Automated Diagnosis of Retinal Diseases from Optical Coherence Tomography Images and Volumes using Deep Learning

Vineeta Das

Automated Diagnosis of Retinal Diseases from Optical Coherence Tomography Images and Volumes using Deep Learning

A

thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

by

VINEETA DAS

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, ASSAM, INDIA

Nov 2021

Certificate

This is to certify that the thesis entitled "**Automated Diagnosis of Retinal Diseases from Optical Coherence Tomography Images and Volumes using Deep Learning**", submitted by **Vineeta Das**, Roll No. 166102012, a research scholar in the *Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati*, for the award of the degree of **Doctor of Philosophy**, is a record of an original research work carried out by her under my supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the institute and, in my opinion, has reached the standard needed for the submission. The results embodied in this thesis have not been submitted to any other university or institute for the award of any degree or diploma.

Prof. Samarendra Dandapat
Professor
Dept. of Electronics and Electrical Engg.
Indian Institute of Technology Guwahati
Guwahati - 781 039, Assam, India.

Prof. Prabin Kumar Bora
Professor
Dept. of Electronics and Electrical Engg.
Indian Institute of Technology Guwahati
Guwahati - 781 039, Assam, India.

To,

My Parents

(*Shri. Madan Mohan Das and Smt. Susmita Das*)

and

Family Members

for their blessings, encouragement and support.

Acknowledgements

I express my deepest gratitude to my supervisors Prof. S. Dandapat and Prof. P. K. Bora for their dedicated mentorship, encouragement and guidance in life. Their insightful feedbacks have helped me very much in improving the quality of my thesis. I greatly admire their attitude towards research, creative thinking and enthusiasm for work.

I am grateful to my doctoral committee chairman Dr. T. Jacob and members Prof. M. K. Bhuyan and Dr. S. Kashyap for the insightful comments and constructive criticisms, which helped me bring my work to the current form. I am also thankful to my former doctoral committee chairman, Prof. S.R.M. Prasanna for his advice, encouragement and assessment of my work. I would also like to thank all other faculty members of the Department of Electronics and Electrical Engineering, IIT Guwahati, for their care and support. I am also very thankful to all the technical, office, security, canteen, and maintenance staff members of the Department for their help when required. Special thanks to Mr. Bhriguraj Borah, technical staff of the Department of Computer Science (CSE) for maintaining the GPU facility well.

I wish to acknowledge Prof. Sina Farsiu of Duke University and Prof. Hossein Rabbani of Isfahan University of Medical Sciences for curating standardized OCT databases that were used in this research work. I extend my sincere thanks to Dr. Mohit Garg from Sri Sankaradeva Nethralaya, Guwahati, India, for finding time from his busy schedule to evaluate my method.

My seniors Ganji Sreeram, Vivek Venugopal and Jiss J. Nallikuzhy deserve a special mention for their uplifting conversations, technical discussions and care. I thank my dear friends Eedara Prabhakararao, Pradipta Sasmal, Suchismita Sasmal, Tilendra Choudhary, Shikha Baghel, Archana Sahu, Sweta Balchandani and Preety Sagar Talukdar for creating memorable experiences. I am also thankful to my lab mates Ato Kapfo, Alex Paul Kamson, Sibasis Sahoo, Debasish Jyotishi, Samarjeet Das, Mousmi Das, Pharvesh Salman Choudhary, Rohan Gupta, Himashree Kalita, James Singh, Omesh Singh and research scholars of Signal Informatics and Signal Processing lab for their unsolicited help and support. A heartful thanks to my friends Dipankar, Vanshali and Gyanendro for their timely help and technical discussions.

I wish to acknowledge the Department of Biotechnology, Government of India for supporting this work under the North East Centre for Biological Sciences and Healthcare Engineering (NECBH) Project BT/COE/34/SP28408/2018.

Lastly, my deepest gratitude goes to my parents and family. Their love, support and understanding were essential for the successful completion of my Ph.D.

Vineeta Das

Abstract

Optical coherence tomography (OCT) is a specialized imaging modality that allows 3D cross-sectional imaging of the retinal tissues at micron-scale resolution. The non-invasive, real-time and in situ image acquisition attributes have made it an essential diagnostic tool in clinical ophthalmology. In practice, the ophthalmologists manually examine the cross-sectional images (B- scans) and the 3D OCT volumes to diagnose retinal diseases. However, the poor resolution of the OCT images and the presence of speckle noise degrade the image quality and bottleneck the quantification of subtle clinical details during diagnosis. Further, the heterogeneity in the characteristics of retinal diseases renders manual evaluation by ophthalmic experts inherently subjective and prone to human error. This dissertation focuses on developing automated classification models that can address the above challenges for the reliable and accurate diagnosis of retinal diseases from OCT B-scans and volumes.

The existing automated methods have mostly used single-scale convolutional neural network (CNN) frameworks to classify the OCT B-scans. However, considering the diversity of the disease manifestations in terms of shape, size, texture, spatial locations and severity stages, these methods may not generalize well to unseen OCT images. In this thesis, the multi-scale behaviour of the retinal lesions is analyzed and explored for extracting powerful discriminative features towards an efficient classification of the OCT B-scan images. To begin with, the multi-scale spatial pyramid decomposition (MSSP) is adopted to obtain the multi-scale views of the images. The obtained images are then fed to the corresponding CNN based feature extractors to obtain the multi-scale features for classification. To further improve the classification performance, we demonstrate the use of a learnable multi-scale feature extraction scheme using dilated convolutions. The proposed learnable multi-scale deep feature fusion (LM-DFF) method uses multiple convolution filters with different dilation rates to extract scale-specific features from the input OCT images. These features are fused to obtain discriminative high-level multi-scale representation for classification. The experimental results verify that the proposed method has superior predictive ability and low test run time compared to the existing methods. These features make it highly suitable for providing a reliable and

fast preliminary diagnosis in eye care centres and hospitals.

The inherent speckle noise and the poor resolution of the OCT B-scans also pose huge challenges in the diagnosis of early and intermediate stages of retinal diseases. In these stages, the diseases are usually asymptomatic in nature. The sparse appearances and small sizes of the lesions at these stages often result in missed detection from the noisy and low resolution (LR) OCT images. Enhancing the diagnostic details using a simultaneous denoising and super-resolution (SR) stage before classification can improve the performance. The traditional denoising and SR frameworks require large amounts of paired noisy LR and clean high-resolution (HR) images for supervised learning. However, the differing standards of the health care industries and the medical data privacy laws hinder the acquisition of large-scale paired LR-HR OCT images for efficient supervised learning. Therefore, the thesis proposes an unsupervised denoising and SR scheme using the generative adversarial network (GAN) that does not demand one-to-one alignment of the LR-HR images during model training. The adversarial learning of the GAN, along with the cycle consistency and the identity mapping priors preserve the spatial correlation, colour and texture details in the generated clean HR OCT images. The experimental results on clinical-grade OCT images show that the proposed framework outperforms the existing methods both in terms of SR performance and test run time. The method reliably reconstructs the clinical details of the OCT images and enhances their diagnostic utility. The results also verify that performing denoising and SR prior to classification can significantly improve the classification of intermediate age-related macular degeneration (AMD) from healthy control subjects.

The OCT volumes are a collection of B-scans sampled from closely spaced retinal locations. As the volumes contain clinical information from a large retinal region, they provide useful insights into the severity of the diseases. The automated classification of OCT volumes can help retinal experts in diagnosing progressive retinal diseases and predict the severity stages. Most existing methods have extended the B-scan classification methods to OCT volume level by aggregating the classification decisions from the individual B-scans using manual threshold-based inference strategies. However, the retinal diseases are mostly local and may not manifest in all B-scans of the volume. It is essential to identify the salient B-scans that exhibit disease characteristics for reliable classification. To address this challenge, a B-scan attentive CNN (BACNN) method is proposed that automatically provides weights of importance to the B-scans in the volumes based on their diagnostic relevance during classification. Specifically, a self-attention mechanism emphasizes the salient B-scan features for learning discriminative features for better classification performance. Experimental results illustrate the efficacy of the proposed method over the existing approaches. An important advantage of the approach is its interpretability since the attention

weights provide meaningful values for the salience of the B-scans in the volume.

Keywords: Optical coherence tomography (OCT), classification, super-resolution, deep learning, convolutional neural network (CNN), attention.

Contents

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

List of Figures

List of Tables

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

List of Acronyms

ACC	Accuracy
AMD	Age related macular degeneration
AUC	Area under the curve
BACNN	B-scan attentive convolutional neural network
BN	Batch normalization
BoW	Bag-of-words
CADNet	Convolutional attention to diabetic macular edema network
CAM	Class activation map
CNN	Convolutional neural network
CNR	Contrast to noise ratio
CNV	Choroidal neovascularization
DL	Deep learning
DME	Diabetic macular edema
DUD	Duke University denoising
DUIA	Duke University intermediate AMD
DUSR	Duke University super-resolution
FC	Fully connected
ELM	External limiting membrane
EZ	Ellipsoid zone
FA	Fluorescein Angiography
GA	Geographic atrophy
GAP	Global average pooling
GAN	Generative adversarial network
GCL	Ganglion cell layer

GMM	Gaussian mixture model
GMP	Global max pooling
GPU	Graphical processing unit
HOG	Histogram of oriented gradients
HR	High resolution
IFCNN	Iterative fusion convolutional neural network
IPL	Inner plexiform layer
ILM	Inner limiting membrane
INL	Inner nuclear layer
IS	Image sharpness
LACNN	Lesion-Aware convolutional neural network
LBP	Local binary pattern
LCI	Low coherence interferometry
LGCNN	Layer guided convolutional neural network
LM-DFF	Learnable multi-scale deep feature fusion
LR	Low resolution
LRSOTTV	Low rank second order tensor-based total variation
LSTM	Long short term memory
MDFF	Multi-scale deep feature fusion
MH	Macular hole
ML	Machine learning
MSE	Mean squared error
MSSP	Multi-scale spatial pyramid decomposition
NEH	Noor eye hospital
NWSR	Non-local weighted sparse representation
OCT	Optical coherence tomography
OS	Overall sensitivity
OP	Overall precision
OA	Overall accuracy
OF1	Overall F1 score
OPL	Outer plexiform layer

ONL	Outer nuclear layer
PRL	Photo-receptor cell
PR	Precision
PSNR	Peak signal and noise ratio
RPE	Retinal pigment epithelium
RBF	Radial basis function
RNN	Recurrent neural network
ReLU	Rectified linear unit
RNFL	Retinal nerve fiber layer
SD-OCT	Spectral domain optical coherence tomography
SR	Super-resolution
SIFT	Scale invariant feature transform
SURF	Speeded up robust features
SVM	Support vector machine
SE	Squeeze and excitation
SERI	Singapore Eye Research Institute
SBSDI	Sparsity based simultaneous denoising and interpolation
SSR	Segmentation based sparse representation
SOTTV	Second order tensor-based total variation
SRGAN	Super-resolution generative adversarial network
SE	Sensitivity
SS-CNN	Scale-specific convolutional neural network
SC-CNN	Single-scale convolutional neural network
TD-OCT	Time domain optical coherence tomography
t-SNE	t-distributed stochastic neighbour embedding
USA	United States of America
UCSD	University of California San Diego
VEGF	Vascular endothelial growth factor

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

1

Introduction

Contents

Clinical examination of the retina can play a crucial role in the diagnosis and the management of prevalent ocular diseases like age-related macular degeneration (AMD) and diabetic macular edema (DME) [1, 2]. Until recently, the color fundus photography and the fluorescein angiography (FA) [3] were used as gold standards for the *in vivo* imaging of the eye. Both the imaging modalities provide only a two-dimensional (2D) en face view of the retina, making it difficult for clinicians to accurately assess disease severity and progression [4]. With the development of optical coherence tomography (OCT), ophthalmologists are now able to capture three-dimensional (3D) *in vivo* images of the retina at a micron-scale resolution [5, 6]. The OCT provides the visualization of the layered retinal morphology as cross-sectional images called B-scans [7]. These scans enable qualitative and quantitative assessment of the structural changes in the retina [8]. Multiple B-scans (3D volume) can also be acquired to monitor disease progression, diagnose severity stages and understand the effect of therapy [9, 10].

In clinical practice, the ophthalmologists manually analyze the OCT images and volumes to provide a diagnosis decision. However, the inherent speckle noise in the OCT images degrades the image quality and affects its diagnostic utility [11, 12]. Often the low resolution (LR) images, captured to eliminate the effects of involuntary patient movements, pose challenges in the diagnosis of early and intermediate stages of the diseases [13, 14]. Moreover, the retinal disease characteristics widely vary in terms of shapes, sizes, texture and spatial locations [15]. The heterogeneity in the pathological manifestations often results in poor reproducibility among OCT assessors. Consequently, automated and reproducible analysis of the OCT images can effectively represent the disease features, improve diagnostic accuracy and reduce intra- and inter-observer variabilities.

Furthermore, the commercialization of OCT has led to the generation of large amounts of data, which is often difficult to be manually analyzed by experts. The advancement in OCT technology and the high prevalence of retinal diseases [16] in the present population have intensified the research needs for the development of automated diagnosis systems. These systems have the potential to save numerous clinical hours of medical practitioners, reduce false-negative cases due to fatigue and improve diagnostic accuracy.

This dissertation focuses on developing efficient and reliable classification methods for the automated diagnosis of retinal diseases using OCT images and volumes. The chapter begins with a brief introduction to OCT imaging of the retina, followed by a review of the existing automated methods for OCT classification. We then highlight the motivation of the research work. Following that, we present the contributions and outline of the thesis.

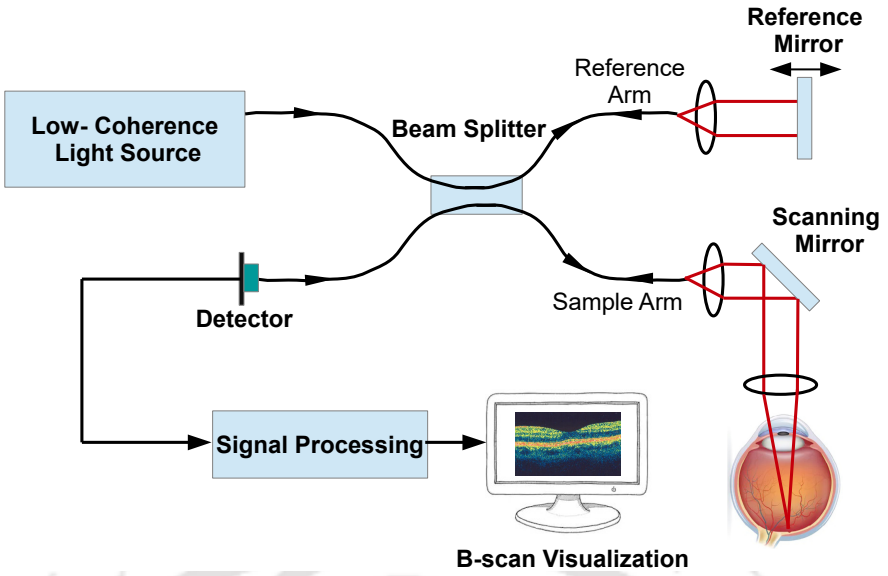


Figure 1.1: Schematic set up of an OCT system.

1.1 Optical Coherence Tomography Imaging of the Retina

OCT is a non-invasive imaging technique that allows cross-sectional and volumetric imaging of the retinal microstructures by measuring the echoes of the backscattered light from the tissues [17]. Developed in 1991 by David Huang and colleagues at the Massachusetts Institute of Technology [18], the OCT is based on the classical optical measurement technique known as the low-coherence interferometry (LCI) [19].

Figure 1.1 shows the schematic setup of an LCI based OCT system. As can be seen, the incident optical power of the input low-coherent light source is evenly split into the sample and the reference arms by the beam splitter. The light traveling through the reference arm is incident on the reference mirror and is redirected back through the same path. The light exiting from the sample arm is incident upon the scanning mirror that focuses the beam onto a spot in the target retinal region. The backscattered light from the desired sample location interferes with the reflected light from the reference arm. A large interferometric signal is generated when the two signals interfere constructively (i.e., the delay mismatch between the signals is nearly zero). The interfering signals add up to nearly a flat line if destructive interference happens between them (i.e., the interfering waveforms vary widely in phase) [20]. The detector detects the interference waves and consequently generates an electrical signal. The reference mirror is displaced laterally to obtain the interferometric peaks at different depths. The electrical signals are then processed to obtain an A-scan (see Figure 1.2). When the OCT probe beam is scanned laterally, multiple A-scans are obtained to form cross-sectional images called B-scans (see Figure 1.2). The B-scans highlight the

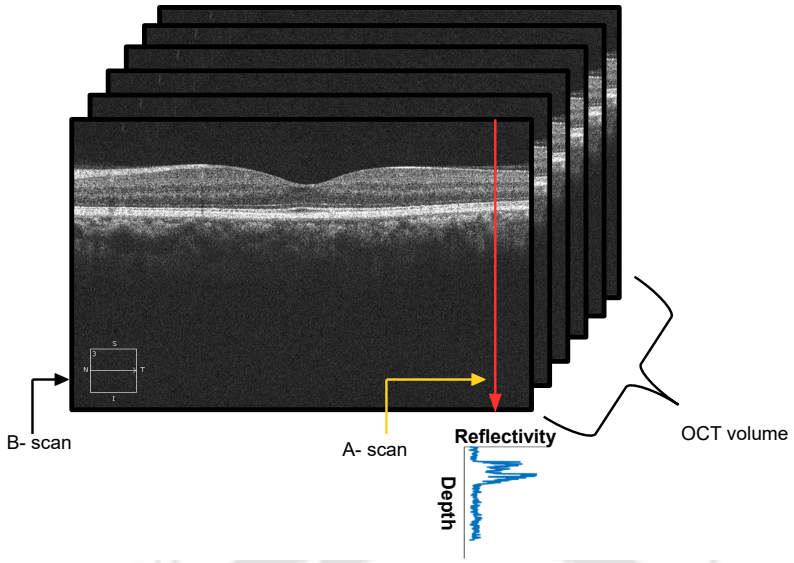


Figure 1.2: A-scan, B-scan and 3D-volume representation in OCT.

thickness and reflectivity profiles of the ten different layers of the retina. The details of the retinal physiology and the interpretation of the layers in the OCT B-scans are discussed in Appendix A. A collection of B-scan images is acquired by lateral scanning at different spatial locations of the retina to form the 3D OCT volume (see Figure 1.2).

The OCT imaging system discussed above is popularly known as the time-domain OCT (TD-OCT). Such systems are relatively slow in acquiring the images as the reference arm has to be mechanically translated to capture tissue details at different depths [21]. Aimed at increasing the imaging speed, advanced spectral-domain OCT (SD-OCT) systems have been developed [22]. The SD-OCT has a stationary reference arm and the depth information is obtained by the Fourier transform of the interference fringes by using a spectrometer at the detector [22]. The image analysis and models developed in this thesis are based on the images and volumes acquired from the SD-OCT imaging systems. In the following paragraphs, we present few challenges that affect the diagnostic accuracy of the computer-aided methods during diagnosis from the retinal SD-OCT images.

