning? 16

2.3 Models based on Deep Learning 17

2.3.1 Introduction to Deep Learning-based Models 17

2.3.2 Region based Convolutional Neural Network Methods 18

2.3.2.1 Regions with CNN Features (R-CNN) 19

2.3.2.2 Simultaneous Detection and Segmentation(SDS) 19

2.3.2.3 Fast R-CNN and Faster R-CNN 20

2.3.2.4 Learning Hierarchical Features 21

2.3.2.5 Recurrent Convolutional Neural Networks 22

2.3.3 Methods Inspired from CNN Classification Models 23
2.3.3.1 Fully Convolutional Neural Network 23
2.3.3.2 SegNet 25
2.3.3.3 UNet 26
2.3.3.4 Reseg 26
2.4 Segmentation Research on Medical Data 27
2.5 Introduction on Convolutional Neural Networks and Evaluation Metrics for Segmentation 31
2.5.1 Architecture of Convolutional Neural Network 32
2.5.1.1 Convolution Layer 32
2.5.1.2 Pooling Layer 33
2.5.1.3 Activation Function 33
2.5.2 Hardware Required to Train a CNN 34
2.5.3 Deep Learning Frameworks 34
2.5.4 Evaluation Metrics 35
2.6 Summary 36

3 A Hybrid CNN-based Model for Skin Lesion Analysis towards Melanoma Detection 37
3.1 Data Preparation 37
3.2 The Proposed Hybrid Model for Tasks of Skin Lesion Segmentation and Dermoscopic Feature Segmentation 39
3.2.1 Reducing the Filter Size of Pooling Layer or Even Removing Pooling Layers to Increase the Accuracy 43
3.2.2 Data Augmentation to Overcome Overfitting 43
3.2.3 Drop out Layers to Prevent Overfitting 44
3.2.4 Early Stopping to Handle Overfitting 45
3.3 Experiments and Results 46
3.4 Summary 50

4 Proposed Contourlet-Convolutional Neural Network 53
4.1 Proposed Method 53
4.1.1 Pre/Post Processing 54

4.1.2 Network Architecture . . . 54
4.1.2.1 Contourlet Transformation . . . 54
4.1.2.2 Contourlet-driven CNN . . . 55
4.1.2.3 Lesion Segmentation . . . 57
4.1.2.4 Lesion Attribute Segmentation . . . 58
4.2 Experiments and Results . . . 59
4.2.1 Data Preparation . . . 60
4.2.2 Implementation . . . 60
4.2.3 Results for the Task of Lesion Segmentation . . . 60
4.2.4 Results for Task of Lesion Dermoscopic Feature Segmentation . . . 65
4.3 Discussion . . . 66
4.4 Summary . . . 69

5 Left Ventricular Segmentation **73**
5.1 Introduction . . . 73
5.2 Proposed Method . . . 75
5.2.1 Data Preperation . . . 75
5.2.2 Proposed CNN based Model for Task of LV Segmentation . . . 76
5.2.3 Proposed Algorithms to Identify the Region of Interest (RoI) . . . 77
5.2.4 U-net based Model Improved by Optical Flow Motion Estimation . . . 79
5.3 Experiments and Results . . . 82
5.4 Discussion . . . 86
5.5 Summary . . . 87

6 Conclusion **89**
6.1 Introduction . . . 89
6.2 Contributions . . . 89
6.3 Future Work . . . 91

Acronyms **93**

References **94**

List of Figures

1.1 Sample images from the ISIC dataset that show various issues such as variety in scales and colour, the existence of hair and other artifacts, and dark corners on images. 3

1.2 The number of publications that used deep learning in medical imaging. Data is from Scopus and search is defined as the search phrases are deep learning and medical imaging within search areas of title, keywords, and abstract. 6

2.1 Classification of segmentation methods 12

2.2 Comparison of traditional machine learning and deep learning, (a)Traditional machine learning flow, (b) deep learning flow 17

2.3 R-CNN: Region-based Convolutional Network (Girshick et al., 2015) 19

2.4 The pipeline of SDS method (Arbeláez et al., 2014) 20

2.5 The Diagram of Fast R-CNN (Girshick, 2015) 21

2.6 The pipeline of the scene parsing system (Farabet et al., 2012) 22

2.7 System considering one (f), two (f o f) and three(f o f o f) instances of the network. In all three cases, the architecture produces labels (1 1 output planes) corresponding to the pixel at the center of the input patch (Pinheiro and Collobert, 2014) . . 22

2.8 Transforming fully connected layers into convolution layers. Fully convolutional networks can efficiently learn to make dense predictions for per-pixel tasks like semantic segmentation. (Long et al., 2015) 24

2.9 The skip connection: combining coarse, high layer information with fine, low layer information. 25

2.10 (a) F1 is 1-dilated convolution with a receptive field of 3×3. (b) F2 is produced from F1 by a 2-dilated convolution and the receptive field increased to 7×7. (c) F3 is produced from F2 by a 4-dilated convolution with receptive field of 15×15 Yu and Koltun (2015) 25

2.11 SegNet and FCN decoders. a, b, c, d correspond to values in a feature map (Badrinarayanan et al., 2017) 26

2.12 Rectifier Linear Unit activation function 34

3.1 (a)Sample images from ISIC dataset (b)The groundtruth for skin lesion segmentation (c) The mask for feature segmentation (Globules) (d)The mask for feature segmentation (streaks). 38
3.2 The outline of the proposed deep network architecture for lesion segmentation and dermoscopic feature segmentation. 40
3.3 Outline of proposed deep network architecture for the tasks of lesion segmentation and dermoscopic feature segmentation. 41
3.4 The lesion border mask is shown in blue. Globules and streaks are red and yellow respectively. 42
3.5 (a) Standard Neural Net with 2 hidden layers, (b) An example of a thinned net produced by applying dropout to the standard network. Crossed units have been dropped (Srivastava et al., 2014) 44
3.6 (a) Standard Neural Network, (b) Dropout Network (Srivastava et al., 2014) . . . 45
3.7 Early Stopping. 46
3.8 Stopping criteria with training error and validation error for an FCN-based network fine tuned on FCN8s for skin lesion analysis 46
3.9 Loss for network fine-tuned on FCN-8s with flipped and cropped images as augmentation 48
3.10 Output error, using FCN8s as pre-trained model without drop out layer 48
3.11 Output error, using FCN8s as pretrained model with drop out layer 49
3.12 (a) Segmented image produced by the model pre-trained with Alex-net and data set augmented by just flipping, (b),(c) are output of the model in which, the input segmented images augmented by cropping and these outputs produced by a deeper network (16 convolution layers) using pre-trained model VOC-FCN8s, (d) Ground truth test image. 49
3.13 (a) Image from test dataset. (b) Test ground truth for globule feature. (c) Feature segmentation by our network 51

4.1 Architecture of proposed model, deep convolutional neural network proposed for lesion segmentation and image representations from various levels of contourlet transform. 53

4.2 (a) The original LP decomposition from (Burt and Adelson, 1987), (b) LP decomposition proposed in (Do and Vetterli, 2005) 55
4.3 Contourlet transform composed of Laplacian Pyramid and Directional filter bank. 56
4.4 Original image and Mask (b) Multiscale image representations of contourlet . . 63
4.5 Training curves for models 1-4 64
4.6 a) Original image b) Segmented output from model 3 3 c) Segmented output from model 4 d) Output of model that is fine-tuned on a pretrained model (Pour, Seker, Shao, 2017) E) Test mask 66
4.7 Histogram of Jaccard index values for proposed method compared to top result of the challenge ISIC 2017 67
4.8 (a), (b), (c), are the original images with globule and streak groundtruth respectively. (d) is the predicted globule groundtruth and (e) is the predicted streak . . . 68
4.9 (a), (b), (c), (d) are the outputs of 4th,8th,9th and 11th convolution layer respectively. Images in the first raw are from model 3 and second raw from model 4. (e) Origin image and the groundtruth 69

5.1 A sample image from Sunnybrook dataset and the converted ground truth. 75
5.2 Images (numbers 0, 5, 13, 20 respectively from the same sequence-patient 1) and the corresponding groundtruths from LV segmentation challenge dataset (Suinesiaputra et al., 2014) 76
5.3 Outline of the proposed method for task of LV segmentation. 77
5.4 The Procedure of Detecting the Region of Interest. 78
5.5 (a)original adjacent frames, (b) and (c) are U and V respectively, and (e) vector representation optical flow 80
5.6 The proposed system with integrating optical flow feature maps 81
5.7 Training and validation loss for model 1, The Left plot is for using pretrained model from previous chapter. 84

List of Tables

3.1 Parameters of the network FCN 40

3.2 Result of fine-tuning the network on Alex-net, Data augmentation is conducted by flipping 47

3.3 Results of fine-tuning the network on pre-trained model of FCN32s, Data augmentation is conducted by flipping and cropping 48

3.4 Evaluation results of lesion segmentation compared to best challenge result(ISIC2016) (Gutman et al., 2016) 50

3.5 Evaluation results for dermoscopic feature segmentation compared to the best result of challenge (ISIC2016) 50

4.1 Hyperparameters of Convolutional Neural Network- Model 1 61

4.2 Evaluation metrics for different architectures compared to the best result of challenge (Gutman, et al., 2016). Model 1 refers to the basic architecture composed of 7 convolution layers and 6 deconvolution layers, Model 2 is the model 1 incorporated with representations of transform domain, Model 3 is the model 1 but deeper, and Model 4 is model 3 with integrated features of transform domain. . . 62

4.3 Training time comparison in different models 64

4.4 Evaluation metrics for ISIC2017 dataset 65

4.5 Evaluated metrics for the task of dermoscopic feature segmentation 66

5.1 Evaluation metrics from the proposed model 1 compared to the recent papers including (Tan et al., 2017), (Khened et al., 2019), (Tran, 2016) 83

5.2 The performance of both models with cross entropy and dice loss 84

5.3 The effect of increasing convolution layers and number of filters on dice metric from proposed model on LVSC dataset 85

5.4 Comparing with the original U-net 85

5.5 Comparing the results with applying IN, BN, LN normalization methods with two recent models 86

Acknowledgements

I would like to thank the supervision team, Professor Huseyin Seker, Professor Ling Shao and Dr Longzhi Yang for their support during the course of this research. I would also like to express my sincere thanks to my husband Saeid, without his understanding and encouragement, it would have been impossible to complete my PhD.

Declaration

I declare that the work contained in this thesis has not been submitted for any other award and that it is all my own work. I also confirm that this work fully acknowledges opinions, ideas and contributions from the work of others.

Any ethical clearance for the research presented in this thesis has been approved. Approval has been sought and granted by the *University Ethics Committee* on *03/03/2016*.

.

I declare that the Word Count of this thesis is 32300 words.

Name: Mansoureh Pezhman Pour

Signature:

Date: 26 July 2021

Published Contributions

Pour, M.P. and Seker, H., 2020. Transform domain representation-driven convolutional neural networks for skin lesion segmentation. Expert Systems with Applications, 144, p.113129.

Pour, M.P., Seker, H. and Shao, L., 2017, July. Automated lesion segmentation and dermoscopic feature segmentation for skin cancer analysis. In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 640-643). IEEE.

Chapter 1

Introduction

1.1 Background

Most research on medical image analysis encompasses similar steps including preprocessing, feature extraction, segmentation, classification, interpretation and measurement (Deserno, 2011). Due to each stage's reliance on its previous step, a system's performance will be considerably affected if one of these steps; is not performed properly, particularly early stages such as segmentation.

Image segmentation aims to localise a specific part or object of an image in order to generate a useful representation for the system to analyse. This is an important stage of many applications in computer vision including traffic control systems and video surveillance, recognition tasks such as face or fingerprint detection, autonomous driving, medical image/video analysis etc.

When it comes to the medical domain, image segmentation is a fundamental stage in various medical diagnosis systems because often it is the initial part, so any error would be propagated in the following analyses. Medical image segmentation is often applied to deal with separating an organ or tissue in pathology, or other relevant structures, such as tumor detection and computer assisted surgery (Pereira et al., 2016), (Cireşan et al., 2013), (Soler et al., 2001). To generate images of different parts of body, various devices and modalities have been produced to improve diagnostic systems. The development of imaging equipment technology has led to an increase in medical data too. Nowadays, several types of medical images and videos are available to researchers includ-

ing Magnetic Resonance Imaging (MRI), X-rays, Computed Tomography (CT) scans, ultrasound and nuclear medicine imaging such as Positron-Emission Tomography (PET). However, certain challenges with medical data limit the effectiveness of medical analysis systems:

- Dataset scarcity is a major problem in medical image segmentation as providing labelled data is costly and requires clinician experience.
- Irregular border and complex structures in medical images for instance, overlapping cells in images (Qi et al., 2011) or certain tumours which mostly form within the tissue and very much resemble normal tissue, lead to further complications in segmentation.
- Low quality of medical images also reduces accuracy. Medical images are mostly noisy and affected by various artifacts while imaging.
- Not only the resolution of medical images has improved over the years but the dimension has also expanded, which requires the development of methods on 3D and 4D images.

Over the years, a wide variety of methods deriving from computer vision and machine learning have been developed for the task of image segmentation. These have been applied and adopted for the medical domain too. Traditional segmentation techniques encompass methods based on thresholding, clustering, region and edge-based techniques, as well as methods powered by machine learning which have seen dramatic improvements. Extracting discriminant features is the key to advancing these segmentation methods and extensive research into feature extraction and selection algorithms has been carried out. Moreover, the methods that researchers use for a specific type of medical image may not work well on other types. For several decades, researchers have considered finding a solution that can be applied to various types of images and is suitable for different medical segmentation applications, yet this is still a controversial topic.

Lesion segmentation as the first step of a melanoma diagnosis system, aims to separate the relevant pixels to melanoma tumors. Skin cancer is a prevalent kind of cancer worldwide with fast increment in incidence and number of deaths over the past decade (Siegel et al., 2016). Nevertheless, there is a high chance of a cure if the cancer is diagnosed in a primary stage before other tissues of the body are invaded. Dermoscopic imaging has significantly assisted dermatologists in the detection of malignant melanoma, the deadliest form of skin cancer. However, expert clinicians are still needed to distinguish the disease. Research on automated computer-based detection

systems for melanoma cancer have increased in the past few years to assist dermatologists, lessen workload, continue monitoring high risk patients and more to reduce the costs of diagnosis and treatment (Alamdari et al., 2017). Moreover, such algorithms can be used to improve embedded systems, robots or even mobile software to create an easy user interface as part of an automated diagnosis system. Lesion border segmentation is still a challenging and complex task due to several problems in dermoscopic images including:

- The lesions being of various shapes, colours and sizes
- Low contrast
- In some images, the region of interest is very small
- Fuzzy borders
- Illumination variation
- Variety of artifacts including ruler marks, hair, air bubble, etc.
- Dark corners

Figure 1.1 shows samples of dataset.

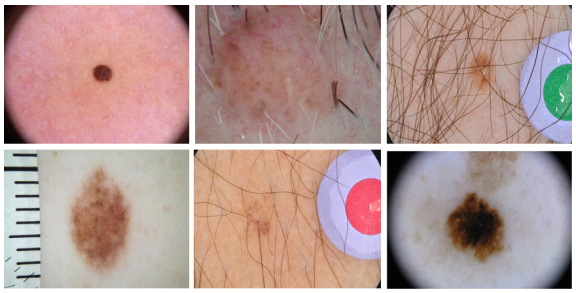

Figure 1.1: Sample images from the ISIC dataset that show various issues such as variety in scales and colour, the existence of hair and other artifacts, and dark corners on images.

Various algorithms have been developed in the literature to tackle the aforementioned issues, for example to remove hair or illumination correction algorithms (Liu and Zerubia, 2015), (Jaisakthi et al., 2018). Moreover, the lack of extensive public datasets has impeded the development of computer-aided systems for melanoma detection. Recently, ISIC (International Skin Imaging Collaboration) has provided public collections of dermoscopic images of skin lesions, and ISBI challenges (IEEE International Symposium on Biomedical Imaging) have been held to improve skin cancer diagnosis (Gutman et al., 2016), (Codella et al., 2019). Deep Convolutional Neural

Networks (CNN) have seen great success in computer vision and machine learning, and surpassed conventional methods in several challenges (Cireşan et al., 2013), (He et al., 2016). CNN learns more complex features with more layers compared to a shallow network but it requires a large amount of data for training, which is a critical issue in the medical domain. Dataset enlargement by adding other available datasets has been generally considered by researchers, whereas preparing labeled data is still time consuming and expensive. Common classical augmentation techniques such as flipping, rotation, and scaling are often applied to produce adequate information to feed the deep neural network (Harangi, 2018), (Hussain et al., 2017) (Kwasigroch et al., 2017). However, augmentation techniques are specific to datasets and require attention to avoid losing information or increasing the share of irrelevant data. Augmentation techniques have therefore expanded into a field widely used in deep learning particularly medical analysis. It has become an active academic area within which many studies are being conducted to develop data-generation methods (Zhang et al., 2017), (Liang et al., 2018), (Frid-Adar et al., 2018).

When it comes to dermoscopic feature segmentation, unbalanced data is the main challenge to training the model. Nearly half of the images in the dataset do not contain any dermoscopic features and the detection of empty masks can improve the performance (Chen et al., 2018). Moreover, among those which hold dermoscopic features, the number of pixels that belong to the classes of globules or streaks are far fewer than background pixels. Nearly 42 percent of images contain pixels of globules while the number of images involving streaks is less than 8 percent of the dataset. Techniques such as weighted loss function and customized augmentation will be considered to tackle this issue.

The task of left ventricular segmentation was chosen to expand the research from image to video segmentation. A cardiac MRI scan is a non-invasive test and an MRI machine is used to generate magnetic and radio waves to show detailed pictures of the inside of the heart. Over the years, cardiovascular research has developed to improve the early identification of cardiac diseases. The left ventricular (LV) is the most investigated chamber in cardiac segmentation due to its key role in pumping blood through the human body. The LV is the thickest of the heart's chambers and pumps oxygenated blood to tissues all over the body. A cardiac magnetic resonance image is a critical part of cardiac function analysis/calculations, such as left ventricular volume and ejection fraction, wall thickness and wall motion abnormality detection and stroke volume. Cardiovascular diseases

(CVD) are a significant cause of disability and death around the world. The Global Burden of Disease research reported that CVD was the main cause of 29.6% of all deaths worldwide. It is still the main reason for over 4 million deaths per year, which constitutes half of all deaths in Europe in 2010 (Nichols et al., 2014). It is estimated that by 2035, nearly 45 % of adults living in the US will be diagnosed with some form of CVD (Benjamin et al., 2018). The number of LV segmentation studies based on CNN is increasing with the recent advances in deep learning (Wu et al., 2020), (Shoaib et al., 2019). However, there is still potential improvement for LV segmentation thanks to advances in deep models. LV segmentation faces several challenges including small region of interest compared to whole slide, intensity issues, and weak boundaries between myocardium and surrounding tissue. This is what motivated this research to improve a CNN-based method with adding regions of interest and edge information to the network to improve the performance.

1.2 Motivation

Extracting effective features in medical images is still a challenging issue. Not only is it a tedious and time consuming task. it also demands medical professionals' experience, hence scientists have been strongly motivated to design fully automatic computer-based diagnostic systems. Furthermore, deep learning with the superiority of automatic feature learning from raw input has received particular attention in medical analysis. Such an automated system not only benefits from not being dependent on expert knowledge to find efficient feature extraction/selection methods but it has also demonstrated considerable accuracy and robustness. Recent advances in image processing demonstrates striking progress relying on deep learning algorithms and the increasing rate of research in this area demonstrates the importance of this subject in the medical domain too. Figure 1.2 highlights the increase in the number of publications on medical imaging with deep learning methods over the last few years.

Although CNN based models have surpassed traditional segmentation techniques, most studies have addressed natural images and not the medical domain where possible procedures on natural images may not be easily applicable. There has been a jump-start in CNN-based models for image segmentation in 2015, inspired from the Fully Convolutional Networks (FCN) study (Long et al., 2015) (which adopted the CNN based classification model for task of segmentation), along with technological advances in hardware including GPUs. This motivates the present study to

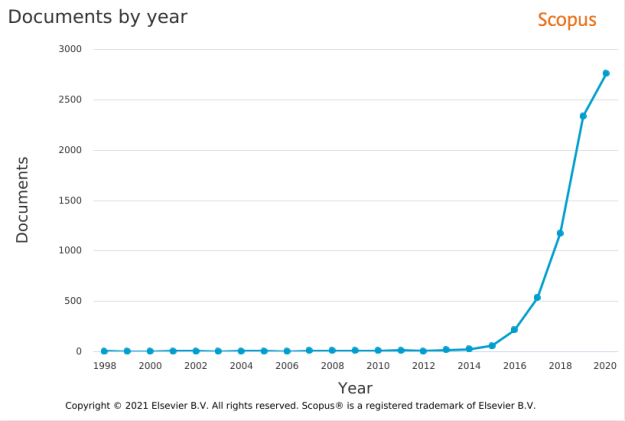

Figure 1.2: The number of publications that used deep learning in medical imaging. Data is from Scopus and search is defined as the search phrases are deep learning and medical imaging within search areas of title, keywords, and abstract.

investigate the challenges that FCN confronts within the medical domain and improve upon a new model for medical image/video segmentation. Although FCN based models have been presented in the literature in computer vision with remarkable success, there is still research potential to improve detection accuracy in image segmentation, specifically in the medical area, which still faces challenges due to the lower quality of the medical images, overlapping organs, their complex structure, noise and artifacts.

Although studies have generated several FCN-based pretrained models on object recognition and segmentation tasks, there are no pretrained models on medical images. The use of a pretrained model made of natural images on medical data has not yet been assessed. Addressing this issue is essential for the application of FCN-based architectures in the medical area. The first phase will therefore investigate transfer learning with a pretrained model on a dataset containing millions of natural images, which is remarkably different from the medical dataset. Data augmentation and fine tuning to adjust the parameters to enhance the model will be considered as well.

Due to the dissimilarity of medical data and data of the pre-trained model in transfer learning, only general information in the first layers can be useful to initially speed up the learning mechanism, but this may negatively affect convergence. Moreover, issues such as local representation on some layers (Yosinski et al., 2015) and the limitation to modify the deep network disparate from pretrained model motivated this study to investigate the training of the deep convolutional network on medical data from scratch. The challenges ahead will be studied to identify an efficient deep architecture for medical image segmentation based on best-known recent segmentation methods

(i.e. FCN and U-net) (Ronneberger et al., 2015). This can be done by going deeper along layers in the deep network to extract higher level features or increasing the inputs. These are common approaches in deep networks, but they are hardly applicable in the medical area where datasets are scarce. Therefore, this study proposes advancing the CNN-based model in a way that improves the network's high-level feature learning without increasing the network's capacity, which causes overfitting.

1.3 Aims and Objectives

This research aims to design a segmentation system based on CNN to address the issues that hinder the progress of deep networks on medical data such as data scarcity, noisy images, lack of pretrained models on medical data and unbalanced datasets. For image segmentation, this study focuses on skin lesion diagnostics. The proposed models will be evaluated for two tasks of lesion border segmentation and attribute detection, (Gutman et al., 2016). The task of left ventricular segmentation to study medical video segmentation is also considered. Based on challenges reviewed in the literature, and to attain the above-mentioned goal, the research objectives are listed as follows:

- To conduct a literature review to understand the state-of-the art in medical segmentation.
- To design a hybrid segmentation method for two tasks of skin lesion segmentation and dermoscopic feature segmentation. As in the literature reviewed, despite of the popularity of transfer learning, there is still the open question of whether transfer learning from a deep model pre-trained on a dataset with significant different images from medical area, can improve performance.
- To improve efficiency in terms of the number of parameters and convergence time, and in particular to ease the training procedure for dermoscopic feature segmentation with issues of complex structure and very small regions of interest and unbalanced data. Although various research on skin melanoma diagnosis has been conducted, these studies mostly address lesion border segmentation. Only a limited number of research papers have focused on dermoscopic feature segmentation where increasing accuracy is still a key challenge.
- To propose a new deep architecture built by integrating spectral domain features into deep

networks to tackle the challenges ahead for training the deep network from scratch on a small dataset. Amending the cost function and using drop out layers will be considered To improve the regularization as well.

- To design a robust segmentation algorithm to deal with critical issues relevant to the skin dataset including low contrast, dark corners, small area, and artifacts such as hair and ruler marks. The network will be improved taking into account minimal pre-processing techniques specific to the dataset, in order to preferably use the raw images as inputs for the network. This is important for the model to be easily applicable on various medical datasets that may need different preprocessing methods.

- To develop the proposed model for video segmentation, specifically automatic left ventricle segmentation. Such a task is still challenging due to the size and intervention of the cardiac area as well as the thorax in slices. Contourlet feature maps will be investigated to handle weak boundaries and provide edge information to the network. To produce a location guide for CNN, the continuity feature of consecutive frames and motion feature will be considered. This would be useful for tackling the issue of the left ventricular area being a small area compared to the whole cardiac image.

1.4 Thesis Structure

The remainder of thesis is comprised of five chapters and is organised as follows:

Chapter 2: In this chapter, the segmentation techniques in the literature are classified into deep learning based models, specifically CNN and traditional methods, and a comprehensive review is conducted. Moreover, this chapter discusses the concept of deep learning beside an architecture of the convolutional neural network as the most common deep learning method in computer vision. This chapter also describes information on the required hardware and the frameworks used in this research.

Chapter 3: A hybrid model inspired by a deep CNN identified in the literature review is proposed for two main segmentation tasks in a melanoma diagnosis system (including a lesion segmentation followed by a dermoscopic feature segmentation). An overview of the model, details on the training and the hyper-parameters are discussed as well. The issue of overfitting as the most

common issue in deep architectures when it comes to medical domain, is discussed and transfer learning is considered along with the effect of applying drop out layers, data augmentation, and early stopping to tackle this issue.

Chapter 4: The diagnosis system proposed in this chapter considers training from scratch, complements the former model and improves it gradually by appending appropriate features and optimisation techniques. Image representations from multi-direction and multi-scale contourlet transform are incorporated into a CNN network and form a novel architecture, which is then compared in terms of efficiency and accuracy to the original model and top research in literature. Various parts of the method including pre-processing, architecture of the segmentation model, and post processing are detailed.

Chapter 5: The proposed U-net based architecture in this chapter addresses the task of video segmentation, specifically automatic left ventricle (LV) segmentation on a short-axis cardiac MRI. Video attributes, including motion features, are considered to improve the model. Data preparation and the design of a new model inspired by models from previous chapters are explained as well.

Chapter 6: Contributions and potential topics for future research are presented.

Chapter 2

Review of Segmentation Methods

2.1 Introduction

Image segmentation as an important initial part of image analysis techniques, is still an active research field in computer vision. Segmentation is the processing of an image to divide it into various regions where each pixel is attributed to a region. It has been used in numerous applications where segmented objects of an image are required, including but not limited to object tracking and video conferencing, medical applications such as detection of cancerous cells and locate tumors, computer assisted surgery, object detection and recognition tasks such as face and iris recognition (Ikonomakis et al., 2000) , (Pereira et al., 2016), (Mukherjee et al., 2016). Image Segmentation techniques can be categorized into two main groups: traditional methods and methods based on deep learning algorithms. Figure 2.1 displays a brief classification of the segmentation methods in this chapter. In next section, from segmentation techniques edge, threshold, region and hybrid methods are briefly described followed by segmentation methods based on Convolutional Neural Network (CNN).

2.2 Traditional Methods

Common conventional techniques of image segmentation can be classified into four major categories i.e. edge based segmentation, region-based segmentation, thresholding and neural network-based segmentation (Figure 2.1). In this section, traditional segmentation methods are briefly

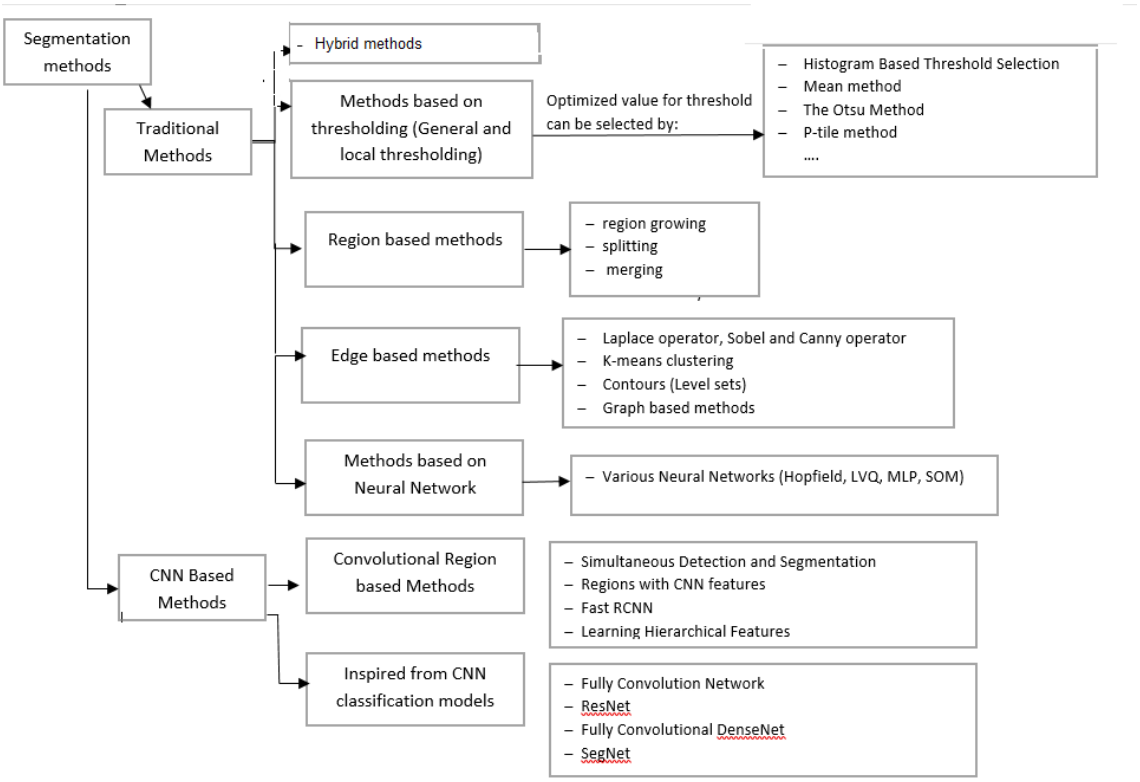

Figure 2.1: Classification of segmentation methods

reviewed followed by segmentation techniques in the medical area that are relevant to this research.