Variability in the pathological manifestations of the retinal diseases: The retinal disease manifestations widely vary in terms of shape, size, texture and spatial location. Figure 1.3 shows few such variations of the pathological SD-OCT B-scans along with the normal B-scans for reference. Figure 1.3 (a) presents the normal images and Figure 1.3 (b) and (c) highlight the retinal lesions of AMD and DME, respectively. AMD is a degenerative retinal disease that affects the elderly and causes progressive loss of the central vision [23]. Clinical manifestations such as drusens, geographic atrophy (GA) and intra-retinal fluid deposits are observed in the OCT B-scans of the AMD subjects [24, 25]. The drusens, shown by

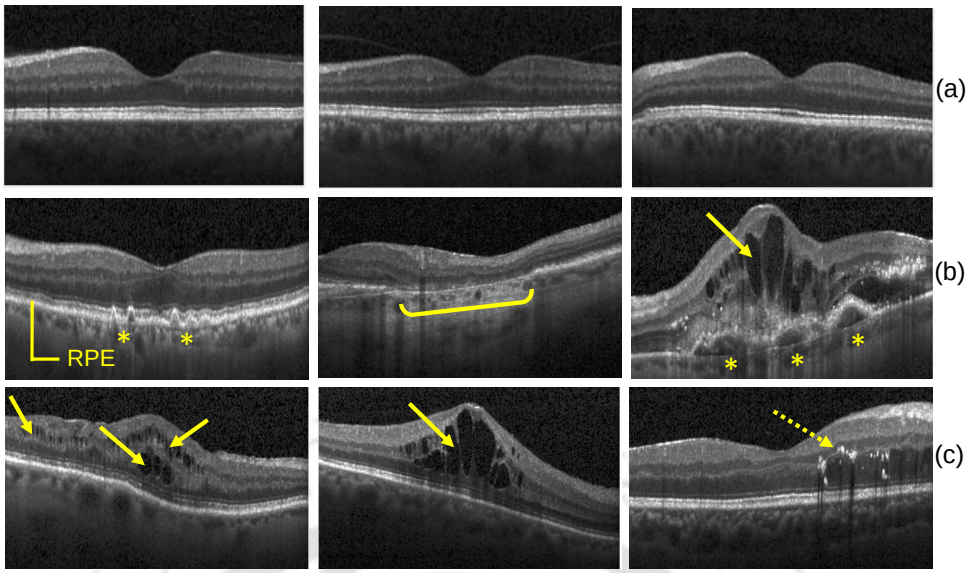


Figure 1.3: OCT B-scans highlighting the wide variations in the retinal lesions. (a) shows the normal images, (b) shows the pathological manifestations in AMD (drusens-asterisk, geographic atrophy-brackets and fluid deposit-solid arrow) and (c) presents the disease characteristics in DME (fluid deposits-solid arrow and exudates-dotted arrow).

asterisk in Figure 1.3 (b), are early manifestations of AMD. They are characterized by the asymptomatic deposition of extra-cellular fluids beneath the retinal pigment epithelium (RPE) [26]. In the intermediate stages, GA (shown by the bracket in Figure 1.3 (b)) is observed, which indicates abnormal thinning of the RPE and loss of the photoreceptor cells [24]. The intra-retinal fluid deposits, shown by the solid arrow in Figure 1.3 (b), result from the exudative damage caused by the blood vessels breaking through the RPE [27].

The DME occurs as a complication of diabetes and results in the accumulation of intra-retinal fluids and exudates (see Figure 1.3 (c)) [28, 29]. It can be observed from Figure 1.3 that the retinal lesions are very diverse and widely vary both within and across the diseases. The effective handling of these variabilities is often challenging for developing reliable OCT-based automated diagnosis systems.

Effect of speckle noise: The speckle noise in OCT imaging arises due to the interference of the photons that undergo multiple scattering in forward and reverse directions within the retinal tissues [11,30]. The speckle noise appears as granular pattern in the OCT B-scans as can be seen in Figure 1.2 and 1.3. The presence of speckle noise often obscures subtle but important morphological details and thus is detrimental to clinical diagnosis. It also affects the performance of automatic analysis methods intended for objective and accurate quantifications [31].

Effect of Motion Artefacts: Motion artefacts appear in the OCT B-scans due to the involuntary eye motion for fixation, head movements and body jitters caused by the cardiorespiratory system [13]. These

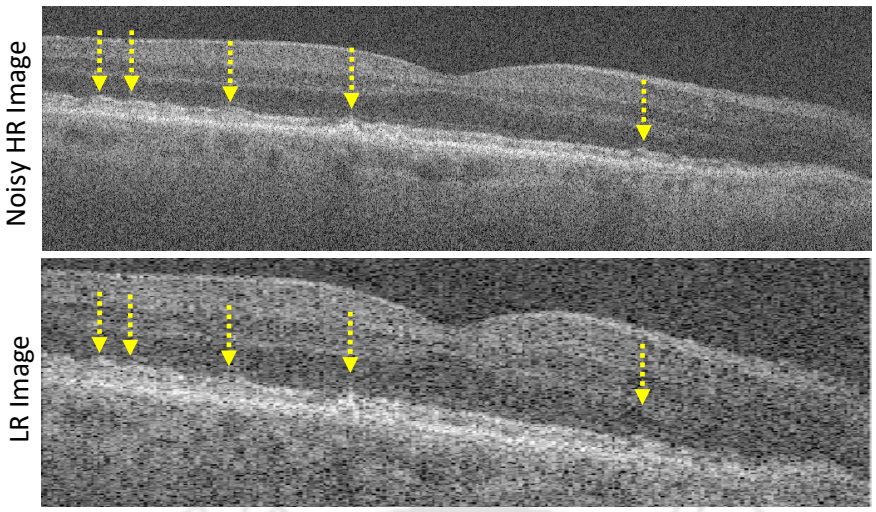


Figure 1.4: Effect of image resolution on the drusen visualization in intermediate AMD.

artefacts can cause inaccurate clinical interpretation of the OCT images [32]. Severe loss of OCT data also happens due to the blinking of the eye. Hardware and software motion correction methods have been proposed in the literature to remove the artefacts [13]. The hardware solutions demand the incorporation of additional hardware to detect the motions. The software-based techniques, on the other hand, take the OCT images with artefacts and estimate the actual OCT images by solving the non-linear and ill-posed inverse problem [13]. However, both these approaches have pertinent drawbacks, i.e., the hardware solutions generate additional overhead costs to the OCT machines and the software solutions are difficult to quantify due to the lack of ground truth. Therefore, in clinical settings, a low spatial sampling rate, i.e., fewer A-scans/B-scan, is preferred to fasten the image acquisition process, thereby reducing the involuntary movements [14, 33]. However, the LR images obtained by the reduced spatial sampling present bottlenecks in the quantification of subtle clinical parameters, especially in the early stages of the diseases, when the retinal lesions are not very prominent. Figure 1.4 shows a high-resolution (HR) OCT B-scan with drusen manifestations (shown by the yellow arrows) acquired at an azimuthal resolution of 1000 A-scans/B-scan and a synthetically generated LR image with a resolution of 500 A-scans/B-scan. It can be observed that the visualization of the drusens has been compromised in the LR image. Using such LR images for diagnosis may lead to high false negatives especially in early and intermediate stages of the diseases, when the lesions are often small in size and sparsely distributed. Therefore, simultaneous denoising and super-resolution (SR) techniques have the potential to improve the quality of the OCT images and improve classification performance.

The following section presents different approaches used in literature for the automated diagnosis of

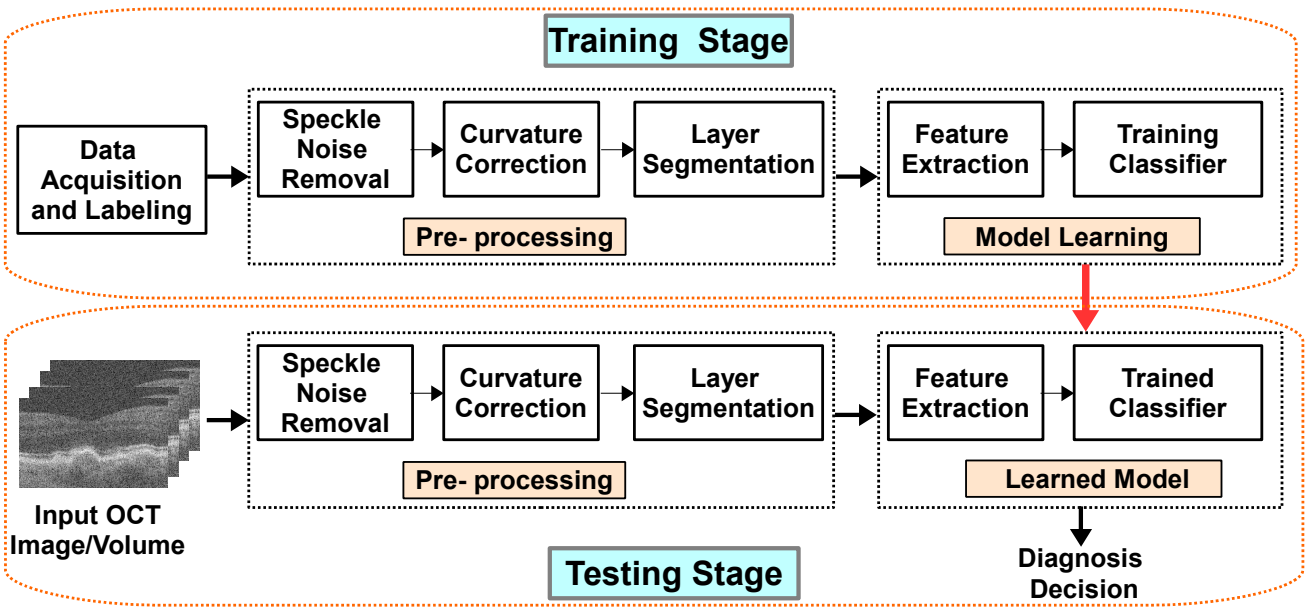


Figure 1.5: General block diagram of an OCT based automated diagnosis system.

retinal diseases from OCT images and volumes. The simultaneous denoising and SR schemes proposed in the literature to improve the diagnostic quality of the OCT images are also discussed.

1.2 OCT based Automated Diagnosis- A Review

Over the last decade, there has been considerable research in the field of automated disease classification from OCT images and volumes. A general block diagram of the OCT based automated diagnosis system is shown in Figure 1.5. As can be seen, there are two stages of operation, i.e., the training and the testing phases. At the training stage, the OCT data are acquired, pre-processed and the classification model is learned. At the testing stages, the OCT images/volumes from unseen patients are pre-processed and provided to the trained model to output the diagnosis decision.

The training phase starts with the acquisition of clinical OCT data from retrospective cohorts of patients. The acquired database then goes through a complex grading system consisting of many trained graders to verify and correct the image labels [34]. The graders in the initial stage conduct quality checks and eliminate the OCT images that have severe artefacts. Then the experienced graders independently label the images based on the observed changes in the anatomical features of the retina. Finally, expert graders verify the correctness of the labels.

Pre-processing of the retinal OCT images is performed to normalize the acquisition distortions and the effects of the natural retinal curvature. Pre-processing steps include speckle noise removal, curvature

correction, layer segmentation and region of interest (ROI) selection [35–37]. The details of each of the steps are given as follows.

- **Speckle noise removal:** The speckle noise inherently exists in the OCT images and affects its visual interpretation [12]. Therefore, OCT image denoising is performed as a pre-processing step to eliminate the effect of noise. The most common despeckling approach adopted in commercial scanners is Bscan averaging [38]. A high-quality image can be obtained by averaging registered multiple Bscans acquired from the same position. However, this approach is currently impractical for 3D scans due to the long acquisition time for overlapping B-scans. Denoising techniques like the non-local means (NLM) [39], sparsity-based block-matching and 3D-filtering (BM3D) [40] and median filtering are generally used. Recently, sophisticated and accurate data-driven denoising methods like the auto-encoder [41] and the SiameseGAN [42] have been proposed.

- **Curvature correction:** The retinal OCT images have a natural curvature that varies across subjects. This curvature is generally removed using curvature correction methods (also known as flattening) [43]. Flattening normalizes the natural retinal curvature and prevents the sensitivity of the classifier to the curvature during model learning [36]. The existing curvature correction methods estimate the RPE layer along each column of the image using pixel intensity and shift the pixels up or down to flatten the RPE [43, 44]. These methods work remarkably well to correct the curvature of the retina.

- **Retinal layer segmentation:** The retinal diseases manifest in specific layers of the retina. Therefore, few automated approaches segment the different layers of the retina to extract layer-specific diagnostic features [45–47]. Chiu *et al.* [43] proposed a graph theory and dynamic programming-based approach to automatically segment the seven different layers of the retina. The method is used in quantifying retinal thickness parameters in intermediate AMD [48] and DME [49]. Few other graph construction based retinal layer segmentation methods can be found in [50–52]. With the advent of deep learning, semantic segmentation based approaches have recently been proposed for retinal layer segmentation. The popularly used methods are ReLayNet [53], DeepRetina [54], branch residual U-net [55], etc.

The retinal B- scans are then cropped to focus on the region of the retina that contains the morphological structures and also to reduce the dimensions of the image. The model learning block takes the pre-processed OCT images/volumes and the corresponding labels as input and learns the model parameters in a supervised manner. Over the past few decades, researchers have focused on developing

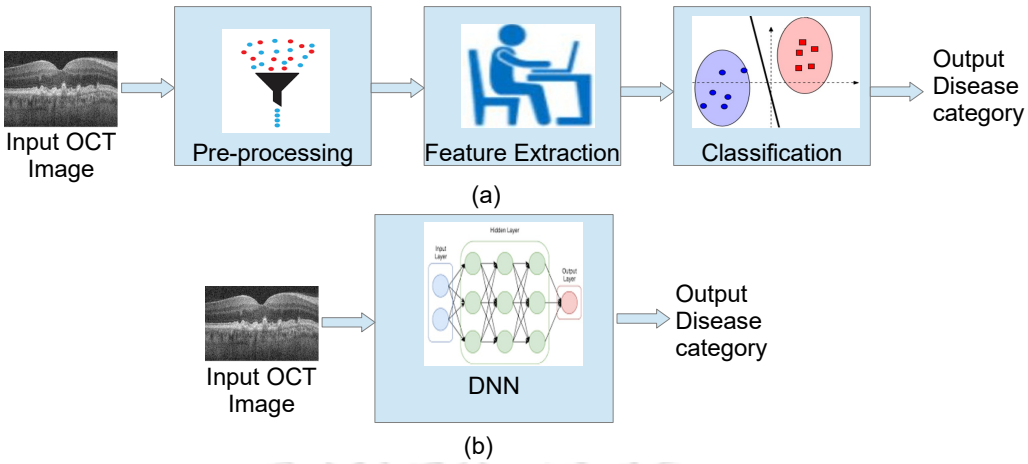


Figure 1.6: Block diagram highlighting the different stages of machine learning and deep learning models.

machine learning (ML) models that can represent the pathological characteristics well and thereby generating improved classification accuracy [45, 46]. Recently, the deep learning (DL) based classification schemes are being increasingly used for the OCT classification tasks [34, 35, 37, 56–58]. The DL based methods automatically learn patterns or representations from the images for efficient classification and thereby completely eliminate the burden of hand designing diagnostic features. Figure 1.6 shows the block diagram highlighting the different steps involved in the ML and the DL based classification models. It can be seen that an ML model majorly involves three steps i.e., pre-processing, feature extraction and classification. The DL models on the other hand do not rely on pre-processing of the images and explicit feature extraction for classification. The models have also shown to generalize well for unseen test data. Based on the type of models used for classification, the current literature can be broadly categorized into two approaches: the feature engineering and the DL based methods.

1.2.1 Feature Engineering based Methods

The works in this category mostly rely on extracting informative features from the pathological biomarkers of the OCT images and volumes, followed by classification using ML classifiers. Sisternes *et al.* [45] computed the area, volume, height and reflectivity of the drusens in the OCT volumes to train a linear regression model for the quantification of the risk of AMD progression. Similarly, the authors in [46] segmented the retinal layers to extract the RPE-drusen complex thickness and the total retinal thickness for detecting AMD patients from healthy subjects. It is worth emphasizing that the design of such hand-crafted features has high dependence on cross-domain experts like ophthalmologists and data scientists. It is also sometimes difficult to find a fixed set of informative, discriminative and independent features for training

an ML model considering the variabilities in the disease manifestations. Therefore, instead of explicitly quantifying disease features, researchers were inclined to explore the use of image feature descriptors like the local binary patterns (LBP) [15, 59, 60], histogram of oriented gradients (HOG) [36], bag-of-words (BoW) [47], scale-invariant feature transform (SIFT) [61] and speeded up robust features (SURF) [62] that can implicitly encode disease-specific information from the pathological OCT scans. These encoded features are then classified using the support vector machine (SVM) [15, 36, 52, 59], Bayesian classifier [60], linear regression [45, 46] and random forest classifier [47, 63] to obtain the diagnosis decision. It is to be noted that the feature descriptors are highly sensitive to the speckle noise in the OCT images. Therefore, sophisticated pre-processing for noise removal is necessary prior to feature extraction.

Sidibé *et al.* [64] approached the classification task as an outlier (anomaly) detection problem. The method modeled the appearance of the normal OCT images with the Gaussian mixture model (GMM) and identified the pathological images as outliers. The diagnosis result for the OCT volume depended on the number of detected outliers.

Of late, the DL based methods have captured the interest of the OCT research community [65–68]. These methods automatically learn features relevant for classification, thus alleviating the need for hand-crafted or fixed descriptors for feature extraction [69]. The DL models have shown better generalization and improved performance than the feature-based methods [70]. The following subsection presents the related literature of the DL based methods for the OCT classification task.

1.2.2 Deep Learning based Methods

Early DL based works exploited the advantages of transfer learning for classification. Transfer learning is a technique where a model learned in one setting is exploited to improve generalization in another setting [71]. Rather than training a neural network from scratch, transfer learning allows using a network with fixed weights adapted from a particular domain at the low levels and retrained at the high levels to learn distinguishing features of the specific task. Such frameworks are useful for learning DL models for domains with limited data. Inspired by the advantages of the transfer learning, Karri *et al.* [37] fine-tuned the GoogLeNet [72] for the classification of AMD, DME and healthy subjects. Similarly, Kermany *et al.* [34] and Li *et al.* [73] fine-tuned the InceptionV3 [74] and the ResNet50 [75] architectures for OCT image classification. Similar explorations of fine-tuning the state-of-the-art DL architectures like the AlexNet [76], the VGG [77] and the ResNet [75] are proposed in [78–85].

As the pre-trained networks are learned for natural images, understanding the behaviour of the filter outputs is inscrutable for medical images. To mitigate this issue, researchers proposed end-to-end convo-

lutional neural networks (CNNs) for the reliable classification of retinal diseases. The CNNs are typically used for recognition tasks from images [86]. The learnable convolution kernels, the multiple layers of representation and the non-linear modules in these networks automatically discover features from raw 2D or 3D data for classification [69]. The initial few stages of the CNNs use multiple convolution layers, non-linear activations and pooling layers to encode features from the input images. The convolution layers comprise multiple learnable convolutional filters to encode local spatial information from the input images [66]. The output feature maps of the convolution layers are provided to the activation functions to add non-linearity to the features. The pooling layers merge the neighbourhood spatial features to reduce the size of the feature maps and improve translation invariance. Traditionally, a series of convolution, activation and pooling layers are used to extract powerful non-linear features, which are transformed to 1D vectors and provided to fully connected (FC) layers for classification. The details of each of the components of the CNN are presented in Appendix B.

Rasti *et al.* [35] proposed a convolutional mixture of experts approach for the classification of AMD, DME and normal OCT images. A new cost function with correlation penalty was employed to enable the experts to learn non-overlapping discriminative features for accurate classification. The method also proposed a new OCT database known as the Noor Eye Hospital (NEH) database. Considering the limited size of the dataset, the authors have used data augmentation methods to obtain generalizable results. An iterative feature fusion CNN (IFCNN) was proposed in [56] for improving the classification performance of the OCT B-scans. The method fused the features of the previous convolutional layer with the current layer and jointly utilized the features for improving the classification performance. The method achieved an overall accuracy of 87.3% for the classification of OCT images on the University of California San Diego (UCSD) database. The authors in [87] not only classified retinal pathologies using an end-to-end CNN model but also presented the interpretation of the designed network by the class activation map (CAM). The CAM identifies the regions of the images contributing most to the network's assignment of the labels at the testing stages [88].

Instead of visualizing the attentive regions of the OCT images during the testing phases, it would be imperative to constrain the network to pay attention to the clinically significant regions during training for improved learning of the disease features. An exploration of this was presented in the lesion-aware convolutional neural network (LACNN) [58]. In this work, the authors utilized the retinal lesions within the OCT images to guide the CNN for learning disease-specific features. The network made use of a CNN based lesion detection framework to obtain a soft attention map for the coarse representation of the retinal lesions. The attention map was then incorporated into the CNN framework to weigh the

Table 1.1: Automated feature-based methods in literature for OCT classification.

Method	# classes	Preprocessing			Approach		Database	Evaluation		Result*
		Denoising	Layer Seg. †	Flattening	Features	Classifier		Scan level	Volume level	
Sisternes *et al.* [45]	2(low risk, high risk of AMD progression)	✓	✓		area, volume, height and reflectivity of drusens	Linear regression	private		✓	Se=80.95%
Farsiu *et al.* [46]	2 (AMD, normal)		✓		Thickness of RPE and RPE-drusen complex	Linear regression	DUIA [46]		✓	AUC=0.99
Liu *et al.* [15]	4 (AMD, DME, MH, normal)			✓	Multi-scale LBP	SVM-RBF	private	✓		AUC=0.93 (10 fold CV)
Srinivasan *et al.* [36]	3 (AMD, DME, normal)	✓		✓	HOG	SVM- Linear	Duke [36]	✓	✓(majority voting)	OA=95.5%
Albarrak *et al.* [60]	2 (AMD, normal)	✓		✓	LBP,HOG + PCA	Bayesian classifier	private		✓	AUC=0.94
Venhuisen *et al.* [63]	2 (AMD, normal)				Unsupervised feature extraction using BoWs	Random forest	DUIA [46]		✓	AUC=0.98
Lemaitre *et al.* [59]	2 (DME, normal)	✓		✓	LBP-TOP+BoW	SVM-RBF	Private		✓	Se=81.2%
Venhuisen *et al.* [47]	5 (No AMD, Early AMD, Intermediate AMD, GA, CNV)		✓		Detection of AMD affected regions, BoWs	Random forest	Private (Eugenda)			AUC=0.98
Sun *et.al* [61]	3 (AMD, DME, normal)			✓	SIFT, sparse coding	SVM-Linear	Duke [36]	✓	✓(majority voting)	OA=95.5%
Dash *et al.* [62]	2 (DME, normal)	✓		✓	SURF	SVM- Polynomial	Private	✓		Acc=99%
Sidibé *et.al* [64]	2 (DME, normal)	✓		✓	vectorized B-scans	GMM	SERI [90], Duke [36]	✓	✓(majority voting)	Se=93.75% (SERI), 80% (Duke)

†:Layer Seg.=Layer Segmentation, *: results taken from the corresponding paper.

convolutional features, thereby guiding the CNN to pay attention to the lesion related regions of the image. In another work, Huang *et al.* [89] suggested the use of a layer guided CNN (LGCNN) to meaningfully represent the lesions specific to the retinal layers for the efficient classification of the B-scans. Particularly, the fully convolutional ReLayNet [53] architecture was used to delineate the lesion-related layers. The probability maps of these layers then softly weighted the CNN feature maps during the classification process. These attention based classification frameworks have shown better performance over the plain CNN based methods. The LACNN and LGCNN methods have achieved improved overall accuracies of 90.1% and 88.4% on the UCSD database.

The attention-based networks discussed above utilize additional frameworks, such as lesion detection or layer segmentation, to obtain the soft attention maps. The performance of the attention networks depends on how efficiently the additional frameworks generate the attention maps. To mitigate this issue, Mishra *et al.* [57] presented a multi-level dual attention network that employed a self-attention mechanism to automatically attend to the higher entropy regions of the OCT images. Similarly, a convolutional attention-to-DME network (CADNet) was presented by Rasti *et al.* [96] to predict the efficacy of the

Table 1.2: Automated DL based methods in literature for OCT classification.

Method	# classes	Preprocessing				Approach	Database	Evaluation		Result*
		Denoising	Layer Seg. †	Flattening	Data Aug. ‡			Scan level	Volume level	
Karri *et.al* [37]	3 (AMD, DME, Normal)	✓		✓		Fine-tuning GoogLeNet	Duke [36]	✓	✓(Majority voting)	OA=100%
Kermany *et.al* [34]	4 (Drusen, CNV, DME, Normal)					Fine-tune InceptionV3	UCSD [34]	✓		AUC=0.99
Rasti *et.al* [35]	3 (AMD, DME, Normal)			✓	✓	End-to-end CNN ensemble	NEH [35], Duke [36]	✓	✓(Threshold based)	AUC=0.99
Fang *et.al* [56]	4 (Drusen, CNV, DME, Normal)					Iterative feature fusion CNN	UCSD [34]	✓		OA=87.3%
Perdomo *et.al* [91]	2 (DME,Normal)		✓			Deep end-to-end CNN	SERI [90]	✓		Acc=93.75%
Motozawa *et.al* [87]	3 (dry AMD, wet AMD, normal)					Deep end-to-end CNN	Private	✓		OA=93.9%
Fang *et.al* [58]	4 (Drusen, CNV, DME, Normal)					Attention (lesions) +CNN	UCSD [34]	✓		OA=90.1%
Huang *et.al* [89]	4 (Drusen, CNV, DME, Normal)		✓			Layer guided CNN	UCSD [34]	✓		OA= 88.4%
Mishra *et.al* [57]	3 (AMD, DME, normal)					Dual attention CNN	NEH [35] , Duke [36]	✓		OA=99.6% (NEH), 95.5% (Duke)
He *et.al* [92]	4 (Drusen, CNV, DME, Normal)					Label smoothing GAN	UCSD	✓		F1=87.11%
Apostolopoulos *et.al* [93]	2 (AMD, normal)			✓	✓	End-to-end CNN	DUIA [46]		✓	AUC=0.99
Wang *et al.* [94]	3 (AMD, DME, normal)					ResNet (feature extraction) and LSTM (feature aggregation), FC-Softmax (classification)	Private, Duke [36]		✓	AUC=0.95 (private), 0.97 (Duke)
Romo-Bucheli *et.al* [95]	prediction of treatment requirement			✓		DenseNet (feature extraction), RNN (feature aggregation), FC-Softmax (classification)	Private		✓	Recall=80%

†:Layer Seg.=Layer Segmentation, ‡: Data Aug.=Data Augmentation, *: results taken from the corresponding paper.

anti-vascular endothelial growth factor (VEGF) treatment for DME patients. The CADNet used the self-attention based squeeze-and-excitation (SE) unit [97], that attended to the informative CNN feature maps for efficient classification.

It is well known that the DL based methods tend to overfit when trained with limited data [98]. To address the overfitting issue under limited training data environments, He *et al.* [92] explored the use of the generative adversarial network (GAN) to augment the limited dataset with synthetically generated OCT images. The mixture of synthetic and real images was used as training data to improve the classification of drusens, choroidal neovascularization (CNV), DME and normal B-scans.

Most of the existing methods have performed the classification of the OCT B-scans only [34, 56–58, 89, 92]. However, a few of these approaches have extended their B-scan based methods to the volume level by aggregating the scan-level scores using majority voting or threshold-based inference strategies [35–37, 64]. To improve the robustness of such classifiers, Qiu *et al.* [99] proposed a self-supervised learning scheme to iteratively refine the model and the label set to improve the classification of the individual B-scans in the OCT volume. On the contrary, a few works have employed only the volume level labels for the classification. The authors in [93] presented a semi-supervised model called

RetiNet that mosaiced the B-scans in the volume to form a 2D-matrix and applied a CNN classifier for diagnosis. With a similar idea, Wang *et al.* [94] proposed a weakly deep supervised learning framework for the classification of macular diseases like DME and AMD from OCT volumes. The method extracted spatial features from each of the B-scan images using a pre-trained ResNet architecture followed by feature aggregation using long short-term memory (LSTM) units. The classification scores of the macular diseases were obtained by an FC layer with softmax activation in the end. Recently, Romo-Bucheli *et al.* [95] proposed an end-to-end DL architecture for prediction of the requirement of anti-VEGF treatment for neovascular AMD patients. In this method, OCT volumes were obtained at different time points of the treatment. B-scan images were sampled from the OCT volumes to form multi-tile images. Spatial features from these multi-tile images were extracted using a shared densely connected neural network. A recurrent neural network (RNN) followed by a tanh activation integrated the clinical information of the OCT images across multiple time-points. An FC layer finally integrated the spatio-temporal information from the RNN to generate a category prediction.

The feature-engineering and DL based explorations are summarized in Tables 1.1 and 1.2, respectively for a ready reference. The tables list the methods based on the following attributes: the retinal diseases addressed, the pre-processing techniques used (denoising, layer segmentation, flattening and data augmentation), the classification approach, the database used, the evaluation scheme (B-scan level or volume level) and the results obtained. It can be observed from Table 1.1 that the feature-based methods heavily rely on pre-processing to obtain the desired performance. On the other hand, the DL based methods have minimal dependence on pre-processing (see Table 1.2, column 3). Also, the DL based methods have mostly been explored for B-scan classification (see Table 1.2, column 6).

In the following section, we present a detailed review of the existing approaches employed for the simultaneous denoising and SR of the OCT images.

1.3 Simultaneous Denoising and Super-resolution of OCT Images- A Review

The conventional single image SR frameworks demand aligned pairs of LR and high-resolution (HR) images during training [100–104]. The creation of such datasets for OCT images faces inevitable challenges. The OCT images are inherently affected by speckle noises. Therefore, even the acquired HR images are noisy. To obtain the clean HR counterparts for the noisy LR images, repeated HR images are acquired from the same location and averaged [14]. Such a cumbersome process requires patient's con-

sent, preparedness and medical supervision. Therefore, research in this direction was quite limited until recently. In 2013, Fang *et al.* [14] in collaboration with the Duke University Medical Center, proposed an OCT dataset with noisy LR images and their clean HR counterparts. The database provided an impetus to the progress of research in this field.

The authors in [14] proposed a sparsity-based simultaneous denoising and interpolation (SBSDI) method that utilized the sparse representation based learned dictionaries to reconstruct the clean HR OCT images. During the training stages, many image patches were simultaneously extracted from the LR-HR image pairs to learn the LR and the HR dictionaries jointly. The method proposed a coupled orthogonal matching pursuit algorithm during training to obtain sparse codes that can preserve the spatial correlation between the LR and the HR patches for reliable reconstruction. The authors extended their work by proposing a segmentation based sparse reconstruction (SSR) [33] method that learns layer-specific structural dictionaries to better represent the anatomical and pathological features within the retinal layers. The method provided speedy reconstruction compared to the SBSDI as the sparse codes for the LR patches now had to be searched in the corresponding layer-specific dictionary rather than from an extensive global dictionary.

Abbasi *et al.* [105] proposed a non-local weighted sparse representation (NWSR) method to denoise and interpolate the B-scan images. The method utilized the self-similarity between image patches to average the sparse codes of non-local similar patches obtained from the corresponding denoised images to estimate the sparse codes for the noisy image patches. The denoised images were obtained from an off-the-shelf denoising method. A dictionary-based HR reconstruction scheme [101] was then applied to the obtained sparse codes to generate the clean HR patches. Two other sparsity-based methods were presented in [106, 107].

Recently, Daneshmand *et al.* [108] proposed a mixed low-rank approximation and second-order tensor based total variation approach for the denoising and SR of the OCT images. The low-rank regularization is used to explore the non-local self-similarity for effective noise suppression. Further, tensor-based total variation regularizers were also incorporated to preserve the retinal layers and suppress the artefacts in the OCT images.

It is worth mentioning that the sparsity regularizers and the sophisticated constraints render the above approaches computationally complex and inflexible. The patch processing framework exploited by these methods also results in the smoothing of the reconstructed images. The smoothing causes the loss of the sharp boundaries between the retinal layers, which leads to inaccurate measurement of the layer thickness and, in turn, hampers diagnosis. To address the above challenges, a DL based OCT SR method called the

SDSR-OCT [109] was recently proposed. The method exploited the super-resolution GAN (SRGAN) [104] for the task. Specifically, the model was divided into two parts, i.e., the generator that took the noisy LR images as input to learn LR features. These features were then utilized to reconstruct the HR images. The discriminator part of the network decided whether the generated clean HR images were similar to the real HR image distribution. The network had an inherent capability to simultaneously denoise and super-resolve the LR OCT images without patch processing and averaging. It showed promising performance compared to the sparsity-based methods.

1.4 Motivation for the Research Work

Although several methods have been proposed for the OCT classification, a few research challenges have not been addressed and are given as follows.

- As discussed earlier, the pathological manifestations of retinal diseases have diverse shapes, sizes and textures. To automatically detect the abnormalities from the OCT images, an algorithm must implicitly recognize the distinct multi-scale pathological symptoms and discern the complex relationships between them. At present, the DL based methods employ single-scale CNNs for feature extraction and classification of the OCT B-scans [34,37,57,58,89]. However, it is challenging to represent the variabilities in the pathological morphologies of different diseases efficiently by ignoring potentially useful information on different scales.

- The drusens are the pathological manifestations of the early and intermediate stages of AMD. Often, the drusens are small in size and sparsely distributed. The automated classifiers may not characterize well the subtle clinical details from the noisy LR OCT images. This may result in high false negatives. We hypothesize that incorporating a simultaneous denoising and SR stage prior to classification can improve the diagnosis results. The denoising and SR of the OCT images can enhance the drusen-related diagnostic details in the noisy LR images. This is expected to improve disease detection accuracy. However, considering the limited size of the available denoising and SR databases, the existing methods are not well generalizable. Less generalizable models may not reconstruct the pathological symptoms reliably, considering their diversity in appearance. It is challenging to develop generalizable denoising and SR methods that can effectively reconstruct the clinical details in the OCT images.

- The existing B-scan classification schemes [35, 37] have extended their methods to OCT volume

level classification by aggregating the classification decision from the individual B-scans. Manual threshold-based inference strategies are used to perform the aggregation of decisions. However, it is argued that such approaches have limited usability in real-world applications, as the pre-set threshold can be influenced by various factors like the azimuthal resolution of the OCT scanner (number of B-scans per volume) and the severity stages of the diseases (number of pathological B-scans in the volume). It would be interesting if the algorithm can automatically identify the salient pathological B-scans from the volume and use those scans to make a diagnosis decision. The threshold-based methods also require fine-grained expert annotations of each B-scan of the volume (during training), which is usually quite expensive and challenging [110]. It would be appealing to leverage only the coarse volume-level labels for the classification of OCT volumes.

1.5 Contributions of the Work

Motivated by the issues mentioned above, the thesis aims at developing DL-based classification methods for reliable and efficient handling of these challenges. The salient contributions are as follows.

- Deep multi-scale feature extraction has been explored to capture the diverse pathological manifestations from the OCT B-scans effectively. Two different approaches have been explored for the task. In the first approach, a pyramidal decomposition technique is used to obtain images at different scales and CNN based feature extractors are employed to extract the scale-specific features. These features are fused to further mine the cross-scale discriminative information for efficient classification. In the second approach, various dilated convolutional filters with different receptive fields extract the scale-specific features. These features are fused attentively to obtain discriminative features for classification. The experimental analysis shows that multi-scale feature extraction and fusion can provide high-level discriminative features for improved classification.

- To enhance the visualization of the drusens, a simultaneous denoising and SR framework is adopted. Recently, the GANs have shown great success in the generation of realistic images [111]. Therefore, we have explored the use of GANs in an unsupervised scenario to build a generalizable network with unpaired noisy LR and clean HR images. The option of using unpaired images provides flexibility to leverage the already available OCT data for better generalization of the denoising and SR performance. The experimental results on clinical-grade OCT images confirm that the proposed method outperforms the existing methods both in terms of SR performance and computational time.

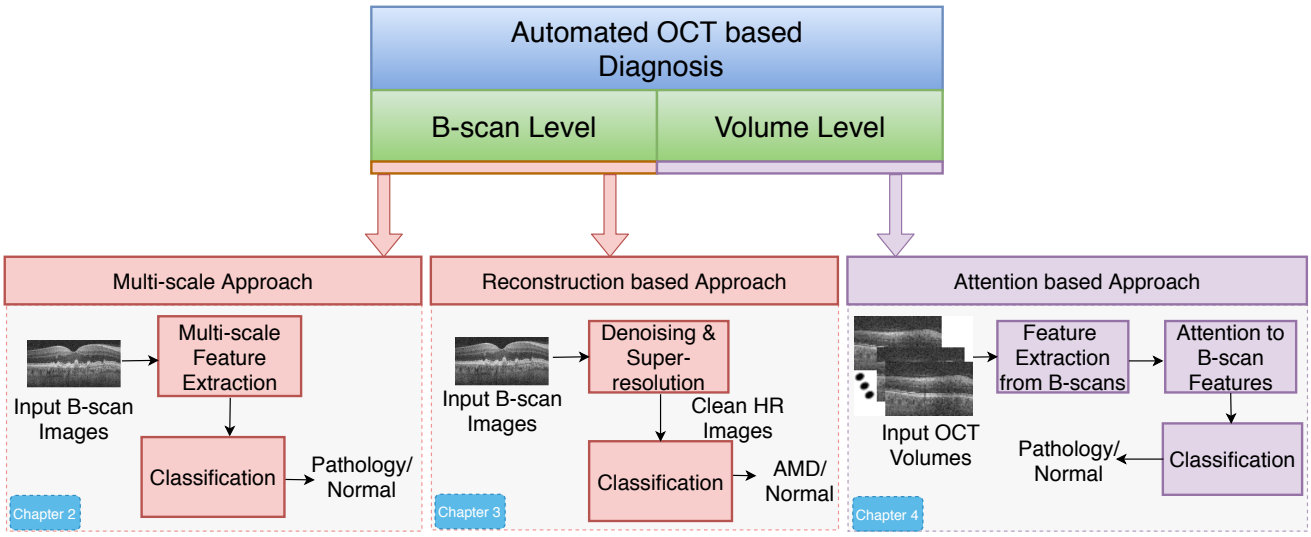


Figure 1.7: Graphical representation of the working chapters of this dissertation.

The denoised and super-resolved images are utilized for the diagnosis of AMD using a baseline classification model. The experimental results verify that performing SR prior to classification can potentially improve the diagnosis performance.

- A B-scan attentive CNN framework is explored for the classification of the OCT volumes. The network automatically assigns weights of importance to the salient B-scans in the volume during feature aggregation to obtain discriminative feature representation for reliable classification. The self-attention mechanism also enables the utilization of only the volume level labels during classification. Additionally, the attention weights demonstrate the correlation of the weights with the B-scans manifesting the disease characteristics. This analysis imparts diagnostic transparency to the proposed model which makes it reliable for clinical applications.

1.6 Organization of the Thesis

Figure 1.7 shows the graphical representation of the proposed methods that are presented in the rest of the chapters. As shown in the figure, **Chapter 2** explores multi-scale deep CNN networks for the classification of OCT B-scan images. It presents the qualitative analysis of the multi-scale behaviour of the retinal pathologies. Following that, the two proposed deep multi-scale feature fusion methods are presented. To combat the imbalance in the OCT databases, a cost-sensitive loss function is explored that ensures the proper learning of the classifiers. The performance of the models is evaluated and compared with the existing methods. In **Chapter 3** the importance of denoising and SR of the OCT images is

discussed for the improved classification of intermediate AMD and healthy subjects. Following that, the proposed unsupervised GAN based simultaneous denoising and SR scheme is presented. The qualitative and quantitative analysis of the reconstruction performance is carried out for the diagnostically significant regions of the B-scans for reliable AMD evaluation. **Chapter 4** investigates the applicability of the attention mechanism for emphasizing the pathological B-scan features for the reliable classification of the OCT volumes. A detailed description of the B-scan attention framework for the OCT volume classification is presented. The importance of the self-attention mechanism for obtaining the weights for the B-scans is discussed. The performance evaluation and comparison with existing methods are also carried out. A summary of the research work and the future directions are discussed in **Chapter 5**.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

2

Fusion of Deep Multi-scale Features for OCT B-scan Classification

Contents

In the previous chapter, it was discussed that the clinical manifestations of retinal diseases are highly heterogeneous. Figure 2.1 (a) and (b) present the retinal disease characteristics at early and advanced stages of AMD, respectively. Figure 2.1 (d) shows a few normal OCT B-scans for reference. As can be seen from Figure 2.1 (a), the early AMD manifestations, i.e., the drusens (shown by the asterisk symbol) have different heights and areas. Similarly, in the advanced stages, the length of the retina affected by GA (shown by the brackets in Figure 2.1 (b)) differs across the B-scans. In most advanced cases, a combination of large drusens, GA and intra-retinal fluid deposits is also observed. Figure 2.1 (c) shows a few variations of the DME affected B-scans. It can be observed that the retinal areas affected by the intra-retinal fluid deposits (shown by solid arrows in Figure 2.1 (c)) and the exudates (dotted arrow in Figure 2.1 (c)) are very diverse. It is difficult to effectively extract discriminative features from the highly variable and multi-scaled disease characteristics. Most of the recent methods use single-scale approaches, where local spatial features are extracted using CNN based feature extractors followed by classification using FC and softmax layers [34, 37, 57, 58, 89]. Very deep neural networks with hierarchical arrangement of convolutional layers have shown to capture the multi-scale activities in natural images [77]. However, considering the wide variabilities in the retinal disease characteristics, the low contrast and noise conditions of the retinal images, such approaches may not be adequate for effectively representing the multi-scale features. Also, very deep networks require large amounts of labeled training data for learning a well generalizable model. Acquisition and labeling of large-scale medical datasets have always been challenging considering the differences in imaging protocols, privacy issues and lack of medical integration [112]. Therefore, we presume that customized multi-scale feature learning using multiple shallow CNNs capturing scale-specific features and a late fusion of these features can provide improved and generalizable classification results.

This chapter proposes two multi-scale approaches: the multi-scale deep feature fusion (MDFF) and the learnable multi-scale deep feature fusion (LM-DFF) for powerful feature extraction and efficient classification. The MDFF method employs the multi-scale spatial pyramid (MSSP) decomposition to obtain views of the B-scans at different scales. These multi-scale images are then fed to different CNNs to extract scale-specific features. These features are fused to obtain high-level discriminative representation for classification. Aimed at improving the classification performance, the LM-DFF method is proposed. The method uses multiple convolution filters with different receptive fields to obtain multi-scale features. The learnable convolution kernels with varied receptive fields capture multi-scale information for efficient feature encoding and classification.