2.2.1 Methods based on Thresholding

Thresholding the image is the oldest and most commonly used segmentation method that can simply be shown as:

$$g(x,y) = \begin{cases} 1 \text{ if f(x,y)>T} \\ 0 \text{ Otherwise} \end{cases} \tag{2.1}$$

Where T stands for threshold value, and g(x,y) is the threshold version of f(x,y).

Image thresholding can be categorized into global threshold and local threshold. Global threshold considers a single threshold value for the whole image and each pixel's gray value will be compared with this threshold value while in the local threshold, the value of threshold appoints to each pixel due to the gray scale information of the neighbouring pixels (Al-Amri et al., 2010). Global techniques are really fast and they show great results for common scanned documents. Statistical methods such as clustering approaches have been the main technique for global thresholding.

However, these methods are inappropriate for complex documents, particularly when the document does not have uniform illumination. The Otsu algorithm is a common used global thresholding technique that seeks for a threshold to make the intra-class variances of the segmented image minimum. Over the years, various methods have been developed by researchers for global and local thresholding to select automatic optimized value for threshold such as histogram based threshold selection, edge maximization technique, mean method, P-tile method (Al-Amri et al., 2010). Peaks and valleys in the histogram of an image can be considered in finding the appropriate value for the threshold. While the mean value of the pixels is set as the value of threshold in mean method, in p-tile method the information of the area size of the desired object is used to conduct image thresholding. A fuzzy technique was also applied in order to select a threshold. Based on the concept of fuzzy sets and the description of membership function, a thresholding method was proposed to employ the measure of fuzziness to determine an adequate threshold value. In (Chaudhuri and Agrawal, 2010), a method based on a fuzzy framework reported that a thresholding histogram has been conducted according to the similarity between gray levels.

2.2.2 Edge based Segmentation

Since edges in an image involve important information, segmentation techniques can be applied to the edges retrieved from edge detection techniques and object contours are produced by connecting the edges. A typical procedure of edge detection based-segmentation includes calculating edge, processing the edge image to keep the border of the objects closed, and transforming the output to a typical segmented image by filling inside of the predicted object borders. Edge detection methods are used regularly as part of segmentation techniques particularly for images that consist of various objects, because the edges efficiently express the object's boundaries. Compared to region-based technique, this method has the benefit of not necessarily needing closed boundaries s. The edge-based methods have been applied to distinguish the discontinuity in intensity level, that constitute the image boundary. Image Segmentation based on edge detection methods use an edge detection operator i.e. Laplace operator, Sobel and Canny operator (Razali et al., 2014). (Fabijańska, 2011) introduced a new approach to define an edge position using a variance filter. The output of applying a variance filter has been clustered by K-means clustering into two classes of edge pixels and non-edge pixels. As a result, a wide edge is obtained to enter skeletonization algorithm to define one-pixel width edge.

2.2.3 Region Based Segmentation

In region based segmentation, the pixels with more similar characteristics are located in the same group to generate homogeneous segments. Various segmentation algorithms are available for region growing, splitting and merging Chaudhuri and Agrawal (2010), Tang (2010), Adams and Bischof (1994), Fan and Lee (2014). Various methods have been proposed for region growing. Fan and Lee (2014) introduced three techniques to enhance the performance of the Seeded Region Growing (SRG) method, which is a fast and powerful algorithm for region based image segmentation. In a simple common SRG method, a set of seeds such as a set of connected pixels are placed in the image to be segmented, followed by growing these seeds into regions with the idea of consecutively adding neighbouring pixels to them until the image is partitioned into segments that come from seeds (Shih and Cheng, 2005). Region splitting as another region based method, considers the whole image as a region that is subdivided iteratively due to homogeneity criteria (Ohlander et al., 1978). However in many studies, splitting has been applied as a first stage of split-merge algorithm that generally contains a decomposition method such as Quadtrees that applies for a splitting phase followed by a merging phase. (Chaudhuri and Agrawal, 2010) proposed an automated thresholding method that stands on bimodality detection, and is followed by rechecking through merging technique between the two neighboring regions. They proposed an algorithm that uses the density ratio of the neighbor pair regions to handle the issues caused by the splitting method.

2.2.4 Neural Network Based Methods

Artificial Neural Networks (ANN) have been primarily inspired by human being's central nervous systems. Over the years, neural networks have shown their incredible ability to solve problems in several applications in prediction, recognition and detection systems. The models stand on neural networks to analyze a small region of an image by applying an ANN. Finally, the regions of an image are classified by the neural network. Neural networks involved in image segmentation are Hopfield, Multi-Layer Perceptron (MLP), Self-Organizing Map (SOM), and Pulse Coupled Neural Network (PCNN) (Cheng et al., 1996), (Cuevas et al., 2009), (Xiao et al., 2009), (Iscan et al., 2009). This section will discuss studies that relied on ANN for image segmentation. (Zhao et al., 2010) designed a segmentation technique that uses textural features and neural network. They

applied Gray Level Co-occurrence Matrix (GLCM), and three parameters of textual features i.e. uniformity, energy and diagonal moment to dental micro-CT images. A back propagation neural network with two hidden layers was also applied as well as the pre-processing technique of de-noising, filtering and image sharpening. Results have demonstrated that this technique is superior in segmentation than other methods such as the thresholding and the region growing methods in terms of speed and accuracy of segmentation. (Teimouri et al., 2014) presented a segmentation algorithm for discriminating almond images which belonged to different classes including normal almond, broken and split almond, shell of almond, wrinkled almond and double or twin almond. Sensitivity analysis has been practiced to select the best features set of colour features from images. Finally, ANNs that consisted of an input layer, one hidden layer and three neurons as output were adopted to classify the images into object, shadow and background. They compared their result to the results obtained from otsu, dynamic thresholding and watershed methods and reported outperforming results. Cuevas et al., explored the implementation of Learning Vector Quantization (LVQ) network and a decision function for the face segmentation task (Cuevas et al., 2009). They showed by applying LVQ network to the pixels of an image and without using any dynamic model or probability distribution, not only did processing speed improve but also performance has enhanced particularly while images with various illumination were practiced. Probabilistic neural networks (PNNs) contained four layers: an input layer, a pattern layer, a summation layer, and an output layer. During learning, the biggest probability of each pixel to the category that it belongs, will be considered as the category of the pixel. Their experimental results showed that PNNs were more accurate in image segmentation in contrast to back propagation neural networks and MLPs.

2.2.5 Hybrid Methods

There are segmentation methods based on a hybrid of the methods discussed in the literature. (Rout et al., 1998) applied a method based on generating a threshold surface by interpolation of edge points using a Hopfield neural network for multi modal image segmentation. Another study that can be considered as a hybrid technique is a research that proposed an image segmentation algorithm that merges information derived from edges and regions with spectralmethods through a morphological algorithm of watersheds (Monteiro and Campilho, 2008). In the first step, their technique includes a rainfalling watershed algorithm that was used to divide images into primitive

regions. Finally, a region-based segmentation method relying on similarity graph representation of the image regions was applied. (Haider et al., 2012) also reported a method based on Pixel Neighborus Pattern Analysis (PNPA) to eliminate the influence of edge discontinuity. Firstly, a canny edge detector was performed followed by PNPA operation for edge enhancement. Then, an image segmentation method was developed using Kohonen's SOM artificial neural network to detect the main features of image and genetic algorithm to cluster the image into homogeneous regions.

2.2.6 Why Deep Learning?

A number of issues in traditional segmentation methods that deep learning successfully address, are outlined below:

- Feature engineering is very important stage in traditional methods. The extraction of effective features is not only a difficult and time consuming task, it also demands specialist attempts and experience. However, by increasing the number of features, a feature selection technique is also required to select the discriminative features. On the other hand, the main advantage of deep learning is automatic feature learning. Thus, there is no need to feature engineering. The network scans the data and learns the hidden patterns from the data by itself and with more data, a higher performance is achieved.

- Designing a segmentation method depends largely on data type, for instance edge-based algorithms do not perform well on images with smooth transitions and low contrasts. Edge based techniques are also highly sensitive to noise. Region based methods require initialization, and finding good start points. Moreover, good similarity criteria directly affect performance. However, a deep learning specifically convolutional neural network automatically learns the features and is not programmed, thus less expert analysis is required and the system is less dependent on data type.

- The traditional methods are not fully automated and mostly are comprised of various algorithms and supplementary processing steps while deep learning, specifically CNN solves the problem on an end-to-end basis, from general features such as edges and blobs to high level features like shapes are extracted with gradient-based learning applied to the whole system.

Figure 2.2(b) shows the workflow for deep learning and a traditional computer vision system in Figure 2.2(a). Deep learning methods have received a considerable attention in recent years and

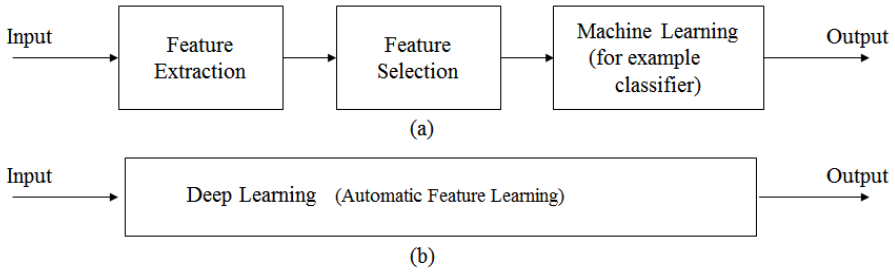

Figure 2.2: Comparison of traditional machine learning and deep learning, (a)Traditional machine learning flow, (b) deep learning flow

various techniques have been developed in this area. Deep networks have demonstrated remarkable success compared to the traditional methods in applications of medical image analysis (such as computer-aided diagnosis and medical image segmentation) (Wang et al., 2014),(Zhang et al., 2015).

2.3 Models based on Deep Learning

2.3.1 Introduction to Deep Learning-based Models

Generally, the methods that researchers use for a specific type of image may not works well on other types of images. Finding a solution that can be applied to different images and that is suitable for all the segmentation applications has been considered by researchers for several decades and it is still a challenging subject in image processing and computer vision. The main aim of deep learning is to automatically learn abstractions from low-level features to high-level representations with less dependency on hand-engineering features. Many studies in recent years have shown that not only deep learning techniques alleviate the need for manually engineered features, but they also produce a powerful representation that captures texture, shape and contextual information (Bengio and Lee, 2015). As the main aim of this study is working on deep learning algorithms, this section discusses recent models in which segmentation methods relied on deep learning. Deep learning methods have attracted considerable attention in recent years and various architectures have been developed. The first deep neural network, “Convolutional Neural Network” (CNN), was introduced in 1998 by LeCun (LeCun et al., 1998). Restricted Boltzmann Machines (RBMs)

introduced by Paul Smolensky in 1986 (Smolensky, 1986) were once again in spotlight after Geoffrey Hinton suggested fast learning algorithms for them in 2006. Moreover, by stacking RBMs and fine-tuning the resulting network another deep architecture was performed which is called deep belief networks((Hinton et al., 2006)), (Hinton, 2009). In recent years, several pattern recognition and machine learning contests have been organized and various high dimensional image datasets have become publicly available on the Internet. CNNs have demonstrated excellent performance on visual recognition issues. Researches demonstrated that deep CNNs have surpassed the traditional algorithms in these competitions, for instance, deep learning won MICCAI 2013 Grand Challenge and ICPR 2012 competition on mitosis prediction (Cireşan et al., 2013) and deep neural networks won the ISBI'12 challenge on segmenting neuronal structures. CNNs had originally been proposed for classification problems and various models based on convolutional layers have been proposed for object recognition, anomaly detection, image restoration, natural language processing and speech recognition. When it comes to segmentation, several methods have been proposed to apply convolutional layers of classification models to segmentation problem mostly based on patch-wise training. Extracted image patches are fed to CNN and center pixels will be classified. Consequently, various improved deep neural network models that were originally applied to classification, have been used for segmentation problems too. Furthermore, studies have been done to transform common convolutional classification models into segmentation tasks. In 2015, a Fully Convolutional Neural network (FCN) was designed to adapt CNN to perform the task of segmentation that consists of encoder layers/downsampling and corresponding decoder layers/upsampling (Long et al., 2015). Various studies applied the same downsampling layers as FCN that was composed of convolution, max-pooling and sub-sampling to provide the feature maps, and their novelties stand on enhancing upsampling path or used similar deconvolution layers and an enhanced encoder part with recently improved convolutional architectures. Convolutional neural network-based segmentation methods can be categorised into two major classes: region-based convolutional models and encoder/decoder models which are mostly inspired by FCN.

2.3.2 Region based Convolutional Neural Network Methods

Algorithms that work on patches of image instead of the whole image are very common in computer vision. As CNN was initially proposed for the task of classification, the main idea of the

using region based algorithms is to perform classification on a proposed region to which an object might belong. Various methods have been applied to generate the region proposals. Moreover, a number of studies motivated to generally use regions or multiscale version of input images to handle the scarcity of dataset with employing many applicable patches since deep learning needs large amounts of data. The following sections briefly describe some recent research relevant to patch-wise and multi-scale segmentation methods based on CNNs.

2.3.2.1 Regions with CNN Features (R-CNN)

Regions with CNN features (R-CNN) is an object recognition and segmentation system that combines multi-layer convolutional networks with region proposal method (Girshick et al., 2015). As Figure 2.3 indicates, the architecture contains category-independent region proposals generated using selective search method (Uijlings et al., 2013) to discern an available candidate for detection system. All pixels of each region candidate will be wrapped in a bounding box around to the required size, then a convolutional neural network is used to extract a fixed-length feature vector from each region. Finally, a class-specific linear Support Vector Machine (SVM) classifies the regions. They used various strategies to compute the features and showed that merging features of full region and the foreground will enhance the results.

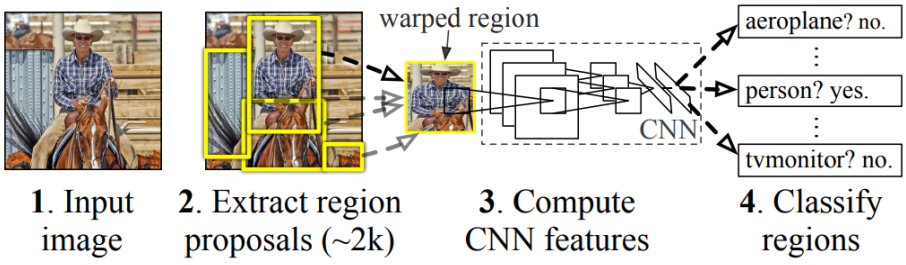

Figure 2.3: R-CNN: Region-based Convolutional Network (Girshick et al., 2015)

A supervised pretrained model for classification task is used for the training of this architecture and any of the CNN structures such as VGG, ResNet and AlexNet can be used for this purpose.

2.3.2.2 Simultaneous Detection and Segmentation(SDS)

The idea of SDS is similar to R-CNN but this method benefits of detecting all objects which place in a same class in an image. It includes the multiscale combinatorial grouping method to extract region proposals (Arbeláez et al., 2014) followed by feature extraction phase with CNN

from both region of interest and the foreground as shown in Figure 2.4. Compared to R-CNN, in this architecture two joint convolutional networks are used for feature extraction from the RoI and the foreground (Figure 2.4). SVM is trained on top of the CNN features to define a score for each class to every object. Finally, region refinement has been conducted using a Non-Maximum Suppression (NMS) technique on the scored candidates and the features from the CNN to produce category-specific coarse mask predictions to fine the candidates.

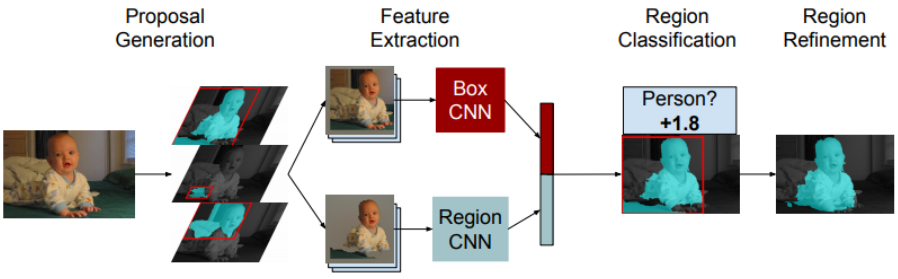

Figure 2.4: The pipeline of SDS method (Arbeláez et al., 2014)

2.3.2.3 Fast R-CNN and Faster R-CNN

The author of R-CNN proposed Fast R-CNN which improves the training speed and increases the detection accuracy (Girshick, 2015). The architecture shown in Figure 2.5 is similar to R-CNN but the whole input image is forwarded to the CNN unlike R-CNN, whereby a multitude of region proposals constitute the input of the network. In this method, the region proposals are produced by CNN and are resized by a RoI max pooling layer to enter into a fully connected layer. The spatial the spatial pyramid pooling layer used in (He et al., 2015) is applied for the RoI layer. A multi-task loss is used since Fast R-CNN is an end-to-end learning method which learns the class of object and the bounding box position and size. Each training RoI has two labels from ground-truth class (u) and from ground-truth bounding-box regression (v). Thus the multi-task loss L is defined as:

$$L(p, u, t^u, v) = L_{\text{cls}}(p, u) + \lambda[u \geq 1]L_{\text{loc}}(t^u, v) \tag{2.2}$$

in which $L_{cls}(p, u) = logp_u$ shows the error for true class and the error of second class. L_{loc}, is determined over a tuple of true bounding-box regression targets for class $u, v = (v_x, v_y, v_w, v_h)$, and a predicted tuple $t^u = (t^u_x, t^u_y, t^u_w, t^u_h)$, again for class u. The function $[u \geq 1]$ is equivalent to

1 when $u \geq 1$ and 0 otherwise. By convention the catch-all background class is labeled u = 0. The loss for the bounding-box regression is:

$$L_{\text{loc}}(t^u, v) = \sum_{i \in \{\text{x,y,w,h}\}} \text{smooth}_{L_1}(t_i^u - v_i), \tag{2.3}$$

in which

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise,} \end{cases} \tag{2.4}$$

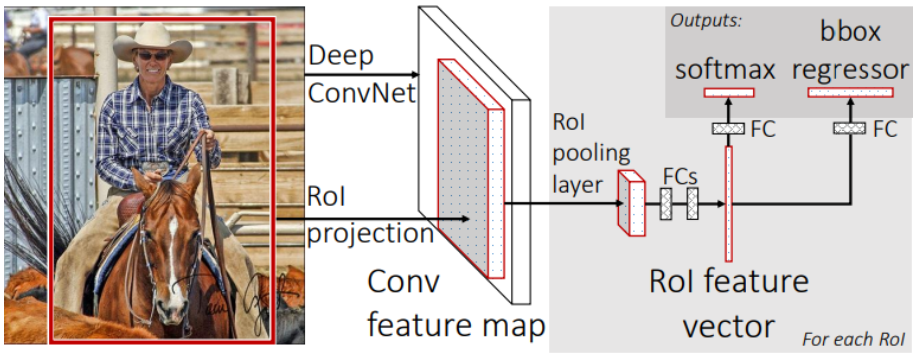

Figure 2.5: The Diagram of Fast R-CNN (Girshick, 2015)

Fast R-CNN benefits of end-to-end learning compared to R-CNN with multi stage architecture that was also much slower because the CNN had to process thousands of region objects for each input image. However, both methods are still slow overall, due to using an external method to generate region proposals. A further research called Faster R-CNN solved this issue by changing the architecture, whereby the network learns the region proposals (Ren et al., 2015). They introduced a Region Proposal Networks (RPNs) that shares the convolution layers with the object detection network. The region proposal networks are placed after the last convolution layer to generate region proposals from a convolutional feature map followed by RoI pooling layer and finally classification and regression blocks similar to fast R-CNN.

2.3.2.4 Learning Hierarchical Features

(Farabet et al., 2012) proposed the idea of multi-scaling convolutional representation in their paper for scene learning, a parallel architecture including CNN to generate feature maps and graph-based classification. A laplacian pyramid is a common method in computer vision, and is applied

to decompose image to multiple different scales, then each scaled image is forwarded through a convolutional network that generates a group of feature maps. The feature maps of all scales are merged after upsampling so that the coarser-scale maps to match the size of the finest-scale map. A parallel processing for over-segmentation such as training conditional random fields over super pixels and multilevel cut with class purity criterion has also been considered. A set of segmentations such as a segmentation tree is created over the image to process the picture at various levels and finally design a method to automatically restore from a group of segmentation pieces to produce the final image labelling.

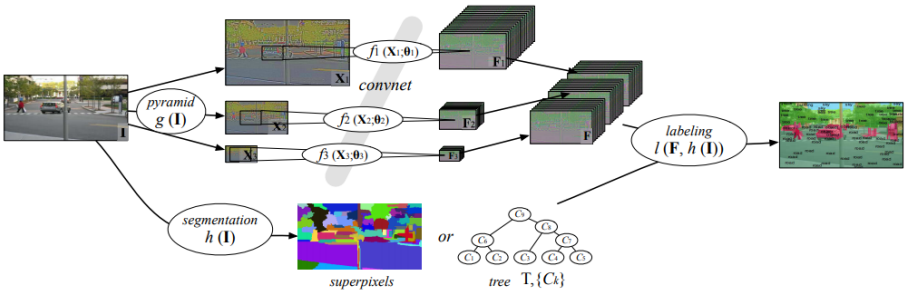

Figure 2.6: The pipeline of the scene parsing system (Farabet et al., 2012)

2.3.2.5 Recurrent Convolutional Neural Networks

(Pinheiro and Collobert, 2014) introduced a recurrent convolutional network that contains a composition of P instances of the convolutional network *f()*. As Figure 2.7 shows, a convolution network takes an image as the input and predicts low resolution patches. Then the process will repeat with the output predictions of the previous instance of the network and the down sampled version of the original image. In this architecture, a sequential series of networks share the same set of parameters. Three convolutional networks with shared parameters make the network which

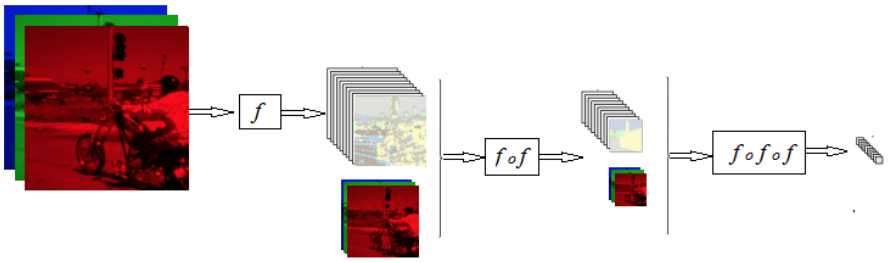

Figure 2.7: System considering one (f), two (f o f) and three(f o f o f) instances of the network. In all three cases, the architecture produces labels (1 1 output planes) corresponding to the pixel at the center of the input patch (Pinheiro and Collobert, 2014)

is trained by maximizing the likelihood:

$$L(f) + L(f \circ f) + L(f \circ f \circ f) \tag{2.5}$$

where L(f) is a shorthand for the likelihood:

$$L_f(\mathbf{W}, \mathbf{b}) = -\sum_{I_{(i,j,k)}} \ln p(l_{i,j,k}|I_{i,j,k}; (\mathbf{W}, \mathbf{b})) \tag{2.6}$$

the parameters (W, b) of the network f(·) are learned in an end-to-end supervised way, by minimizing the negative log-likelihood over the training set (2.6). $l_{i,j,k}$ defines the correct pixel label class at position(i, j) in image I_k.

2.3.3 Methods Inspired from CNN Classification Models

Deep CNN was initially proposed for classification and CNN based models soon found their way into many research proposals including object recognition,localization and segmentation. Pioneer models were mostly region based methods mainly seeking to use a convolution network to generate feature maps from region proposals or classification of the regions followed by localization and refinement. A Fully Convolutional Network (FCN) contributes to adapting the standard deep CNN classifier into a segmentation model by changing fully connected layers of classification models to convolutional layers followed by upsampling with deconvolutional (fractionally strided convolutions) layers (Long et al., 2015). The FCN can be considered a keystone of semantic segmentation that uses deep neural networks. Thereafter, studies inspired by FCN were conducted to convert CNN-based classification models to segmentation. Consequently, new architectures were proposed which consisted of encoder recent classification models in the downsampling line and an FCN inspired upsampling model for the decoder part. Some of these models are briefly described below.

2.3.3.1 Fully Convolutional Neural Network

The Berkeley Vision and Learning Centre reported using convolutional neural network for semantic image segmentation (Long et al., 2015). The idea was to convert fully connected layers to convolution layers besides merging features with outputs of corresponding former layers named

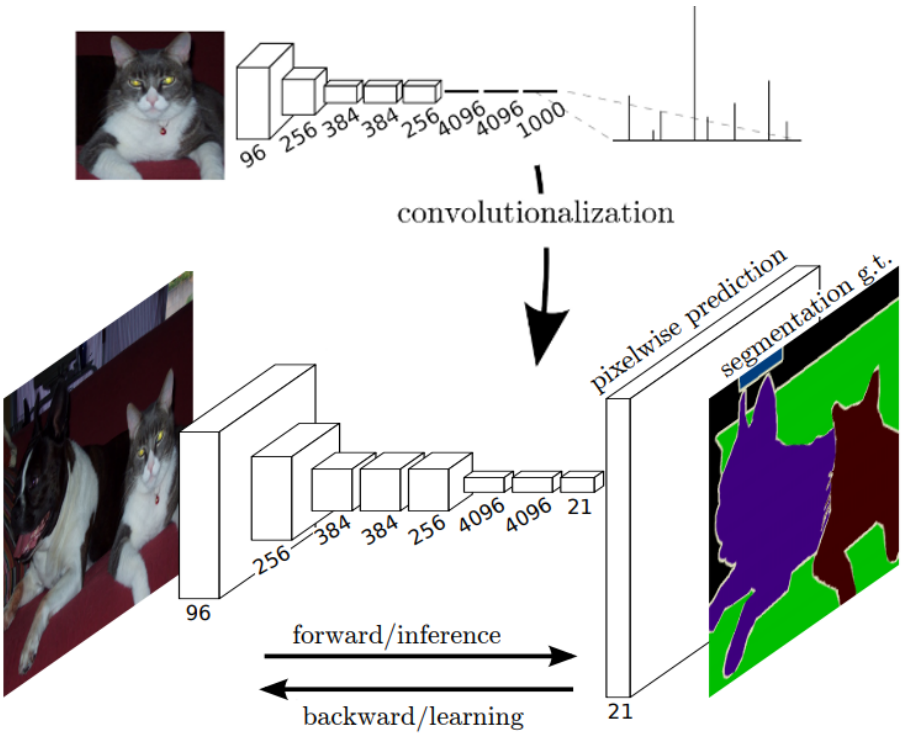

Figure 2.8: Transforming fully connected layers into convolution layers. Fully convolutional networks can efficiently learn to make dense predictions for per-pixel tasks like semantic segmentation. (Long et al., 2015)

skip connections (Figure 2.9). Down-sampling and upsampling ways are linked, thus expanding the path, retrieve more information from encoder layers. This architecture takes a 2D image as input and a 21-class semantic segmentation of that images provided as an output. In order to move from classification to segmentation, they decapitate the net by ignoring the fully connected classifier layer, and replaced these fully connected layers to convolution layers. The 1×1 convolution layer was also changed to a channel dimension of 21 in order to estimate scores for each class including background, followed by deconvolution layer which performs bilinear upsampling to make pixel-wise output from the coarse outputs. Learnable deconvolution filters were introduced to upsample feature maps. The deconvolutional filters of final layer are fixed to bilinear interpolation, while intermediate upsampling layers are initialized to bilinear upsampling, and then learned. They adapt recent common classification networks (AlexNet,the VGG net, and GoogLeNet) into fully convolutional networks and transfer their learned representations by fine-tuning to the segmentation task. Although with going deeper, the more detailed deeper features are provided, the output of convolution is also smaller and spatial location of shallower layers will lost.The skip connections enhanced the result by fusing coarse, high layer information with fine, low layer in-

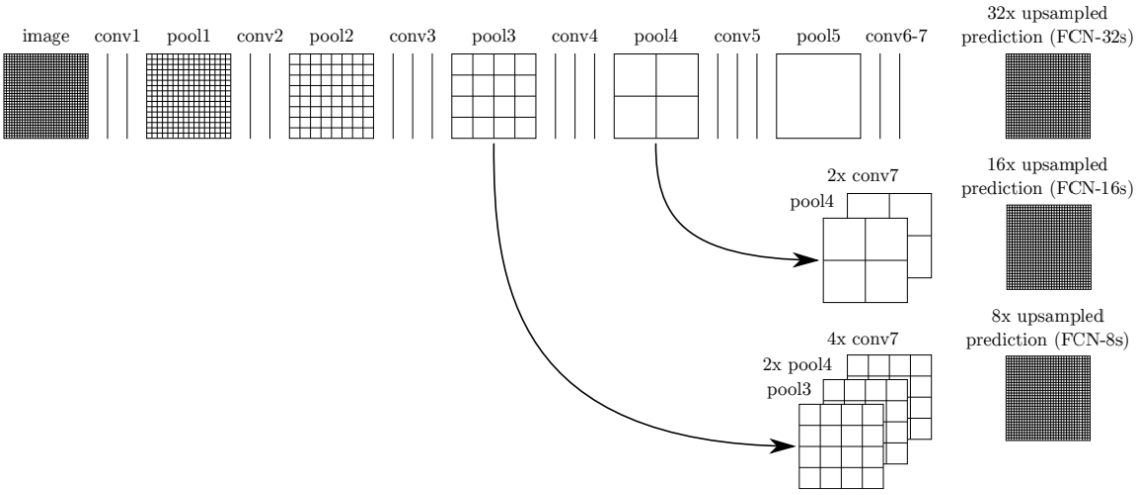

Figure 2.9: The skip connection: combining coarse, high layer information with fine, low layer information.

formation. The possibility of making the receptive field grow while the number of parameters are

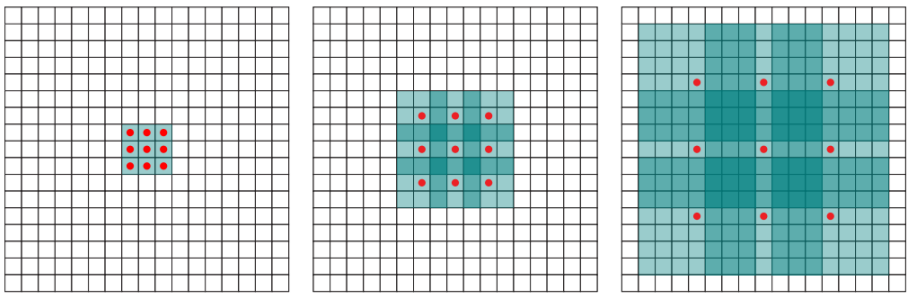

Figure 2.10: (a) F1 is 1-dilated convolution with a receptive field of 3×3. (b) F2 is produced from F1 by a 2-dilated convolution and the receptive field increased to 7×7. (c) F3 is produced from F2 by a 4-dilated convolution with receptive field of 15×15 Yu and Koltun (2015)

not increasing reveals the strength of this architecture specifically in medical analysis where the datasets are often scarce and training a network with many parameters is not feasible.