The rest of the chapter is organized as follows. Section 2.1 and 2.2 present the proposed framework

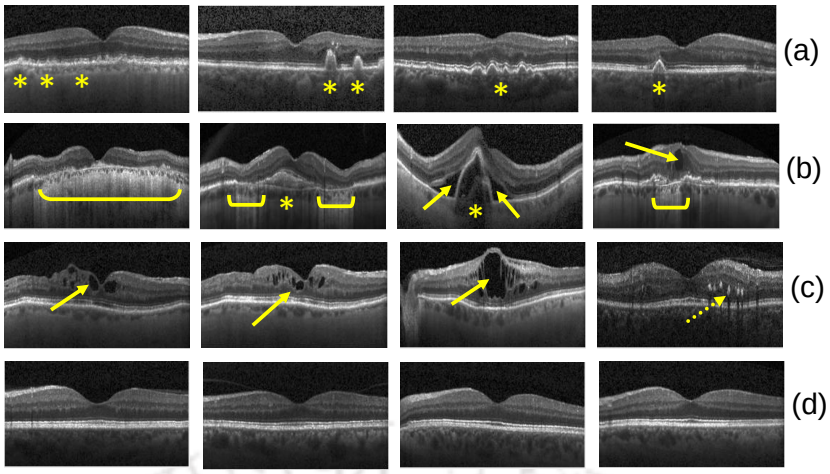


Figure 2.1: OCT images highlighting the wide variations in the retinal lesions. (a) shows the early AMD manifestations (drusens-asterisk) of different sizes, (b) presents the diverse pathological manifestations of advanced AMD (GA-brackets, large drusens-asterisk and intra-retinal fluids-solid arrows), (c) highlights the different lesions of DME (fluid deposits-solid arrow and exudates-dotted arrow) and (d) Normal images.

and the experimental results for the MDFF method, respectively. The LM-DFF method and its experimental evaluation are discussed in sections 2.3 and 2.4, respectively. The summary of the chapter is presented in section 2.5.

2.1 Multi-scale Deep Feature Fusion for OCT B-scan Classification

This section presents the MDFF approach for the classification of retinal diseases from OCT B-scans. The block diagram of the method is shown in Figure 2.2. The method is implemented in three steps. First, in the pre-processing step, the natural curvature of the retina is removed [44, 113] followed by cropping of the regions of interest. In the second step, the pre-processed image is decomposed into various scales to obtain multi-scale views of the image. The deep CNN features extracted from the multi-scale images are fused to classify the input OCT image into one of the disease categories in the final step. The details of the different stages in the block diagram are discussed as follows.

2.1.1 Pre-processing

The retinal layers in the OCT images have a natural curvature that varies among patients. To prevent the sensitivity of the classifier to the curvature and obtain a consistent shape of the layers, the curvature is flattened out [113]. In this work, the flattening method proposed by Chiu *et al.* [43] is employed. The flattening process starts with the estimation of the RPE. It is estimated by finding out the location of the

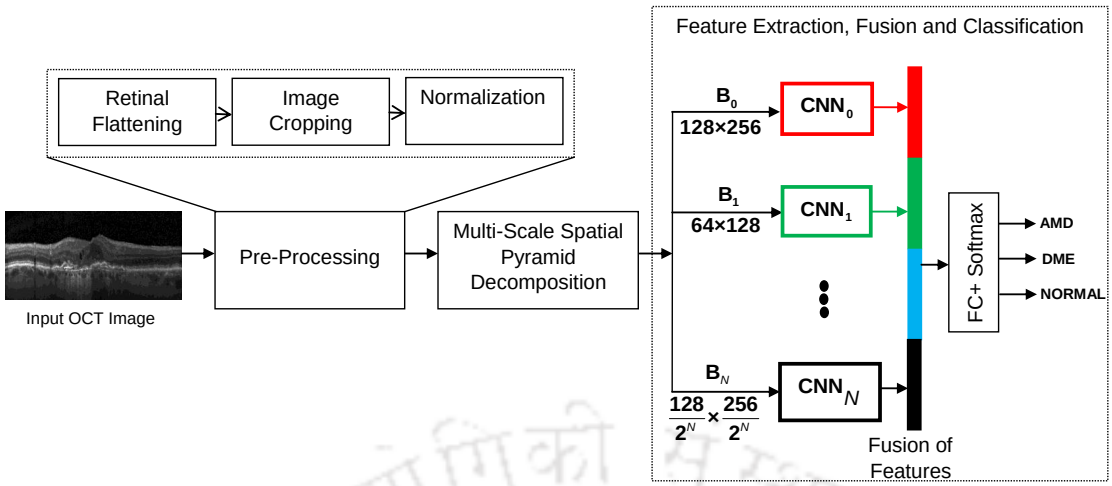


Figure 2.2: Block diagram of the proposed MDFF method.

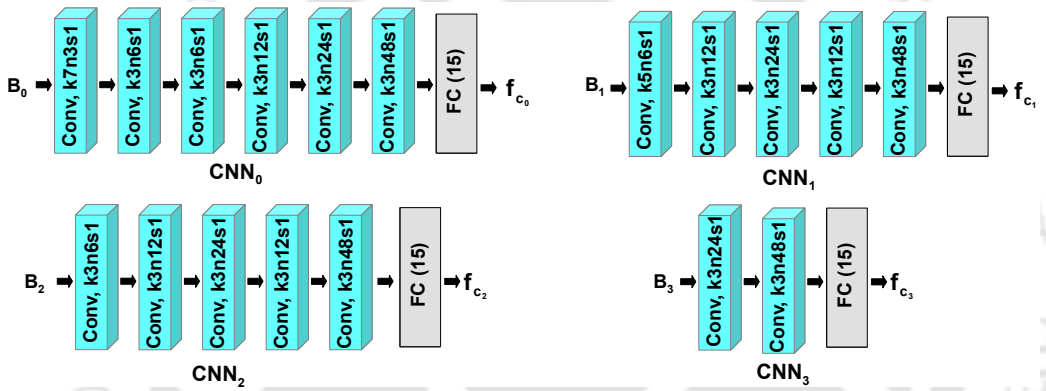


Figure 2.3: Detailed architecture of the CNNs in the MDFF classifier.

brightest pixel (as the RPE is the most hyper-reflective layer of the retina) in each column of the B-scan. A second-order polynomial is fitted on the obtained locations and each column of the image is shifted up or down to obtain a flat RPE.

The diagnostic information in the OCT images is confined to the retinal layers. Therefore, each B-scan is cropped horizontally to remove portions of the vitreous and sclera regions. The OCT B-scans are then resized to 128×256 pixels for further processing. The B-scans are then normalized to have zero mean and unit standard deviation. Figure 2.4 shows the original images (top row) and the corresponding pre-processed images (bottom row). It can be seen that the flattening process has removed the curvature of the retina. Some portions of the vitreous and the sclera have also been cropped, as they may not contribute much to the classification process. The pre-processed OCT images are then fed to the MSSP block to obtain the multi-scaled representations.

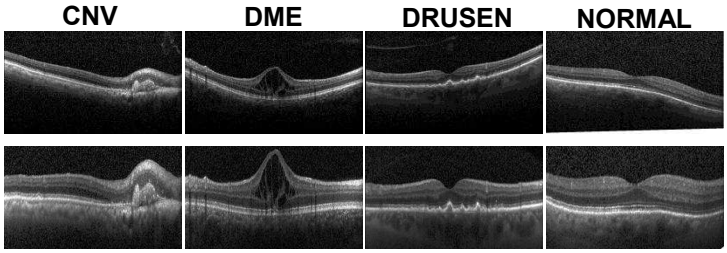


Figure 2.4: Original Images (top row) and pre-processed images (bottom row).

2.1.2 Multi-Scale Spatial Pyramid Decomposition

As shown in Figure 2.1, the pathological manifestations of DME and AMD have multi-scaled characteristics. The drusens at the early stages of AMD are very small in size and need to be analyzed at a finer scale. The intra-retinal fluid accumulation in DME can be visualized on a coarse scale as their homogeneous texture within the retinal layers stands prominent. In this work, the MSSP decomposition is used to create multi-scale views of the input image to capture key multi-scale information for better classification. It creates image pyramids by the sub-sampling and Gaussian low pass filtering the images at the previous levels [114]. An input OCT B-scan $\boldsymbol{B} \in \mathbb{R}^{d_m \times d_n}$ is considered to be at level $i = 0$. Here d_m and d_n are the dimensions of the image. Given the image, at level $(i-1)$, the image at level i is obtained as

$$\boldsymbol{B_i}(m_1, m_2) = \sum_{p=-2}^{2} \sum_{q=-2}^{2} \boldsymbol{w_d}(p, q) \boldsymbol{B_{i-1}}(2m_1 + p, 2m_2 + q) \tag{2.1}$$

where $\boldsymbol{B_i}$ is the image obtained at scale i and $\boldsymbol{w_d}$ is the kernel function. Each value of the matrix $\boldsymbol{B_i}$ is computed as the weighted average of the pixel values in $\boldsymbol{B_{i-1}}$ within a 5×5 window. The weights for the averaging process is provided by $\boldsymbol{w_d}$. In this study, the separable filter $\boldsymbol{w_d}(p, q) = \boldsymbol{w_1}(p)^T \boldsymbol{w_1}(q)$ with $\boldsymbol{w_1} = [1/4 - a_1/2, 1/4, a_1, 1/4, 1/4 - a_1/2]$ and $a_1 = 0.375$ are used [35].

2.1.3 Feature Extraction, Fusion and Classification

The MSSP decomposed images are fed to the feature fusion and classification stage to obtain the diagnosis output. As the exact scale of the pathological manifestations is unknown, multiple scaled representations of the input image are fed to different CNNs to effectively represent the multi-scaled disease features. Figure 2.2 shows that the decomposed images, $\boldsymbol{B_i}$, $i \in \{0, 1, ..., N\}$ are fed to the corresponding CNNs ($\mathrm{CNN_i}$, $i \in \{0, 1, ..., N\}$) for extracting scale-specific information. Here $N+1$ represents the number of decomposed image scales and hence the number of CNNs used for feature extraction. Mathematically,

the output feature vector ($\boldsymbol{f_{c_i}}$) for the $\mathrm{CNN_i}$ can be represented as

$$\boldsymbol{f_{c_i}} = f(\boldsymbol{B_i}; \boldsymbol{\theta_{c_i}}) \tag{2.2}$$

where $f(.)$ is a composite function representing the multiple linear and non-linear operations of the CNN and $\boldsymbol{\theta_{c_i}}$ represents the learnable parameters of the $\mathrm{CNN_i}$. The number of decomposed image scales and the number of CNNs for feature extraction are set through experimentation. In this work, three image scales are used. The network architecture of the three CNNs, i.e., $\mathrm{CNN_i}$, $i \in \{0, 1, 2\}$ are provided in Figure 2.3. The network architecture for $\mathrm{CNN_3}$ is also provided in Figure 2.3 as it will be used for ablation study in section 2.2.3. The variables k, n and s represent the kernel size, the number of filters and the stride for the convolution operations. Each of the convolutional layers in the CNNs is followed by batch normalization (BN), rectified linear unit (ReLU) activation and pooling layers. The BN after each convolution layer ensures fast and stable training [115]. The ReLU layer maintains the sparsity of the convolution kernel [116, 117]. The sub-sampling of the feature maps using max-pooling with a kernel size of two makes the features position and rotation invariant. More details on the different components of the CNN are provided in Appendix B. As shown in Figure 2.3, the number of convolution filters doubles with the increasing depth for each of the CNNs. This design is inspired by the architecture of the VGG16 [77], which is one of the best performing classifiers for object detection. A two-fold increment is also introduced in the number of the convolutional filters as we move across scales at a particular convolution layer. This helps the classifier to learn distinct features across scales. In the end, FC layers with 15 neurons are employed to summarize the convolutional feature maps at different scales.

The outputs of the CNNs ($CNN_i, i \in \{0, 1, 2\}$) are then fused through concatenation to encode the multi-scale feature information. The fused feature vector can be represented as $\boldsymbol{f_{mf}} = [\boldsymbol{f_{c_0}}, \boldsymbol{f_{c_1}}, ..., \boldsymbol{f_{c_N}}]$. The fusion of features from multiple scales can capture the inter-scale variations introducing complementary information to the classifier. It also transforms the feature representation to higher dimensions with increased non-linearity, thereby assuring better discrimination capability of the classifier. In the end, $\boldsymbol{f_{mf}}$ is fed to the output FC layer with softmax activation to obtain the class probabilities for the input image $\boldsymbol{B}$. The probability distribution, p_{mf} of the output categories obtained at the output layer is given as

$$p_{mf}(c|\boldsymbol{B}) = Softmax(\boldsymbol{W_0 f_{mf}} + \boldsymbol{b_0}) \tag{2.3}$$

where $p_{mf}(c|\boldsymbol{B})$ represents the probability of $\boldsymbol{B}$ belonging to class c, $c \in \{1, 2, ..., C\}$, C is the number of output categories, $\boldsymbol{W_0} \in \mathbb{R}^{C \times d_{mf}}$ and $\boldsymbol{b_0} \in \mathbb{R}^{C \times 1}$ are the weights and biases of the output layer. d_{mf} is

the dimension of the fused feature vector $\boldsymbol{f_{mf}}$. The final predicted label for the image $\boldsymbol{B}$ is obtained as

$$class(\boldsymbol{B}) = \underset{c=1,2,...,C}{\arg\max}\, p_{mf}(c|\boldsymbol{B}). \tag{2.4}$$

Training Loss Optimization: Conventionally, the categorical cross-entropy loss between the probabilistic output and one hot encoded label for a set of training samples is minimized to learn the network parameters of the model [118]. Mathematically, the categorical cross-entropy loss is formulated as

$$L = -\frac{1}{M}\sum_{m=1}^{M}\sum_{c=1}^{C}\mathcal{I}(y_m = c)\, log\, p_{mf}(c|\boldsymbol{X_m}) \tag{2.5}$$

where $(\boldsymbol{X}_j, y_j), j \in \{1, 2, ..., M\}$ are the training examples, M is the number of training examples and $\mathcal{I}(.)$ is an indicator function which is equal to one if y_m equals to c.

The clinical OCT datasets are often imbalanced as some diseases occur less frequently than others. The imbalance in the datasets can bias the classification result towards the majority classes, thereby leading to poor detection of the minority class samples. To tackle this issue, a cost-sensitive loss function is employed instead of the categorical cross-entropy loss while training the classifier [119–121]. A class weighted categorical cross-entropy loss is used that penalizes the class errors differently. Specifically, the minority class errors are privileged with a higher weight value than the majority class samples' errors. The class weighted categorical cross-entropy loss function (L_w) is defined as

$$L_w = -\frac{1}{M}\sum_{m=1}^{M}\sum_{c=1}^{C} w_c\, \mathcal{I}(y_m = c)\, log\, p_{mf}(c|\boldsymbol{X_m}) \tag{2.6}$$

where w_c is the weight assigned to the classification errors of samples from class c. The weight for a particular class c is computed as

$$w_c = \frac{\text{number of training samples in the majority class}}{\text{number of training samples in class c}}. \tag{2.7}$$

It can be observed from Eq. 2.7, that the weights are inversely proportional to the class distribution. L_w assigns a higher weight to the training examples of the class with low sample size. The reason for choosing weights in this manner is to treat one instance of class c as w_c instances of the majority class. The obtained weights equalize the degree of the impact of the losses from the individual classes towards the total loss.

Table 2.1: Details of the databases used for evaluation.

Database	Acquisition System	Scan type	Axial Resolution	No. of Subjects	Classes	Volumes/class	B-scans/class
UCSD [34]	Spectralis OCT, Heidelberg Engineering, Germany	Foveal B-scan	-	4686	Drusen	-	8866
					CNV		37455
					DME		11598
					Normal		26565
NEH [35]	Heidelberg SD-OCT imaging systems	Volume	3.5μm	148	AMD	48	1296
					DME	50	1086
					Normal	50	1578

2.2 Experimental Results for the MDFF Method

In this section, the details of the databases used for evaluation, the performance measures and the ablation study are presented. Following that, the details of the existing methods used for performance comparison and the results are discussed.

2.2.1 Clinical Database

The proposed method is evaluated on two publicly available databases: the UCSD dataset [34] and the NEH dataset [35]. The UCSD dataset contains 84,484 OCT B-scan images from 4686 retrospective cohorts. The database contains labeled OCT B-scans from two AMD stages (drusens-early and CNV-advanced), DME and healthy controls. Precisely, the dataset consists of 8866, 37455, 11598 and 26565 OCT B-scan images from drusen, CNV, DME and normal categories, respectively. The NEH dataset comprises 148 SD-OCT volumes (48 AMD, 50 DME and 50 normal subjects). A total of 1296, 1086, and 1578 B-scans are extracted from the OCT volumes for AMD, DME, and normal classes, respectively. Additional details about the databases are presented in Table 2.1.

2.2.2 Evaluation Scheme and Performance Measures

In this work, 10-fold and 5-fold cross-validation protocols are considered for the UCSD and the NEH databases, respectively. Cross-validation ensures generalizable results and prevents overfitting or bias to specific portions of the databases. Considering the limited sample size of the NEH dataset, the training B-scan images are augmented by horizontally flipping for efficient learning. It is worth mentioning that no augmentation is carried out for the testing folds of the NEH dataset. The UCSD database exhibits class imbalance with its varied class distribution. Therefore, the class weighted categorical cross-entropy loss is used for reliable training. The conventional categorical cross-entropy loss is used for training the model

Table 2.2: A multi-class confusion matrix

		Predicted output			
		1	2	...	C
Actual output	1	u_{11}	u_{12}	...	u_{1C}
	2	u_{21}	u_{22}	...	u_{2C}
	.	.	.	...	.
	.	.	.	...	.
	.	.	.	...	.
	C	u_{C1}	u_{C2}	...	u_{CC}

on the NEH dataset. The classification performance is evaluated using the class-wise and the overall measures derived from the confusion matrix. Table 2.2 shows the representation of a multi-class confusion matrix with entries u_{ab} representing the number of samples from the test set originally belonging to class a and classified as class b where $a, b \in \{1, 2, ..., C\}$. The class-wise measures include the sensitivity (SE), the precision (PR) and the F1 score. The mathematical formulation of these measures are given in Eq. 2.8 where SE_a, P_a and the $F1\ score_a$ are the class-wise sensitivity, precision and F1 score for class $a, a \in \{1, 2, ..., C\}$. The overall measures used are the overall sensitivity (OS), the overall precision (OP), the overall F1 score (OF1) and the overall accuracy (OA) and are computed as shown in Eq. 2.9.

$$SE_a = \frac{u_{aa}}{\sum_{e=1}^{C} u_{ae}} \qquad P_a = \frac{u_{aa}}{\sum_{e=1}^{C} u_{ea}} \qquad F1\ score_a = 2 \times \frac{SE_a \times P_a}{SE_a + P_a} \tag{2.8}$$

$$OS = \frac{1}{C}\sum_{e=1}^{C} SE_e \qquad OP = \frac{1}{C}\sum_{e=1}^{C} P_e \qquad OF1 = \frac{1}{C}\sum_{e=1}^{C} F1\ score_e \tag{2.9}$$

The area under the curve (AUC), the Cohen's Kappa [122] and the run time are also analyzed for the proposed and the existing methods.

2.2.3 Network Parameters and Ablation Study

The pre-processed OCT images are provided to the MSSP module for extracting multi-scale representations from the input OCT B-scans. These representations are provided as input to different CNNs to extract scale-specific features. The network details of the CNNs are provided in Figure 2.3. The number of CNNs ($CNN_i, i \in \{0, 1, ..., N\}$) used in the method is fixed through experimentation. Table 2.3 shows the classification performance for the increasing number of CNNs in the MDFF method for the UCSD database. An improvement in the classification performance is observed for the fusion of three scales over two scales. It is also observed that the number of trainable parameters increases as the features

Table 2.3: The effect of number of CNNs on the overall classification performance for the MDFF method on the UCSD database.

Scale combinations	OP (%)	OS (%)	OA (%)	# parameters
CNN_0-CNN_1	91.92	91.98	94.37	40984
CNN_0-CNN_1-CNN_2	92.51	93.01	94.74	63187
CNN_0-CNN_1-CNN_2-CNN_3	92.91	93.05	94.79	82846

The bold entity indicates the selected scale combination for the proposed method.

Table 2.4: The effect of multi-scale features on the overall classification performance for the MDFF method.

Database	Configuration	OP (%)	OS (%)	OA (%)
UCSD	CNN_0	91.13	92.60	93.94
	CNN_1	90.73	90.23	93.15
	CNN_2	88.49	87.64	91.57
	MDFF	**92.51**	**93.03**	**94.74**
NEH	CNN_0	93.35	93.03	93.19
	CNN_1	96.17	96.35	96.23
	CNN_2	95.10	94.43	94.84
	MDFF	**97.09**	**97.01**	**97.10**

The bold values show the performance of the proposed method.

from different scales are appended. Therefore, it is essential to find a trade-off between the classification accuracy and the computational burden. It can be seen from Table 2.3, that performance improvement of the fusion of four scales over the three scales is minimal. However, the computational load increases from learning 63187 to 82846 parameters. Hence, to reduce the computational burden and still obtain a reliable classification performance, we consider the fusion of three scales, $i \in \{0, 1, 2\}$ for the proposed framework.

The proposed network is implemented using Keras with a TensorFlow backend. The model is trained with an Adam [123] optimizer with a learning rate of 10^{-3} on mini-batches of size 32 for 69080 iterations. An NVIDIA Tesla V100 graphical processing unit (GPU) is used to perform the experiments.

To analyze the contribution of the different components of the model, we evaluate the proposed MDFF method with some variations.

Significance of multi-scale feature fusion: To verify the effectiveness of multi-scale feature extraction and fusion, we compare the classification performance with the single-scale frameworks. The single-scale frameworks classify the macular pathologies by considering input images from only one scale. For this experiment, we define each $CNN_i, i \in \{0, 1, 2\}$ as the single-scale CNNs learned with the OCT images at scale i for classification. The network architecture of the CNN_i is the same as employed for the MDFF classifier until the FC layer. An output layer with a softmax activation function is used after the FC layer

Table 2.5: The effect of cost-sensitive learning on classification performance for the MDFF method on the UCSD database.

Method	Class	Sensitivity (%)	Specificity (%)	G-mean (%)
MDFF (categorical cross entropy)	CNV	96.92	97.15	97.03
	DME	92.51	98.59	95.50
	DRUSEN	83.08	98.59	90.50
	NORMAL	96.81	98.84	97.82
MDFF (class weighted categorical cross entropy)	CNV	95.25	97.89	96.56
	DME	91.78	98.93	95.29
	DRUSEN	**87.28**	98.10	**92.53**
	NORMAL	97.74	97.85	97.79

for classification. A dropout factor of 0.5 is used prior to the output layer to prevent overfitting. The single-scale frameworks are trained individually with similar hyper-parameter settings (optimizer, learning rate, iterations) as the proposed method. Table 2.4 shows the overall classification performance of the single-scale CNNs and the MDFF method for the UCSD and the NEH datasets. As can be seen, the proposed multi-scale approach outperforms the single-scale methods. This verifies that the multi-scale features improve the discrimination ability of the classifier, thereby providing better performance.

Significance of the cost-sensitive learning: The effects of the class weighted categorical cross-entropy loss function on classification are analyzed. The loss function is supposed to improve the detection sensitivity for the minority class samples. We compare the classification results obtained by using the conventional and the class weighted categorical cross-entropy losses for the UCSD database and the results are shown in Table 2.5. The sensitivity, specificity and G-mean metrics are used for evaluation. The G-mean is a popularly used measure to validate the imbalanced datasets [124, 125]. It can be observed that there is an improvement in the sensitivity of the drusen class from 83.08% to 87.28% while learning the MDFF classifier using the class weighted categorical cross-entropy loss over the conventional categorical cross-entropy loss. Similar improvements are also observed for the G-mean. Therefore, it can be inferred that the class weights aid in improving the detection rate of the minority class without affecting the detection rate of the majority class significantly.

2.2.4 Existing Methods used for Performance Comparison

To verify the effectiveness of the proposed method, we compare it with the following existing methods.

- *Feature based method:* The proposed method is compared with the popularly used feature-based HOG+SVM method [36]. The HOG features for the pre-processed input OCT images are computed

for a cell size of 16 × 16 with 9 histogram bins. The obtained feature vectors have 3780 dimensions for input images of size 128 × 256 for a block size of 2 × 2 cells with 50% overlap. The obtained feature descriptors are provided to the SVM classifier for obtaining the class predictions.

- *Transfer learning based methods:* Transfer learning models like the VGG16 [77] and the InceptionV3 [34] are used in this study. The fine-tuning of these models is carried out by resizing the input OCT images to 224×224 pixels and replacing the network's output layer neurons with the number of classes (four and three for the UCSD and NEH dataset, respectively). All the network parameters except the output layer are frozen. The Adam [123] optimizer is used for fine-tuning on mini-batches of size 32 for 69080 iterations. The fine-tuned models are then employed to obtain the class predictions for the test images.

- *State-of-the-art OCT classification methods:* Recent state-of-the-art OCT classification methods like the IFCNN [56], the LGCNN [89] and the multi-scale convolutional mixture of experts (MCME) [35] are used for comparison. The IFCNN and the LGCNN are single-scale approaches originally proposed and evaluated on the UCSD database. The results for these methods are directly taken from the research articles, as a similar evaluation strategy on the same dataset is adopted in this chapter. The MCME is a multi-scale CNN framework originally proposed for the NEH database. As the evaluation strategy of the proposed method is slightly different from that presented in the [35], the network parameters of the MCME method are learned on the NEH database with the same hyper-parameters as the proposed method for a fair comparison.

2.2.5 Results on the UCSD Database

Table 2.6 shows the quantitative classification performance of the proposed and the existing methods on the UCSD database using the 10-fold cross-validation scheme. As can be seen, the DL based methods achieve better performance compared to the feature-based HOG+SVM method. This is because the HOG can only extract limited gradient and texture features from the OCT images, which are not sufficient for reliable classification. On the other hand, the learnable convolution kernels, the multiple layers of representation and the non-linear modules of the deep CNN frameworks encode high-level discriminative feature representations leading to improved performance [71]. Although the deep models perform well, the transfer learning frameworks (VGG16 and InceptionV3) show limited performance compared to the proposed method (see Table 2.6). This is because these networks are pre-trained on natural images and may not be suitable for capturing clinical features from medical images like OCT.

Table 2.6: Performance evaluation of the proposed MDFF and the existing methods on the UCSD database using 10-fold cross-validation.

Method	Class	SE (%)	PR (%)	F1 score (%)	OS/OP/OF1/OA (%)	AUC/Kappa	Avg. Test Time/image (ms)
HOG+SVM [36] (feature based)	CNV	89.35 ± 0.41	91.33 ± 0.33	90.28 ± 0.45	80.81 ± 0.49	0.88 ± 0.003 0.79 ± 0.005	0.4 ± 0.001
	DME	81.19 ± 1.51	78.79 ± 1.03	79.97 ± 1.12	80.20 ± 0.53		
	Drusen	61.99 ± 1.10	61.06 ± 1.28	61.52 ± 1.09	80.45 ± 0.49		
	Normal	90.57 ± 1.21	89.34 ± 0.57	89.99 ± 0.42	85.78 ± 0.36		
VGG16 [77] (Transfer learning)	CNV	86.26 ± 5.14	95.09 ± 1.76	90.34 ± 2.09	84.26 ± 1.00	0.890 ± 0.55 0.788 ± 0.04	34.9± 0.1
	DME	84.53 ± 6.73	79.47 ± 8.63	81.29 ± 2.38	80.98 ± 2.77		
	Drusen	77.83 ± 6.29	56.53 ± 9.48	68.66 ± 10.88	81.69 ± 2.31		
	Normal	88.41 ± 5.11	92.85 ± 1.68	90.42 ± 2.15	86.03 ± 2.72		
InceptionV3 [34] (Transfer learning)	CNV	94.84 ± 1.12	95.7 ± 0.67	95.18 ± 0.49	90.26 ± 0.47	0.939 ± 0.003 0.896 ± 0.004	39.48 ± 0.5
	DME	89.33 ± 1.54	79.47 ± 1.43	89.88 ± 0.49	90.14 ± 0.67		
	Drusen	80.82 ± 2.11	79.47 ± 2.99	80.07 ± 0.93	90.17 ± 0.28		
	Normal	96.07 ± 0.74	94.95 ± 0.60	95.50 ± 0.29	93.01 ± 0.27		
IFCNN [56] (Single-scale)	CNV	94.80 ± 1.90	87.90 ± 4.30	90.90 ± 2.90	82.50 ± 3.00	-	-
	DME	79.20 ± 8.90	81.90 ± 6.80	97.20 ± 1.00	84.70 ± 2.40		
	Drusen	64.40 ± 8.4	76.80 ± 7.20	97.30 ± 0.80	-		
	Normal	91.50 ± 3.00	92.20 ± 4.70	96.40 ± 2.00	87.30 ± 2.20		
LGCNN [89] (Single-scale)	CNV	93.30 ± 0.90	91.50 ± 2.40	93.30 ± 1.70	84.60 ± 2.10	-	-
	DME	85.70 ± 2.60	79.40 ± 3.70	96.80 ± 0.50	82.90 ± 1.80		
	Drusen	71.00 ± 6.60	65.20 ± 5.10	96.00 ± 0.60	-		
	Normal	88.50 ± 2.70	95.50 ±2.40	97.90 ± 1.10	88.40 ± 1.30		
MDFF (Proposed)	CNV	**95.91 ± 1.31**	**96.65 ± 0.92**	**96.27 ± 0.31**	**92.56 ± 1.06**	**0.954 ± 0.005** **0.919 ± 0.005**	**5.03 ± 0.13**
	DME	**91.56 ± 1.46**	**92.49 ± 1.96**	**92.00 ± 0.61**	**92.43 ± 1.23**		
	Drusen	**86.62 ± 3.59**	**85.23 ± 4.05**	**85.28 ± 0.96**	**92.48 ± 0.43**		
	Normal	**96.92 ± 1.01**	**95.78 ± 0.91**	**96.34 ± 0.25**	**94.57 ± 0.35**		

The bold values show the performance of the proposed method.

Compared to the single-scale state-of-the-art OCT classification methods like the IFCNN and the LGCNN, the proposed MDFF method performs favorably well with a minimum improvement of nearly 2.5% in OS, 1.6% in OA and 3.3% in kappa values. The method attains an OS, OP and OA of 92.56%, 92.43%, and 94.57%, respectively, without compromising on the class-wise detection rates. A slight drop in the SE, PR and F1 score is observed for the results of the proposed method on the drusen class. This is because the drusens are early manifestations of AMD having a small size and sparse distribution. These subtle clinical changes are often indistinguishable from normal OCT images. The drusens also show high visual similarity to the CNV class as drusens also manifest in CNV. However, compared to the existing methods, the detection rate for the drusen class is higher for the proposed approach.

Table 2.7: Performance evaluation of the proposed MDFF and the existing methods on the NEH database using 5-fold cross-validation.

Method	Class	SE (%)	PR (%)	F1 score (%)	OS/OP/OF1 (%)	OA (%)	AUC/Kappa	Avg. Test Time/image (ms)
HOG+SVM [36] (Feature based)	AMD	81.87 ± 3.23	84.23 ± 1.58	82.99 ± 1.83	84.53 ± 1.04	84.85 ± 0.90	0.883 ± 0.007	0.4 ± 0.001
	DME	83.52 ± 3.54	88.15 ± 0.86	85.74 ± 1.90	85.25 ± 0.75		0.769 ± 0.01	
	Normal	88.21 ± 1.82	83.35 ± 1.89	85.69 ± 1.03	84.81 ± 0.92			
VGG16 [77] (Transfer learning)	AMD	90.19 ± 1.47	91.57 ± 2.48	90.87 ± 1.81	92.67 ± 1.18	92.75 ± 1.20	0.945 ± 0.009	36.04 ± 0.36
	DME	93.46 ± 2.89	93.93 ± 1.62	93.67 ± 1.19	92.82 ± 1.21		0.889 ± 0.02	
	Normal	94.24 ± 1.85	93.00 ± 0.84	93.64 ± 1.11	92.45 ± 1.25			
InceptionV3 [34] (Transfer learning)	AMD	95.37 ± 0.98	94.82 ± 1.65	95.08 ± 0.53	96.02 ± 0.38	96.06 ± 0.41	0.969 ± 0.003	42.8 ± 1.63
	DME	96.13 ± 1.06	97.41 ± 1.26	96.75 ± 0.56	96.15 ± 0.31		0.94 ± 0.006	
	Normal	96.58 ± 1.09	96.22 ± 0.89	96.39 ± 0.80	96.08 ± 0.36			
MCME [35] (Multi-scale)	AMD	88.96 ± 3.69	87.66 ± 3.03	88.24 ± 2.09	89.57 ± 1.98	89.79 ±1.9	0.921 ± 0. 015	6.82 ± 0.01
	DME	88.12 ± 2.21	93.84 ± 2.22	90.84 ± 1.85	90.26 ± 1.70		0.844 ± 0. 029	
	Normal	91.64 ± 1.95	89.35 ± 4.69	90.42 ±2.48	89.83 ± 1.85			
MDFF (Proposed)	AMD	**97.45 ± 0.85**	**97.23 ± 0.77**	**97.34 ± 0.69**	**97.51 ± 0.46**	**97.57 ± 0.43**	**0.982 ± 0. 004**	**6.34 ± 0.43**
	DME	**97.05 ± 0.76**	**97.59 ± 1.08**	**97.33 ± 0.83**	**97.56 ± 0.45**		**0.963 ± 0. 007**	
	Normal	**98.03 ± 0.34**	**97.85 ± 0.34**	**97.84 ± 0.37**	**97.54 ± 0.45**			

The bold values show the performance of the proposed method.

2.2.6 Results on the NEH Database

Table 2.7 shows the classification results of the proposed and the existing methods on the NEH database. A performance trend similar to the UCSD database is observed for the NEH dataset as well. As can be seen, the proposed MDFF method outperforms the existing methods. It can be observed that the method achieves an OS, OP and OA of 97.51%, 97.56% and 97.57%, respectively. Although the MCME presents a promising multi-scale approach, its low performance might be attributed to the choice of network architecture. The method uses only a few filters in the convolutional layers, which may not be sufficient to effectively capture diverse clinical features.

The proposed method outperforms the feature-based (HOG+SVM) method, transfer learning approaches (VGG16 and InceptionV3) and MCME with a large margin of nearly 14.9%, 5.2%,1.6% and 8.6% in OA, respectively. Similar improvements in performance can be observed for other measures as well. It can also be observed from Table 2.7 that the proposed method has a small variance in measures across the 5 folds, thereby highlighting the excellent generalization of the model.

Tables 2.6 and 2.7 also present the average time taken by the method to generate the classification outputs. These measurements are taken from a system with 16 GB RAM and Quadro K600 graphics processor. It can be observed that bulky networks like the VGG16 and the InceptionV3 take the largest amount of time (more than 34 ms) to generate a class prediction. Similarly, the less complex HOG+SVM takes the least test run time of 0.4 ms. The MCME also has a significantly low test run time as it has a very

light-weight architecture. The MDFF method takes an average test run time of nearly 6 ms to produce the diagnosis decision.

The improved classification performance of the MDFF method can be attributed to the multi-scale feature extraction and fusion strategy that enabled the mining of cross-scale discriminative features for efficient classification. The method employs a fixed multi-scale transform method to obtain the multi-scale representations. It would be interesting to study the performance of the classifier when the network is allowed to learn the multi-scale representations automatically. Therefore, in the following section, we propose the LM-DFF method that explores learnable convolutional kernels with different receptive fields to obtain the multi-scale representations for efficient classification of the OCT images.

2.3 Learnable Multi-scale Deep Feature Fusion for OCT B-scan Classification

In this approach, the multi-scale features from the OCT B-scans are extracted using multiple convolution filters with different receptive fields. These convolutional filters encode the varied scale-specific features, which are aggregated in the end to capture cross-scale information for improved classification. A joint multi-loss optimization approach is then utilized to enable the network to effectively learn the scale-specific and complementary cross-scale features from the OCT images. The pipeline of the proposed framework is shown in Figure 2.5. As can be seen from the figure, the method consists of 4 modules, namely, the CNN backbone, the multi-scale feature extraction, the feature fusion and classification and the joint multi-loss optimization. The details of each of the modules are described as follows.

2.3.1 CNN Backbone

In this module, low-level feature representations are extracted from the input B-scan image $\boldsymbol{B}$. As can be seen from Figure 2.5, a stack of learnable convolutional kernels is used to encode the local spatial features from the OCT image. The convolutional layers are followed by a max-pooling layer to reduce the spatial dimensions by fusing the neighborhood spatial information. Mathematically, the output feature maps of the CNN backbone $\boldsymbol{F_c}$ can be represented as

$$\boldsymbol{F_c} = f(\boldsymbol{B}; \boldsymbol{\theta_b}) \tag{2.10}$$

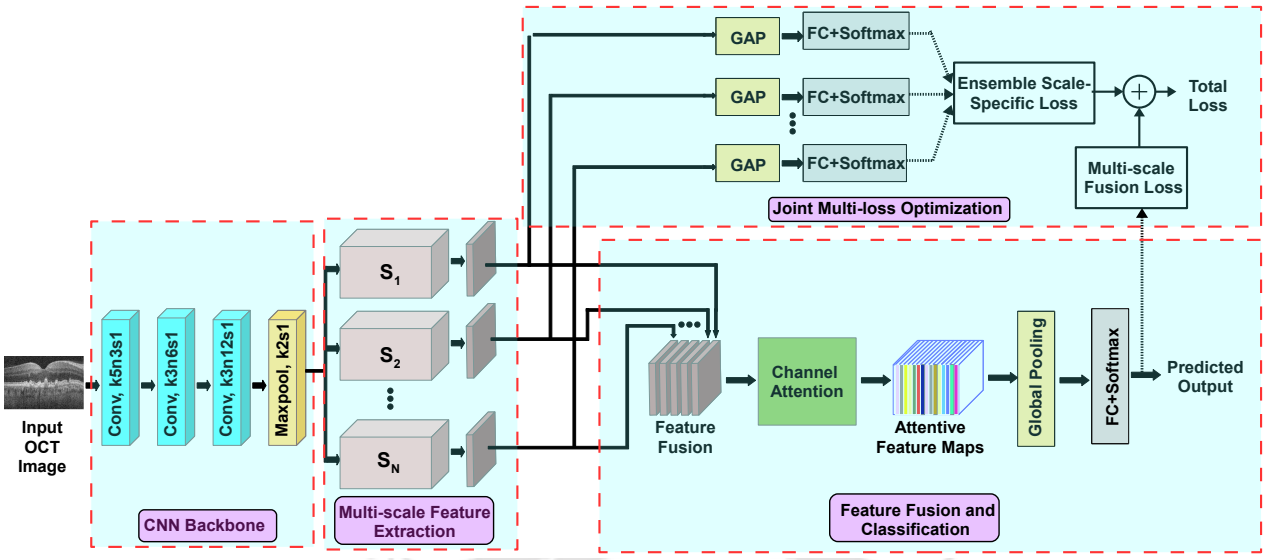


Figure 2.5: Pipeline of the proposed LM-DFF classification method.

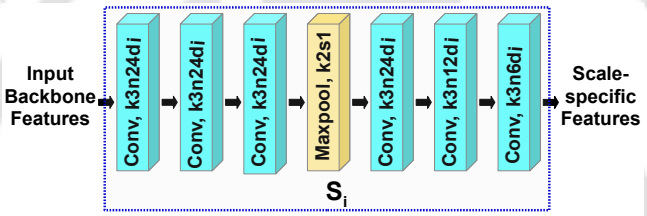


Figure 2.6: Network architecture of the SS-CNN.

where $f(.)$ is the composite function representing the multiple linear and non-linear operations of the CNN backbone, θ_b represents the learnable parameters of the backbone network. The feature maps obtained from the CNN backbone are provided to the multi-scale feature extraction module to extract scale-specific features.