2.3.3.2 SegNet

Another segmentation method called SegNet also consists of encoder layers and corresponding decoder layers followed by a soft-max classifier (Badrinarayanan et al., 2017). In this architecture, for each sample, the indices of the max locations calculated in the pooling are stored. The novelty of this technique is that the decoder upsamples the feature maps by using the stored pooled indices from the corresponding encoder. The input image is reconstructed by convolution of the sparsed upsampled maps with trainable filter banks. The final feature maps in the decoder path are entered

into a soft-max classifier for pixel-wise classification.

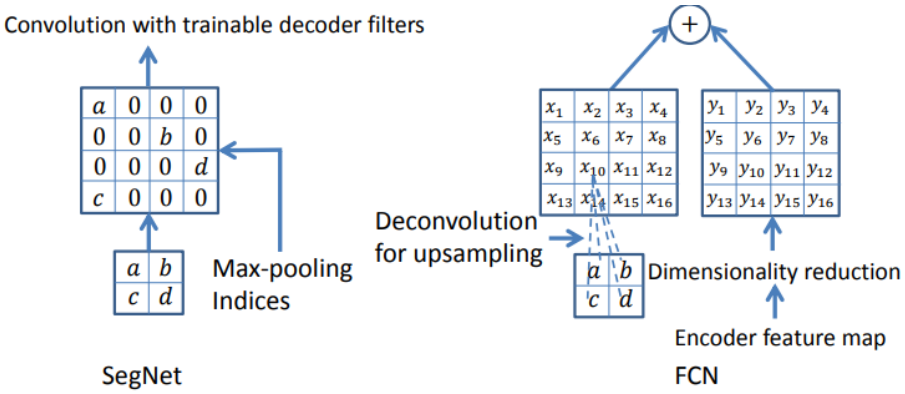

Figure 2.11: SegNet and FCN decoders. a, b, c, d correspond to values in a feature map (Badrinarayanan et al., 2017)

Figure 2.11 displays the decoder parts of FCN and SegNet. SegNet uses the max pooling indices to upsample the feature maps and convolves with a trainable decoder filter bank while in FCN upsampling is conducted, then the provided feature map fused to the output of max pooling layer in the encoder part. This feature map was resized to the size of upsampling output before fusion.

2.3.3.3 UNet

U-net architecture comprises of convolution layers followed by a rectifier linear unit and max pooling operation for downsampling that provide the feature map for upsampling path (Ronneberger et al., 2015). In this method a convolution layer comes after each upsampling layer, then this layer will concatenate to the corresponding feature map from the downsampling path that is cropped to have the same size of this layer. The architecture is similar to FCN but in the upsampling there are a large number of feature channels, which allow the network to propagate context information to higher resolution layers. So, the encoder part is completely symmetric to the decoder part that makes a u-shaped model. Feature maps transferring via skip connections are also concatenate to corresponding upsampled feature map instead of fusion as would be the case in an FCN.

2.3.3.4 Reseg

ReSeg is based on an image classification model called ReNet. Three ReNet layers composed of four recurrent neural network have been proposed that are followed by upsampling layers. They

have applied ReNet layers on top of the first layers in FCN, to have generic local features and ReNet layers that get local features by first sweeping the image horizontally, followed by sweeping the output of hidden states vertically.

First, the initial layers of VGG-16 network receives the input image and initialized with pre-trained on ImageNet then without fine tuning in this stage, they keep the resolution of image. ResNet layers receive this output feature maps. In the last stage, the last feature maps will be resized by upsampling layers to the same resolution of the input and a softmax non-linearity used to predict the probability distribution over the classes for each pixel. In this architecture, the recurrent layer constitute the main part of architecture which is made by multiple RNNs. Each recurrent layer is composed by 4 RNNs coupled together (Visin et al., 2016).

2.4 Segmentation Research on Medical Data

This section briefly discusses segmentation methods on medical data relevant to chapter three, four and five. Various algorithms of image analysis have been proposed to assist clinicians in early diagnosis of skin cancer over years. Dermoscopic feature based algorithms such as ABCD rule that includes Asymmetry,Border, Color, and Dermoscopic structure (Nachbar et al., 1994) and CASH (Color, Architecture, Symmetry, and Homogeneity) (Henning et al., 2007) are primary methods that have been used for many years.

Moreover, common segmentation techniques including edge or region-based methods (Wong et al., 2011), (Tajeddin and Asl, 2018), (Jaisakthi et al., 2018), (Lau et al., 2018), thresholding techniques (Zortea et al., 2017) and methods based on features from transform domain such as wavelet and Fourier transform (Garnavi et al., 2012) have been developed for the task of skin lesion segmentation. (Jaisakthi et al., 2018) proposed a technique including edge and region-based algorithm for skin lesion segmentation. Illumination enhancement and artifact removal such as hair and air bubbles constitute the initial stage as the pre-processing phase. The Grabcut algorithm that uses both edge and boundaries information was applied for segmentation followed by further stages including K-means clustering and the flood-fill technique, to segment the lesion area with enhanced boundaries. The result (Jaccard index) was lower than that of the winners of challenge 2017 (Codella et al., 2019). Another popular region-based method termed watershed was developed

in various algorithms for medical segmentation tasks. (Masoumi et al., 2012) proposed using the watershed algorithm and MLP neural network for feature extraction in an iterative process. The extracted features from both techniques were compared and the error was computed in each iteration to sequentially adjust the required parameters of the algorithm. In addition, morphological smoothing, Gaussian filtering and morphological gradients were used at this preprocessing stage but no post-processing was conducted. A popular method that has been widely developed for image segmentation is active contour composed of deformable contours that adjust to variety of shapes. The method includes an energy maximization procedure built on region or edge based models have been employed for segmenting several kinds of medical images including ultrasound imaging, CT, and MRI of different organs in the body (Ciecholewski, 2016),(Riaz et al., 2018). In research (Riaz et al., 2018), to generate the initial curve, adaptive thresholding was applied and an optimization problem was proposed to maximize the Kullback–Leibler divergence of gray level distribution between background and lesion. A recent research proposed saliency map generated by improved discriminative regional feature integration (mDRFI) (Jahanifar et al., 2018). This method was also composed of multiple stages including pre-processing such as colour constancy and hair removal, generating an initial mask by thresholding the saliency map based on the DRFI method and a final mask formation using a distance regularized level set evolution (DRLSE) framework. They extended regional property descriptors and proposed a pseudo-background region to improve the DRFI method but the result was still lower than in the highly ranked papers of both ISBI2016 and 2017 challenges. (Tajeddin and Asl, 2018) proposed adding new texture features of peripheral regions for classification and a segmentation technique composed of estimating initial contour and propagating it with an iterative process based on dual component speed function. Otsu's method is followed by morphological process conducted to generate the threshold initial mask. They used a level set framework and proposed two component speed function for the image gradient and the color probability distribution of pixels to generate the final mask. General shape features based on common ABCD features, colour-based features and texture related features from luminance channel of Labcolour space implied for feature extraction phase. They also proposed a textural feature set from peripheral region that is based on masks from the segmentation phase. A variety of pre-processing methods were applied to remove hair, marks and eliminate dark corners, correct image illumination and crop images regarding to the masks. This segmentation method ranked 5th in ISBI challenge 2016. Techniques based on super-pixels also

have been extensively used in medical image segmentation. The super-pixel is an efficient method to segment images by partitioning the image into groups of connected pixels that have similarities (Nguyen et al., 2018), (Navarro et al., 2018). In the research (Navarro et al., 2018), the common Simple Linear Iterative Clustering (SLIC) method is improved by focusing on segmenting the ROI precisely instead of segmenting whole image accurately. Firstly, feature points are detected in the image by SIFT operation and then Gaussian distribution is applied to place initial centres followed by applying SLIC to these centres. The result showed marginally higher Jaccard index compared to the top results of ISIC 2017 challenge. All traditional methods that were mentioned in previous section, are composed of multiple stages such as pre-processing, initialization, edge/region extraction or various techniques for feature extraction while deep learning methods benefit from receiving the input as a raw image and generate the output via an end-to-end learning process. Another drawback of traditional models is that the discriminative features play an important role in success of these models. Extracting effective features is a difficult task which requires high level knowledge. A variety of algorithms have been suggested to extract features regarding the image structure of medical images, but these algorithms mostly deal with particular features of an image that may not work for all kinds of images. For instance, the low contrast between the lesion and the background would not contribute to an accurate thresholding method, weak or noisy edges deter the performance of edge-based segmentation models and active contours build upon an initial contour that limits the efficiency of model. However, the deep convolution network has the advantage of learning the features automatically from general features such as edges and lines extracted in first convolutional layers to high level features like shapes extracted in higher layers. Deep learning algorithms have shown remarkable progress in various computer vision tasks. Notably, CNN has outperformed conventional methods in several pattern recognition and machine learning domains. CNN is introduced as a deep neural network architecture composed of more layers in comparison to shallow conventional neural networks. More convolution layers enable the network to learn more complex features. CNN was originally proposed for the task of classification and extensive research was conducted to design efficient deep architectures. Recently, several studies modified CNN-based networks that are designed for classification problems to be applied to segmentation tasks. Early proposed CNN models for segmentation were based on classifying super-pixels or the region surrounding a pixel (Farabet et al., 2012), (Ciresan et al., 2012).

In 2015, FCN was designed to adapt the classification model to perform segmentation. An end-to-end pixelwise learning architecture was proposed in which fully connected layers are transformed to convolution layers so the network has spatial output maps (Long et al., 2015). In FCN, the fractionally strided convolution is introduced as upsampling, also called deconvolution. Various subsequent studies were conducted to improve FCN for different segmentation problems. U-net ("Convolutional Networks for Biomedical Image Segmentation") is a popular model used in medical area. It has symmetric encoder-decoder architecture that includes deconvolution layers with a larger number of feature channels, each followed by a concatenation of feature maps from the corresponding layer in a contraction path (Ronneberger et al., 2015). A Residual Network (ResNet) is proposed to efficiently increase the depth of a convolution network by introducing shortcut connections of identity mapping that connects the output of each layer to a higher layer (He et al., 2016). The encoder part of Refine-net (Lin et al., 2017) is based on this model and the decoder contains multi-level Refine-net blocks that fuse the features received from the encoder as well as the features from the previous Refine-Net block. A very deep convolutional network composed of fifty layers is used in (Yu et al., 2016) and residual learning is applied to deal with overfitting. (Yu et al., 2016) proposed a fully convolutional residual network (FCRN) for task of lesion border segmentation and their experiments ranked second in segmentation task of ISBI 2016 challenge. (Bi et al., 2017) designed a model based on FCN that contains FCNs in multi-stage structure (mFCN). In each stage, the FCN receives inputs including the original input image and the estimated output of previous stage. They also integrate the segmentation results of all stages in a parallel way and their result slightly outperformed the best results of the challenge (Codella et al., 2019). The full resolution convolutional networks (FrCN) proposed by (Al-Masni et al., 2018), removed all sub-sampling layers in the encoder part and considered each pixel as a training sample. Although this method benefits from not including any pre/post processing techniques or artifact removal, the high computation load for this technique is due to not using pooling layers, which requires computational resources. Resizing the input images is an alternative that they considered but decreasing the resolution leads to loose information as well.

A large amount of research on medical video segmentation is inspired from image segmentation methods applied on video frames. Thresholding and edge detection methods were applied in (Huang et al., 2011) to segment the left ventricle from short axis cine cardiac MR images. In

(Jolly, 2006), localization techniques such as maximum discrimination and thresholding is used followed by region segmentation and active contours. Thresholding, region growing and active contours constitute major part of LV segmentation studies which are not based on machine learning models, (Codella et al., 2008), (Queirós et al., 2014), (Kaus et al., 2004). The KNN classifier is applied to predict pixel class after feature extraction (gradient magnitude, the largest eigenvalue, the output of median filter, and the gray value) in (Hadhoud et al., 2012). Recent studies have applied CNN-based methods, specifically FCN and Unet on frames of videos as well. (Yan et al., 2018), proposed an Optical Flow Feature Aggregation sub-network which is integrated into the Unet and is further developed by dilated convolution. A method proposed by (Khened et al., 2019), employed densely connected convolutional neural networks for LV segmentation. They have applied Fourier analysis and circular Hough transform to detect the region of interest and in the deep network they proposed short-cut and long connections similar to skip connections. (Dong et al., 2018) proposed an FCN based architecture with feature fusion across different layers and a residual module for LV segmentation from three-dimensional echocardiography. By converting 3D dataset to 2D image slices, the input samples for training a deep network increased and further data augmentation was conducted with the rotate and resize functions. Finally, they conducted segmentation with transfer learning on a pretrained VOC model and based on the coarse segmentation results, they proposed a fine segmentation method based on 3D initialization and the 3D snake model.

2.5 Introduction on Convolutional Neural Networks and Evaluation Metrics for Segmentation

In recent years, Artificial Intelligence has grown dramatically through CNNs. The proposed methods in this research are focused on developing CNN based architectures for tasks of detection and segmentation. After a short introduction on deep learning concept, a CNN architecture was the most common deep learning method in computer vision that has been discussed. In order to train deep convolutional networks, various frameworks have been created and special hardware for large computational operations in terms of memory are required. Information on hardware and frameworks used in this research, are described in the following sections.

2.5.1 Architecture of Convolutional Neural Network

A CNNs overall architecture looks like a typical neural network that includes neurons that have learnable weights and biases. The inputs of neurons update with weights (dot product) and followed by a non-linearity function . The whole network still expresses a single differentiable score function: from the raw image pixels on one end to class scores on the other. The last fully-connected layer includes a loss function (e.g.Softmax) and all the learning process is the same as for regular neural networks. However, CNNs benefits from getting images as input and the architecture is improved so that the neurons in layers of a Convolution Network are placed on three dimensions.

Building a typical CNN involves stacking various layers including convolution, pooling, and fully connected layers. Convolution layers are connected to a local area of the previous layer instead of connecting to all neurons to which fully connected layers are connected. CNNs therefore have fewer parameters due to the local connectivity compared to regular feed forward networks with similar sized layers. Pooling is a downsampling operation along with the spatial dimensions and are fully-connected layer with an output of the class scores. The parameters of the convolution layer and fully connected layers are trained with gradient descent. In the following sections, convolution and pooling layers and the activation function are described.

2.5.1.1 Convolution Layer

CNN has the advantage of receiving an input layer in a shape of 2D information layer and a neuron within any layer is linked to a small region of the previous layer named receptive field where the filter applies, and convolution is computed. The convolution layer is the most important layer with highest computation in the network. This layer includes a group of learnable kernels that slide over the width and height of the image to conduct a convolution operation simply as dot products between the kernel and the input at each location. The result is a 2D feature map for each filter sliding over the image and finally, the set of these feature maps is concatenated along the depth dimension to make the output. The output of convolution operation for position (i,j) in a feature map is:

$$\sum_{m=0}^{k=1}\sum_{n=0}^{k-1} x_{si+m,sj+n}K_{m,n} + b, \tag{2.7}$$

where x is the location in the preceding layer, K represents the kernel and k the size of kernel. The stride (s) is the number of pixels that the filter skips while sliding over the image. There are three parameters to define the size of the output volume: stride, depth that is the number of filters and zero-padding which is applied to put zero values around the border of the input volume and can be considered as a parameter to set the spatial size of the output volumes. The spatial size of the output volume is $(W - F + 2P)/S + 1$ in which S is stride, W is the size of input, F is the filter size and P is the zero padding parameter.

2.5.1.2 Pooling Layer

The pooling layer also termed downsampling, is usually applied after convolution layers to decrease the spatial size of the feature map and consequently decrease the number of parameters that lead to less computation. Pooling applies on each depth slice of the input and resizes it. Generally, there are three forms of pooling: max pooling, average pooling and L2-norm pooling. The output of max pooling is the maximum number in each area that the filter convolves while the average pooling computes the average in the region. Assuming a 2D matrix X_{i_j} as the output of the specific patch of convolution layer, for each item in feed forward or backward pass the L2-norm is calculated as (Rezaei et al., 2017):

$$
\begin{aligned}
Forward :\mid x_{i,j} \mid &= \sqrt[2]{\sum {x_{i,j}}^2} \\
Backward : \partial \mid x_{i,j} \mid &= \frac{n\partial(\sum x_{i,j})}{2\sqrt[2]{\sum {x_{i,j}}^2}}
\end{aligned}
\tag{2.8}
$$

2.5.1.3 Activation Function

Rectifier Linear Unit (ReLU) is a type of activation function that comes after the convolution layer and it is the most commonly used activation function in CNN. Mathematically, it is determined as $y = max(0, x)$ and no exponential calculation is needed compared to tanh and sigmoid activation function, it is simply implemented (LeCun et al., 2015).

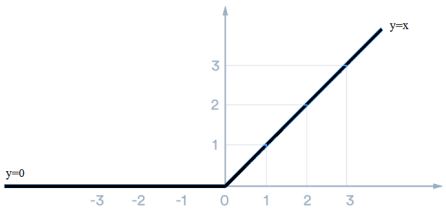

Figure 2.12: Rectifier Linear Unit activation function

2.5.2 Hardware Required to Train a CNN

A Graphics Processing Unit (GPU) is considered a necessary part of deep learning. Training the deep CNN and storing the trained model are computationally intensive tasks because of the large number of parameters. A GPU provides more logical cores and a higher bandwidth to retrieve a larger amount of memory at once compared to CPU. Huge parallel computing is another capability that makes a GPU useful for training deep neural network Li et al. (2016). GPU GeForce GTX TITANX was used to implement deep architectures in this research.

2.5.3 Deep Learning Frameworks

A number of libraries have been designed for implementing deep CNNs on GPUs. Common machine learning frameworks such as , Tensorflow (Abadi et al., 2015), Torch (Collobert et al., 2011), Theano (Al-Rfou et al., 2016), Caffe (Jia et al., 2014) have their own GPU libraries for CNNs. Caffe is a very common deep learning framework that has been extensively used by machine learning experts. In this research Caffe and Tensorflow libraries are used for coding. Caffe (Convolutional Architecture for Fast Feature Embedding) is a deep learning framework that is established by the Berkeley Vision and Learning Center (BVLC) and by community contributors that is released under the BSD 2-Clause license (Jia et al., 2014). It is originally written in C++, with command line, Python, and MATLAB interfaces and supports both GPU and CPU-based acceleration computational kernel libraries. TensorFlow is developed by the brain team of Google's intelligence research division for machine learning and deep learning research.It is an end-to-end open source platform for machine learning. It has many flexible tools and libraries which researchers can use to develop machine learning applications[1].

[1] https://www.tensorflow.org

2.5.4 Evaluation Metrics

The following five different metrics have been used to compare the performance of the methods studied in the paper;

$$Sensitivity(SE) = \frac{TP}{TP + FN} \tag{2.9}$$

$$Accuracy(AC) = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.10}$$

$$Specificity(SP) = \frac{TN}{FP + TN} \tag{2.11}$$

$$Dicecoefficient(DI) = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{2.12}$$

where TP, TN, FP, FN, are true positive, true negative, false positive, and false negatives respectively and performance metrics are computed at the level of single pixels. A true positive shows a pixel that is correctly predicted to be in a class (according to the target mask) while a true negative defines a pixel that is correctly recognized as not to be in the given class. The results of the challenge ISIC are ranked based on the Jaccard index. The Jaccard index also known as Intersection over Union measures the similarity and diversity of the sample sets. The Jaccard coefficient shows similarity between sample sets, and is determined as the size of the intersection divided by the size of the union of the sample sets, by considering A as test groundtruth and B as output groundtruth:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{2.13}$$

The Jaccard index can also be determined based on TP, FP and FN as shown below:

$$Jaccardindex(JA) = \frac{TP}{TP + FP + FN} \tag{2.14}$$

2.6 Summary

This chapter reviewed traditional and deep learning based methods. Traditional methods are composed of multiple stages such as initialization, edge/region extraction or various techniques for feature extraction while deep learning methods benefit from receiving the input as a raw image and generate the output via an end to end learning process. Another drawback of traditional models is that discriminative features play an important role in the success of these models. Extracting effective features is a tedious task that requires expert attempts too. A variety of algorithms have been suggested to extract features regarding the image structure of medical images, but these algorithms deal mostly with particular features of the image that do not necessarily work for all kinds of images. For instance, low contrast between lesion and the background would not contribute to an accurate thresholding method, weak or noisy edges deter the performance of edge-based segmentation models and active contours build upon an initial contour that may limit the efficiency of model. However, the deep convolution network has the advantage of learning the features automatically from general features such as edges and lines extracted in first convolutional layers, to high level features like shapes extracted in higher layers. This automatic feature learning makes deep learning an efficient solution for medical diagnostic tasks that are highly dependent on specialist knowledge. Moreover, deep learning methods with the idea of using a large amount of raw data as input, have the ability to successfully address the issue that medical systems often have with various input types of 2D/3D/4D medical images taken by different instruments and experts. However, training the deep learning models with medical data meet challenges that were discussed in section background, thus this study focuses on the deep learning segmentation model in medical area.

Chapter 3

A Hybrid CNN-based Model for Skin Lesion Analysis towards Melanoma Detection

3.1 Data Preparation

The data analyzed in this project comes from the challenges of ”Skin Lesion Analysis towards Melanoma Detection” (Gutman et al., 2016), (Codella et al., 2019) that leverage datasets of annotated skin lesion images from the International Skin Imaging Collaboration (ISIC) archive. The ISIC archive contains the largest publicly available collection of quality controlled dermoscopic images of skin lesions.

The dataset contains a representative mix of images of both malignant and benign skin lesions. The dataset was randomly partitioned into both training and test sets, with 900 JPEG colored images in the training set and 379 images in the test set. The size of both train and test images varies from 722*542 to 4288*2848. The masks for both training and test sets was generated in PNG format and are the same size as their corresponding lesion image. Ground truths are 8-bit PNGs and each pixel is either: 0: belongs the background class, 255: belongs the image foreground, or the region inside the lesion. Moreover, 5-Fold cross-validation was used. So, the training dataset was randomly divided into five exclusive subsets. The algorithm ran 5 times and each turn, one of

the five parts was considered the validation set and the remaining part was the training set.

The data set for dermoscopic feature segmentation includes 807 lesion images, each paired with two binary masks (1614 binary mask images, separate masks for the dermoscopic features of ”globules” and ”streaks”) that present locations of the globules and streaks. Dermoscopic features were provided in the ISIC 2016 for task of lesion dermoscopic feature segmentation. 335 images are supplied as test data. A sample image from the dataset with all masks is provided in (3.1).

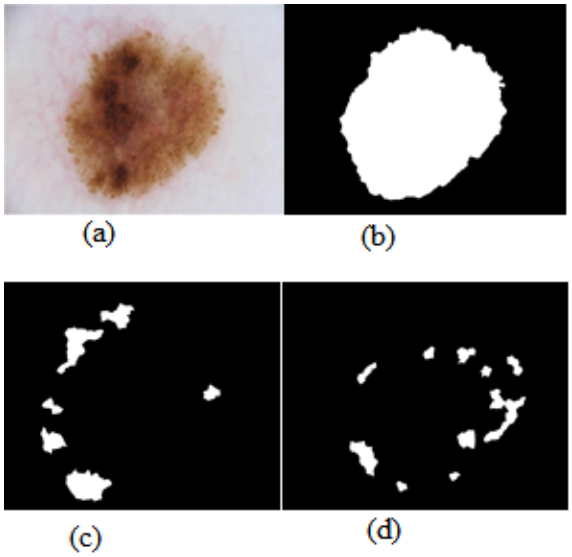

Figure 3.1: (a)Sample images from ISIC dataset (b)The groundtruth for skin lesion segmentation (c) The mask for feature segmentation (Globules) (d)The mask for feature segmentation (streaks).

The Caffe framework defined in Chapter 2, requires datasets in various formats including Lightning Memory-Mapped Databases(LMDB) and HDF5. Although HDF5 is a simpler format to read/write, LMDB is more common due to better I/O performance and good performance with large datasets Jia et al. (2014). In this research, LMDB format is used. Caffe stores, and employs the information as blobs. Data layers load input and save output by converting to and from Blob to other formats. The conventional blob dimensions for batches of image data are number N $\times$ channel K $\times$ height H $\times$ width W. For instance, with batch of 8 and using RGB images with size 500$\times$500, the images stored in LMDB are: 8$\times$3$\times$500$\times$500. To make data ready for training, image data for train and test/validation are converted into separate LMDB datasets.

3.2 The Proposed Hybrid Model for Tasks of Skin Lesion Segmentation and Dermoscopic Feature Segmentation

Fully Convolutional Networks (FCN) with the idea to transform fully connected layers into convolution layers following by upsampling layers allow for a state-of-the-art configuration for semantic image segmentation (Long et al., 2015). As identified in the literature, FCN benefits from inputs of any size and generating correspondingly-sized outputs. This chapter proposes, a FCN-based deep convolutional neural network to address two main segmentation tasks in melanoma diagnosis, a lesion border segmentation followed by a lesion dermoscopic feature segmentation including the dermoscopic features of "globules" and "streaks". Details of the proposed model and its evaluation on a database from the 2016 ISBI challenge (Gutman et al., 2016) are presented below.

Generally, a deep convolutional network can be trained from scratch when the data set is large or by using a pretrained network and applying for transfer learning when dataset is scarce. In this part of the research, transfer learning and training from scratch have been used, this will be described in next chapter. Transfer learning consists of using a pretrained model and fixing some layers and retraining the rest of the network, for example keeping fixed the earlier layers that contain general features, such as edge, and only fine-tune the later layers which extract higher level features, such as shapes (Donahue et al., 2014). The proposed model in this chapter is inspired from FCN. As the dataset only contains 900 images, transfer learning from a model pretrained on millions of natural images is investigated. However, due to the dissimilarity between the skin lesion dataset and ImageNet (natural images), fine-tuning is conducted to keep the earlier layers and to retrain the later layers. The pre-trained models applied in this study were namely (i) FCN-AlexNet and (ii) VOC-FCN8s.

- FCN-AlexNet with 7 convolutional layers. The first work that popularised Convolutional Networks in computer vision was the AlexNet (Krizhevsky et al., 2012). This model was based on semantic segmentation research done by the UC Berkeley Vision and Learning Center (BVLC) (Long et al., 2015).
- Pascal VOC-FCN8s with 15 convolutional layers: a standard recognition model that was benchmarked with detection and semantic segmentation research by BVLC. (Long et al., 2015).

Architecture of the network that is fine tuned on a pre-trained model called FCN8s is provided in Figure 3.2. The parameters of the network include stride, filter size and zero padding as shown in Table 3.1.

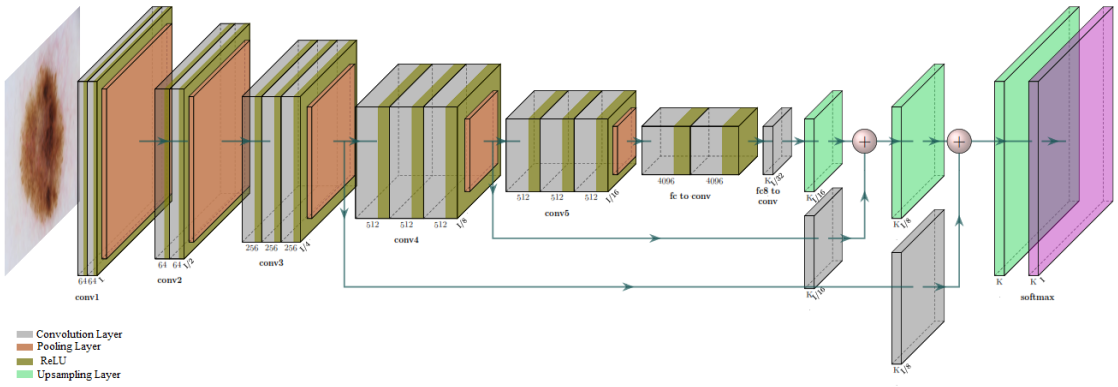

Figure 3.2: The outline of the proposed deep network architecture for lesion segmentation and dermoscopic feature segmentation.

Table 3.1: Parameters of the network FCN

Layer	Filter size	Stride	Number of Filters	Pad	Layer	Filter size	Stride	Number of Filters	Pad
Conv1_1	3×3	1	64	100	Conv4_3	3×3	1	512	1
Conv1_2	3×3	1	64	1	Pool4	2×2	2	-	-
Pool1	2×2	2	-	-	Conv5_1	3×3	1	512	1
Conv2_1	3×3	1	128	1	Conv5_2	3×3	1	512	1
Conv2_2	3×3	1	128	1	Conv5_3	3×3	1	512	1
Pool2	2×2	2	-	-	Pool5	2×2	2	-	-
Conv3_1	3×3	1	256	1	Conv6	7×7	1	4096	0
Conv3_2	3×3	1	256	1	Conv7	1×1	1	4096	0
Conv3_3	3×3	1	256	1	Score (Conv)	1×1	1	2	0
Pool3	2×2	2	-	-	Deconv1	4×4	2	2	-
Conv4_1	3×3	1	512	1	Deconv2	4×4	2	2	-
Conv4_2	3×3	1	512	1	Deconv3	16×16	8	2	-

The deep network includes convolution and pooling layers followed by deconvolutional layers. A normalization layer termed Local Response Normalization (LRN) is applied since the activation function after the convolution layer is a Rectified Linear Unit (ReLU), which is non-saturating (see Chapter 2). ReLU significantly reduces training time compared to the hyperbolic tangent function ($y = tanh(x)$) which is more common in neural networks. The inter channel LRN across

the channel is demonstrated in (Krizhevsky et al., 2012). For $a^i_{x,y}$ as the activity of a neuron is computed by applying kernel i at position (x, y) and then applying the ReLU nonlinearity, the response-normalized activity $b^i_{x,y}$ is:

$$b^i_{x,y} = a^i_{x,y} / \left(k + \alpha \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} (a^j_{x,y})^2 \right)^{\beta} \tag{3.1}$$

where N is the total number of kernels in the layer and the constants (k,α,β,n) are hyper-parameters. The effect of reducing the filter size or stride of pooling layer to improve the accuracy was investigated and a larger stride was selected in convolutional layers to decrease the size of the feature map. Furthermore, the stride of the last pooling layer was reduced to provide higher resolution inputs for upsampling layers. This led to a higher computational load but no considerable improvement in accuracy.

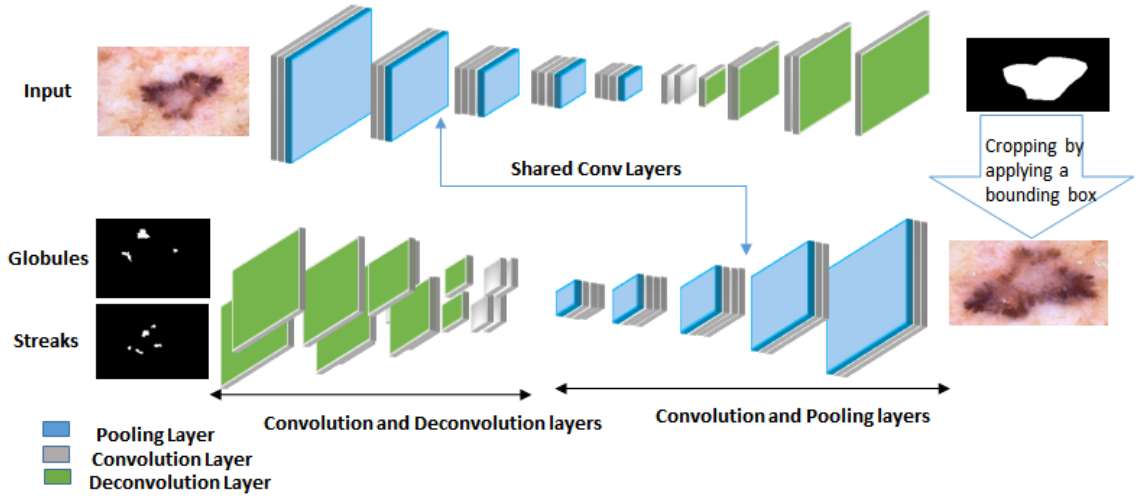

Figure 3.3: Outline of proposed deep network architecture for the tasks of lesion segmentation and dermoscopic feature segmentation.

Lesion dermoscopic pattern detection is the second part of the model. Finding particular clinical dermoscopic features increases the accuracy of melanoma diagnosis. Globules and streaks are dermoscopic features used by clinicians to distinguish melanoma from benign skin lesions. Binary masks were provided as training data to identify where features are present in lesions, and the goal was to automatically generate these masks for the test data as shown in Figure 3.1. Ground-truth images from the task of lesion border segmentation were used to extract the RoI of images by

applying a bounding box to the predicted lesion region. The images are cropped by a factor of 1.1 times, the size of the RoI provided by binary masks, because streak features can be ejected at the periphery of the lesion (shown in yellow in Figure 3.4).

Binary masks, which are detected by lesion border segmentation, and in the testing phase these were used to crop test images before entering the network. Convolutional layers were initialized with a pre-trained model from the previous part that helped the network converge fast. All layers of the AlexNet model were fine tuned because the pretrained model was trained on a large dataset of natural images, which was very different to our dataset. For the deeper model though, early convolution layers (conv1_1 to conv1_3) were fixed, because initial layers are supposed to retrieve more general data. This architecture is followed by two parts, each contains two convolutional layers and four deconvolutional layers to predict masks for both streaks and globules features. Moreover, fusion, convolutional and crop layers were used to concatenate upsampling layers with previous shallower pooling layers that are not shown in Figure 3.3 for the sake of simplicity. For this purpose, the idea in (Long et al., 2015) was used to combine the high layer information with the low layer information.

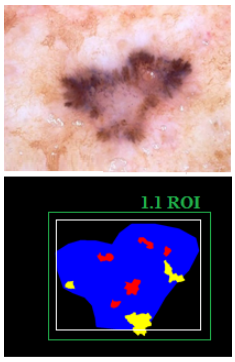

Figure 3.4: The lesion border mask is shown in blue. Globules and streaks are red and yellow respectively.

Cross entropy and the dice loss function were investigated. By considering pixel i in image X with n pixels, the Dice loss (which is computed for each class mask c and averaged to calculate the final score), is defined as:

$$L_{dice} = 1 - 2\frac{\sum\limits_{i \in X_n^{(c)}} P_n^{(i)(c)} Y_n^{(i)(c)}}{\sum\limits_{i \in X_n^{(c)}} (P_n^{(i)(c)} + Y_n^{(i)(c)})} \tag{3.2}$$

where $P_n^{(i)(c)}$ presents the probability of pixel i belonging to the class c, $Y_n^{(i)(c)}$ is the groundtruth of pixel i. The multiply and sum are element-wise and Σ is calculated over all pixels of image. Cross entropy is the most common loss in deep learning, is defined as:

$$L_{mce} = -\sum_{i \in X_n} Y_n^{(i)} \log(P_n^{(i)}) \tag{3.3}$$

L_{mce} push the prediction $P_n^{(i)}$ from segmentation network to be close to the groundtruth label $Y_n^{(i)}$ (Huang et al., 2016).

3.2.1 Reducing the Filter Size of Pooling Layer or Even Removing Pooling Layers to Increase the Accuracy

The all Convolutional Net, which proposed a network architecture that would consist of convolutional layers and pooling layers, was discarded (Springenberg et al., 2014). It is recommended to use a larger stride in the convolution layer to decrease the size of the representations. Moreover, the effect of widening the receptive field was investigated and authors achieved the same result with layer composition rather than increasing the kernel size that led to a decrease in the number of parameters. A three stacked 3×3 also provides a 7×7 receptive field while a 7×7 layer had 81% more parameters than three stacked 3×3 layers. In this practice, the pooling layers after the two last convolutional layers were discarded in order to provide higher resolution inputs for upsampling layers.

3.2.2 Data Augmentation to Overcome Overfitting

The pixel size of both train and test images varied from 722×542 to 4288×2848. A resizing of the input images was considered due to memory limitation. This resizing to 500×500 may lead the loss of some information and deteriorate the accuracy, as some images are very large, (up to 2000×1500 pixels). To solve this problem, the images were cropped, operated on separately, and were merged with the sub-images together for evaluation after training the network. This can also be considered an augmentation technique that prevents overfitting and improves generalization. Flipping also was used to increase the size of the dataset. The dataset included both malignant and benign skin lesions and was randomly partitioned into both training and test sets, with 900 JPEG

images as the training data and 379 images as the test data. The number of images to be used for training increased to 7200 coloured images by cropping in two and flipping vertically, horizontally and both.

3.2.3 Drop out Layers to Prevent Overfitting

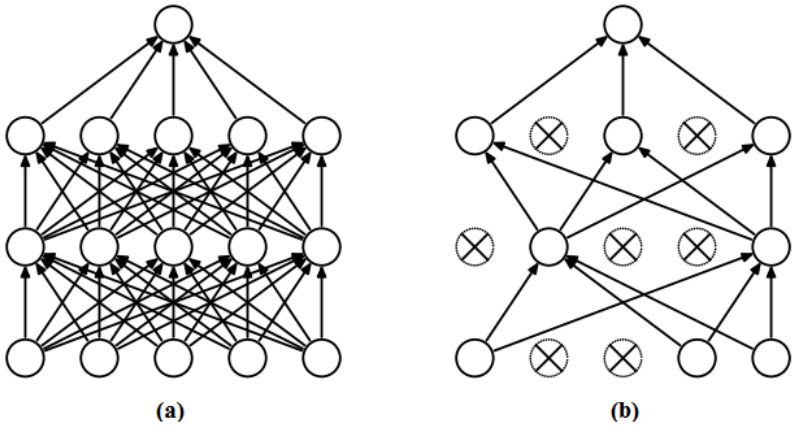

Figure 3.5: (a) Standard Neural Net with 2 hidden layers, (b) An example of a thinned net produced by applying dropout to the standard network. Crossed units have been dropped (Srivastava et al., 2014)

Another idea that was practiced to improve the training of CNN was called drop out, which applied to fully connected layers or after the max pooling layer (Srivastava et al., 2014). A thinner network would provided by dropping out units randomly in a neural network (i.e. temporarily removing it from the network), along with all its incoming and outgoing connections. Learning a fraction of the weights in the network, in each training iteration, offers a very computationally cheap and significantly effective regularization technique to delay overfitting and reduce generalization error in deep neural networks (Srivastava et al., 2014). By considering a neural network with L hidden layers whereby $l \in \{1, \ldots, \mathrm{L}\}$ shows the index of hidden layers of the network, the feed-forward operation with dropout is:

$$\begin{aligned} \widetilde{\mathbf{y}}^{(l)} &= \mathbf{r}^{(l)} * \mathbf{y}^{(l)}, \\ z_i^{(l+1)} &= \mathbf{w}_i^{(l+1)} \widetilde{\mathbf{y}}^{l} + b_i^{(l+1)} \\ y_i^{(l+1)} &= f(z_i^{(l+1)}). \end{aligned} \tag{3.4}$$

where f is any activation function and $*$ denotes an element-wise product. z is considered the vector of inputs into layer l and r is a vector of independent Bernoulli random variables each of

which has probability p of being 1. This vector is sampled and multiplied element-wise with the outputs of that layer, y, to create the thinned outputs $\widetilde{y}$. The thinned outputs are then used as input to the next layer and this process is applied at each layer as shown in Figure 3.6 (Srivastava et al., 2014).

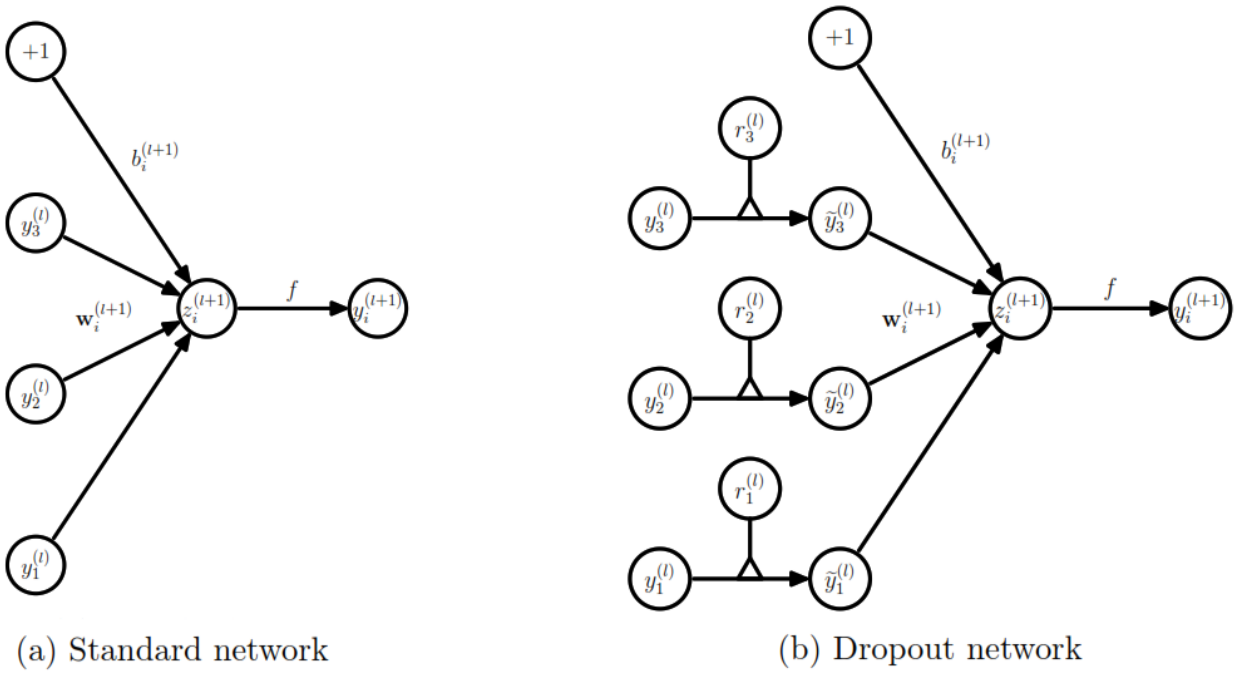

Figure 3.6: (a) Standard Neural Network, (b) Dropout Network (Srivastava et al., 2014)

3.2.4 Early Stopping to Handle Overfitting

Early stopping is a technique that helps avoid overfitting. The idea is to train the network to learn the pattern of input to output and measure the performance in training and validation set every few iterations until the validation set error reaches its lowest level and starts going up, while training error continues to decrease as shown in Figure 3.7. A stopping criterion from (Orr and Müller, 2003) was used to stop the training. By considering $E_{tr}(t)$ as the training set error and $E_{va}(t)$ as validation error, the value E_{opt}^t is defined to be the lowest validation set error obtained in epochs up to t:

$$E_{opt}(t) := min_{t<t'} E_{va}(t') \tag{3.5}$$

And generalisation loss to stop the training as soon as it exceeds a certain threshold (a) defines as:

$$GL(t) := 100(\frac{E_{va}(t)}{E_{opt}(t)} - 1) \tag{3.6}$$

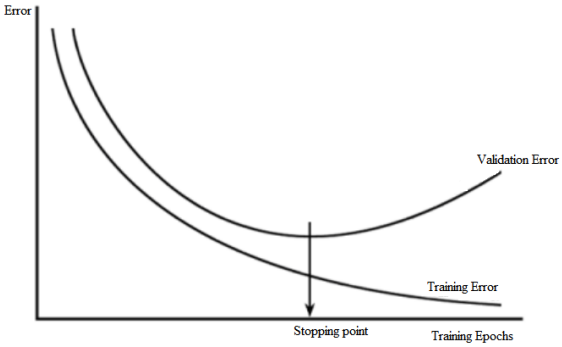

Figure 3.7: Early Stopping.

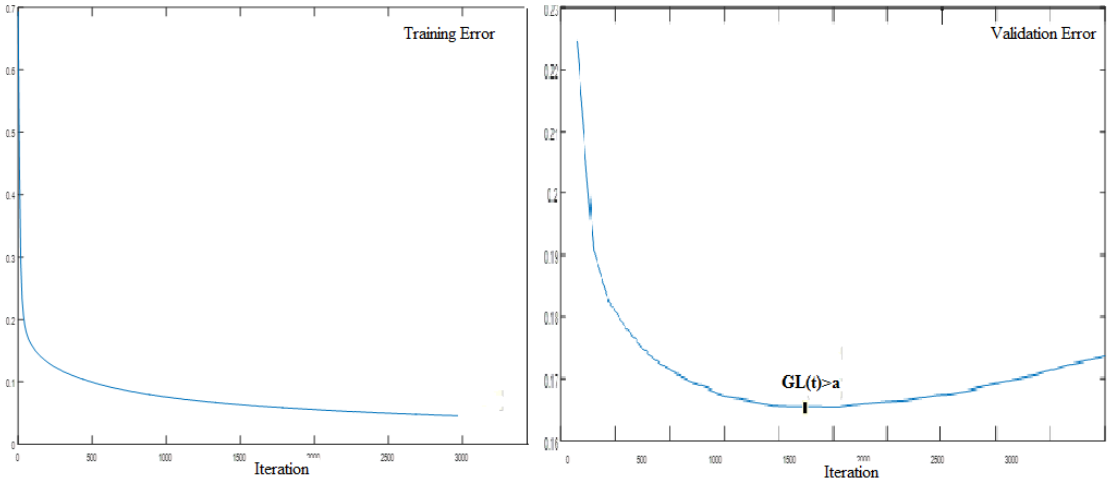

Figure 3.8: Stopping criteria with training error and validation error for an FCN-based network fine tuned on FCN8s for skin lesion analysis

Figure 3.8 shows the early stopping point for an FCN-based network fine tuned on FCN8s for task of lesion border segmentation.

3.3 Experiments and Results

In order to implement the deep learning-based concept for the detection of melanoma, a deep learning based network on lesion segmentation was utilized. This was further extended to a hybrid convolutional neural network designed for both tasks of dermoscopic feature segmentation and lesion border segmentation. The networks were trained with the stochastic gradient descent (SGD) method and hyper-parameters including learning rate, weight decay and momentum set at 0.001, 0.0005 and 0.9, respectively (although a large learning rate leads to faster learning, may deteriorate the convergence). On the other hand, with a very small learning rate, not only the training is slower, but it may lead to being permanently stuck with a high training error (Goodfellow et al., 2016).

Thus, as the most important hyper-parameter, the learning rate can not be very small or large and generally prior calculation to choose the best learning rate is not possible. (Bengio, 2012) recommended values between 1 and 10^{-6} for a neural network with inputs mapped to the (0,1) interval. Typical momentum values in practice are 0.5, 0.9, and 0.99 (Goodfellow et al., 2016). In this research, an adaptive learning rate, was used which is tracking the learning process and choose a smaller learning rate when the loss plateaus. Therefore, the training phase started with a learning rate of 0.001 and was reduced throughout the training. The kernel size was set at 3×3 and 2×2 for convolutional and pooling layers, respectively. Moreover, a 5-Fold cross-validation test was applied, which randomly split into five exclusive subsets. The algorithm runs 5 times subsequently, each time taking one of the 5 splits as the validation set and the rest as the training set. Evaluation metrics including Jaccard index and Dice were the most common metrics for segmentation. These are described in Chapter 2. Table 3.3 reports the results on model FCN32s which doesn't have the skip connections. Compared to the Alex-Net model in Table 3.2 (which is a shallower network and flipping conducted as augmentation), shows that by using cropping images in two to increase the data and resolution, and applying a deeper model, all metrics improved from 2% for accuracy to 10% for Jaccard index. Moreover, by using a model (FCN8s) that has skip connections (second row in Table 3.4), the Jaccard index improved by 6% compared to the model(FCN32s) (Table 3.3), that shows the benefit of skip connections as corroborated by the results in (Long et al., 2015) who reported that fusing the feature maps from primary layers with upsampling layers improves the result.

When it comes to overfitting issue, techniques including drop out layers and augmentation are investigated by comparing the Figures, 3.9 to 3.11. Studying Figure 3.9 and 3.10 shows that adding a drop out layer not only advanced the convergence time but also helped the network in terms of overfitting. Figure 3.11 depicts how data augmentation helps tackle overfitting.

Figure 3.12-a was produced by FCN-AlexNet model using original images as training data whereas

Table 3.2: Result of fine-tuning the network on Alex-net, Data augmentation is conducted by flipping

Metrics	Sensitivity	Specificity	Accuracy	Jaccard	Dice
Average Result	0.822	0.928	0.911	0.67	0.79

Table 3.3: Results of fine-tuning the network on pre-trained model of FCN32s, Data augmentation is conducted by flipping and cropping

Metrics	Sensitivity	Specificity	Accuracy	Jaccard	Dice
Average Result	0.887	0.949	0.932	0.77	0.855

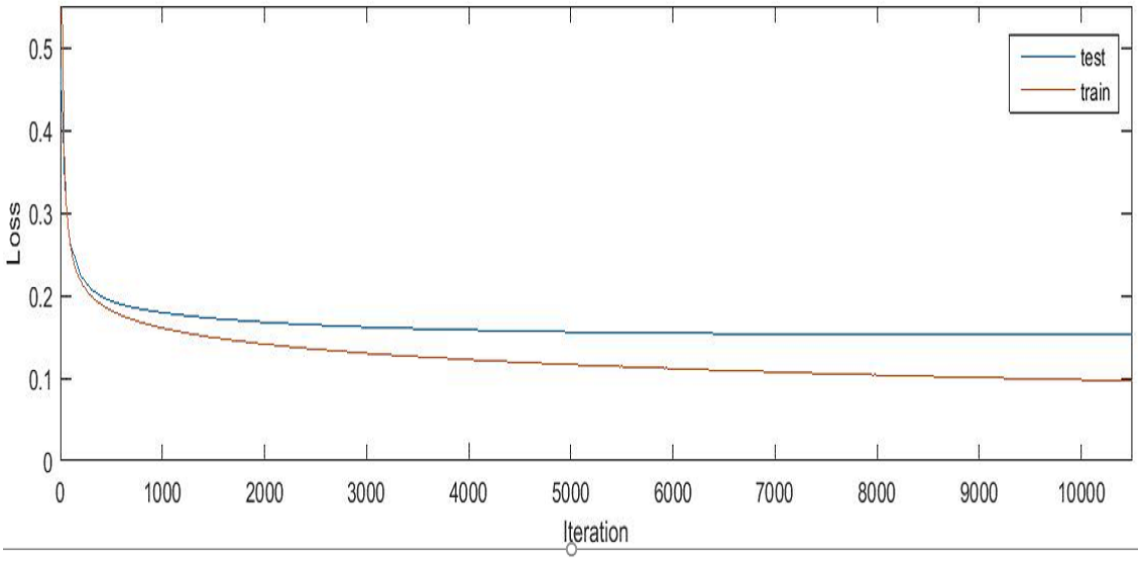

Figure 3.9: Loss for network fine-tuned on FCN-8s with flipped and cropped images as augmentation

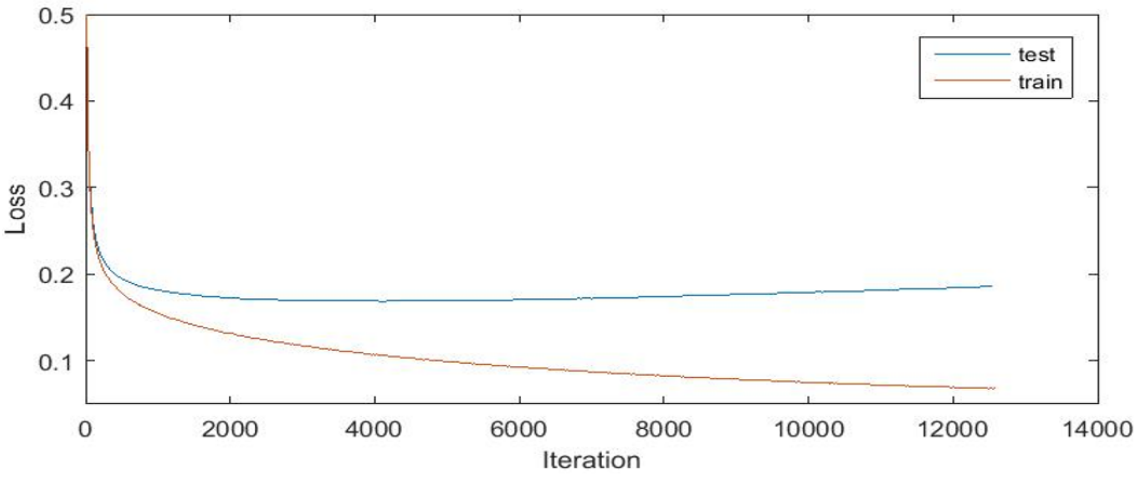

Figure 3.10: Output error, using FCN8s as pre-trained model without drop out layer

Figure 3.12-b and Figure 3.12-c are produced by the deeper model that employed the augmented training dataset which was created by cropping images. Thus, the generated masks for the cropped images of the network (b,c) make the output by merging. Image (d)is the original mask. This figure shows significant improvement in the generated output mask by deeper network which also

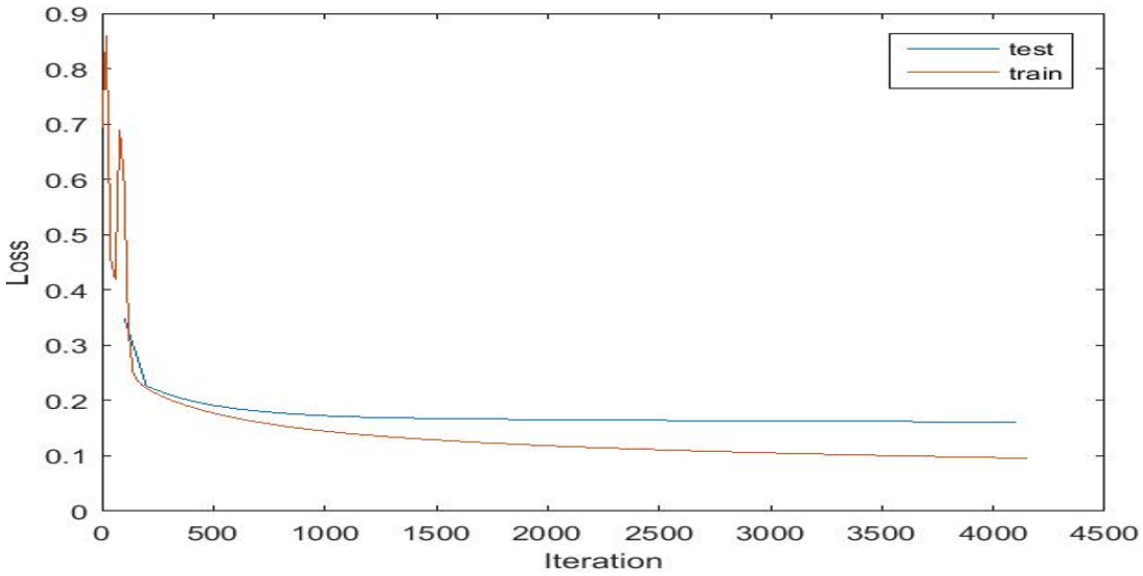

Figure 3.11: Output error, using FCN8s as pretrained model with drop out layer

benefited from augmenting the database by cropping the image.

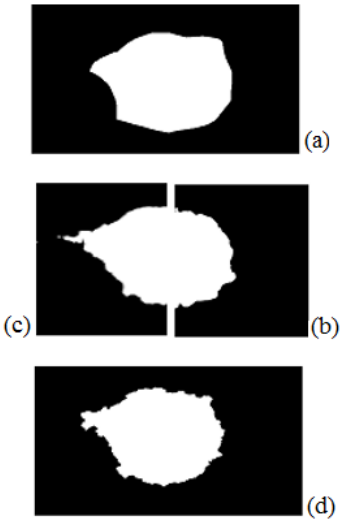

Figure 3.12: (a) Segmented image produced by the model pre-trained with Alex-net and data set augmented by just flipping, (b),(c) are output of the model in which, the input segmented images augmented by cropping and these outputs produced by a deeper network (16 convolution layers) using pre-trained model VOC-FCN8s, (d) Ground truth test image.

The evaluation results from Table 3.4 indicates the importance of augmentation (from 900 input images to 3600) which considerably improved the Jaccard index from 0.61 to 0.67. In addition to the augmentation, by going deeper (from 7 convolution layers to 16 convolution layers), the Jaccerd index score increased by 0.16 which was a significant improvement and was comparable

Table 3.4: Evaluation results of lesion segmentation compared to best challenge result(ISIC2016) (Gutman et al., 2016)

METRIC / METHOD	SE	SP	AC	JA	DI
Gutman et al.,2016	0.91	0.96	0.95	0.84	0.91
Deeper model with 16 conv layers augmentation by flipping and cropping (7200 training images)	0.91	0.95	0.94	0.83	0.89
Model with 7 conv layer augmentation by flipping (3600images)	0.82	0.93	0.91	0.67	0.79
Model with 7 conv layer without augmentation (900images)	0.75	0.91	0.89	0.61	0.74

to the top result of the ISBI challenge (Gutman et al., 2016). This is important as later layers extracts more complex features, however the augmentation technique is required to tackle early overfitting. The same indices have been applied to evaluate the dermoscopic feature segmentation task as well. The metrics calculated for this task are shown in Table 3.5. The results are comparable with the results of the second place winner in the 2016 ISBI challenge. The sensitivity of proposed method is slightly higher while other metrics are close to the 2nd best results. Figure 3.13 shows test ground truth for the globule feature and the globule feature segmented by our network.