2.3.2 Multi-scale Feature Extraction

This module automatically learns the disease-related features using a series of scale-specific CNNs (SS-CNNs), S_i, $i \in \{1, 2, ..., N\}$. Here, N represents the number of SS-CNNs used in the framework. Each of the SS-CNNs has different receptive fields to effectively capture the multi-scaled pathological characteristics. To achieve the varied receptive fields, dilated convolution [126] is adopted in this work. Mathematically, the 2D dilated convolution can be defined as

$$\boldsymbol{z}(p_1, p_2) = \sum_{l_1} \sum_{l_2} \boldsymbol{x}(p_1 + d.l_1, p_2 + d.l_2)\boldsymbol{w}(l_1, l_2). \tag{2.11}$$

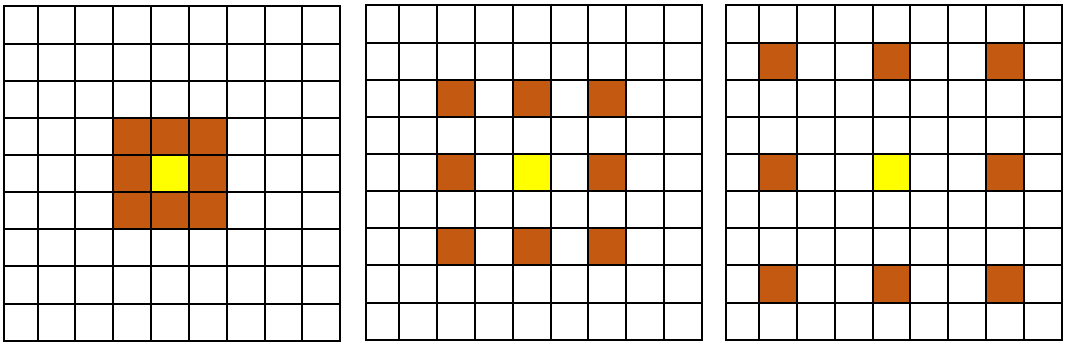


Figure 2.7: Visualization of the convolution kernels with different dilation rates.

Here, $\boldsymbol{x}$ is the input, $\boldsymbol{w} \in \mathbb{R}^{k \times k}$ is the convolution kernel, d is the dilation rate and $\boldsymbol{z}$ is the output of the dilated convolution. Eq. 2.11 can be interpreted as the convolution of input $\boldsymbol{x}$ with a dilated filter. The dilated filter is obtained by inserting holes (zeros) between the kernel elements of $\boldsymbol{w}$ based on the dilation factor. The effective kernel size k' of the dilated filter for a dilation rate d is given as $k' = (k-1)d + 1$. Figure 2.7 shows the visualization of a 3×3 convolution kernel with different dilation rates. As can be seen from the figure, $(d-1)$ zeros are inserted between the kernel elements for a dilation rate of d. It can be observed that the effective receptive field of the convolutional kernel increased from 3×3 to 5×5 and 7×7 for the dilation rates of 2 and 3, respectively. The obtained dilated filters with holes reduce the redundant feature encoding and can help in learning discriminative features. The spacing between the kernel elements introduced by the dilation rate increases the receptive field without much increase in the network parameters compared to the traditional convolutions.

The SS-CNNs, S_i, $i \in \{1, 2, ..., N\}$ in the proposed framework employ 2D dilated convolutions with dilation factors of i to extract high-level multi-scale representative features. The network architecture of S_i is provided in Figure 2.6. Multiple S_is, $i \in \{1, 2, ..., N\}$ are used to extract discriminative feature representations from the disease regions effectively.

The output convolutional feature map ($\boldsymbol{F_{s_i}}$) for each S_i is given as

$$\boldsymbol{F_{s_i}} = f_i(\boldsymbol{F_c}; \boldsymbol{\theta_{s_i}}) \tag{2.12}$$

where $f_i(.)$ is the function representing S_i, $\boldsymbol{\theta_{s_i}}$ represents the trainable parameters of the convolution filters of S_i. Based on experimentation, three S_is, $i \in \{1, 2, 3\}$ are considered to obtain the classification results. The experimental analysis for the choice of the number of SS-CNNs is provided in Section 2.4.1.

2.3.3 Feature Fusion and Classification

The convolutional feature maps obtained by the SS-CNNs are fused to mine the cross-scale complementary clinical information for robust feature extraction. The fusion is performed using the concatenation operation and the multi-scale fused features are obtained as

$$\boldsymbol{F_{mf}} = [\boldsymbol{F_{s_0}}, \boldsymbol{F_{s_1}}, ..., \boldsymbol{F_{s_N}}]. \tag{2.13}$$

To strengthen the representation power of the fused multi-scale features, the channel attention mechanism is explored. It improves the quality of the feature representations by explicitly modeling inter-dependencies between the channels. In this work, the SE unit [97] is adopted to improve attention to the informative feature maps generated by the feature fusion [96]. The SE unit selectively emphasizes the informative features and suppresses the less useful ones [97]. The weights of importance for the feature maps are obtained by first squeezing the global spatial information using the global average pooling (GAP). The aggregated information in the squeeze operation is then used as input to FC layers to obtain the recalibration weights. The channel attentive feature maps ($\boldsymbol{F_{att}}$) are obtained by rescaling $\boldsymbol{F_{mf}}$ with the obtained weights.

Finally, global pooling is applied to summarize the channel attentive convolutional feature maps by squeezing the spatial dimensions. Here, both GAP and global max-pooling (GMP) [127] are used for feature summarization. The GAP can encode the overall average information of the spatial feature maps and the GMP can effectively extract discriminative features from the spatial dimensions. The final integrated feature vector $\boldsymbol{h_m} \in \mathbb{R}^{d_h \times 1}$ is obtained as

$$\boldsymbol{h_m} = P_a(\boldsymbol{F_{att}}) + P_m(\boldsymbol{F_{att}}) \tag{2.14}$$

where $P_a(.)$ and $P_m(.)$ are the GAP and GMP operations, respectively. Finally, $\boldsymbol{h_m}$ is fed to an FC output layer with softmax activation to obtain the probability distribution of the output categories as

$$p_f(c|\boldsymbol{B}) = Softmax(\boldsymbol{W}\boldsymbol{h_m} + \boldsymbol{b}) \tag{2.15}$$

where $p_f(c|\boldsymbol{B})$ represents the probability of $\boldsymbol{B}$ belonging to class c, $c \in \{1, 2, ..., C\}$, C is the number of output categories, $\boldsymbol{W} \in \mathbb{R}^{C \times d_h}$ and $\boldsymbol{b} \in \mathbb{R}^{C \times 1}$ are the weights and biases of the output layer. The final

predicted label for the image $\boldsymbol{B}$ is obtained as

$$class(\boldsymbol{B}) = \underset{c=1,2,...,C}{\arg\max}\, p_f(c|\boldsymbol{B}). \tag{2.16}$$

2.3.4 Joint Multi-Loss Optimization

To effectively learn the network parameters for encoding scale-specific and cross-scale information, two loss functions are used, namely, the multi-scale fusion loss (L_f) and the ensemble scale-specific loss (L_e). L_f optimizes the complementary information across the scales. It is obtained by minimizing the categorical cross-entropy loss (Eq. 2.17) between the probabilistic output obtained at the feature fusion and classification module and one hot encoded labels for a set of training samples $(\boldsymbol{X}_j, y_j), j \in \{1, 2, ..., M\}$ and M is the number of training examples. It is given as

$$L_f = -\frac{1}{M}\sum_{m=1}^{M}\sum_{c=1}^{C} w_c\, \mathcal{I}(y_m = c)\, log\, p_f(c|\boldsymbol{X_m}) \tag{2.17}$$

where w_c are the weights assigned to the errors of training samples of class c and $\mathcal{I}(.)$ is an indicator function which is equal to one if y_m equals to c.

To enable the SS-CNNs to learn discriminative scale-specific features, individual supervision is provided to each of the S_is. The convolutional feature maps $\boldsymbol{F_{s_i}}$ of each S_i is summarized using the GAP and an output FC layer with softmax activation (see Figure 2.5) to obtain the scale-specific predictions p_{s_i}. The scale-specific loss L_{s_i} for the *CNN backbone* $\rightarrow S_i \rightarrow$ *GAP* $\rightarrow$ *FC+Softmax* sub-network is given as

$$L_{s_i} = -\frac{1}{M}\sum_{m=1}^{M}\sum_{c=1}^{C} w_c\, \mathcal{I}(y_m = c)\, log\, p_{s_i}(c|\boldsymbol{X_m}). \tag{2.18}$$

The ensemble scale-specific loss L_e is computed as the sum of the all the scale-specific losses and is given as

$$L_e = \sum_{i=1}^{N} L_{s_i}. \tag{2.19}$$

A joint loss optimization strategy is adopted to simultaneously balance the scale-specific and complementary cross-scale learning. The final objective function is given as follows.

$$L_t = L_f + \lambda L_e \tag{2.20}$$

where λ is the constant factor weighting the losses. In this work, λ =1 is chosen.

At the end of the training process, the CNN backbone, multi-scale feature extraction, feature fusion and classification modules are utilized for generating class labels for the unseen test OCT B-scans. In the following section, the performance of the LM-DFF method is analysed and the comparison results with the MDFF method are presented.

2.4 Experimental Results for the LM-DFF Method

In this section, we present the details of the network parameters, the ablation study and the comparison results.

2.4.1 Network Parameters and Ablation Study

Similar to the MDFF method, the input OCT images are pre-processed by flattening of the retinal curvature, cropping and resizing the images to 128×256 pixels [35]. The pre-processed images are provided as input to the proposed LM-DFF framework. The details of the network architecture of the proposed method are provided in Figure 2.5 and 2.6. The variables k, n, s and d represent the kernel size, the number of filters, the stride and the dilation rate for the convolution operations. It is to be noted that each convolution layer in the network is followed by BN and ReLU activation layers. In this work, the number of the SS-CNNs ($S_i, i \in \{1, 2, ..., N\}$) are set through experimentation. Table 2.8 shows the classification performance of the proposed framework on the UCSD database with the increasing number of the SS-CNNs. The performance is compared in terms of OP, OS, OA and number of parameters. It can be observed that classification performance increases as the number of SS-CNNs increases. The scale combinations $S_1 - S_2 - S_3$ generate improved OA over the $S_1 - S_2$ combination. It is also observed that the number of parameters in $S_1 - S_2 - S_3 - S_4$ is quite high compared to the $S_1 - S_2 - S_3$ combination with only minimal improvement in the classification performance. Therefore, in this work, the number of SS-CNNs is set to three with $i \in \{1, 2, 3\}$.

The network is developed using Keras with a TensorFlow backend. The model is trained with an Adam [123] optimizer with a learning rate of 10^{-3} on mini-batches of size 32 for 69080 iterations. An NVIDIA Tesla V100 GPU is used to perform the experiments.

The ablation study for the proposed LM-DFF method is given as follows.

Significance of the multi-scale feature extraction: The classification performance of the proposed LM-DFF method is effectively improved by the multi-scale feature extraction that can encode complementary

Table 2.8: The effect of number of SS-CNNs on the overall classification performance for the LM-DFF method on the UCSD database.

Scale combinations	OP (%)	OS (%)	OA	# parameters
S_1-S_2	93.45	94.08	95.41	45168
S_1-S_2-S_3	94.01	95.56	96.17	67363
S_1-S_2-S_3-S_4	94.68	95.01	96.25	89630

The bold entity indicates the selected scale combination for the LM-DFF method.

Table 2.9: The effect of multi-scale features on the overall classification performance for the LM-DFF method.

Database	Configuration	OP (%)	OS (%)	OA (%)
UCSD	S_1	89.34	93.07	93.00
	S_2	93.65	94.14	95.57
	S_3	92.29	94.60	95.07
	LM-DFF	**94.01**	**95.56**	**96.17**
NEH	S_1	98.51	98.50	98.49
	S_2	98.52	98.40	98.48
	S_3	98.56	98.41	98.49
	LM-DFF	**99.57**	**99.66**	**99.62**

The bold values show the performance of the proposed method.

cross-scale information. To substantiate this claim, we compare the classification performance of the proposed LM-DFF method with the single-scale frameworks. The single-scale frameworks employ the CNN backbone, an SS-CNN followed by GAP and FC layer with softmax activation to obtain the classification output. The network architecture of a single-scale network can be given as *CNN backbone* $\rightarrow S_i \rightarrow$ *GAP* $\rightarrow$ *FC+Softmax*. For this study, three such networks with different S_is, $i \in \{1, 2, 3\}$ are considered. The proposed LM-DFF also employs these three SS-CNNs for constructing the multi-scale network. Each of these single-scale networks is trained individually with the same hyper-parameter settings as the proposed LM-DFF method.

Table 2.9 shows the overall classification performance of the single-scale approaches and the proposed LM-DFF method on the UCSD and the NEH databases. As can be seen, the LM-DFF method achieves improved classification results compared to the single-scale frameworks. This verifies that the multi-scale framework improves the discrimination capability of the network by encoding disease-related features well.

Significance of the attention-based feature fusion: To verify the usefulness of channel attention in the feature fusion module, the results for the method without attention (w/o attention) are obtained. For this experiment, the channel attention operation before the global pooling is removed, keeping the rest of the network unchanged in the training stage. The experimental results are shown in Table 2.10. Compared to the model w/o attention, the proposed method attains an improvement of 1.4% and 1% in terms of the

Table 2.10: The effect of channel attention on the overall classification performance for the LM-DFF method.

Database	Configuration	OP (%)	OS (%)	OA (%)
UCSD	w/o Attention	93.78	94.28	95.66
	LM-DFF	**94.01**	**95.56**	**96.17**
NEH	w/o Attention	98.76	98.64	98.74
	LM-DFF	**99.57**	**99.66**	**99.62**

The bold values show the performance of the proposed method.

Table 2.11: The effect of joint loss optimization on the overall classification performance for the LM-DFF method.

Database	Configuration	OP (%)	OS (%)	OA (%)
UCSD	$\lambda = 0$	93.49	95.22	95.77
	$\lambda = 1$	**94.01**	**95.56**	**96.17**
NEH	$\lambda = 0$	98.87	98.79	98.87
	$\lambda = 1$	**99.57**	**99.66**	**99.62**

The bold values show the performance of the proposed method.

OP on the UCSD and the NEH datasets, respectively. Similar improvements are also observed for the OS and OA on both the databases. The impressive results verify that the attention module helps to mine discriminative features and improves classification performance.

Significance of the joint multi-loss optimization: The joint multi-loss optimization adopted in the LM-DFF method combines the fusion and the scale-specific losses as given in Eq. 2.20. The contribution of L_e to the total loss is dependent on the hyper-parameter λ. Hence, we evaluate the sensitivity of the model's performance to the variation of λ. Table 2.11 presents the classification results for the proposed model trained with $\lambda = 0$ (no contribution of L_e) and $\lambda = 1$ (complete contribution of L_e). Experimental results show that improved performance is obtained when λ is set to 1. This verifies the importance of combining the fusion and the scale-specific losses to effectively learn within and across scale discriminative features for reliable classification.

2.4.2 Results on the UCSD and NEH Databases

In this section, the performance comparison of the LM-DFF method with the previously proposed MDFF method is presented. Table 2.12 and 2.13 show the classification results of the LM-DFF and the MDFF methods on the UCSD and the NEH databases, respectively. As can be seen, the LM-DFF method provides improved classification results compared to the MDFF approach. Specifically, the proposed LM-DFF method achieves an OS, OP, OF1 and OA of 94.37%, 94.64%,94.43% and 96.03%, respectively, on the UCSD database. It can be observed from Table 2.13, that the proposed LM-DFF method attains

Table 2.12: Performance comparison of the MDFF and the LM-DFF methods on the UCSD database using 10-fold cross-validation.

Method	Class	SE (%)	PR (%)	F1 score (%)	OS/OP/OF1/OA (%)	AUC	Kappa
MDFF	CNV	95.91 ± 1.31	96.65 ± 0.92	96.27 ± 0.31	92.56 ± 1.06	0.954 ± 0.005	0.919 ± 0.005
	DME	91.56 ± 1.46	92.49 ± 1.96	92.00 ± 0.61	92.43 ± 1.23		
	Drusen	86.62 ± 3.59	85.23 ± 4.05	85.28 ± 0.96	92.48 ± 0.43		
	Normal	96.92 ± 1.01	95.78 ± 0.91	96.34 ± 0.25	94.57 ± 0.35		
LM-DFF	CNV	**97.33 ± 1.05**	**97.05 ± 1.19**	**97.18 ± 0.32**	**94.37 ± 1.16**	**0.965 ± 0.007**	**0.941 ± 0.006**
	DME	**93.22 ± 3.22**	**96.26 ± 2.17**	**94.65 ± 1.09**	**94.64 ± 0.9**		
	Drusen	**89.29 ± 3.59**	**87.73 ± 3.84**	**88.34 ± 1.27**	**94.43 ± 0.59**		
	Normal	**97.62 ± 1.11**	**97.49 ± 1.30**	**97.55 ± 0.49**	**96.03 ± 0.43**		

Table 2.13: Performance comparison of the MDFF and the LM-DFF methods on the NEH database using 5-fold cross-validation.

Method	Class	SE (%)	PR (%)	F1 score (%)	OS/OP/OF1 (%)	OA (%)	AUC	Kappa
MDFF	AMD	97.45 ± 0.85	97.23 ± 0.77	97.34 ± 0.69	97.51 ± 0.46	97.57 ± 0.43	0.982 ± 0. 004	0.963 ± 0. 007
	DME	97.05 ± 0.76	97.59 ± 1.08	97.33 ± 0.83	97.56 ± 0.45			
	Normal	98.03 ± 0.34	97.85 ± 0.34	97.84 ± 0.37	97.54 ± 0.45			
LM-DFF	AMD	**99.62 ± 0.27**	**99.54 ± 0.17**	**99.58 ± 0.16**	**99.58 ± 0.23**	**99.60 ± 0.21**	**0.997 ± 0. 002**	**0.994 ± 0. 003**
	DME	**99.45 ± 0.59**	**99.45 ± 0.38**	**99.45 ± 0.35**	**99.59 ± 0.20**			
	Normal	**99.68 ± 0.22**	**99.75 ± 0.41**	**99.71 ± 0.20**	**99.60 ± 0.22**			

a near accurate classification performance with an OS, OP, OF1, and OA of 99.58%, 99.59%, 99.60% and 99.60%, respectively. The improved performance of the LM-DFF method can be attributed to the learnable multi-scale feature extraction and the joint loss optimization that can effectively capture within and cross-scale discriminative feature representations for efficient classification.

Figure 2.8 compares the OA and the average test run times of the proposed and the existing methods on the UCSD and the NEH databases. The measurements for the average run times are taken from a system with 16 GB RAM and Quadro K600 graphics processor. As can be seen, the HOG+SVM method provides the fastest prediction among all the methods. However, the low OA of the method makes it unacceptable for clinical applications. The VGG16 and InceptionV3 have seemingly high run times with moderate improvements in OA. On the contrary, both the proposed methods, i.e., the MDFF and the LM-DFF, have significantly low average run times. It can be seen that the LM-DFF approach achieves a significant improvement in the OA with only a slight increase in the test run time than the MDFF method. The acceptable test run time and the impressive OA of the proposed LM-DFF method make it best suitable for the preliminary screening of retinal diseases in eye clinics.

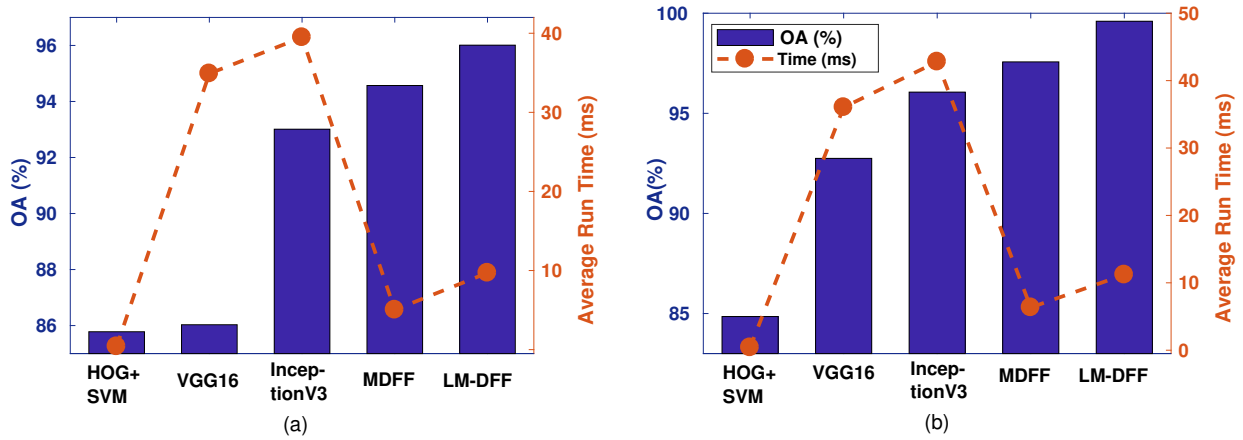


Figure 2.8: Comparison of the OA and the average run time for the methods on (a) UCSD and (b) NEH databases.