3.4 Summary

Two main segmentation tasks in melanoma diagnosis systems were developed in this part of the research. An FCN-based deep network and a pre-trained model from the semantic image dataset ImageNet in (Krizhevsky et al., 2012)) were investigated for this medical application. Several

Table 3.5: Evaluation results for dermoscopic feature segmentation compared to the best result of challenge (ISIC2016)

METRIC / METHOD	SE	SP	AC	JA	DI
(Gutman et.al, 2016)	0.117	0.997	0.989	0.063	0.118
Proposed deep neural network	0.119	0.997	0.991	0.060	0.108

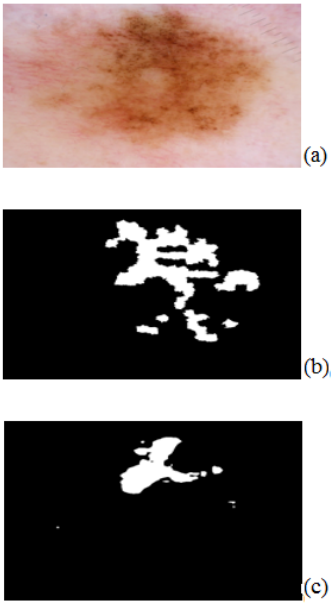

Figure 3.13: (a) Image from test dataset. (b) Test ground truth for globule feature. (c) Feature segmentation by our network

papers have used deep learning for lesion segmentation but not on dermoscopic feature segmentation. The proposed model yields promising results. Transfer learning was applied due to the small dataset but later layers were retrained because of the dissimilarity between skin dataset and ImageNet. This issue is particularly important for the deeper model for which the initial convolution layers could be fixed and the rest of the network be retrained. The earlier layers of a Convolution network contain more generic features such as edges that are generally helpful in many tasks while the later layers extract more specific details relevant to the dataset and pertaining to the classes, such as shapes. This research investigated to improve the FCN model but not designing a very deep network due to the overfitting problem. In many previous studies on skin lesion segmentation, considering data scarcity, massive data augmentation techniques were applied, or the models were fine-tuned on a model that was pretrained on irrelevant data (i.e. natural images and not medical images). Data Augmentation improved the Jaccard index and the efficiency of techniques. Adding drop out layers and early stopping were investigated to prevent overfitting. The computation load for dermoscopic feature segmentation was significantly decreased because of a shared convoultional network from lesion segmentation. In addition, the masks from the task of lesion segmentation were used to extract the region of interest to fed into the network in dermoscopic feature segmentation. In the validation phase, the masks provided by first task (lesion

segmentation) used to crop the RoI. In terms of augmentation, cropping was found to be very effective since the images were large and needed to be resized before forwarding to the deep network due to the memory issues. To solve this problem, images were cropped into two parts and were fed to the network. In test/validation phase, images are also cropped and then sub-images merged together to calculate the metrics. The final proposed method was composed of 15 convolution layers, which is significantly less complicated compared to FCRN with 50 layers (Yu et al., 2016), or mFCN (Bi et al., 2017) that contains FCN architecture in each stage.

Chapter 4

Proposed Contourlet-Convolutional Neural Network

4.1 Proposed Method

Training from scratch and challenges were investigated and a modified FCN model was proposed for lesion border segmentation and attribute detection tasks. An overview of the proposed segmentation architecture is depicted in Figure 4.1. Details of the method, including pre-processing, architecture of segmentation model, and post processing are explained below.

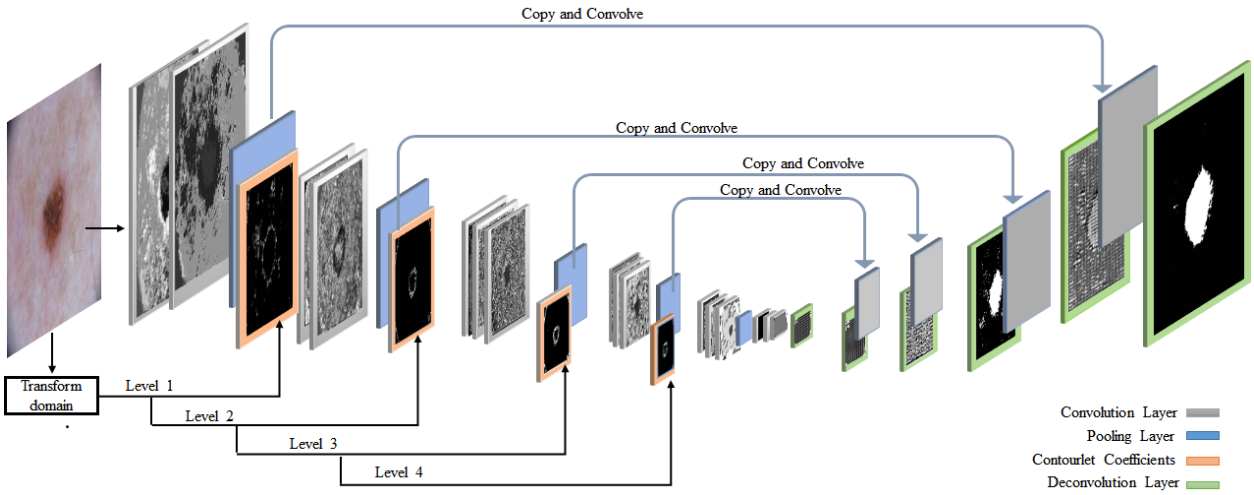

Figure 4.1: Architecture of proposed model, deep convolutional neural network proposed for lesion segmentation and image representations from various levels of contourlet transform.

4.1.1 Pre/Post Processing

Images of the surface of the skin mostly involve dark corners, hairs, ruler marks, variation in color, and uneven illumination. Removing noisy artifacts has extensively been investigated as an essential preprocessing task to enhance image quality (Oliveira et al., 2016). In this study, no preprocessing procedure for the task of lesion segmentation was conducted, but the dataset size was slightly increased by flipping the images. The effect of dark corners on generated masks is compensated in post processing and the model performs very efficiently on images with hairs and other noisy artifacts. For the task of lesion dermoscopic feature segmentation, the images are cropped first by applying a bounding box and using masks from the previous task, then the images are flipped to balance the dataset. In the post-processing phase, the masks are resized to the original size, then thresholding and morphological dilation are used to extract the objects in a predicted mask, choose an object closer to the center of image, and finally remove unwanted components, such as corner's effect, and cover the small holes (Gonzalez and Wintz, 1977).

4.1.2 Network Architecture

The modified CNN model proposed in this chapter is inspired from FCN and U-net and is trained from scratch. The model is improved by feature maps from the frequency domain as well as spatial feature maps of the original FCN. Contourlet as a multiscale transform was considered as it is multidirectional and sparser compared to wavelet, which is the most widely used transformation in image analysis. The proposed method consists of three parts: contourlet transformation which generates features in the frequency domain, concatenation of frequency and spatial domain features and convolutional neural network which is the main part.

4.1.2.1 Contourlet Transformation

Contourlet transform, with the distinctive attributes of directionality, anisotropy and localisation has demonstrated itself as an excellent performance in computer vision problems (Do and Vetterli, 2005). Contourlet represents notable features of image such as edges, curves and contours more efficiently than wavelet. Wavelet is based on point singularities and works adequately to detect edges but not smooth contours. In this research, multiscale and multidirectional coefficients from contourlet transform are integrated into the network, which added distinctive feature representa-

tions to the CNN. In contourlet transform, multiscale and directional decomposition is achieved by applying a combination of a Laplacian Pyramid (LP) and a Directional Filter Bank (DFB) (Do and Vetterli, 2005). The architecture is illustrated in Figure 4.3. The laplacian pyramid generates a down sampled low-pass sketch of an image and the difference between origin and the prediction that produces the band pass image. In (Do and Vetterli, 2005), a LP decomposition is proposed, inspired from the research (Burt and Adelson, 1987), (Figure 4.2) in which, the output of each LP level is a down-sampled low-pass version of the original (a[n]) and the difference between the original and the prediction that produces a band-pass image (b[n]). H and G are low-pass analysis and synthesis filters, respectively, and M shows the sampling matrix.

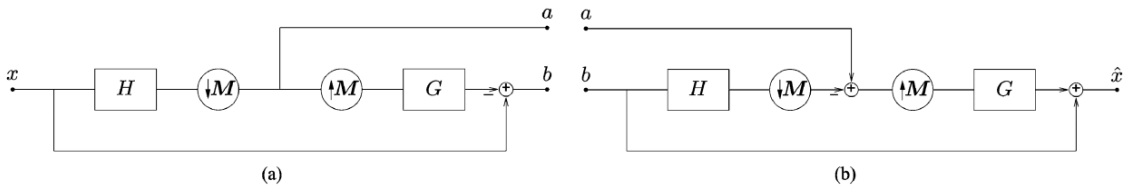

Figure 4.2: (a) The original LP decomposition from (Burt and Adelson, 1987), (b) LP decomposition proposed in (Do and Vetterli, 2005)

The directional filter bank receives the band-pass image as input and produces final directional decomposition. Consecutively, the Laplacian pyramid applies to the low pass image and the process is repeated to reach the desired level of decomposition. As demonstrated in Figure 4.3, the multidirectional representations generated by the directional filter bank in each scale are considered to concatenate with the equivalent pooling layer. By considering $x^0[n]$ as input image, the output of LP stage consists of J band-pass images $b^j[n]$, $j = 1, 2, ...J$, in the fine-to-coarse order and a low-pass image $a^j[n]$. In other words, the j^{th} level of the LP with the input image $a^{j-1}[n]$, generates $a^j[n]$ and band-pass $b^j[n]$ which is further decomposed by l_j-level DFB into 2^{l_j} band-pass directional images $C_d^j[n]$, d=1,2,..$2^{l_j} - 1$.

4.1.2.2 Contourlet-driven CNN

The proposed model is inspired from FCN and U-net, which are both composed of an encoder and a decoder path (Long et al., 2015), (Ronneberger et al., 2015). FCN has been designed to use the CNN of the classification task as in the supervised pre-training, and to fine tune the fully convolution network to perform the task of segmentation. In the FCN architecture, fully connected layers are converted to convolution layers besides adding the feature map from lower layers in the

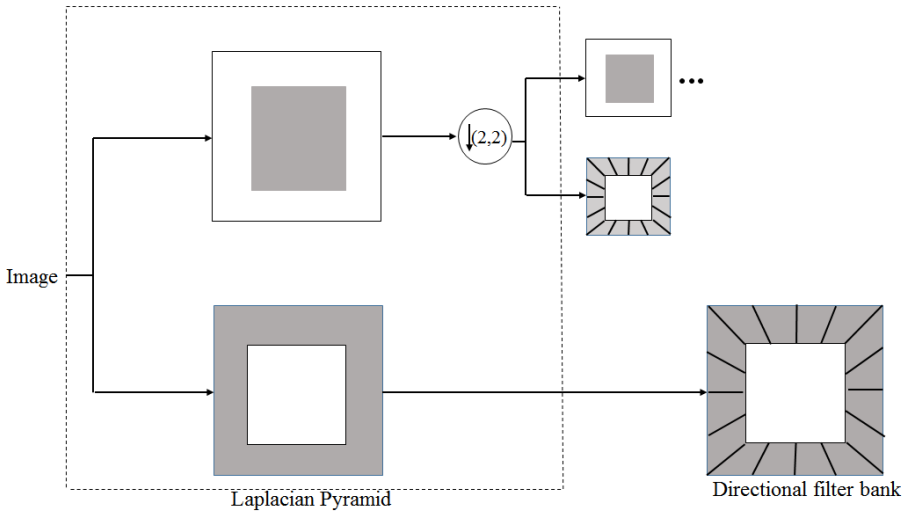

Figure 4.3: Contourlet transform composed of Laplacian Pyramid and Directional filter bank.

encoder path to the corresponding layer in decoder path that termed skip connections. The U-net (Ronneberger et al., 2015) modified the FCN by using the multitude of feature channels in the expanding path compared to the FCN, so that the number of kernels in the upsampling path is limited to the number of classes. Moreover, learnable filters are utilised and skip connections are concatenated to the corresponding upsampling layer instead of fusing in FCN. Thus, using the Unet architecture, requires more memory due to many feature maps being created because of the numbers of kernels. In Equation 2.7, the z^{th} feature map in the convolution layer *l* can be defined as:

$$h_z^l = \sum_{k=1}^{K} h_k^{l-1} * w_{kz}^l, \tag{4.1}$$

where $*$ denotes the convolution and *K* is the kernel size. $h^l$and w show the feature maps in convolution layer *l* and the kernel respectively. From the previous section, the output of j^{th} level of LP and DFB in contourlet can be described as:

$$F_{LP}^j(a^{j-1}[n]) = b^j[n], a^j[n] \tag{4.2}$$

$$F_{DFB}(b^j[n]) = C_d^j[n], d = 1, 2, ..2^{l_j} - 1 \tag{4.3}$$

By considering g as an activation function such as ReLU and p as the pooling function such as max/average pooling, the new feature maps generated by merging l^{th} convolution layer and the

contourlet transform can be described as:

$$h_{new}[n] = p(g(h^l[n]))©C_d^j[n] \tag{4.4}$$

The symbol © represents the technique of integrating CNN and contourlet representations. In contourlet, various scales and directional decomposition are generated depending on parameters set for LP decomposition and the number of directional decomposition in DFB. Also, in CNN, the size of feature maps is defined by the parameters of the convolution layer and the pooling layer, including stride and kernel size. Concatenation was considered as in this study, the size of the feature maps in CNN and contourlet were set to be the same in each scale. Alternatively, methods such as resizing or extracting the RoI could also have been considered. Moreover, feature fusion can be used instead of concatenating to reduce the computation load.

4.1.2.3 Lesion Segmentation

A basic model inspired by Unet and FCN is considered and gradually improved by injecting features from the transform domain and adding CIElab color model of input images. The initial model is composed of encoder and decoder elements. The encoder part of this model includes a series of convolution layers followed by max pooling layer. ReLUs are applied after convolution layers to accelerate the training (Krizhevsky et al., 2012). The number of feature channels starts from 16 in the first convolution layer and duplicates in each subsequent convolution layer. To deal with overfitting, a drop out layer is applied to the last two convolutional layers (Srivastava et al., 2014). In decoder side, a series of deconvolution layers operate as learnable upsampling layers. The number of kernels in the decoder part is considered equal to the number of classes: two. This model is trained from scratch and the learnable weights in both convolutional layers and deconvolutional layers are initialised with Xavier filters. As recommended in FCN, feature maps from pooling layers in the encoder path are connected to later deconvolution layers in the decoder path but in this study, concatenation is used similar to U-net. The number of deconvolution layers is also different from Unet and FCN. There are 6 upsampling layers to get the output with the same resolution as the input image. Moreover, the number of filters considered to be 16 for the first convolution layers compared to 64 in FCN and U-net, to decrease the depth and number of the training parameters since the data set is small. To improve performance, representations

from contourlet transform in four levels are concatenated to the pooling layers that make model2. Discrete contourlet transform with 4 decomposition levels and 4 directions in each level are applied to three color channels of the input images, in total they provided 12 representations in each level. In another experiment, changing the Model 1 to a deeper model (increasing the number of convolution layers in encoder part from 7 to 15) was investigated instead of integrating with image representations of the transform domain. This model was named Model 3. The architecture of the Model 2 was improved by using a deeper model, this was termed Model 4 (which had 15 convolution layers in the encoder part instead of 7 convolution layers) and due to high correlation between the red, green and blue colors in RGB, CIELAB color channels were also applied to the input of final final model (model 4 with CIELAB). CIELAB contains a lightness component (L), and two-color components (A and B) and has the benefit of being device independent compared to RGB.The 4 models are briefly defined as:

- Model 1: The model with 7 convolution layers in the encoder part.
- Model 2: The model with 7 convolution layers the in encoder part and that is integrated with representations of the transform domain.
- Model 3: The model with 15 convolution layers in the encoder part.
- Model 4: The model with 15 convolution layers in the encoder part and that is integrated with representations of the transform domain.

4.1.2.4 Lesion Attribute Segmentation

Segmentation of dermoscopic features including globules and streaks helps clinicians diagnose melanoma from benign skin lesions. The goal of this task is to automatically generate two masks for each lesion that reveal the location of streaks and globules. For this task, the segmentation model contains two parts including the encoder and the decoder. The encoder part is the same as for the task of lesion border segmentation. Therefore, transfer learning was considered from task1. The encoder path consists of two parts for globules and streaks localization each of which includes two convolution layers and four learnable upsampling layers, and so two losses are added up to generate the final loss of the network. The main issue with training the model for this task is unbalanced data. Nearly half of the images in this dataset do not contain any dermoscopic features

and the detection of empty masks can improve performance. Thus, classification (to classify empty and non empty masks) was added to the network and the corresponding loss added to the segmentation loss. The idea was to use a max pooling so that if the network detection was wrong and predicted an empty mask as a mask that included any of features (non empty mask), the loss of classification would be high and this would force the network to predict the right empty mask as presented in (Chen et al., 2018). The final loss of the network is the sum of classification and two segmentation losses for predicting globules and streaks. Moreover, among those which hold dermoscopic features, the number of pixels that belong to the classes of globules or streaks are far fewer than the background pixels and each skin image does not necessarily contain both streaks and globules features. Almost 42 percent of images contain pixels of globules while the number of images involving streaks is limited to less than 8 percent of the dataset. In the previous chapter, a bounding box was applied to separate the lesion region by using ground truth images. Thus, the network will look to a larger region of interest as input and the number of background pixels falls, lessening class imbalance. Images are cropped by a factor of 1.1 because in a few cases the streak and globule pixels are located slightly outside of the border of the lesion (from the mask of lesion border segmentation). Accordingly, the output masks generated by lesion segmentation from task1 were used to crop the test images before entering the network. To deal with imbalanced data, images containing streaks were flipped over the vertical axis, the horizontal axis, and both vertical/horizontal axes. So, the number of images with streaks increased to 23 percent and images with globules increased to 48 percent as most images with streaks contained globules too. In the following, the network was extended and deconvolution layers were used with a larger number of feature channels followed by concatenation of feature maps from the corresponding layer in a contraction path similar to the Unet model.

4.2 Experiments and Results

To validate the model, the method is applied to two databases provided by Skin Lesion Analysis towards Melanoma Detection challenges (ISIC 2016 and 2017) (Gutman et al., 2016), (Codella et al., 2019). The performance of the proposed method is compared to the results of the winners of both challenges. Further information on datasets, implementation details and results are presented below.

4.2.1 Data Preparation

The challenges of "Skin Lesion Analysis towards Melanoma Detection" provided datasets of annotated skin lesion images from the ISIC Archive. Two publicly available datasets from the ISIC 2016, 2017 challenges (Gutman et al., 2016), (Codella et al., 2019) are utilised in this research. Details are described in Chapter 3. The proposed model takes advantage of not using common methods of preprocessing, such as removing hair or artifacts in the images. Instead, raw images are the input of networks in all experiments of this research. Furthermore, only flipping was used in terms of data augmentation. The images and corresponding ground truths are flipped vertically to expand training data for task of lesion segmentation and flipping vertically and horizontally to balance the dataset for task of dermoscopic feature segmentation.

4.2.2 Implementation

Initially, a basic architecture composed of 7 convolution layers followed by 6 deconvolution layers, named Model 1, was considered in this study, and a series of comparative experiments was conducted to improve the results. Training was performed using stochastic gradient descent (SGD), such that weight decay and momentum were set at 0.0005 and 0.99, respectively. An initial learning rate of 0.001 was considered, which was then reduced manually by a factor of 10 when the error reaches the plateau (details are discussed in Chapter 3). In experiments where the convolutional neural network is trained with the Adam optimization method, parameters were determined as $\alpha = 0.001, \beta 1 = 0.9$ and $\beta 2 = 0.999$ as recommended in (Kingma and Ba, 2014) for good default settings for machine learning problems. To ensure a fair comparison, similar values for parameters such as filter size, stride and learning rate are considered in all models. The hyperparameters of the proposed network are provided in 4.1. For the number of epochs, the maximum was set to 10000 but early stopping technique was applied to deal with overfitting (details are discussed in Chapter 3). The Caffe framework with GPU GeForce GTX TITANX was used to implement the deep architectures. More details about Caffe are described in Chapter 2.

4.2.3 Results for the Task of Lesion Segmentation

Evaluation metrics are described in detail in Chapter 2. Since the database is scarce and many parameters in a deep neural network stand on heuristics, such a simple model was used to pre-

Table 4.1: Hyperparameters of Convolutional Neural Network- Model 1

Layer	Filter size	Stride	Number of Filters	Size of Output
Conv1	3×3	1	16	1×16×698×698
Pool1	2×2	2	-	1×16×349×349
Conv2	3×3	1	32	1×32×349×349
Pool2	2×2	2	-	1×32×175×175
Conv3	3×3	1	64	1×64×175×175
Pool3	2×2	2	-	1×64×88×88
Conv4	3×3	1	128	1×128×88×88
Pool4	2×2	2	-	1×128×44×44
Conv5	3×3	1	256	1×256×44×44
Pool5	2×2	2	-	1×256×22×22
Conv6	7×7	1	512	1×512×16×16
Conv7	1×1	1	512	1×512×16×16
Deconv1	7×7	1	2	1×2×22×22
Deconv2	4×4	2	2	1×2×44×44
Deconv3	4×4	2	2	1×2×88×88
Deconv4	3×3	2	2	1×2×175×175
Deconv5	3×3	2	2	1×2×349×349
Deconv6	4×4	2	2	1×2×698×698

vent early overfitting, and find appropriate weight initialization and an optimization algorithm to train the network from scratch. The model was gradually improved and the results reported the compromise made between computational complexity and accuracy. Model 1 includes a series of convolution layers followed by a pooling layer that takes the input image with size of 3*698*698 in first layer ended to feature representation of size 512*16*16 in last convolution layer. It was empirically found that the network converges more slowly when filters are initialized with a Gaussian distribution rather than Xavier. Therefore, all convolution layers were initialized with a Xavier distribution (Glorot and Bengio, 2010), and were learned from scratch in all subsequent experiments. A Rectified Linear Unit along with normalization layer were used after each convolution layer. Moreover, drop out layers were applied after both of the latest layers in decoder i.e. conv6 and conv7 (Srivastava et al., 2014). Inspired from (Long et al., 2015), skip connections were used to concatenate information from primary layers to later deconvolution layers. In addition, early

stopping was applied. The results derived from this model are provided in Table 4.2.

Table 4.2: Evaluation metrics for different architectures compared to the best result of challenge (Gutman, et al., 2016). Model 1 refers to the basic architecture composed of 7 convolution layers and 6 deconvolution layers, Model 2 is the model 1 incorporated with representations of transform domain, Model 3 is the model 1 but deeper, and Model 4 is model 3 with integrated features of transform domain.

Method	SE	SP	ACC	DI	JA
(Bi et al., 2017)	0.922	0.965	0.955	0.912	0.846
(Jahanifar, Tajeddin, & Gooya, 2018)	0.901	**0.982**	0.943	0.907	0.838
Best result of challenge (Gutman, et al., 2016)	0.910	0.965	0.953	0.910	0.843
Second ranked in challenge, (Yu, Chen, Dou, Qin, & Heng, 2017)	0.911	0.957	0.949	0.897	0.829
Model in (Pour, Seker, & Shao, 2017)	0.911	0.950	0.943	0.893	0.826
Model 1	0.892	0.879	0.885	0.761	0.634
Model 2	0.927	0.913	0.918	0.849	0.752
Model 3	0.934	0.891	0.915	0.817	0.699
Model 4	0.948	0.922	0.939	0.881	0.803
Model4 with added images from CIElab color space model	0.952	0.931	0.947	0.895	0.816
Model 4(Augmented with flipped vertically)	0.974	0.949	**0.961**	**0.921**	**0.852**

Further improvement was investigated by considering two options. The first option consisted of combining the transform domain representations of input images into convolutional layers, this was referred to as Model 2. The motivation behind this consisted of integrating these proper features leads convolutional layers to understand the input better. An alternative common option was to make the network deeper so that it could learn more complex representations. In Model 2, the contourlet transform in four levels and four directions was applied to different color channels of images, which provided 12 images for each level. These representations have the same size of the outputs of pooling layers with which they are concatenated at various levels and increased the depth. The evaluation metrics for this architecture in Table 4.2 demonstrated improvements for all metrics with a significant rise of 12 percent on the Jaccard index compared to Model 1. Figure

4.4 shows a sample image with relevant groundtruth and representations of the image derived from contourlet transform. The consequence of making the network deeper instead of applying contourlet coefficients was explored in Model 3. In this model, convolution layers were extended in an encoder part from 7 to 15 layers by adding one convolution layer after conv1 and conv2, and adding 2 conv layers after conv3 to conv5. The performance metrics showed 6 percent increase in the Jaccard index compared to Model 1, but still had a lower performance than Model 2.

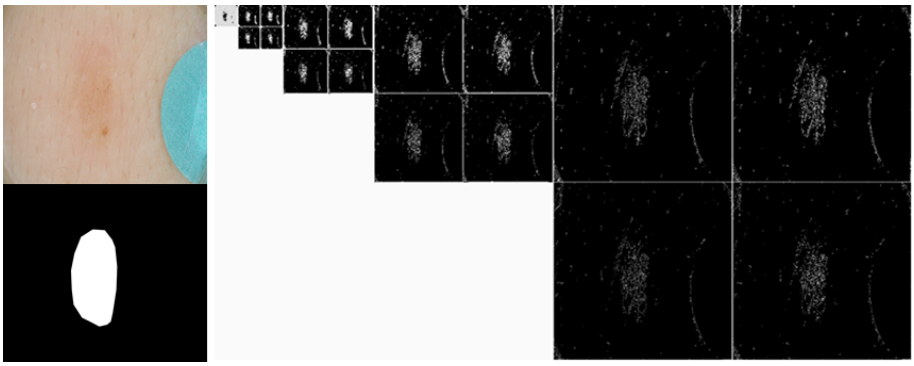

Figure 4.4: Original image and Mask (b) Multiscale image representations of contourlet

To investigate the system performance in terms of training time, forward and backward execution time averaged over 50 iterations per image reported in Table 4.3. The model to which transform domain features were added (Model 2) showed higher performance and less inference time compared to Model 3 that was made by more convolution layers. Figure 4.5 demonstrates the training error curves. The network converged faster if either feature maps from contourlet transform were added, or the number of convolution layers increased. As extending convolution layers also yielded higher performance, the final model (Model 4) was built by increasing convolution layers in the Model 2. 15 convolution and 6 deconvolution layers were applied under this architecture. Furthermore, performing optimization with Adam led to a significant reduction in convergence time compared to SGD in the deeper architecture apart from slightly improving the results. Training error curves in Figure 4.5 confirmed that Model 4 with Adam optimization converges two times faster.

As the deeper model got more complex, the training data was also increased by flipping to tackle the overfitting issue. At this point, transformed images from RGB to CIELAB color space also added to the network that concatenated with the input feature maps. The results were compared to the model in (Pour et al., 2017) which had a similar architecture but trained using transfer learning.

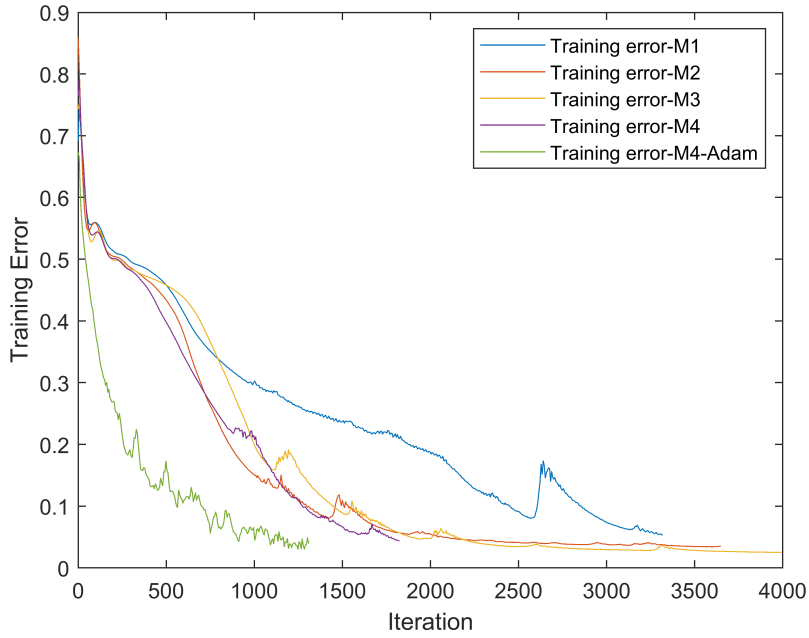

Figure 4.5: Training curves for models 1-4

Table 4.3: Training time comparison in different models

Method	Average Forward pass	Average Backward pass
Model 1 with 7 convolution and 6 deconvolution layers	56.05ms	86.19ms
Model 2 that is Model 1 with contourlet coefficients combined to the network	70.45ms	106.50ms
Model 3 that is Model 1 with more layers (15 convolution and 6 deconvolution layers)	95.61ms	175.46ms

The dataset was also expanded eight times by augmentation techniques like cropping images to two and flipping horizontally, vertically, and both. The Six segmented lesion cases generated by deep convolutional network in Model 3 and similar architecture improved by transform domain features in Model 4 were compared to corresponding masks in Figure 4.6. For these instances, masks produced by a deep convolutional network fine tuned with a pre-trained model were also compared to emphasize the advantages of gradually improving the performance by training from scratch that is not easily feasible when limiting the model to be tuned from a pretrained model.

(Pour et al., 2017) explored, fine tuning the network using a pretrained model that was trained on millions of natural images (ImageNet) were explored. The results of this study outperformed the former particularly for noisy images which contains artifact or hair. For further validation, the model was also evaluated on ISIC 2017 dataset (Codella et al., 2019) for the task of lesion segmentation. The performance metrics in Table 4.4 indicate that the proposed model outperformed the (Codella et al., 2019) by 7% improvement in Sensitivity ,1.1% improvement in Accuracy and 2.2% improvement in the Jaccard index. The Jaccard index was also evaluated without post processing that was 0.778 and was still higher than the winner of the challenge (Codella et al., 2019) and other models in Table 4.4. A histogram of Jaccard Index values is shown in Figure 4.7. Although the number of images with Jaccard index higher than 0.9 in our model is lower than the top challenge result, more images with Jaccard index between 0.75 to 0.9, besides fewer segmented images with Jaccard under 0.05 were achieved.

Table 4.4: Evaluation metrics for ISIC2017 dataset

Method	SE	SP	ACC	DI	JA
1- (Krizhevsky, Sutskever, & Hinton, 2017)	0.810	0.981	0.930	0.839	0.749
2- (Navarro, Escudero Vinolo, & Bescos, 2018)	-	-	0.955	0.854	0.769
3- Best Result of Challenge (Codella et al., 2018)	0.825	0.975	0.934	0.849	0.765
4- Proposed method	0.883	0.981	0.945	0.871	0.782
%improvement (Proposed method compared to 3)	7	0.6	1.1	2.5	2.2

4.2.4 Results for Task of Lesion Dermoscopic Feature Segmentation

Two binary masks were used to identify the position of dermoscopic features (globules and streaks) in lesions. Evaluation metrics were the same as in task 1 and the aim was to automatically generate two masks (globules and streaks) for each test image. Transfer learning was applied using the model trained in previous part that helped the network converge fast. The encoder included the similar convolutional layers as pretrained model and this architecture was followed by two parts, each contained two convolutional layers and four deconvolution layers to predict masks for both streak and globule features. The earlier layers (encoder part) retained freeze for the first 40 epochs then the whole network was retrained with the learning rate decreasing by a factor of ten. The segmentation metrics were calculated over the entire test data set and the results are demonstrated in Table 4.5. Compared to the best result of the challenge, a 17 % improvement in the Jaccard index

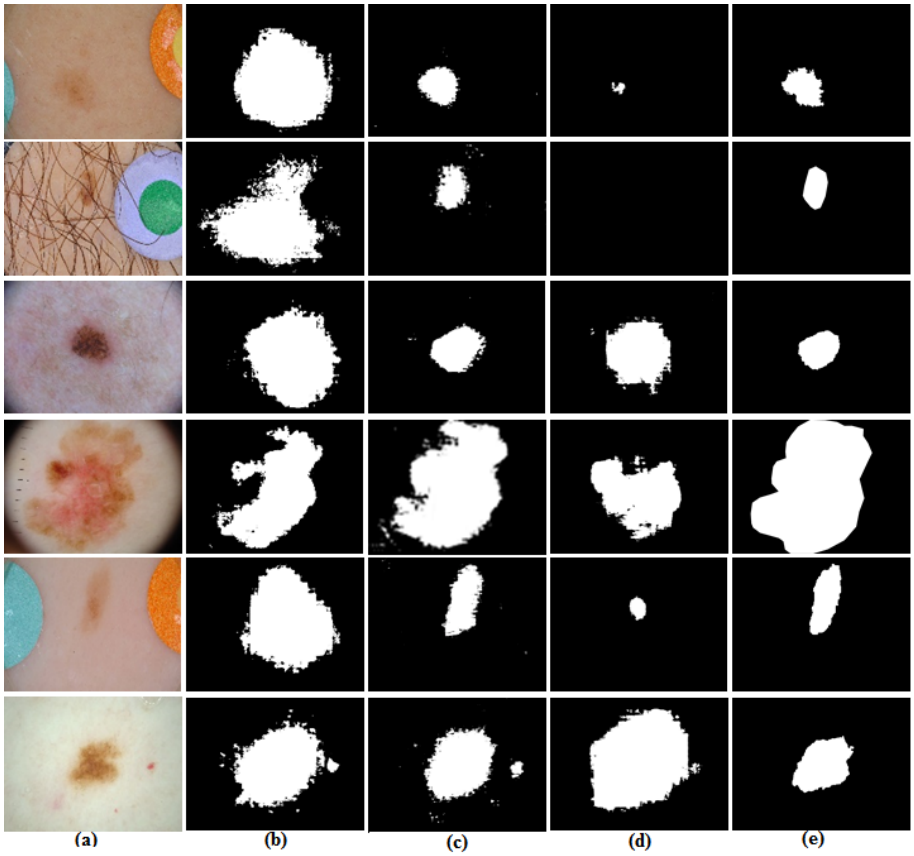

Figure 4.6: a) Original image b) Segmented output from model 3 3 c) Segmented output from model 4 d) Output of model that is fine-tuned on a pretrained model (Pour, Seker, Shao, 2017) E) Test mask

is observed and samples of images from test dataset with predicted groundtruth are presented in 4.8.

Table 4.5: Evaluated metrics for the task of dermoscopic feature segmentation

Method	SE	SP	ACC	DI	JA
Best Result of Challenge (Gutman, et al., 2016)	0.396	0.968	0.962	0.128	0.070
Proposed method	0.368	0.979	0.971	0.150	0.082
%improvement	-	1.1	1	17.2	17.1

4.3 Discussion

Deep convolutional neural networks have widely improved various kinds of tasks solved by classical algorithms in machine learning over the past few years. A lack of appropriately sized dataset is a major issue when it comes to medical analysis. Although going deeper led to higher per-

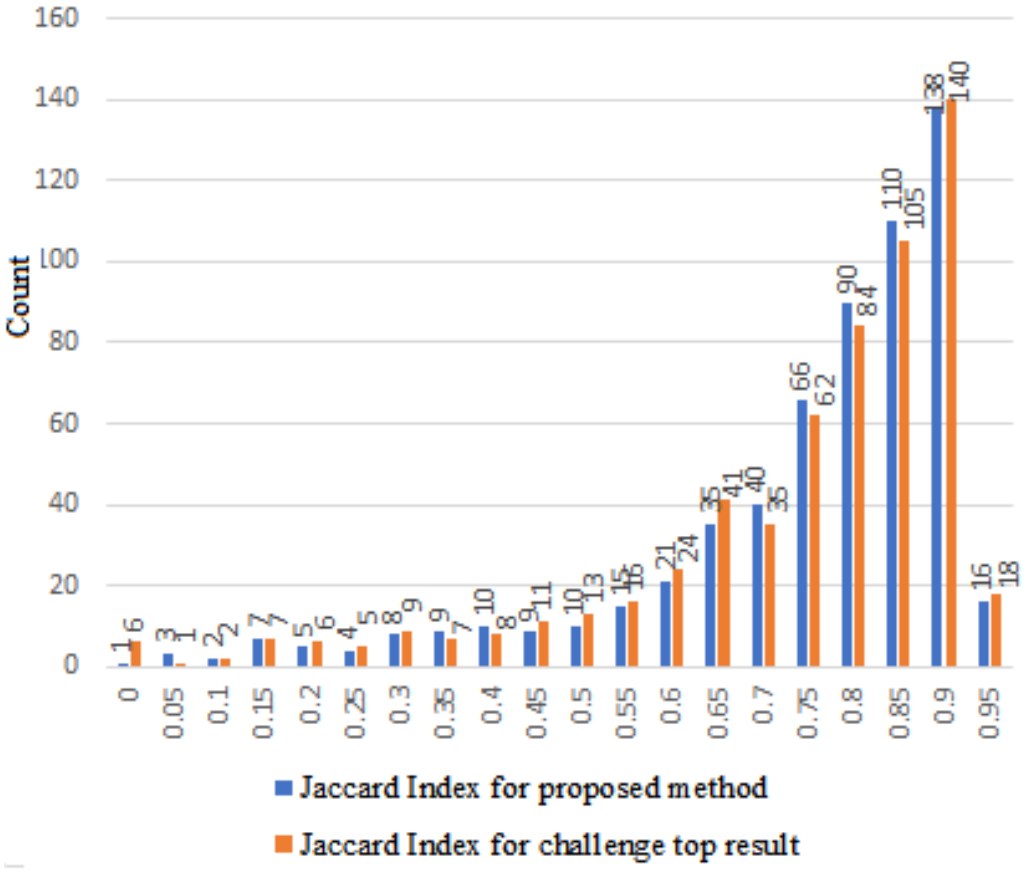

Figure 4.7: Histogram of Jaccard index values for proposed method compared to top result of the challenge ISIC 2017

formance, it is more prone to overfitting. Applying transfer learning is a solution for this issue . However, a model trained on medical data can barely be found to be used as a pretrained model. Moreover, the model can hardly be improved by modifying the architecture, as the architecture is limited to be designed similar to the pretrained model and fixing a layer is not simply applicable due to local distribution representations may found in some layers as discussed in (Yosinski et al., 2015). (Pour et al., 2017) investigated using a pretrained model from semantic image dataset for the task of skin lesion segmentation. In this research, training the network from scratch and improving the model by inserting appropriate features to the network were explored for the task of skin lesion segmentation and dermoscopic attribute detection. A simple model based on convolutional neural network considered and was improved gradually by appending appropriate features and optimization techniques. The model benefits from not applying excessive data augmentation techniques, and instead adds multiscale and multidirectional representation of input images from transform domain to a convolutional network. This led to a considerable rise of 12% in the Jaccard

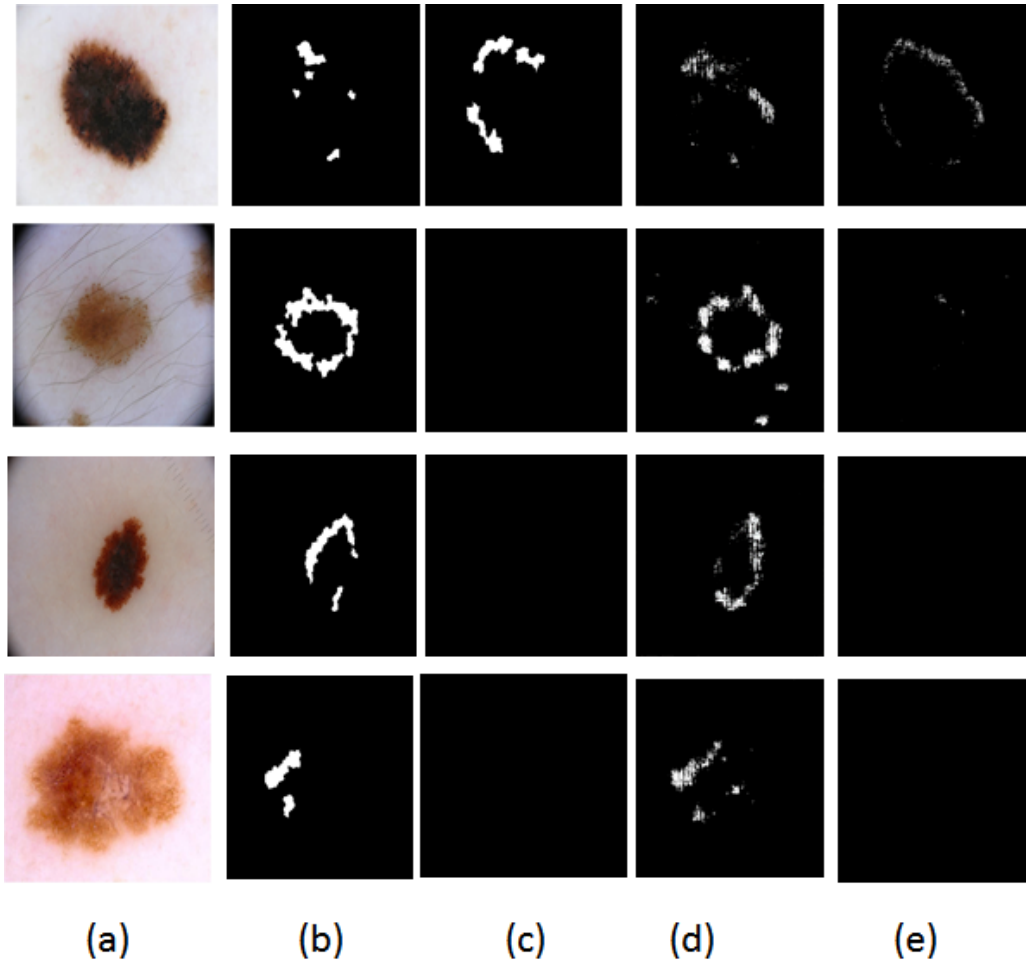

Figure 4.8: (a), (b), (c), are the original images with globule and streak groundtruth respectively. (d) is the predicted globule groundtruth and (e) is the predicted streak

index in Model 2 compared to a 6% raise in Jaccard index in Model 3, modified by making the network deeper by increasing the number of convolution layers from 7 to 15. The training time of these models are compared in Table 4.3, which shows that the training time increased 128.83 ms by increasing the number of layers from 7 to 15, compared to just 34.71ms for model with 7 convolution layers and integrated with the transform domain features. Figure 4.9 compares the output of the 4th, 8th, 9th, and 11th convolution layers in Models 3 and 4. Model 4, which includes image representations of contourlet transform is learning the pattern more effectively while Model 3 is a deep model without features of the transform domain showing noisier patterns. The proposed model with incorporated representations (Model 4) shows 10 percent improvement in the Jaccard index compared to a similar model without adding features (Model 3). This confirms the idea of integrating features of the frequency domain to the network particularly when the dataset

is scarce and going deeper can not really improve the results. In comparison with the model in (Pour et al., 2017) that is fine-tuned on pretrained model on natural images and data augmentation that is conducted to increase the data 8 times, the average Jaccard index has improved 3% and the proposed model indicates significantly higher performance in noisy images such as images that contains hair or artifacts.

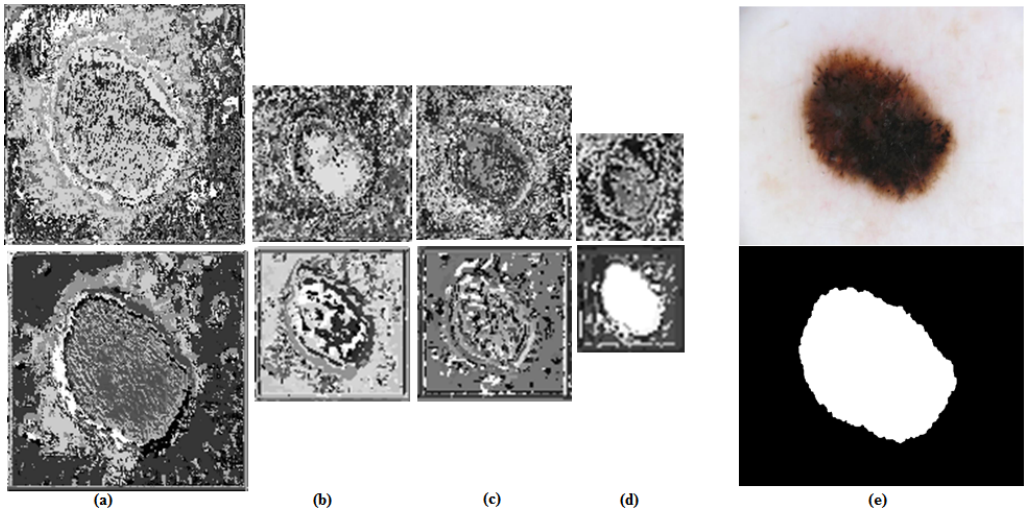

Figure 4.9: (a), (b), (c), (d) are the outputs of 4th,8th,9th and 11th convolution layer respectively. Images in the first raw are from model 3 and second raw from model 4. (e) Origin image and the groundtruth

The advantages of the proposed model, in short, contain not using preprocessing and excessive data augmentation, improving the performance of a not very deep and complex CNN-based model by integrating multi-scale and multi-directional contourlet representations, high performance of the model on small dataset and significant improvements in segmenting noisy images.

4.4 Summary

In this work, a segmentation model based on CNN was proposed for the tasks of lesion segmentation and dermoscopic feature segmentation. While adding more layers and increasing the depth are common ways to improve the accuracy of a CNN. This may not be applicable to medical data, because the network requires more training data, which is a major issue in the medical domain. To deal with this issue, many studies use excessive augmentation algorithms that may add irrelevant data as well. Furthermore, providing the labelled data in medical domain is expensive and requires expert skills, also entails privacy issues of medical records. This contributes to limited data ac-

cess in this area. A further solution to use deep architectures for scarce data is transfer learning, but this is also limited because a pretrained model on a medical dataset can hardly be found. In this study, training the network from scratch was investigated and increasing the depth of input to convolutional layers by concatenating efficient feature maps from transform domain and using CIELAB colour space in addition to RGB colour channels instead of excessive augmentation or using a pretrained model on natural images. A basic CNN model was initially proposed and was progressively improved, and the results were compared to the common techniques such as adding more layers to the network, transfer learning or data augmentation. In the first stage, incorporating multiscale image representations from transform domain improved the Jaccard index by 12%, while adding layers to the network increased it by 6% compared to our basic model. The model improved by combining these two models, boosting with the CIELAB colour model and flipping as a trivial augmentation that outperformed the winner of 2017 challenge by a 2.2%improvement in the Jaccard index and 7% in sensitivity. A summary of the achievements of the proposed model includes:

- A CNN is designed to do automatic learning from training data with a deep architecture that applied to extract low level to high level features in various layers.
- The relevant feature maps are concatenated to the network by inserting image representations from the transform domain that provide a superior understanding of the input into the model.
- CIELAB colour space is applied in addition to RGB colour channels that provide more information for the network.
- This architecture benefits from not applying pre-processing methods as well as not using excessive data augmentation techniques.
- Despite the small dataset, the proposed architecture is trained from scratch and improved the results, particularly for noisy images, compared to the model that was fine-tuned with a model pretrained on natural images.
- Improves accuracy by 12% while adding layers to the network increases the Jaccard index score by 6% compared to our basic model.

- The final proposed model outperformed the results of both 2017 and 2016 challenges with 2%and 7% improvements in the Jaccard index and Sensitivity for 2017 and an increase of 1%in Jaccard index with 6%in sensitivity for 2016.
- The model with integrated transform domain features (Model 2) shows less inference time compared to model 3 that is made by more convolution layers.

Chapter 5

Left Ventricular Segmentation

5.1 Introduction

A cardiac magnetic resonance image (MRI) scan is a non-invasive test and an MRI machine is used to generate magnetic and radio waves to show the detailed pictures of inside of the heart. Cardiovascular research has improved over the years to improve early identification of cardiac diseases. The left ventricular (LV) is the most investigated chamber in cardiac segmentation due to its key role in the blood pumping in human body. It is the thickest of the heart's chambers and is responsible to pump oxygenated blood to tissues all over the body. Moreover, cardiac MRI is a critical part of cardiac function analysis such as left ventricular volume and the ejection fraction, wall motion abnormality and stroke volume. Cardiovascular diseases (CVD) is a significant cause of disability and death around the world. The Global Burden of Disease study estimated that 29.6% of all deaths worldwide were caused by CVD. It is still responsible for over 4 million deaths per year, close to half of all deaths in Europe in 2010 (Nichols et al., 2014).

Many studies on medical video segmentation are inspired from image segmentation methods and have been applied on video frames too. These methods include traditional segmentation techniques such as thresholding, region growing and active contours (Codella et al., 2008), (Queirós et al., 2014), (Kaus et al., 2004) as well as recent deep learning based models such as FCN and specifically Unet, which has shown significant improvement in biomedical image segmentation (Wang et al., 2014), (Moradi et al., 2019). Several research papers on medical video segmentation, have incorporated prior information including motion and shape, to improve robustness and

accuracy (Petitjean and Dacher, 2011). Moreover, the size of ventricles is small compared to the whole cardiac MRI image and detecting the region of interest (RoI) is an essential primary step in LV segmentation to decrease the intervention of surrounding tissues. Frequent movement property of LV is an important feature that can be applied to detect the RoI, where intensity varies significantly in a cardiac cycle. The circular shape of LV in the short-axis MR images and the center position of LV are further factors that have been widely used in the literature too.

(Alba et al., 2014), considered the center position of the heart in the image and the endocardial and papillary muscles were identified by optimal threshold method of Otsu, watershed, and 1D fast Fourier transform. The final epicardial contour was detected by applying a multiple seed, region-growing on pixels that were mapped from Cartesian to polar coordinates. The motion characteristic was used to find the RoI in (Lu et al., 2019). Cumulative Difference Variation was defined as sum of absolute difference of adjacent slice images followed by making a binary image by using an Otsu threshold method. The center of the image was considered as the center of the rectangle with a predefined side length to generate the RoI. An edge detection method and Circle Hough Transform (CHT) were applied to estimate the initial contours of LV and finally, a minimum distance constraint was used to specify the proper circle. (Nambakhsh et al., 2013) used shape and intensity information that were derived from manually segmentation of the first frame in each sequence for 2D and a single point per target regions (cavity or myocardium) in 3D were applied followed by an optimization of distribution measures to achieve the optimal cavity and myocardium regions. Although various image segmentation systems based on image processing techniques have shown encouraging results, the generalization problem and the issue that these techniques are mostly customized to the data, have led to the further growth of machine learning-based methods and specifically, recent deep convolutional neural network models with the ability of automatic end to end feature learning from raw data. (Yan et al., 2018), proposed an optical flow feature aggregation sub-network which was integrated to the Unet and was further developed by dilated convolution. A method proposed by (Khened et al., 2019), employed Fourier transform and circular Hough-transform to detect the RoI and used densely connected CNNs for LV segmentation. Further details of deep CNN-based models are provided in Chapter 2. To detect the RoI in this research, motion and continuity of the frames in a sequence beside shape and position of LV are considered by using optical flow and absolute difference of frames in a sequence

(Lu et al., 2019). Finally, a CNN-based network is designed to segment the LV. Data preparation, proposed model inspired of U-net as a leading deep learning model for medical domain and details of the experiments will be presented.

.

5.2 Proposed Method

5.2.1 Data Preperation

The dataset of Sunnybrook Cardiac Data (SCD) has 45 cine-MRI images from a range of patients and pathologies: healthy, hypertrophy, heart failure with infarction and heart failure without infarction. The data became publicly accessible as part of the MICCAI 2009 challenge on automated LV segmentation from short-axis cardiac magnetic resonance imaging (MRI). [1] The MRI images are in the DICOM (Digital Imaging and Communications in Medicine) format that consists of various metadata parameters related to the patient and the image. The contours format is in text files which include contour points which required to be converted into the groundtruth.

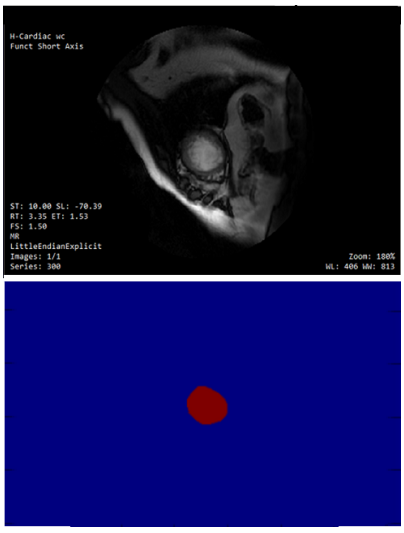

Figure 5.1: A sample image from Sunnybrook dataset and the converted ground truth.

Another database that is used for this task comes from the Left Ventricle Segmentation Challenge (LVSC)[2] that is accessible from STACOM 2011 challenge on automated LV myocardium segmentation from short-axis cine MRI (Suinesiaputra et al., 2014). The dataset includes 100 sets

[1] http://www.cardiacatlas.org/studies/sunnybrook-cardiac-data/
[2] http://www.cardiacatlas.org/challenges/lv-segmentation-challenge/

of cardiac MRI images of patients with coronary artery disease and myocardial infarction. The images have the resolutions from 0.7 to 2.1 mm/pixel with sizes from 156 $\times$ 192 to 512 $\times$ 512. Cardiac cycle contains 18 to 35 frames. A few samples of dataset are provided in Figure 5.2.

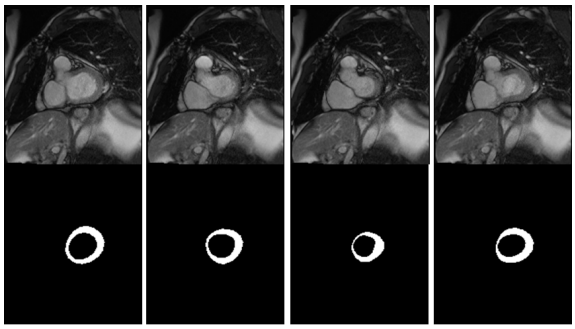

Figure 5.2: Images (numbers 0, 5, 13, 20 respectively from the same sequence-patient 1) and the corresponding groundtruths from LV segmentation challenge dataset (Suinesiaputra et al., 2014)

The MRI images are in various sizes and orientations which are taken by different experts and belong to various patients, thus data normalization is required. As 2D slices from all sequences are extracted, the normalization can be conducted on each frame of sequence.

5.2.2 Proposed CNN based Model for Task of LV Segmentation

In this research, the feature of the constant movement of LV was used to find the region that it belongs to. This feature was concatenated to the network feature map as a location guide to render the output sensitive to the region of interest. The procedure to design the location guide module includes: applying an edge detection (canny edge detector or contourlet transform); calculating the absolute difference of two frames in a sequence; and the employment of adoptive thresholding and morphological operations. More details of the location module are given in the next section that identifying the region of interest is described.

The output of the location module is fed to the deep convolutional neural network after having been down-sampled to reflect the corresponding sizes of the various layers in the network. The deep network consists of encoder (downsampling) and decoder (upsampling) parts and skip connections as proposed in the U-net architecture (Ronneberger et al., 2015). The proposed architecture is depicted in Figure 5.3. This architecture is referred as 'model 1' in the experiment section. The encoder comprises 15 convolutional layers followed by RELU activation and pooling layers, and the upsampling path includes convolution and deconvolution (fractional convolution) layers. The

sum of cross-entropy terms over each pixel of output map is used as the loss function. In each iteration, two frames from a sequence are fed to the network and location module. To alleviate the class imbalance between segmented and the surrounding background particularly in the first architecture where images are not cropped, the Dice loss function described in the previous chapter is also applied to improve accuracy. This is important due to the problem of imbalanced classes, because less than 2 per cent of all pixels belong to the LV class in the Sunnybrook dataset.

Moreover, the model is compared to the architecture in which the RoI is identified in the first step followed by extracting the RoI. The extracted RoI is then fed to the deep CNN and the output mask is generated. Details of the method used to predict the RoI are presented in the next section, but the architecture of deep CNN is the same as in the previous model, and this architecture is referred as model 2 in the experiment section.

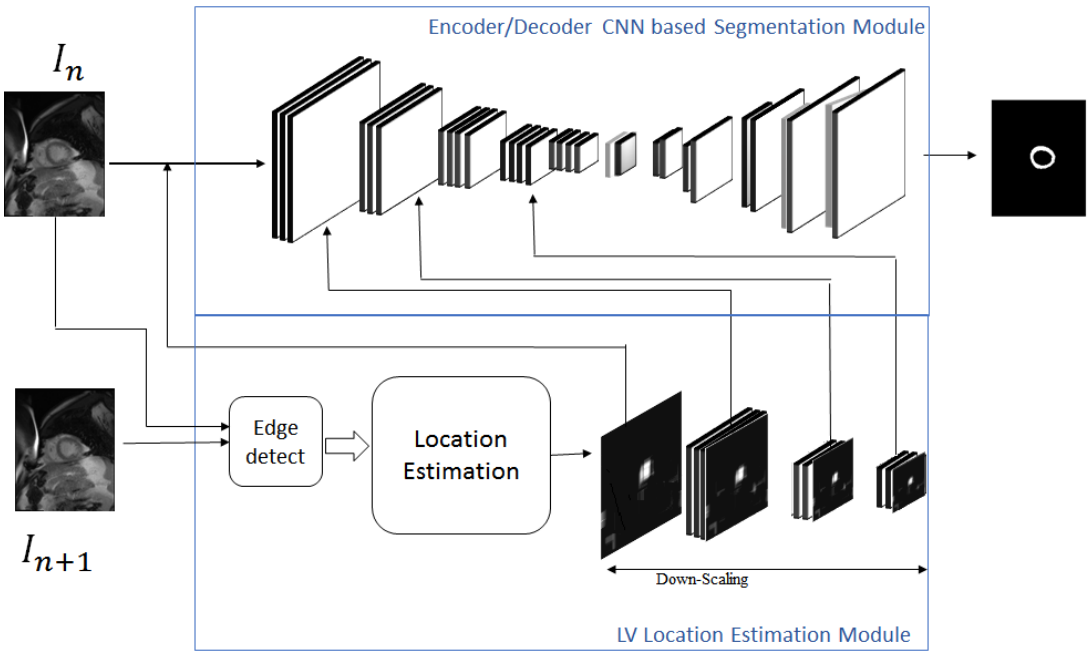

Figure 5.3: Outline of the proposed method for task of LV segmentation.

5.2.3 Proposed Algorithms to Identify the Region of Interest (RoI)

The heart is located in the thoracic cavity between the lungs, and extends as far as the diaphragm. Hence the cardiac MR images often include the heart and the surrounding chest cavity such as the lungs and diaphragm. The region of interest was extracted to enhance accuracy by focusing on the area in which the left ventricular is located and by decreasing the noise. Furthermore, capturing

the RoI alleviated the imbalance between classes by reducing the number of background pixels. To capture the RoI, two factors relevant to the shape and function of the LV were investigated: its motion and the fact that it is mostly located close to the center of the MRI image. Firstly, a canny edge detector was applied to the images in a sequence (Canny, 1986). Detecting motion in the neighboring images in a sequence, allowed for an estimation of the region that the LV belongs to. In each image sequence, the absolute difference of two frames were calculated after operating the canny edge detector, and local thresholding was conducted at the end. Adaptive thresholding computes a locally adaptive threshold that chooses the threshold based on the local mean intensity (first-order statistics) in the neighborhood of each pixel.

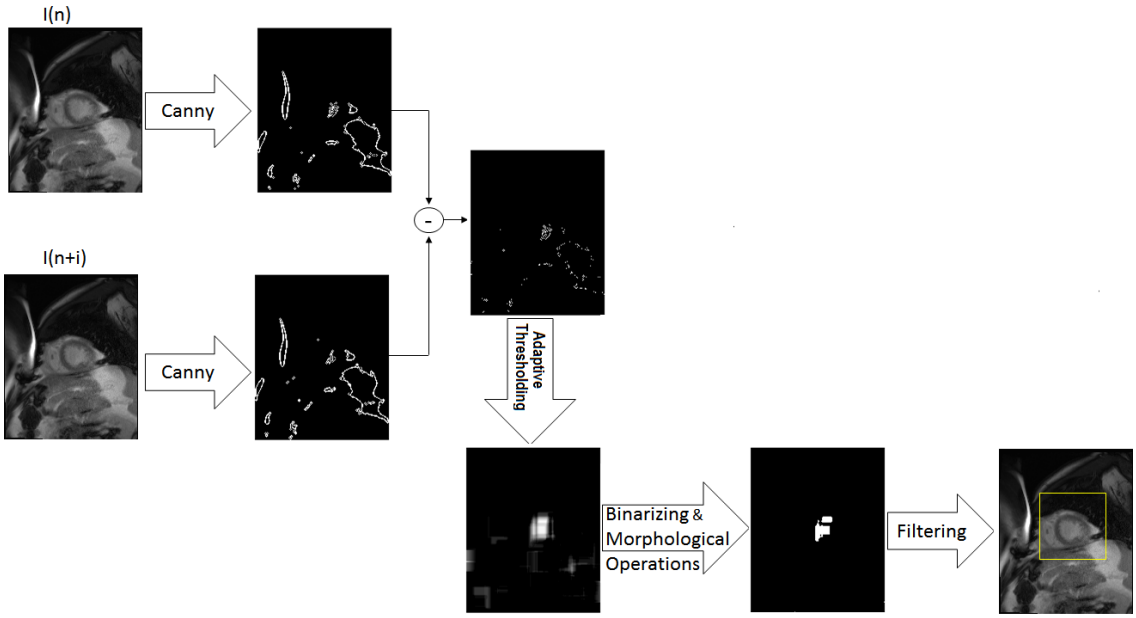

Figure 5.4: The Procedure of Detecting the Region of Interest.