2.4.3 Visualization of the Learned Features

In this section, we qualitatively evaluate the proposed methods by visualizing the discrimination ability of the features. The feature visualization for the existing methods like the HOG+SVM and the VGG16 are also presented for comparison. Specifically, for the VGG16, the MDFF and the LM-DFF methods, the learned features prior to the output layer are extracted and dimensionality reduced using the t-Distributed Stochastic Neighbor Embedding (t-SNE) to facilitate visualization [128]. Similar dimension reduction is also applied for the HOG features. Figure 2.9 shows the feature visualization of the reduced features for the different methods. As can be seen, the features for the different classes are highly overlapping for the HOG and the VGG16, providing little or no discrimination. The feature separability is better for the MDFF and LM-DFF methods compared to the HOG and the VGG16. It is evident from Figure 2.9 (d), that the LM-DFF method provides enhanced discrimination as the between-class distance is seemingly large compared to that in the MDFF method. This implies that the scale-specific feature extraction using the dilated convolutions and the attention based feature fusion have resulted in high-level discriminative features for classification.

We also present the Grad-CAM visualizations [129] for the LM-DFF method to provide a transparent diagnostic basis. The Grad-CAM highlights the image regions that the network focuses on during the classification. Figure 2.10 shows the Grad-CAM visualizations as heat-maps for the CNV, DME and drusen affected B-scans. It can be observed that the method focuses precisely on the regions having pathological manifestations to make a diagnosis prediction. This implies that the model has learned domain-specific clinical knowledge from the OCT images and can provide a reliable diagnostic decision.

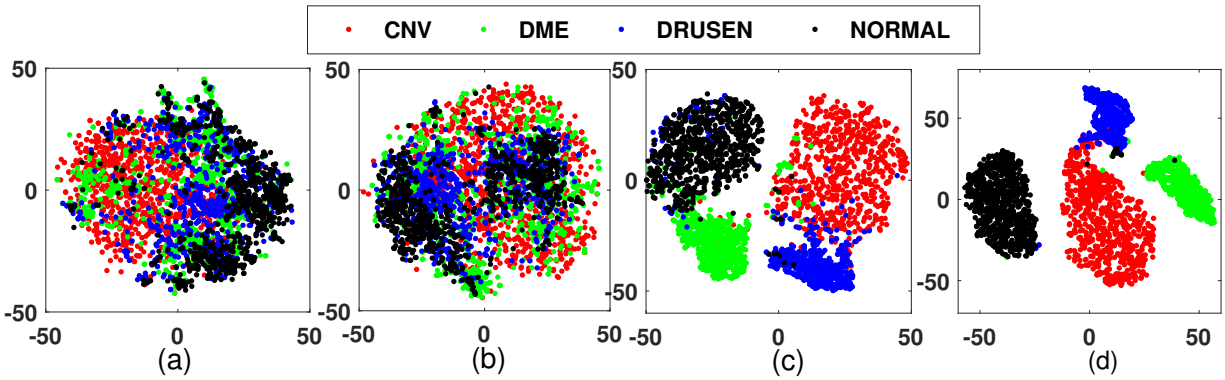


Figure 2.9: Visualization of the features: (a) HOG, (b) VGG16, (c) MDFF and (d) LM-DFF.

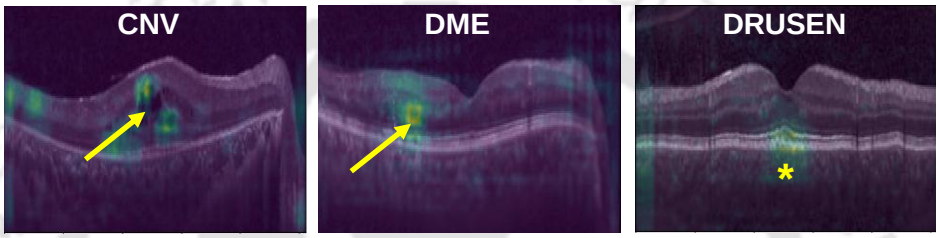


Figure 2.10: Grad-CAM visualizations of the OCT images for the LM-DFF method. The arrows and the asterisk indicate the location of retinal fluid deposits and drusens, respectively.

2.5 Summary

The automated OCT B-scan classification methods can be useful for preliminary diagnosis in eye clinics, large-scale screening programs and remote health care applications. To this end, in this chapter, we presented two new multi-scale approaches for the automated classification of retinal disorders from OCT B-scans. The main contribution presented in this chapter is the exploration of the multi-scale feature representations for the efficient classification of the OCT images. Specifically, fixed multi-scale image decomposition using MSSP and learnable multi-scale feature representations using dilated convolutions were experimented with to obtain improved classification performance. It was observed that the learnable multi-scale feature representations using the dilated convolution provided better performance. In addition, both the methods did not rely on any denoising, layer segmentation, or lesion detection step during classification. The superior prediction ability and low test run time of the proposed LM-DFF method make it highly suitable for reliable and fast diagnosis of retinal diseases from OCT B-scans.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

3

Denoising and Super-resolution of OCT B-scans for Improved Diagnosis of Intermediate AMD

Contents

In clinical practice, the presence of the drusens and the abnormality in the retinal layers [23] are observed for the diagnosis of AMD at early and intermediate stages. Specifically, the size of the drusens [130], the discontinuity in the external limiting membrane (ELM) [131] and the thickness of the inner nuclear layer (INL) [132] are assessed (see Figure 3.1). However, the sparse appearance of the small-sized drusens, limited visualization of ELM and the narrow width of the INL present inevitable challenges in the quantification of the clinical parameters [133]. HR B-scan images have the ability to represent the subtle clinical features well. The HR OCT images require more samples of the target, which is achieved by acquiring more A-scans per B-scan. This increases the image acquisition time, thereby causing motion artefacts due to involuntary patient movements [13]. The artefacts result from the involuntary eye motion for fixation, head movements and body jitters caused by the cardiorespiratory system. The artefacts can cause inaccurate clinical interpretation of the OCT images [32, 134]. The inherent speckle noise also degrades the image quality and affects its diagnostic utility [12,135]. Therefore, in the clinical setting, noisy LR images are acquired from the OCT machines and computer-aided algorithms are utilized to recover the denoised HR counterparts [14]. The reconstructed HR images can then be used for diagnosis.

The existing SR methods rely on supervised frameworks that require aligned pairs of noisy LR and clean HR images for training. It is well known that creating a generalizable supervised model using DL requires massive amounts of data [98]. However, the differing standards of health care industries and medical data privacy laws hinder the acquisition of large-scale paired LR-HR OCT images for efficient supervised learning [136]. This problem becomes even more severe for AMD because of its asymptomatic nature, where most cases are reported at the advanced stages [137]. Therefore, unsupervised frameworks that do not rely on one-to-one alignment between the LR and the HR images during training are appropriate for the denoising and SR of the OCT images. The option of using unpaired images provides flexibility to leverage the already available OCT data for better generalization of the denoising and SR performance.

Recently, the GANs [111] have been popularly used for the SR task. The adversarial learning of the GANs has been successful in reconstructing realistic and visually appealing images [104,138–140]. In this chapter, we present an unsupervised GAN based SR framework that can reliably and swiftly reconstruct the clean HR images without the requirement of aligned LR-HR pairs.

The rest of the chapter is organized as follows. The existing GAN based SR framework is discussed in section 3.1. The proposed method is presented in section 3.2. The database and network details are described in section 3.3. The experimental results are analyzed in section 3.4. Finally, the chapter is summarized in section 3.5.

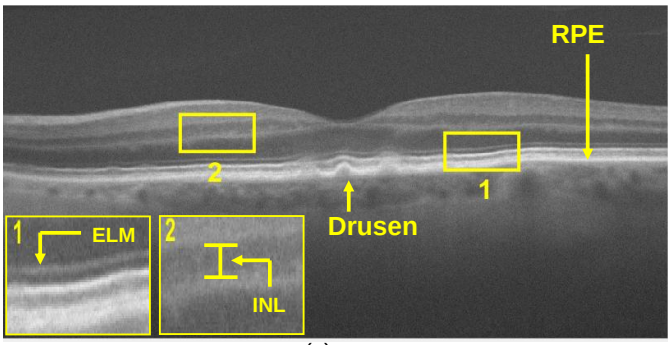


Figure 3.1: OCT image highlighting retinal structures crucial for AMD diagnosis.

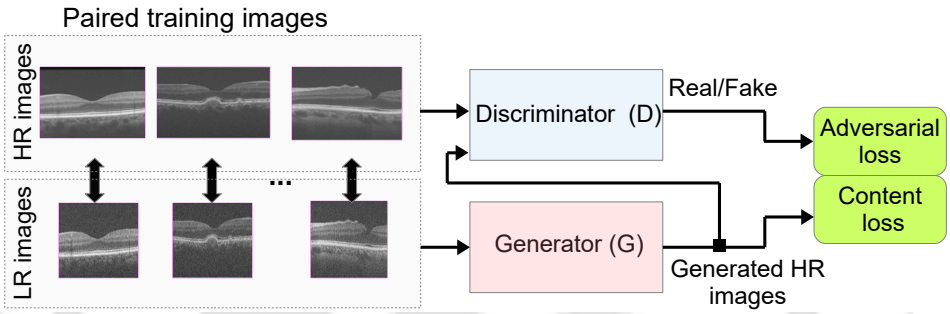


Figure 3.2: Block diagram of the SRGAN.

3.1 Generative Adversarial Network for SR

The GANs are an approach for generative modeling using deep neural networks. Generative modeling involves learning patterns from the input data in such a way that it can generate new examples plausibly from the input data distribution [111]. The GANs employ two neural networks, namely, the generator and the discriminator for the task. The generator is trained to generate new examples, and the discriminator tries to classify the examples as either real (from the data distribution) or fake (generated by the generator). The two models are trained together in a zero-sum game until the generator is trained to generate plausible examples. The adversarial learning of the networks enables the generation of images with high perceptual quality. GANs have been successfully explored in the supervised SR frameworks like the super-resolution GAN (SRGAN) [104] and its variants such as enhanced SRGAN (ESRGAN) [138], ESRGAN+ [141] and attentional SRGAN (A-SRGAN) [142]. A few other supervised GAN based SR methods have been proposed like the tempoGAN [143], the fine-grained attention generative adversarial network (FASRGAN) [144], the domain prior GAN [145], the face complexion and SR GAN (FCSR-GAN) [146] and the capsule GAN [147].

Figure 3.2 shows the block diagram of the SRGAN [104]. The framework contains two neural network

components, i.e., the generator (G) and the discriminator (D). G takes the LR images as input and generates the corresponding HR images as output. D, on the other hand, assigns labels (real/fake) to these generated images based on their closeness to the real HR image distribution. G is trained to generate realistic HR images to fool D into believing that the generated images belong to the real HR image distribution. D is trained to discriminate the generated images from the real HR images efficiently. This adversarial learning is performed by minimizing the adversarial and the content losses, as shown in Figure 3.2. The adversarial loss is the negative of the logarithm of the probability that the generated image belongs to the real HR distribution [104]. The content loss is obtained as the feature-based mean square error (MSE) between the generated image and the corresponding true HR image. It should be noted that the applicability of the MSE based loss function is restricted to situations where aligned pairs of training LR-HR images are available. Therefore, these frameworks are highly suitable for the SR of natural images, where a large number of paired images can be easily acquired. However, as already discussed, the acquisition of a sufficiently large number of LR-HR pairs of medical images may be challenging. Hence, the GANs have not been well explored for the SR of the OCT images.

In the following section, we propose an unsupervised GAN based SR framework that eliminates the need for paired LR-HR images for training and builds a generalizable network with already available OCT databases. This framework is inspired by the work of Zhu *et al.* [148], where a well generalizable architecture is presented for unsupervised image translation. The details of the proposed framework are discussed below.

3.2 The Proposed Unsupervised GAN for the Simultaneous Denoising and SR of the OCT Images

Figure 3.3 shows the pipeline of the proposed method. As can be seen, the framework contains two generators (*Generator_HR* (G_{HR}) and *Generator_LR* (G_{LR})) and two discriminators (*Discriminator_HR* (D_{HR}) and *Discriminator_LR* (D_{LR})). Let us denote the set of unpaired training noisy LR and clean HR OCT images as $\{\boldsymbol{L_i}\}_{i=1}^{M}$, $\boldsymbol{L_i} \in X_{lr}$ and $\{\boldsymbol{H_i}\}_{i=1}^{M}$, $\boldsymbol{H_i} \in X_{hr}$, respectively. Here, $\boldsymbol{X_{lr}}$ and $\boldsymbol{X_{hr}}$ are the set of all training noisy LR and clean HR images, respectively. We denote the noisy LR and clean HR data distribution as $\boldsymbol{L} \sim p_{OCT}(\boldsymbol{L})$ and $\boldsymbol{H} \sim p_{OCT}(\boldsymbol{H})$, respectively. As illustrated in Figure 3.3, the G_{HR} takes the noisy LR OCT images as input and maps them to the clean HR output images ($G_{HR} : \boldsymbol{X_{lr}} \rightarrow \boldsymbol{X_{hr}}$). Similar to the discriminator of the SRGAN, the D_{HR} also discriminates these generated clean HR images from the real HR OCT images by assigning labels of fake or real. The other set of generator and discrim-

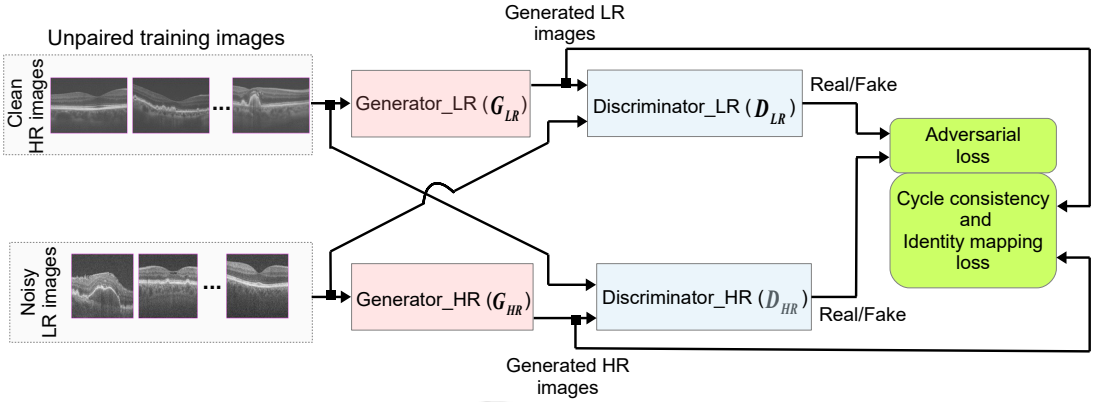


Figure 3.3: Block diagram of the proposed framework for the simultaneous denoising and SR of OCT images.

inator, i.e., G_{LR} and D_{LR} present in the proposed framework, are not directly involved in the generation of HR images. However, these components, in association with the constrained minimization objectives, indirectly help G_{HR} learn appropriate mapping function in the absence of paired training examples. The G_{LR} performs the inverse mapping of converting the clean HR OCT images to the noisy LR domain ($G_{LR} : \boldsymbol{X_{hr}} \rightarrow \boldsymbol{X_{lr}}$). The D_{LR} differentiates the real noisy LR OCT images from the generated LR OCT images $\{G_{LR}(\boldsymbol{H})\}$. Similar to the functionality of G in the SRGAN, G_{HR} and G_{LR} in the proposed framework are learned to generate realistic images close to the true distribution. D_{HR} and D_{LR} are learned to distinguish the generated images from the real images efficiently. As discussed in Section 3.1, the MSE based content loss function is only suitable for learning networks with paired training images. Therefore, to train an unsupervised network with unpaired images, we use the specialized loss functions, i.e., the cycle consistency loss and the identity mapping prior, along with the adversarial loss for reliable reconstruction of the HR images.

As discussed previously, the measurement of the thickness of the retinal layers is crucial to AMD diagnosis. Therefore, it is essential to accurately reconstruct the retinal layers for a reliable diagnosis. It is also important to ensure the faithful reconstruction of the sparsely distributed drusens to quantify the AMD progression. Therefore, we define the cycle consistency loss to preserve the spatial clinical details between the input and the generated images. This loss is inspired by the work of Zhu *et al.* [148] and is given as

$$\begin{aligned} Lg_{cyc} = E_{\boldsymbol{L} \in p_{OCT}(\boldsymbol{L})} \Big[\|G_{LR}(G_{HR}(\boldsymbol{L})) - \boldsymbol{L}\|_2 \Big] + \\ E_{\boldsymbol{H} \in p_{OCT}(\boldsymbol{H})} \Big[\|G_{HR}(G_{LR}(\boldsymbol{H})) - \boldsymbol{H}\|_2 \Big]. \end{aligned} \tag{3.1}$$

Here, $p_{OCT}(\boldsymbol{L})$ and $p_{OCT}(\boldsymbol{H})$ represent the distribution of the real noisy LR and the clean HR OCT im-

ages. As can be seen, the loss function contains two terms. In the first term, the LR counterparts of the generated HR images ($G_{HR}(\boldsymbol{L})$) are obtained through G_{LR}. The dissimilarity between the obtained LR images and the real noisy LR images is computed through the L_2 norm. Similarly, the dissimilarity between the real HR images and the HR counterparts of the generated LR images ($G_{LR}(\boldsymbol{H})$) are computed in the second term of Lg_{cyc}. The minimization of the total dissimilarity ensures that the minute clinical details of the real input images ($\boldsymbol{L}$ or $\boldsymbol{H}$) are well preserved in the generated images.

In clinical practice, the texture and color (intensity) of the retinal layers play a vital role in diagnosis [131]. Therefore, it is essential to preserve the color and texture details in the generated HR OCT images. It can be achieved through identity mapping [149], where the generators should produce the same image if an input image of the target domain is provided. We define identity loss as

$$\begin{aligned} Lg_{idt} = E_{\boldsymbol{H} \in p_{OCT}(\boldsymbol{H})} \left[\| G_{HR}(\boldsymbol{H}_{\downarrow}) - \boldsymbol{H} \|_1 \right] + \\ E_{\boldsymbol{L} \in p_{OCT}(\boldsymbol{L})} \left[\| G_{LR}(\boldsymbol{L}_{\uparrow}) - \boldsymbol{L} \|_1 \right] \end{aligned} \tag{3.2}$$

where $\boldsymbol{H}_{\downarrow}$ denotes the down-sampled version of the HR image $\boldsymbol{H}$ and $\boldsymbol{L}_{\uparrow}$ is the up-sampled version of the LR image $\boldsymbol{L}$.

Finally, the GAN adversarial objective is employed to ensure that the images generated by G_{HR} and G_{LR} are close to the real data distribution. The adversarial objective for G_{HR} in the HR domain can be expressed as

$$Lg_{adv}^{HR} = E_{\boldsymbol{L} \in p_{OCT}(\boldsymbol{L})} \left[\| D_{HR}(G_{HR}(\boldsymbol{L})) - 1 \|_2 \right]. \tag{3.3}$$

Similarly, the adversarial objective for the D_{HR} in the HR domain is given as

$$\begin{aligned} Ld_{adv}^{HR} = E_{\boldsymbol{H} \in p_{OCT}(\boldsymbol{H})} \left[\| D_{HR}(\boldsymbol{H}) - 1 \|_2 \right] + \\ E_{\boldsymbol{L} \in p_{OCT}(\boldsymbol{L})} \left[\| D_{HR}(G_{HR}(\boldsymbol{L})) \|_2 \right]. \end{aligned} \tag{3.4}$$

The minimization of the adversarial generator loss (Lg_{adv}^{HR}) ensures that the G_{HR} generates realistic HR images to fool D_{HR} into believing that the generated images are real. Similarly, the task of adversarial discriminator loss (Ld_{adv}^{HR}) is to update the D_{HR} parameters such that it effectively classifies the real HR images from the generated images. This adversarial learning renders the G_{HR} helpful in providing realistic and diagnostically relevant OCT images.

We define the adversarial losses for G_{LR} and D_{LR} in a similar manner and show them in Eq. 3.5 and

3.6, respectively.

$$Lg_{adv}^{LR} = E_{\boldsymbol{H} \in p_{OCT}(\boldsymbol{H})} \left[\| D_{LR}(G_{LR}(\boldsymbol{H})) - 1 \|_2 \right] \tag{3.5}$$

$$\begin{aligned} Ld_{adv}^{LR} = E_{\boldsymbol{L} \in p_{OCT}(\boldsymbol{L})} \left[\| D_{LR}(\boldsymbol{L}) - 1 \|_2 \right] + \\ E_{\boldsymbol{H} \in p_{OCT}(\boldsymbol{H})} \left[\| D_{LR}(G_{LR}(\boldsymbol{H})) \|_2 \right] \end{aligned} \tag{3.6}$$

Finally, the combined generator objective is defined as

$$L_g = \underset{G_{HR}, G_{LR}}{\arg\min}\ Lg_{adv}^{HR} + Lg_{adv}^{LR} + k_1 Lg_{cyc} + k_2 Lg_{idt} \tag{3.7}$$

where k_1 and k_2 are weights to the loss functions.