Canny Edge Detection is a multi-stage edge detection algorithm that includes the following stages. Firstly the noise is reduced by a Gaussian filter,

$$S(m,n) = G_\sigma(m,n) * I(m,n) \tag{5.1}$$

where

$$G_\sigma = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{m^2+n^2}{2\sigma^2}\right) \tag{5.2}$$

The gradient of the smoothed array S(m,n) is computed in the next step by using any of the gradient

operations (Roberts, Sobel, Prewitt, and so on):

$$M(n,n) = \sqrt{g_m^2(m,n) + g_n^2(m,n)} \tag{5.3}$$

$$\theta(m,n) = tan^{-1}[g_n(m,n)/g_m(m,n)] \tag{5.4}$$

in the next step, non-maximum suppression to the gradient magnitude is followed by double thresholding to detect and link the potential edges investigated (Jain et al., 1995). Binarizing and morphological operations are conducted to select the closest object to the centre, removing small and disconnected objects and creating a rectangular filter that is extended by 20 percent from all sides. Moreover, a contourlet transform similar to that described in the previous chapter is used instead of the canny, and obtained a slightly higher performance but computationally, it is more complicated.

5.2.4 U-net based Model Improved by Optical Flow Motion Estimation

The proposed segmentation method outlined in this section benefits from considering temporal information between cine frames by adding motion analysis to the CNN. Optical flow indicates the motion of image objects between two consecutive frames and can be defined by the following equations (albeit with the assumptions that illumination is constant over time and there is very minor LV displacement between adjacent frames).

$$I(x,y,t) = I(x+dx, y+dy, z+dz) \tag{5.5}$$

$$\frac{dI}{dx}U + \frac{dI}{dy}V + \frac{dI}{dt} = 0 \tag{5.6}$$

I is the image intensity as a function of space and time, U and V are horizontal and vertical velocity components of the pixel in the position (x,y). Figure 5.5 (b,c) shows an example of U, V with full range of colours in the colour-map. The resultant optical flow is indicated in Figure 5.5 (d) where the arrows show directional components specified by (U+du,V+dv).

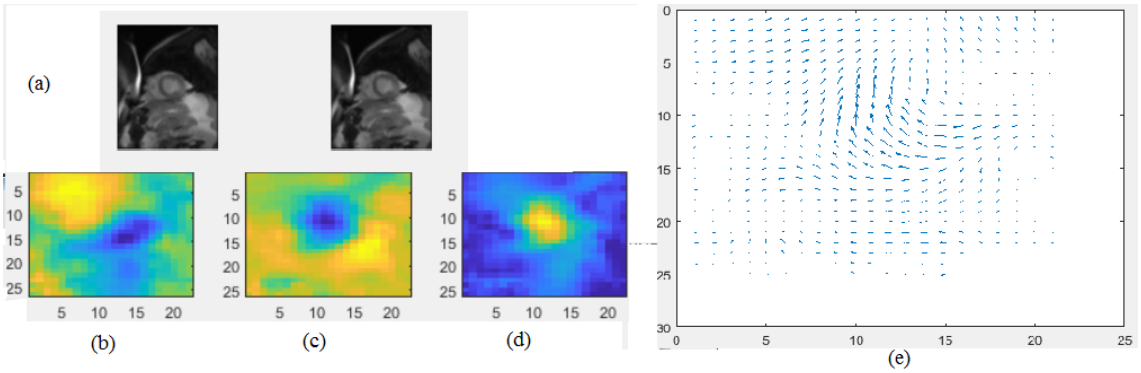

Figure 5.5: (a)original adjacent frames, (b) and (c) are U and V respectively, and (e) vector representation optical flow

The architecture of the main network is the same as in the prior model, and additional convolutional layers receive optical field data as input with multi-dimensional, multi-directional contourlet representations integrated into the corresponding convolution layers. Conv1 to Conv3 layers in Figure 5.6 are a group of 3 convolution layers. A fusion technique is applied to integrate feature maps of optical flow into the fourth convolution layer of the network. As shown in Figure 5.6, feature maps of the contourlet are fused with the convolution layers of the optical flow network. In the method proposed in Section 5.2.2, the contourlet representations were concatenated with the pooling layers. The reason for the use of fusion instead of concatenation was that the number of parameters was increased, since convolution-pooling layers were added for LV frames and optical flow feature maps and concatenating the layers pooling 3 of both networks will double the number of parameters too. Therefor, fusion of feature maps was considered, which includes summations of the corresponding positions of feature maps of the optical flow network and the main convolution network. In order to equalize the size and depth of the feature maps of the contourlet and the corresponding pooling layer, resizing and convolution layers are applied before fusion. To further improve the model, inspired from (Zhou and Yang, 2019) three normalization methods are investigated, including Batch Normalization (BN), Layer Normalization (LN) and Instant Normalization (IN). BN is employed to normalize the input feature map $F_{N\times H\times W\times C}$ by considering a mean of 0.0 and variance of 1.0. H, W and C are height, width and number of channels respectively. N is the number of images in a mini batch in the input layer. It is demonstrated that this technique reduces the number of training epochs required to train deep networks too (Ioffe and Szegedy, 2015). The normalization is performed by computing the mean and variance of each mini batch

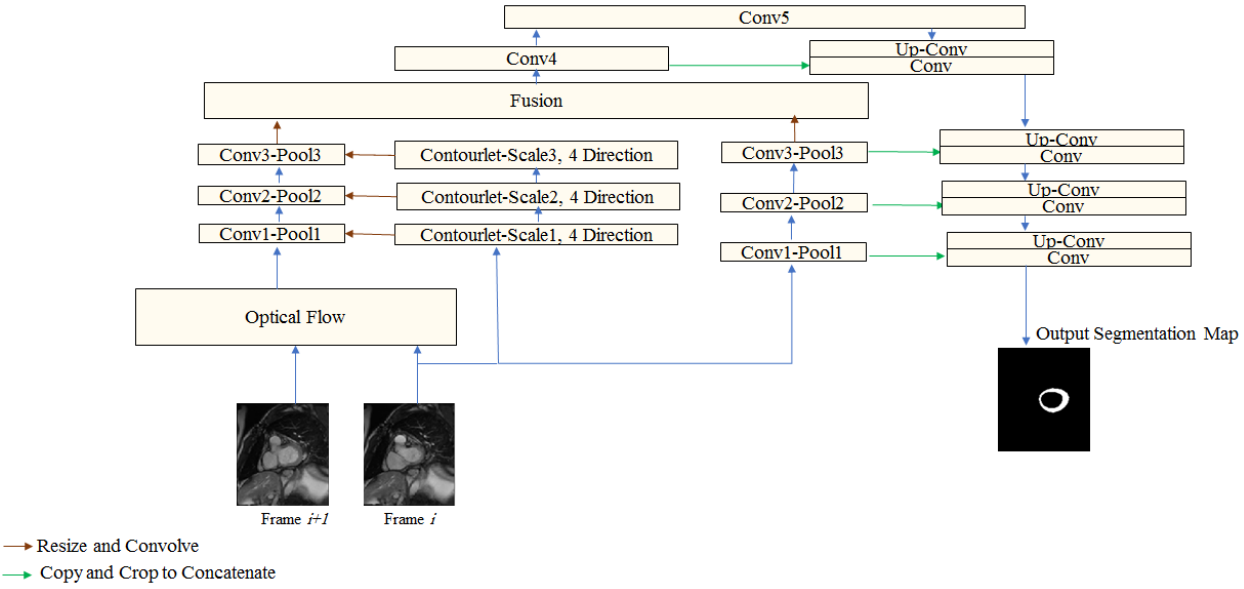

Figure 5.6: The proposed system with integrating optical flow feature maps

input variable along the channel:

$$\mu_c = \frac{1}{\mathrm{N} \times \mathrm{H} \times \mathrm{W}} \sum_{n=1}^{\mathrm{N}} \sum_{h=1}^{\mathrm{H}} \sum_{w=1}^{\mathrm{W}} f_{n,h,w}$$
$$\delta_c^2 = \frac{1}{\mathrm{N} \times \mathrm{H} \times \mathrm{W}} \sum_{n=1}^{\mathrm{N}} \sum_{h=1}^{\mathrm{H}} \sum_{w=1}^{\mathrm{W}} (f_{n,h,w} - \mu_c)^2 \tag{5.7}$$

$$\hat{f}_{n,h,w} = \frac{f_{n,h,w} - \mu_c}{\sqrt{\delta_c^2 + \epsilon}} \tag{5.8}$$

where ϵ is a small value added to increase the stability of the division. Normalization is applied to the feature maps ($f_{n,h,w}$) after the convolution layer and before the activation layer. In LN, the mean and variance are calculated along batch (Equation 5.9), while in IN method they are

calculated along channel and batch (Equation 5.10).

$$\mu_n = \frac{1}{\mathrm{H} \times \mathrm{W} \times \mathrm{C}} \sum_{h=1}^{\mathrm{H}} \sum_{w=1}^{\mathrm{W}} \sum_{c=1}^{\mathrm{C}} f_{h,w,c}$$
$$\delta_n^2 = \frac{1}{\mathrm{H} \times \mathrm{W} \times \mathrm{C}} \sum_{h=1}^{\mathrm{H}} \sum_{w=1}^{\mathrm{W}} \sum_{c=1}^{\mathrm{C}} (f_{h,w,c} - \mu_n)^2$$
$$\hat{f}_{h,w,c} = \frac{f_{h,w,c} - \mu_n}{\sqrt{\delta_n^2 + \epsilon}} \tag{5.9}$$

$$\mu_{n,c} = \frac{1}{\mathrm{H} \times \mathrm{W}} \sum_{h=1}^{\mathrm{H}} \sum_{w=1}^{\mathrm{W}} f_{h,w}$$
$$\delta_{n,c}^2 = \frac{1}{\mathrm{H} \times \mathrm{W}} \sum_{h=1}^{\mathrm{H}} \sum_{w=1}^{\mathrm{W}} (f_{h,w} - \mu_{n,c})^2$$
$$\hat{f}_{h,w} = \frac{f_{h,w} - \mu_{n,c}}{\sqrt{\delta_{n,c}^2 + \epsilon}} \tag{5.10}$$

5.3 Experiments and Results

The proposed U-net based model includes convolutional layers with 3*3 filter, normalization, ReLU activation function followed by 2*2 pooling layers. Drop out layers were also applied to overcome overfitting. Skip connections were used to concatenate the feature maps from shallow layers to deep layers (Long et al., 2015). The model was implemented using GPU TitanX. Both architectures on datasets of Sunnybrook and the dataset from LV segmentation challenge (LVSC) were evaluated. The network was trained using an Adam optimizer with parameters including α = 0.001, $\beta 1$ = 0.9 and $\beta 2$ = 0.999. 15 percent of data was considered for validation and 15 percent for test. 70 percent was also considered for training. After each epoch the model was checked with the validation data and the final model used to evaluate the test data.

Dice and Softmax loss were applied, which were formulated in previous chapters. In terms of the evaluation metrics, Dice, Jaccard, Sensitivity and Specificity were used. The two other metrics,

positive predictive value (PPV), and negative predictive value (NPV) are defined as:

$$PPV = \frac{T_1}{T_1 + F_1}, \quad NPV = \frac{T_0}{T_0 + F_0}, \tag{5.11}$$

where T_1,T_0 are the number of correctly predicted pixels from object and background classes respectively and F_1,F_0 are the number of misclassified pixels as object and background, respectively. Dice loss improved the segmentation metrics, especially Dice, and accordingly the Jaccard index.

Table 5.1: Evaluation metrics from the proposed model 1 compared to the recent papers including (Tan et al., 2017), (Khened et al., 2019), (Tran, 2016)

Method	Dice	Jaccard	Specificity	Sensitivity	PPV	NPV
(1) Tan2017	-	.77	.95	.88	.86	.96
(2) Khened2019	.84	.74	.96	.84	.87	.95
(3) Tran2016	-	.74	.96	.83	.86	.95
The proposed method	.85	.76	.96	.85	.86	.96

Table 5.1 presents the performance of the proposed model 1, which is comparable to the highest performance reported in recent papers on LV segmentation and using the LVSC dataset. The performance of two models were compared, model 1 (in which the location guide module was used and its multi-scale features are integrated into the network) and model 2 (in which identifying RoI and cropping the images were the initial steps). The two loss functions were considered, the Dice layer and Cross Entropy loss, as shown in Table 5.2. In method 2, the cross entropy loss outputs were very close to the output of dice loss since by RoI cropping, the class imbalance was alleviated that led to optimal performance for cross entropy loss too.

Fine tuning the model on the pretrained model from chapter 3 is evaluated and as Figure 5.7 shows that the network converges more quickly. The learning of general information, particularly in the early layers of these two networks, could be similar but transfer learning even on a different medical dataset helps the network to converge faster. However, the later layers were retrained.

The training of the network on the Sunnybrook dataset was divided into three parts: train:31

Table 5.2: The performance of both models with cross entropy and dice loss

Method	Dice	Jaccard
Model1- Dice loss	.85	.76
Model2- Dice loss	.84	.74
Model1- Cross entropy loss	.83	.74
Model2- Cross entropy loss	.84	.74

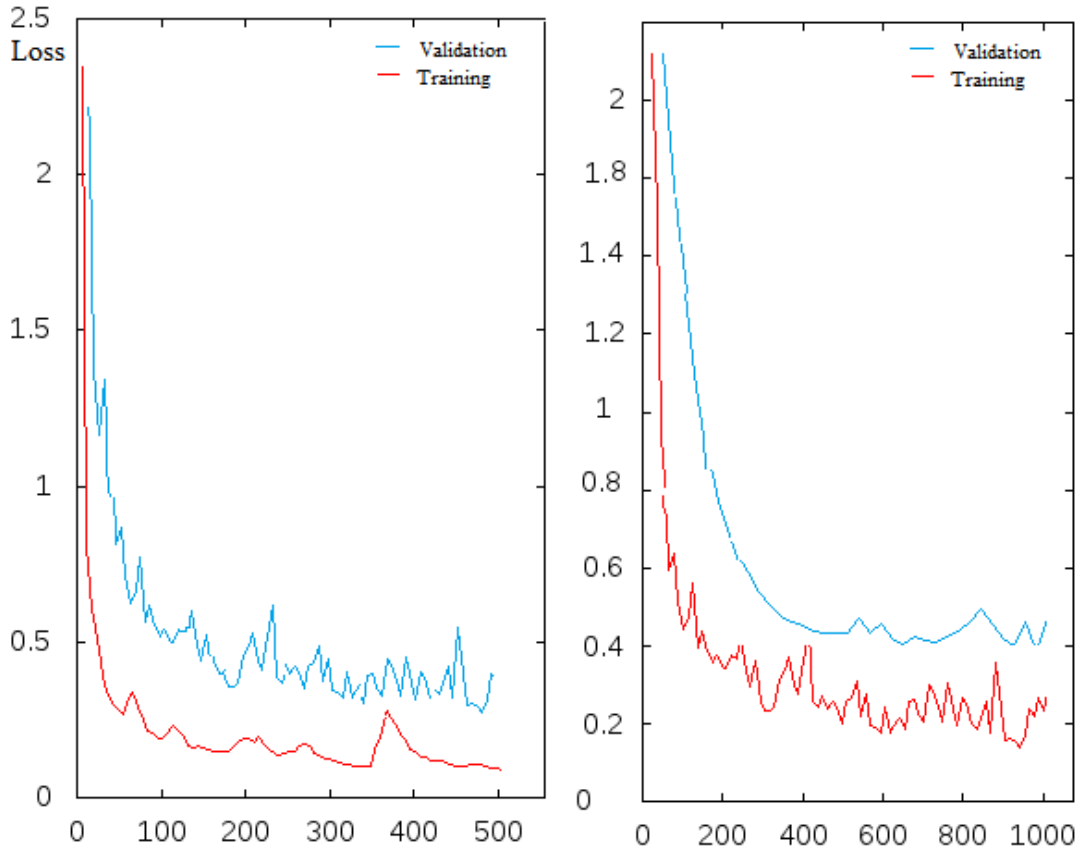

Figure 5.7: Training and validation loss for model 1, The Left plot is for using pretrained model from previous chapter.

cases; validation:7 cases; and test:7 cases. A dice of 0.96 was achieved, thus outperforming the result of (Tran, 2016) which was 0.94. For this part, data augmentation was also conducted with flipping vertically and horizontally to increase the dataset and prevent early overfitting. Transfer learning was used from the model pretrained on the LVSC dataset. The other technique that was considered to handle the overfitting issue was early stopping, as described in Chapter 3. The effect of increasing the number of convolution layers and number of filters to make the network deeper

was investigated to find the efficient architecture. Table 5.3 represents the results showing that increasing the depth does not necessarily lead to improved performance.

Table 5.3: The effect of increasing convolution layers and number of filters on dice metric from proposed model on LVSC dataset

Number of Kernels	8 Conv Layers	15 Conv Layers	20 Conv Layers
(16,32,64,128,256)	0.81	**0.85**	0.82
(32,64,128,256,512)	0.81	0.84	0.80
(64,128,256,512,1024)	0.79	0.81	**0.79**

Table 5.4: Comparing with the original U-net

Method	Dice	Jaccard
U-net based model (Ronneberger et al., 2015)	0.81	0.70
The proposed model with location estimation module (section 5.1.3)	0.85	0.76
The proposed model with contourlet based location module	0.87	0.77

The proposed model is also compared to the model similar to the original U-net architecture that is without concatenating the feature maps of location modules to the CNN. The Jaccard metric in the Table 5.4 shows 6 percent improvement by manipulating the network with the RoI guide. The LVSC dataset set categorized individual images into Apex, Mid, and Base levels, which is conducted in this study but the slices not identified from validation were removed. The results of the proposed model with optical flow and contourlet with IN, LN and BN normalization techniques are reported in Table 5.5. (Yan et al., 2018) reported the results of using max-pooling in the CNN architecture and the improved results by using dilated convolution. The proposed architecture in this research uses max-pooling and outperforms the results in (Yan et al., 2018) when max-pooling is used as well as the results in (Tan et al., 2017). Table 5.5 shows the higher Jaccard index with applying IN for Apex slices and slightly higher result for Mid slices with BN compared to results in (Yan et al., 2018) when they used dilated convolution instead of max-pooling. BN was used in the previous reported results in this chapter. (Yan et al., 2018) and (Tan et al., 2017) applied BN as well.

Table 5.5: Comparing the results with applying IN, BN, LN normalization methods with two recent models

Method \ Jaccard	Apex	Mid	Base
(Tan et al., 2017)	0.71±0.13	0.79±0.09	0.77±0.12
(Yan et al., 2018) (Max Pooling)	81.9±3.2	92.3±3.6	86.3±2.9
(Yan et al., 2018) (Dilated Conv)	84.5±3.7	94.8±3.2	89.3±2.5
Proposed Method with LN	0.83±0.07	0.88±0.02	0.87±0.05
Proposed Method with BN	0.84±0.12	0.91±0.01	0.89±0.11
Proposed Method with IN	0.85±0.09	0.9±0.07	0.88±0.06

5.4 Discussion

In this chapter, the proposed CNN-based models of image segmentation were expanded to video segmentation, aiming to establish a model for the task of automatic left ventricle (LV) segmentation on short-axis cardiac MRI. A fully automated technique was proposed to address LV segmentation. As LV covers a small area in the image, methods to detect the RoI were proposed based on a similarity measure, i.e. absolute differences of images in a sequence, edge detection techniques including contourlet transform and optical flow as a motion detection technique. Further features including the shape and center position of LV were considered too. The experimental results demonstrate the significant improvement of the U-net model by incorporating contourlet coefficients and an optical flow module. Although the pretrained model data from a previous chapter was different from LV frames, the network convergence time was almost halved, as shown in Figure 5.7. In terms of network parameters, a 2% improvement in the Dice metric was achieved in Model 1 by changing the loss function from cross entropy to Dice. Table 5.3 shows a compromise between number of kernels, depth and accuracy. The impact of normalization techniques is depicted in Table 5.5 stating that instant normalization improved the average Jaccard index in Apex and Batch normalization in Mid and Basal images. The proposed method in this chapter that uses max pooling showed a higher Jaccard index score compared to the method in (Yan et al., 2018) using max pooling for Apex and Base slices and surpassed the results in (Tan et al., 2017) too. Moreover, the proposed model with IN normalization outperformed the average Jaccard index reported in (Yan et al., 2018) for Apex frames and stated comparable results for Mid and Base images. Although this research improved U-net by incorporating features, the network elements

remained unchanged and there is still potential for progress as (Yan et al., 2018) replaced the max pooling operation with dilated convolution and blocks of the U-net were updated to res-block to improve the results.

5.5 Summary

In this chapter, an automated U-net based segmentation system was proposed for left ventricle segmentation, which is a crucial task in cardiac disease diagnosis. LV segmentation is still a challenging task due to the small size of RoI, intensity issues and weak boundaries between myocardium and surrounding tissue. The continual movement of the LV is an important feature that has been ignored in most cardiac diagnostics studies, many of which only analyse single frames. In this chapter, the inherent continuity feature of neighbouring frames in the video was considered to propose techniques to highlight the RoI and provide a location guide for the neural network. Recent CNN-based models, specifically U-net have demonstrated encouraging results in medical image segmentation. A U-net based model was improved significantly by integrating temporal and frequency information into the deep CNN.

A further experiment was conducted designing an ROI identification by applying an edge detector and adoptive thresholding to find RoI. So, in the second model the input images are wisely cropped to the ROI in the first step, and the output is fed to the convoultional network. The results are close to the first model but the model is now more efficient in reducing GPU memory required in the training phase since all input images are cropped to a smaller size. Using contourlet transform as edge detector to find the RoI, slightly improved the Jaccard index by 1%. By employing the optical flow and contourlet feature maps to the network led to a significant improvement in results, nearly 9% improvement in Jaccard index.

The effect of making the network deeper by increasing the number of kernels and convolution layers was investigated to determine the more efficient architecture. To further improve the results, normalization techniques including batch, layer and instant normalization was investigated and higher jaccard index achieved with IN for Apex slices and BN for Mid and Base slices.

Chapter 6

Conclusion

6.1 Introduction

This research aimed to address the limitations of applying deep learning methods in the medical area specifically CNN based methods. From the literature review in Chapter 2, top recent architectures including FCN and U-net were considered and two tasks of medical image/video segmentation on skin and heart disease diagnosis systems were identified. For image segmentation, two sub-tasks of skin lesion border segmentation and dermoscopic segmentation were studied and for video segmentation the task of left ventricular segmentation was investigated. Although many models based on CNN have been proposed in recent years, not all models are easily applicable to medical data due to limitations on access to pretrained models and scarcity of medical dataset. In this research, CNN based architectures were proposed to deal with challenges of training a deep network on medical data. A summery of research contributions are presented in the following section.

6.2 Contributions

Contribution1: In Chapter3, a novel hybrid model inspired from fully convolution network was proposed which addressed multiple tasks on skin disease diagnosis, including a skin lesion border segmentation and dermoscopic feature segmentation. The model was efficient in terms of computation load as the convolution layers of both models shared information beside the fact that

second model which addressed feature segmentation, benefited from receiving the cropped region of interest from mask provided by task 1.

Contribution2: The multi-task model in chapter 3 was among the first few models that evaluated transfer learning on a medical CNN based model from a model pretrained on irrelevant data (natural images). The performance was comparable with the winners of the ISIC 2016, 2017 challenge.

Contribution3: A novel CNN based model was proposed in Chapter 4 which outperformed the existing models for two tasks of lesion segmentation and dermoscopic feature segmentation. The proposed model improved well known CNN based models (FCN and U-net) with integrating appropriate feature maps from the frequency domain, providing a superior understanding of the input to the model. Multiscale and multidirectional representations of the input images from the transform domain were incorporated to the convolutional network that led to a considerable increase in the Jaccard index.

This network benefited from training from scratch and presented significant improvement on noisy images compared to the model that was fine-tuned on a pretrained model on natural images. It also found out that rather than adding the depth of network, integrating the feature maps from the transform domain, not only improves the Jaccard index score, but also significantly decreases the inference time.

Furthermore, the convergence time was significantly reduced by using the optimization technique and for the task of attribute detection, by transfer learning with the pretrained model of Task 1.

Contribution 4: The proposed segmentation model in Chapter 4 was extended to medical video segmentation (Left Ventricular segmentation) in Chapter 5. The feature of frequent motion of LV was used and predicted RoI was fed to the network as a location guide. Also, temporal features provided by optical flow estimation along with contourlet representations fused to the network and significant performance was observed with the addition of this element. The effect of using the Dice loss function and the cross entropy beside changing the network hyper-parameters such as number of kernels or layers to improve the model were discussed as well.

6.3 Future Work

More data provided in the future will improve the performance of applying deep architectures on medical data. CNN based data augmentation techniques such as Generative Adversarial Network (Antoniou et al., 2017), (Yi et al., 2019) and natural data augmentation (Goyal et al., 2018) could be applied as well. There is also potential of improvement by dealing with selected efficient relevant features and new deep architectures.

Moreover, due to robustness and practicality, the proposed frameworks will become a gold standard approach to the analysis of similar image data sets, in particular, medical and biological domains where there is always small number of samples available.

The four-class proposed system could be designed as a multi-class segmentation task and improved via weighted loss function to address the data imbalance too. In order to find the more efficient parameters of the network, optimization techniques could be added to the network to choose the best number of layers and filters.

Three medical datasets were applied to different models in this research. However, tuning a deep network on various medical datasets will provide a pretrained model on medical data that is hardly available at present but could be very beneficial in the medical domain that most datasets are scarce. Although the training would be on different kind of medical data, this still helps the network to converge faster than randomly initialization.

The 2D convolutional model was applied but there is recent research that proposed models based on 3D convolutional neural networks and the proposed system could be implemented for embedded systems, robots and even mobiles to make a simple user interface part of diagnosis system.

Acronyms

ABCD	Asymmetry, Border, Color, Diameter
BN	Batch Normalization
CAD	Computer-Aided Diagnosis
CASH	Color, Architecture, Symmetry, and Homogeneity
CHT	Circle Hough Transform
CNN	Convolutional Neural Network
DFB	Directional Flter Bank
FCN	Fully Convolutional Network
GAN	Generative Adversarial Network
GLCM	Gray Level Co-occurrence Matrix
IN	Instance Normalization
KNN	K-Nearest Neighbour
LN	Layer Normalization
LP	Laplacian Pyramid
LSTM	Long Short-Term Memory

LV Left Ventricular

LVQ Learning Vector Quantization

MLP Multi-Layer Perceptron

MRI Magnetic Resonance Image

NMS Non-maximum suppression

PNN Probabilistic Neural Networks

PNPA Pixel Neighbours Pattern Analysis

R-CNN Regions with CNN Features

RBM Restricted Boltzmann Machines

RPN Region Proposal Networks

SDS Simultaneous Detection and Segmentation

SRG Seeded Region Growing

SVM Support Vector Machine

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. and Zheng, X. (2015), 'TensorFlow: Large-scale machine learning on heterogeneous systems'. Software available from tensorflow.org.
URL: https://www.tensorflow.org/

Abuzaghleh, O., Barkana, B. D. and Faezipour, M. (2015), 'Noninvasive real-time automated skin lesion analysis system for melanoma early detection and prevention', *IEEE journal of translational engineering in health and medicine* **3**, 1–12.

Adams, R. and Bischof, L. (1994), 'Seeded region growing', *IEEE Transactions on pattern analysis and machine intelligence* **16**(6), 641–647.

Al-Amri, S. S., Kalyankar, N. V. et al. (2010), 'Image segmentation by using threshold techniques', *arXiv preprint arXiv:1005.4020* .

Al-Masni, M. A., Al-antari, M. A., Choi, M.-T., Han, S.-M. and Kim, T.-S. (2018), 'Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks', *Computer methods and programs in biomedicine* **162**, 221–231.

Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., Ballas, N., Bastien, F., Bayer, J., Belikov, A., Belopolsky, A., Bengio, Y., Bergeron, A., Bergstra, J., Bisson, V., Bleecher Snyder, J., Bouchard, N., Boulanger-Lewandowski, N., Bouthillier, X., de Brébisson,

A., Breuleux, O., Carrier, P.-L., Cho, K., Chorowski, J., Christiano, P., Cooijmans, T., Côté, M.-A., Côté, M., Courville, A., Dauphin, Y. N., Delalleau, O., Demouth, J., Desjardins, G., Dieleman, S., Dinh, L., Ducoffe, M., Dumoulin, V., Ebrahimi Kahou, S., Erhan, D., Fan, Z., Firat, O., Germain, M., Glorot, X., Goodfellow, I., Graham, M., Gulcehre, C., Hamel, P., Harlouchet, I., Heng, J.-P., Hidasi, B., Honari, S., Jain, A., Jean, S., Jia, K., Korobov, M., Kulkarni, V., Lamb, A., Lamblin, P., Larsen, E., Laurent, C., Lee, S., Lefrancois, S., Lemieux, S., Léonard, N., Lin, Z., Livezey, J. A., Lorenz, C., Lowin, J., Ma, Q., Manzagol, P.-A., Mastropietro, O., McGibbon, R. T., Memisevic, R., van Merriënboer, B., Michalski, V., Mirza, M., Orlandi, A., Pal, C., Pascanu, R., Pezeshki, M., Raffel, C., Renshaw, D., Rocklin, M., Romero, A., Roth, M., Sadowski, P., Salvatier, J., Savard, F., Schlüter, J., Schulman, J., Schwartz, G., Serban, I. V., Serdyuk, D., Shabanian, S., Simon, E., Spieckermann, S., Subramanyam, S. R., Sygnowski, J., Tanguay, J., van Tulder, G., Turian, J., Urban, S., Vincent, P., Visin, F., de Vries, H., Warde-Farley, D., Webb, D. J., Willson, M., Xu, K., Xue, L., Yao, L., Zhang, S. and Zhang, Y. (2016), 'Theano: A Python framework for fast computation of mathematical expressions', *arXiv e-prints* **abs/1605.02688**.

Alamdari, N., MacKinnon, N., Vasefi, F., Fazel-Rezai, R., Alhashim, M., Akhbardeh, A., Farkas, D. L. and Tavakolian, K. (2017), Effect of lesion segmentation in melanoma diagnosis for a mobile health application, *in* '2017 Design of Medical Devices Conference', American Society of Mechanical Engineers Digital Collection.

Alba, X., Figueras i Ventura, R. M., Lekadir, K., Tobon-Gomez, C., Hoogendoorn, C. and Frangi, A. F. (2014), 'Automatic cardiac lv segmentation in mri using modified graph cuts with smoothness and interslice constraints', *Magnetic resonance in medicine* **72**(6), 1775–1784.

Antoniou, A., Storkey, A. and Edwards, H. (2017), 'Data augmentation generative adversarial networks', *arXiv preprint arXiv:1711.04340* .

Arbeláez, P., Pont-Tuset, J., Barron, J. T., Marques, F. and Malik, J. (2014), Multiscale combinatorial grouping, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 328–335.

Badrinarayanan, V., Kendall, A. and Cipolla, R. (2017), 'Segnet: A deep convolutional encoder-

decoder architecture for image segmentation', *IEEE transactions on pattern analysis and machine intelligence* **39**(12), 2481–2495.

Baldi, P. and Sadowski, P. J. (2013), Understanding dropout, *in* 'Advances in neural information processing systems', pp. 2814–2822.

Bar, Y., Diamant, I., Wolf, L. and Greenspan, H. (2015), Deep learning with non-medical training used for chest pathology identification, *in* 'Medical Imaging 2015: Computer-Aided Diagnosis', Vol. 9414, International Society for Optics and Photonics, p. 94140V.

Bengio, Y. (2012), Practical recommendations for gradient-based training of deep architectures, *in* 'Neural networks: Tricks of the trade', Springer, pp. 437–478.

Bengio, Y. and Lee, H. (2015), 'Editorial introduction to the neural networks special issue on deep learning of representations', *Neural Networks* **64**(C), 1–3.

Benjamin, E. J., Virani, S. S., Callaway, C. W., Chamberlain, A. M., Chang, A. R., Cheng, S., Chiuve, S. E., Cushman, M., Delling, F. N., Deo, R. et al. (2018), 'Heart disease and stroke statistics-2018 update: a report from the american heart association.', *Circulation* **137**(12), e67.

Bi, L., Kim, J., Ahn, E., Kumar, A., Fulham, M. and Feng, D. (2017), 'Dermoscopic image segmentation via multistage fully convolutional networks', *IEEE Transactions on Biomedical Engineering* **64**(9), 2065–2074.

Burt, P. J. and Adelson, E. H. (1987), The laplacian pyramid as a compact image code, *in* 'Readings in computer vision', Elsevier, pp. 671–679.

Canny, J. (1986), 'A computational approach to edge detection', *IEEE Transactions on pattern analysis and machine intelligence* (6), 679–698.

Chaudhuri, D. and Agrawal, A. (2010), 'Split-and-merge procedure for image segmentation using bimodality detection approach', *Defence Science Journal* **60**(3), 290–301.

Chen, E. Z., Dong, X., Wu, J., Jiang, H., Li, X. and Rong, R. (2018), 'Lesion attributes segmentation for melanoma detection with deep learning', *bioRxiv* p. 381855.

Cheng, K.-S., Lin, J.-S. and Mao, C.-W. (1996), 'The application of competitive hopfield neural network to medical image segmentation', *IEEE transactions on medical imaging* **15**(4), 560–567.

Ciecholewski, M. (2016), 'An edge-based active contour model using an inflation/deflation force with a damping coefficient', *Expert Systems with Applications* **44**, 22–36.

Cireşan, D. C., Giusti, A., Gambardella, L. M. and Schmidhuber, J. (2013), Mitosis detection in breast cancer histology images with deep neural networks, *in* 'International Conference on Medical Image Computing and Computer-assisted Intervention', Springer, pp. 411–418.

Ciresan, D., Giusti, A., Gambardella, L. and Schmidhuber, J. (2012), 'Deep neural networks segment neuronal membranes in electron microscopy images', *Advances in neural information processing systems* **25**, 2843–2851.

Codella, N. C. F., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S. W., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M. A., Kittler, H. and Halpern, A. (2019), 'Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)', *CoRR* **abs/1902.03368**.

Codella, N. C., Weinsaft, J. W., Cham, M. D., Janik, M., Prince, M. R. and Wang, Y. (2008), 'Left ventricle: automated segmentation by using myocardial effusion threshold reduction and intravoxel computation at mr imaging', *Radiology* **248**(3), 1004–1012.

Collobert, R., Kavukcuoglu, K. and Farabet, C. (2011), Torch7: A matlab-like environment for machine learning, *in* 'BigLearn, NIPS workshop', number CONF.

Cuevas, E., Zaldivar, D., Perez, M. and Sanchez, E. N. (2009), 'Lvq neural networks applied to face segmentation', *Intelligent Automation & Soft Computing* **15**(3), 439–450.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. (2009), Imagenet: A large-scale hierarchical image database, *in* '2009 IEEE conference on computer vision and pattern recognition', Ieee, pp. 248–255.

Deserno, T. M. (2011), Fundamentals of medical image processing, *in* 'Springer Handbook of Medical Technology', Springer, pp. 1139–1165.

Do, M. N. and Vetterli, M. (2005), 'The contourlet transform: an efficient directional multiresolution image representation', *IEEE Transactions on image processing* **14**(12), 2091–2106.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E. and Darrell, T. (2014), Decaf: A deep convolutional activation feature for generic visual recognition, *in* 'International conference on machine learning', pp. 647–655.

Dong, S., Luo, G., Wang, K., Cao, S., Li, Q. and Zhang, H. (2018), 'A combined fully convolutional networks and deformable model for automatic left ventricle segmentation based on 3d echocardiography', *BioMed research international* **2018**.

Ercal, F., Chawla, A., Stoecker, W. V., Lee, H.-C. and Moss, R. H. (1994), 'Neural network diagnosis of malignant melanoma from color images', *IEEE Transactions on biomedical engineering* **41**(9), 837–845.

Erkol, B., Moss, R. H., Joe Stanley, R., Stoecker, W. V. and Hvatum, E. (2005), 'Automatic lesion boundary detection in dermoscopy images using gradient vector flow snakes', *Skin Research and Technology* **11**(1), 17–26.

Fabijańska, A. (2011), Variance filter for edge detection and edge-based image segmentation, *in* 'Perspective Technologies and Methods in MEMS Design', IEEE, pp. 151–154.

Fan, M. and Lee, T. C. (2014), 'Variants of seeded region growing', *IET image processing* **9**(6), 478–485.

Farabet, C., Couprie, C., Najman, L. and LeCun, Y. (2012), 'Learning hierarchical features for scene labeling', *IEEE transactions on pattern analysis and machine intelligence* **35**(8), 1915–1929.

Ferlay, J., Steliarova-Foucher, E., Lortet-Tieulent, J., Rosso, S., Coebergh, J.-W. W., Comber, H., Forman, D. and Bray, F. (2013), 'Cancer incidence and mortality patterns in europe: estimates for 40 countries in 2012', *European journal of cancer* **49**(6), 1374–1403.

Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J. and Greenspan, H. (2018), 'Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification', *Neurocomputing* **321**, 321–331.

Gang, L. (2005), 'Remote sensing image segmentation with probabilistic neural networks', *Geospatial information science* **8**(1), 28–32.

Garnavi, R., Aldeen, M. and Bailey, J. (2012), 'Computer-aided diagnosis of melanoma using border-and wavelet-based texture analysis', *IEEE Transactions on Information Technology in Biomedicine* **16**(6), 1239–1252.

Girshick, R. (2015), Fast r-cnn, *in* 'Proceedings of the IEEE international conference on computer vision', pp. 1440–1448.

Girshick, R., Donahue, J., Darrell, T. and Malik, J. (2015), 'Region-based convolutional networks for accurate object detection and segmentation', *IEEE transactions on pattern analysis and machine intelligence* **38**(1), 142–158.

Glorot, X. and Bengio, Y. (2010), Understanding the difficulty of training deep feedforward neural networks, *in* 'Proceedings of the thirteenth international conference on artificial intelligence and statistics', pp. 249–256.

Gonzalez, R. C. and Wintz, P. (1977), 'Digital image processing(book)', *Reading, Mass., Addison-Wesley Publishing Co., Inc.(Applied Mathematics and Computation* (13), 451.

Goodfellow, I., Bengio, Y., Courville, A. and Bengio, Y. (2016), *Deep learning*, Vol. 1, MIT press Cambridge.

Goyal, M., Hassanpour, S. and Yap, M. H. (2018), 'Region of interest detection in dermoscopic images for natural data-augmentation', *arXiv preprint arXiv:1807.10711* .

Gutman, D., Codella, N. C., Celebi, E., Helba, B., Marchetti, M., Mishra, N. and Halpern, A. (2016), 'Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic)', *arXiv preprint arXiv:1605.01397* .

Hadhoud, M., Eladawy, M. I., Farag, A., Montevecchi, F. M. and Morbiducci, U. (2012), 'Left ventricle segmentation in cardiac mri images', *American Journal of Biomedical Engineering* **2**(3), 131–135.

Haider, W., Malik, M. S., Raza, M., Wahab, A., Khan, I. A., Zia, U., Tanveer, J. and Bashir, H. (2012), 'A hybrid method for edge continuity based on pixel neighbors pattern analysis (pnpa) for remote sensing satellite images', *Int'l J. of Communications, Network and System Sciences* **5**(29), 624–630.

Harangi, B. (2018), 'Skin lesion classification with ensembles of deep convolutional neural networks', *Journal of biomedical informatics* **86**, 25–32.

He, K., Zhang, X., Ren, S. and Sun, J. (2015), 'Spatial pyramid pooling in deep convolutional networks for visual recognition', *IEEE transactions on pattern analysis and machine intelligence* **37**(9), 1904–1916.

He, K., Zhang, X., Ren, S. and Sun, J. (2016), Deep residual learning for image recognition, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 770–778.

Henning, J. S., Dusza, S. W., Wang, S. Q., Marghoob, A. A., Rabinovitz, H. S., Polsky, D. and Kopf, A. W. (2007), 'The cash (color, architecture, symmetry, and homogeneity) algorithm for dermoscopy', *Journal of the American Academy of Dermatology* **56**(1), 45–52.

Hinton, G. E. (2009), 'Deep belief networks', *Scholarpedia* **4**(5), 5947.

Hinton, G. E., Osindero, S. and Teh, Y.-W. (2006), 'A fast learning algorithm for deep belief nets', *Neural computation* **18**(7), 1527–1554.

Huang, D.-S., Bevilacqua, V. and Premaratne, P. (2016), *Intelligent Computing Theories and Application: 12th International Conference, ICIC 2016, Lanzhou, China, August 2-5, 2016, Proceedings*, Vol. 9771, Springer.

Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K. Q. (2017), Densely connected convolutional networks, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 4700–4708.

Huang, S., Liu, J., Lee, L. C., Venkatesh, S. K., San Teo, L. L., Au, C. and Nowinski, W. L. (2011), 'An image-based comprehensive approach for automatic segmentation of left ventricle from cardiac short axis cine mr images', *Journal of digital imaging* **24**(4), 598–608.

Hussain, Z., Gimenez, F., Yi, D. and Rubin, D. (2017), Differential data augmentation techniques for medical imaging classification tasks, *in* 'AMIA Annual Symposium Proceedings', Vol. 2017, American Medical Informatics Association, p. 979.

Ikonomakis, N., Plataniotis, K. N. and Venetsanopoulos, A. N. (2000), 'Color image segmentation for multimedia applications', *Journal of Intelligent and Robotic Systems* **28**(1-2), 5–20.

Ioffe, S. and Szegedy, C. (2015), Batch normalization: Accelerating deep network training by reducing internal covariate shift, *in* 'International conference on machine learning', PMLR, pp. 448–456.

Iscan, Z., Yüksel, A., Dokur, Z., Korürek, M. and Ölmez, T. (2009), 'Medical image segmentation with transform and moment based features and incremental supervised neural network', *Digital Signal Processing* **19**(5), 890–901.

Jahanifar, M., Tajeddin, N. Z., Asl, B. M. and Gooya, A. (2018), 'Supervised saliency map driven segmentation of lesions in dermoscopic images', *IEEE journal of biomedical and health informatics* **23**(2), 509–518.

Jain, R., Kasturi, R. and Schunck, B. G. (1995), *Machine vision*, Vol. 5, McGraw-Hill New York.

Jaisakthi, S. M., Mirunalini, P. and Aravindan, C. (2018), 'Automated skin lesion segmentation of dermoscopic images using grabcut and k-means algorithms', *IET Computer Vision* **12**(8), 1088–1095.

Jégou, S., Drozdzal, M., Vazquez, D., Romero, A. and Bengio, Y. (2017), The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops', pp. 11–19.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S. and Darrell, T. (2014), 'Caffe: Convolutional architecture for fast feature embedding', *arXiv preprint arXiv:1408.5093* .

Jolly, M.-P. (2006), 'Automatic segmentation of the left ventricle in cardiac mr and ct images', *International Journal of Computer Vision* **70**(2), 151–163.

Kasmi, R. and Mokrani, K. (2016), 'Classification of malignant melanoma and benign skin lesions: implementation of automatic abcd rule', *IET Image Processing* **10**(6), 448–455.

Kaus, M. R., Von Berg, J., Weese, J., Niessen, W. and Pekar, V. (2004), 'Automated segmentation of the left ventricle in cardiac mri', *Medical image analysis* **8**(3), 245–254.

Khened, M., Kollerathu, V. A. and Krishnamurthi, G. (2019), 'Fully convolutional multi-scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers', *Medical image analysis* **51**, 21–45.

Kingma, D. P. and Ba, J. (2014), 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980* .

Kittler, H., Seltenheim, M., Dawid, M., Pehamberger, H., Wolff, K. and Binder, M. (1999), 'Morphologic changes of pigmented skin lesions: a useful extension of the abcd rule for dermatoscopy', *Journal of the American Academy of Dermatology* **40**(4), 558–562.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012), Imagenet classification with deep convolutional neural networks, *in* 'Advances in neural information processing systems', pp. 1097–1105.

Kwasigroch, A., Mikołajczyk, A. and Grochowski, M. (2017), Deep neural networks approach to skin lesions classification—a comparative analysis, *in* '2017 22nd International Conference on Methods and Models in Automation and Robotics (MMAR)', IEEE, pp. 1069–1074.

Lau, H., Chang, J., Daut, N., Tahir, A., Samino, E. and Hijazi, M. (2018), 'Exploring edge-based segmentation towards automated skin lesion diagnosis', *Advanced Science Letters* **24**(2), 1095–1099.

LeCun, Y., Bengio, Y. and Hinton, G. (2015), 'Deep learning', *nature* **521**(7553), 436.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. et al. (1998), 'Gradient-based learning applied to document recognition', *Proceedings of the IEEE* **86**(11), 2278–2324.

Li, C., Yang, Y., Feng, M., Chakradhar, S. and Zhou, H. (2016), Optimizing memory efficiency for deep convolutional neural networks on gpus, *in* 'SC'16: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis', IEEE, pp. 633–644.

Liang, D., Yang, F., Zhang, T. and Yang, P. (2018), 'Understanding mixup training methods', *IEEE Access* **6**, 58774–58783.

Lin, G., Milan, A., Shen, C. and Reid, I. (2017), Refinenet: Multi-path refinement networks for high-resolution semantic segmentation, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 1925–1934.

Liu, Z. and Zerubia, J. (2015), 'Skin image illumination modeling and chromophore identification for melanoma diagnosis', *Physics in Medicine & Biology* **60**(9), 3415.

Long, J., Shelhamer, E. and Darrell, T. (2015), Fully convolutional networks for semantic segmentation, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 3431–3440.

Lu, J., Feng, C., Li, W. and Zhao, D. (2019), Roi localization and initialization method for left ventricle segmentation, *in* 'Proceedings of the Third International Symposium on Image Computing and Digital Medicine', pp. 12–16.

Masoumi, H., Behrad, A., Pourmina, M. A. and Roosta, A. (2012), 'Automatic liver segmentation in mri images using an iterative watershed algorithm and artificial neural network', *Biomedical signal processing and control* **7**(5), 429–437.

Monteiro, F. C. and Campilho, A. (2008), Watershed framework to region-based image segmentation, *in* '2008 19th International Conference on Pattern Recognition', IEEE, pp. 1–4.

Moradi, S., Oghli, M. G., Alizadehasl, A., Shiri, I., Oveisi, N., Oveisi, M., Maleki, M. and Dhooge, J. (2019), 'Mfp-unet: A novel deep learning based approach for left ventricle segmentation in echocardiography', *Physica Medica* **67**, 58–69.

Mukherjee, J., Shaikh, S. H., Kar, M. and Chakrabarti, A. (2016), A comparative analysis of image segmentation techniques toward automatic risk prediction of solitary pulmonary nodules, *in* 'Advanced Computing and Systems for Security', Springer, pp. 159–179.

Nachbar, F., Stolz, W., Merkle, T., Cognetta, A. B., Vogt, T., Landthaler, M., Bilek, P., Braun-Falco, O. and Plewig, G. (1994), 'The abcd rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions', *Journal of the American Academy of Dermatology* **30**(4), 551–559.

Nambakhsh, C. M., Yuan, J., Punithakumar, K., Goela, A., Rajchl, M., Peters, T. M. and Ayed, I. B. (2013), 'Left ventricle segmentation in mri via convex relaxed distribution matching', *Medical image analysis* **17**(8), 1010–1024.

Nasr-Esfahani, E., Rafiei, S., Jafari, M. H., Karimi, N., Wrobel, J. S., Soroushmehr, S., Samavi, S. and Najarian, K. (2017), 'Dense fully convolutional network for skin lesion segmentation', *arXiv preprint arXiv:1712.10207* .

Nasr-Esfahani, M., Mohrekesh, M., Akbari, M., Soroushmehr, S. R., Nasr-Esfahani, E., Karimi, N., Samavi, S. and Najarian, K. (2018), Left ventricle segmentation in cardiac mr images using fully convolutional network, *in* '2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)', IEEE, pp. 1275–1278.

Navarro, F., Escudero-Viñolo, M. and Bescós, J. (2018), 'Accurate segmentation and registration of skin lesion images to evaluate lesion change', *IEEE journal of biomedical and health informatics* **23**(2), 501–508.

Nguyen, D. C. T., Benameur, S., Mignotte, M. and Lavoie, F. (2018), 'Superpixel and multi-atlas based fusion entropic model for the segmentation of x-ray images', *Medical Image Analysis* **48**, 58–74.

Nichols, M., Townsend, N., Scarborough, P. and Rayner, M. (2014), 'Cardiovascular disease in europe 2014: epidemiological update', *European heart journal* **35**(42), 2950–2959.

Ohlander, R., Price, K. and Reddy, D. R. (1978), 'Picture segmentation using a recursive region splitting method', *Computer Graphics and Image Processing* **8**(3), 313–333.

Oliveira, R. B., Mercedes Filho, E., Ma, Z., Papa, J. P., Pereira, A. S. and Tavares, J. M. R. (2016), 'Computational methods for the image segmentation of pigmented skin lesions: a review', *Computer methods and programs in biomedicine* **131**, 127–141.

Orr, G. B. and Müller, K.-R. (2003), *Neural networks: tricks of the trade*, Springer.

Pereira, S., Pinto, A., Alves, V. and Silva, C. A. (2016), 'Brain tumor segmentation using convolutional neural networks in mri images', *IEEE transactions on medical imaging* **35**(5), 1240–1251.

Petitjean, C. and Dacher, J.-N. (2011), 'A review of segmentation methods in short axis cardiac mr images', *Medical image analysis* **15**(2), 169–184.

Pinheiro, P. H. and Collobert, R. (2014), Recurrent convolutional neural networks for scene labeling, *in* '31st International Conference on Machine Learning (ICML)', number CONF.

Pour, M. P. and Seker, H. (2020), 'Transform domain representation-driven convolutional neural networks for skin lesion segmentation', *Expert Systems with Applications* **144**, 113129.

Pour, M. P., Seker, H. and Shao, L. (2017), Automated lesion segmentation and dermoscopic feature segmentation for skin cancer analysis, *in* '2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)', IEEE, pp. 640–643.

Qi, X., Xing, F., Foran, D. J. and Yang, L. (2011), 'Robust segmentation of overlapping cells in histopathology specimens using parallel seed detection and repulsive level set', *IEEE Transactions on Biomedical Engineering* **59**(3), 754–765.

Queirós, S., Barbosa, D., Heyde, B., Morais, P., Vilaça, J. L., Friboulet, D., Bernard, O. and D'hooge, J. (2014), 'Fast automatic myocardial segmentation in 4d cine cmr datasets', *Medical image analysis* **18**(7), 1115–1131.

Razali, M. R. M., Ahmad, N. S., Hassan, R., Zaki, Z. M. and Ismail, W. (2014), Sobel and canny edges segmentations for the dental age assessment, *in* '2014 International Conference on Computer Assisted System in Health', IEEE, pp. 62–66.

Ren, S., He, K., Girshick, R. and Sun, J. (2015), Faster r-cnn: Towards real-time object detection with region proposal networks, *in* 'Advances in neural information processing systems', pp. 91–99.

Rezaei, M., Yang, H. and Meinel, C. (2017), Deep neural network with l2-norm unit for brain lesions detection, *in* 'International Conference on Neural Information Processing', Springer, pp. 798–807.

Riaz, F., Naeem, S., Nawaz, R. and Coimbra, M. (2018), 'Active contours based segmentation and lesion periphery analysis for characterization of skin lesions in dermoscopy images', *IEEE journal of biomedical and health informatics* **23**(2), 489–500.

Ronneberger, O., Fischer, P. and Brox, T. (2015), U-net: Convolutional networks for biomedical image segmentation, *in* 'International Conference on Medical image computing and computer-assisted intervention', Springer, pp. 234–241.

Rout, S., Srivastava, P., Majumdar, J. et al. (1998), 'Multi-modal image segmentation using a modified hopfield neural network', *Pattern Recognition* **31**(6), 743–750.

Senthilkumaran, N. and Vaithegi, S. (2016), 'Image segmentation by using thresholding techniques for medical images', *Computer Science & Engineering: An International Journal* **6**(1), 1–13.

Shih, F. Y. and Cheng, S. (2005), 'Automatic seeded region growing for color image segmentation', *Image and vision computing* **23**(10), 877–886.

Shoaib, M. A., Lai, K. W., Khalil, A. and Chuah, J. H. (2019), Mask r-cnn for segmentation of left ventricle, *in* 'International Conference for Innovation in Biomedical Engineering and Life Sciences', Springer, pp. 14–22.

Siegel, R. L., Miller, K. D. and Jemal, A. (2016), 'Cancer statistics, 2016', *CA: a cancer journal for clinicians* **66**(1), 7–30.

Smolensky, P. (1986), Information processing in dynamical systems: Foundations of harmony theory, Technical report, Colorado Univ at Boulder Dept of Computer Science.

Soler, L., Delingette, H., Malandain, G., Montagnat, J., Ayache, N., Koehl, C., Dourthe, O., Malassagne, B., Smith, M., Mutter, D. et al. (2001), 'Fully automatic anatomical, pathological, and functional segmentation from ct scans for hepatic surgery', *Computer Aided Surgery* **6**(3), 131–142.

Springenberg, J. T., Dosovitskiy, A., Brox, T. and Riedmiller, M. (2014), 'Striving for simplicity: The all convolutional net', *arXiv preprint arXiv:1412.6806* .

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014), 'Dropout: a simple way to prevent neural networks from overfitting', *The journal of machine learning research* **15**(1), 1929–1958.

Suinesiaputra, A., Cowan, B. R., Al-Agamy, A. O., Elattar, M. A., Ayache, N., Fahmy, A. S., Khalifa, A. M., Medrano-Gracia, P., Jolly, M.-P., Kadish, A. H. et al. (2014), 'A collaborative resource to build consensus for automated left ventricular segmentation of cardiac mr images', *Medical image analysis* **18**(1), 50–62.

Tajeddin, N. Z. and Asl, B. M. (2018), 'Melanoma recognition in dermoscopy images using lesion's peripheral region information', *Computer methods and programs in biomedicine* **163**, 143–153.

Tan, L. K., Liew, Y. M., Lim, E. and McLaughlin, R. A. (2017), 'Convolutional neural network regression for short-axis left ventricle segmentation in cardiac cine mr sequences', *Medical image analysis* **39**, 78–86.

Tang, J. (2010), A color image segmentation algorithm based on region growing, *in* '2010 2nd International Conference on Computer Engineering and Technology', Vol. 6, IEEE, pp. V6–634.

Teimouri, N., Omid, M., Mollazade, K. and Rajabipour, A. (2014), 'A novel artificial neural networks assisted segmentation algorithm for discriminating almond nut and shell from background and shadow', *Computers and electronics in agriculture* **105**, 34–43.

Torbati, N., Ayatollahi, A. and Kermani, A. (2014), 'An efficient neural network based method for medical image segmentation', *Computers in biology and medicine* **44**, 76–87.

Tran, P. V. (2016), 'A fully convolutional neural network for cardiac segmentation in short-axis mri', *arXiv preprint arXiv:1604.00494* .

Uijlings, J. R., Van De Sande, K. E., Gevers, T. and Smeulders, A. W. (2013), 'Selective search for object recognition', *International journal of computer vision* **104**(2), 154–171.

Visin, F., Ciccone, M., Romero, A., Kastner, K., Cho, K., Bengio, Y., Matteucci, M. and Courville, A. (2016), Reseg: A recurrent neural network-based model for semantic segmentation, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops', pp. 41–48.

Wang, H., Roa, A. C., Basavanhally, A. N., Gilmore, H. L., Shih, N., Feldman, M., Tomaszewski, J., Gonzalez, F. and Madabhushi, A. (2014), 'Mitosis detection in breast cancer pathology im-

ages by combining handcrafted and convolutional neural network features', *Journal of Medical Imaging* **1**(3), 034003.

Wang, X., Yang, S., Fang, Y., Wei, Y., Wang, M., Zhang, J. and Han, X. (2021), 'Sk-unet: An improved u-net model with selective kernel for the segmentation of lge cardiac mr images', *IEEE Sensors Journal* **21**(10), 11643–11653.

Wong, A., Scharcanski, J. and Fieguth, P. (2011), 'Automatic skin lesion segmentation via iterative stochastic region merging', *IEEE Transactions on Information Technology in Biomedicine* **15**(6), 929–936.

Wu, B., Fang, Y. and Lai, X. (2020), 'Left ventricle automatic segmentation in cardiac mri using a combined cnn and u-net approach', *Computerized Medical Imaging and Graphics* **82**, 101719.

Xiao, Z., Shi, J. and Chang, Q. (2009), Automatic image segmentation algorithm based on pcnn and fuzzy mutual information, *in* '2009 Ninth IEEE International Conference on Computer and Information Technology', Vol. 1, IEEE, pp. 241–245.

Yan, W., Wang, Y., Li, Z., Van Der Geest, R. J. and Tao, Q. (2018), Left ventricle segmentation via optical-flow-net from short-axis cine mri: preserving the temporal coherence of cardiac motion, *in* 'International Conference on Medical Image Computing and Computer-Assisted Intervention', Springer, pp. 613–621.

Yi, X., Walia, E. and Babyn, P. (2019), 'Generative adversarial network in medical imaging: A review', *Medical image analysis* **58**, 101552.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T. and Lipson, H. (2015), 'Understanding neural networks through deep visualization', *arXiv preprint arXiv:1506.06579* .

Yu, F. and Koltun, V. (2015), 'Multi-scale context aggregation by dilated convolutions', *arXiv preprint arXiv:1511.07122* .

Yu, L., Chen, H., Dou, Q., Qin, J. and Heng, P.-A. (2016), 'Automated melanoma recognition in dermoscopy images via very deep residual networks', *IEEE transactions on medical imaging* **36**(4), 994–1004.

Yuan, X., Situ, N. and Zouridakis, G. (2009), 'A narrow band graph partitioning method for skin lesion segmentation', *Pattern Recognition* **42**(6), 1017–1028.

Yüksel, M. E. and Borlu, M. (2009), 'Accurate segmentation of dermoscopic images by image thresholding based on type-2 fuzzy logic', *IEEE Transactions on Fuzzy Systems* **17**(4), 976–982.

Zhang, H., Cisse, M., Dauphin, Y. N. and Lopez-Paz, D. (2017), 'mixup: Beyond empirical risk minimization', *arXiv preprint arXiv:1710.09412* .

Zhang, W., Li, R., Deng, H., Wang, L., Lin, W., Ji, S. and Shen, D. (2015), 'Deep convolutional neural networks for multi-modality isointense infant brain image segmentation', *NeuroImage* **108**, 214–224.

Zhao, W., Zhang, J., Li, P. and Li, Y. (2010), Study of image segmentation algorithm based on textural features and neural network, *in* '2010 International Conference on Intelligent Computing and Cognitive Informatics', IEEE, pp. 300–303.

Zheng, Q., Delingette, H., Duchateau, N. and Ayache, N. (2018), '3-d consistent and robust segmentation of cardiac images by deep learning with spatial propagation', *IEEE transactions on medical imaging* **37**(9), 2137–2148.

Zhou, H., Schaefer, G., Sadka, A. H. and Celebi, M. E. (2009), 'Anisotropic mean shift based fuzzy c-means segmentation of dermoscopy images', *IEEE Journal of Selected Topics in Signal Processing* **3**(1), 26–34.

Zhou, X.-Y. and Yang, G.-Z. (2019), 'Normalization in training u-net for 2-d biomedical semantic segmentation', *IEEE Robotics and Automation Letters* **4**(2), 1792–1799.

Zortea, M., Flores, E. and Scharcanski, J. (2017), 'A simple weighted thresholding method for the segmentation of pigmented skin lesions in macroscopic images', *Pattern Recognition* **64**, 92–104.