During training, the generators and the discriminators are learned one at a time. First, the parameters of D_{HR} and D_{LR} are updated by minimizing Ld_{adv}^{HR} and Ld_{adv}^{LR} respectively. Then, keeping the parameters of G_{LR} and D_{HR} fixed, the G_{HR} is updated by minimizing L_g. Finally, G_{LR} is updated, keeping G_{HR} and D_{LR} fixed. The details of the network architecture and the experimental setup for training are discussed in the following section. At the end of the training process, G_{HR} is engaged in generating the clean HR OCT images from the noisy LR OCT images. The other components of the network, i.e., D_{HR}, G_{LR} and D_{LR} are discarded as they are only required to assist G_{HR} in better training.

3.3 Clinical Database and Network Parameters

In this section, the details of the databases used for evaluation and the network architecture of the proposed method are presented.

3.3.1 Clinical Database Description

The proposed method is evaluated on three publicly available databases, namely, the Duke University SR (DUSR) database [14], the Duke University denoising (DUD) database [150] and the Duke University Intermediate AMD (DUIA) database [46]. The DUSR and the DUD databases contain 28 and 17 paired noisy and clean OCT B-scan images, respectively. The images of these databases are acquired from subjects with and without non-neovascular AMD using the Bioptigen, Inc. OCT imaging system. Two different scanning protocols are used for obtaining the noisy LR and the clean HR images. Firstly, the

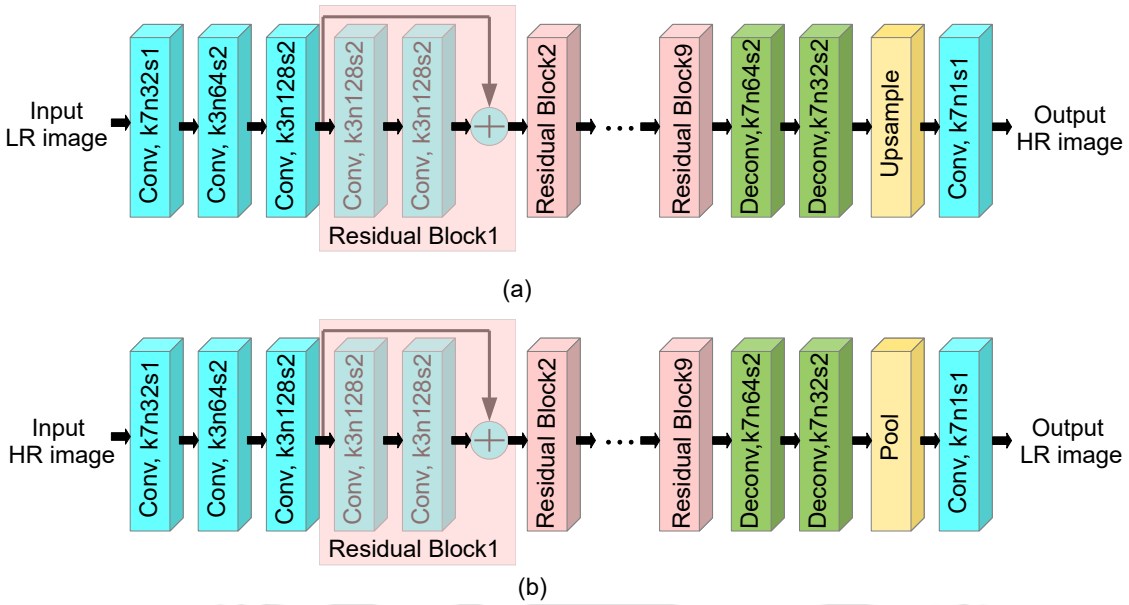


Figure 3.4: Network architectural details for (a) G_{HR} and (b) G_{LR}.

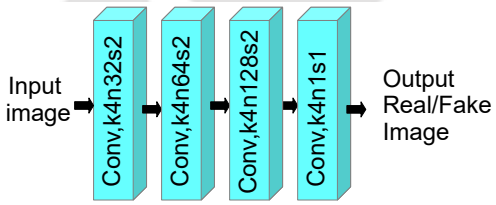


Figure 3.5: Architectural details for D_{HR} and D_{LR} networks.

volume scans are captured with an axial resolution of 4.5μm from a $6.6 \times 6.6\ mm^2$ retinal area centered around the fovea. Each of the volumes contains 100 B-scans with 1000 A-scans per B-scan. The central foveal B-scan of the volumes is sub-sampled to obtain the LR images. The second set of scans is obtained by capturing 40 azimuthally repeated B-scans centered at the fovea. The repeated B-scans are registered and averaged using the ImageJ software to obtain clean HR images [14, 150].

The DUIA dataset is the world's largest SD-OCT dataset of intermediate AMD and control subjects [46]. It contains OCT volumes from 269 AMD and 115 control subjects. Subjects included in this database are between 50 and 85 years of age, exhibiting intermediate AMD with large drusen (>125 mm) in both eyes or large drusen in 1 eligible eye and advanced AMD in the fellow eye, with no history of vitreoretinal surgery or ophthalmologic disease that might affect acuity in either eye. The SD-OCT imaging system by Bioptigen, Inc. is used to capture the OCT volumes with 1000 A-scans per B-scan and 100 B-scans per volume.

The unpaired clean HR and noisy LR images for training the network are created from these three

databases. A total of 27 clean HR images are obtained from the DUSR (10 HR images) and the DUD (17 HR images) databases. For quantitative comparison, the HR images are rescaled to 450×900 pixels. 4000 central foveal noisy B-scan images are extracted from the volumes of the DUIA dataset. The LR images are generated by down-sampling the noisy images by a factor of 2 and 4, thereby rendering low image resolution of size 450×450 and 450×225 pixels, respectively. We can observe that the height of the images is preserved during down-sampling. It is because the height of the OCT images is a property of the low coherence imaging light source [19]. The resolution of the OCT images is dependent on the number of A-scans used for obtaining a B-scan [151], which is given by the width of the images.

For the purpose of training the network, random crops of size 300×600 pixels are extracted from the region of interest of the clean HR images. A total of 4000 random crops are extracted from the 27 clean HR images for training. Similarly, crops of size 300×300 pixels and 300×150 pixels are extracted from the noisy LR images, which can be magnified by factors of 2 and 4, respectively, during the SR process. The LR dataset for training contains 2000 crops each from the AMD and the healthy control classes. The test set for evaluating the SR performance consists of 17 images from the DUSR database that have not been used for training [1].

3.3.2 Network Details

The network architectures for the generators, G_{HR} and G_{LR} are shown in Figure 3.4. The G_{HR} (see Figure 3.4 (a)) takes the noisy LR OCT images as input and outputs the clean HR OCT images. The network contains three "Conv" blocks that perform convolution operation, nine residual blocks for proper gradient flow, two "Deconv" blocks to perform fractionally strided convolution [152] and an "Upsample" block for obtaining the desired resolution at the output. The "Deconv" layers convert the coarse outputs of the convolutional layers to dense pixels through nonlinear upsampling [153]. The residual blocks introduce identity residual or skip connections between layers. These skip connections alleviate the vanishing gradient problem and enable effective learning of the deeper networks [75]. As shown in Figure 3.4, the variables k, n and s represent the kernel size, number of filters and stride for the convolution and strided convolution operations. Each of the "Conv" (except the last) and "Deconv" blocks are followed by instance normalization [154] and ReLU activation layers. The final "Conv" block has a $tanh$ activation function. The "Upsample" block transforms the feature maps to the desired resolution of the target HR domain. The G_{LR} has a similar architecture as the G_{HR} with variations only in the input and the "Upsample" block. The

[1]There are 18 images of the DUSR that have not been used for training. One image is not used for evaluation as it has registration error.

G_{LR} (see Figure 3.4 (b)) takes the HR OCT images as input and generates the LR images. Similarly, the "Upsample" block in G_{HR} is replaced by a "Pool" block in G_{LR}. The "Pool" block utilizes average pooling to transform the intermediate feature maps to the resolution of the LR images.

The discriminators D_{HR} and D_{LR} have the same architecture as shown in Figure 3.5. The first three "Conv" blocks are followed by instance normalization layers [154] and leakyReLU [155] activation functions with a negative slope of 0.2. A sigmoid activation function is applied to the final "Conv" block. The method is implemented using Keras with the Tensorflow backend. The model uses Adam [123] optimizer with a learning rate of 2×10^{-4} for the generators and 2×10^{-5} for the discriminators. For stable training, the discriminators are updated once in three iterations. The model is trained on mini-batches of size 1 for 68,000 iterations. The training is performed on an NVIDIA Tesla V100 GPU.

3.4 Experimental Results

In this section, we qualitatively and quantitatively compare the reconstruction performance of the proposed and the existing methods. We also analyse the effectiveness of the proposed denoising and SR method in improving the automated diagnosis performance of intermediate AMD and healthy subjects.

3.4.1 Results for HR Reconstruction

The effectiveness of the proposed method is verified by qualitative and quantitative comparisons with the state-of-the-art simultaneous denoising and SR methods. The shared sparse representation models like the SBSDI [14], the SSR [33] and the NWSR [105] are used for comparison. The recently proposed low-rank approximation and second-order tensor-based total variation (LRSOTTV) [108] OCT reconstruction method is also included in the study. The proposed method is also compared with the popular DL based SRGAN [104] approach. The qualitative comparison is performed by visually inspecting the reconstruction quality at different regions of the OCT images essential for AMD diagnosis. The quantitative analysis is performed by evaluating the methods using metrics such as the contrast to noise ratio (CNR), the peak signal to noise ratio (PSNR) and the image sharpness (IS).

The CNR measures the contrast of the clinically significant foreground regions with respect to the background regions. Mathematically, the CNR can be represented as

$$CNR = \frac{1}{m}\sum_{i=1}^{m}\left[10\, log_{10}\left(\frac{|\mu_i - \mu_b|}{\sqrt{\sigma_i^2 + \sigma_b^2}}\right)\right] \tag{3.8}$$

where m is the number of region of interest (ROI) image patches used for CNR computation. μ_i and σ_i are the mean and the standard deviation of selected foreground patches (selected from the regions containing the retinal layers). μ_b and σ_b are the mean and the standard deviation of the background ROI. Here $m = 5$ is selected with each patch having a size of 50×100.

The PSNR computes the similarity between the generated and the true HR images. The mathematical definition of the PSNR is given as

$$PSNR = 10\, log_{10}\left(\frac{(max(I_o))^2}{\frac{1}{M_1 \times M_2}\sum_i \sum_j (I_r(i,j) - I_o(i,j))^2}\right) \tag{3.9}$$

where I_o and I_r are the original and the reconstructed HR images, respectively, each of size $M_1 \times M_2$.

The IS measures the sharpness of the denoised and super-resolved images and is given as

$$IS = \frac{1}{M_1 \times M_2}\sum_{i=1}^{M_1}\sum_{j=1}^{M_2} |I_r * S_v| \tag{3.10}$$

where S_v is the Sobel mask in vertical direction and * represents the convolution operation.

3.4.1.1 Qualitative Analysis

In this section, we qualitatively evaluate the OCT images by visualizing the reconstruction quality of the different retinal micro-structures. Figure 3.6 shows the visual reconstruction results of an OCT image for a magnification factor of 2. The highlighted portions in the images (yellow squares) show the key regions for AMD diagnosis. The zoomed versions of these regions are shown on the right-hand side of the respective images for better visualization. Figure 3.6 (a) and (h) show the noisy LR and the original clean HR images, respectively. Figure 3.6 (b)-(g) show the reconstructed images using the existing and the proposed methods. The first highlighted portion (see Figure 3.6 (h1)) aims at analyzing the reconstruction quality of the ELM (shown by the solid white arrow). It is well discussed in the medical literature that the status of ELM is a predictor of visual acuity in AMD [131]. It can be observed that the ELM is completely blurred out by the SBSDI (Figure 3.6 (b1)), the SSR (Figure 3.6 (c1)), the NWSR (Figure 3.6 (d1)) and the LRSOTTV (Figure 3.6 (e1)) methods. The SRGAN (Figure 3.6 (f1)) has been partially successful in reconstructing the ELM. However, the reconstruction quality is not enough to unambiguously measure the layer thickness. As shown in Figure 3.6 (g1), the proposed method has generated the ELM accurately with its boundaries intact.

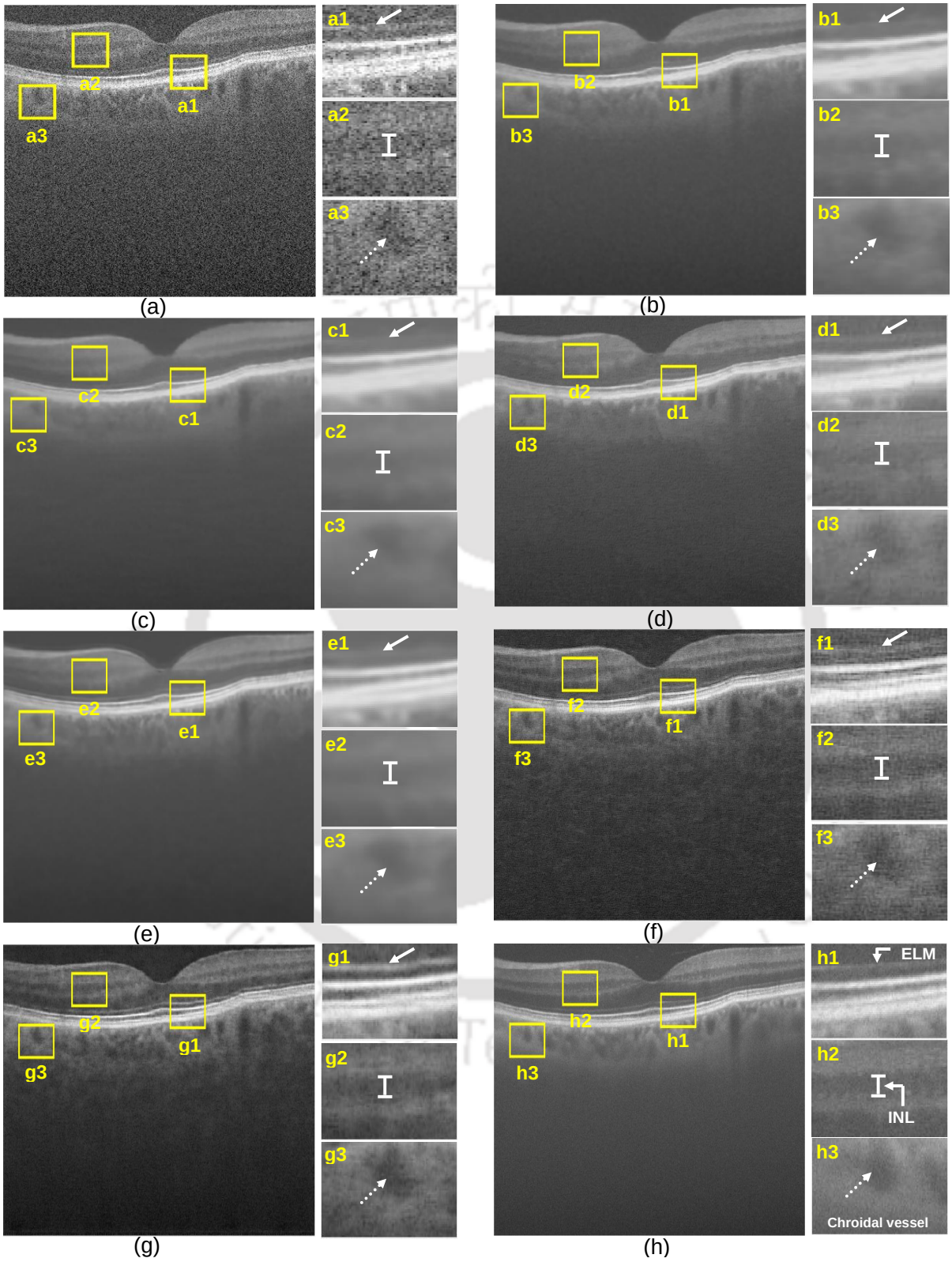


Figure 3.6: Visual comparison of the denoising and SR performance for a magnification factor of 2. (a) LR image, reconstructed images using (b) SBSDI, (c) SSR, (d) NWSR, (e) LRSOTTV, (f) SRGAN, (g) proposed method and (h) original HR image.

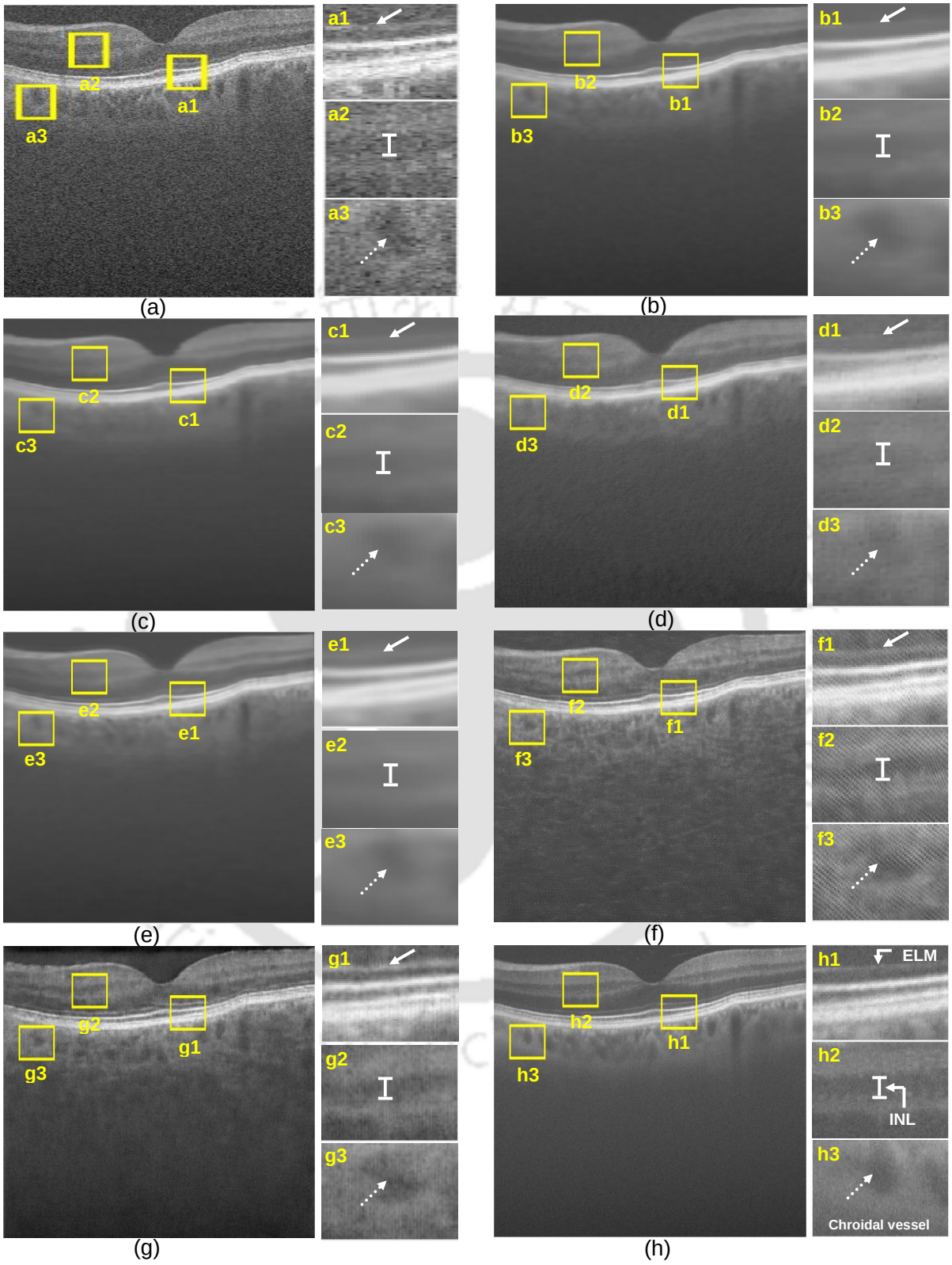


Figure 3.7: Visual comparison of the denoising and SR performance for a magnification factor of 4. (a) LR image, reconstructed images using (b) SBSDI, (c) SSR, (d) NWSR, (e) LRSOTTV, (f) SRGAN, (g) proposed method and (h) original HR image.

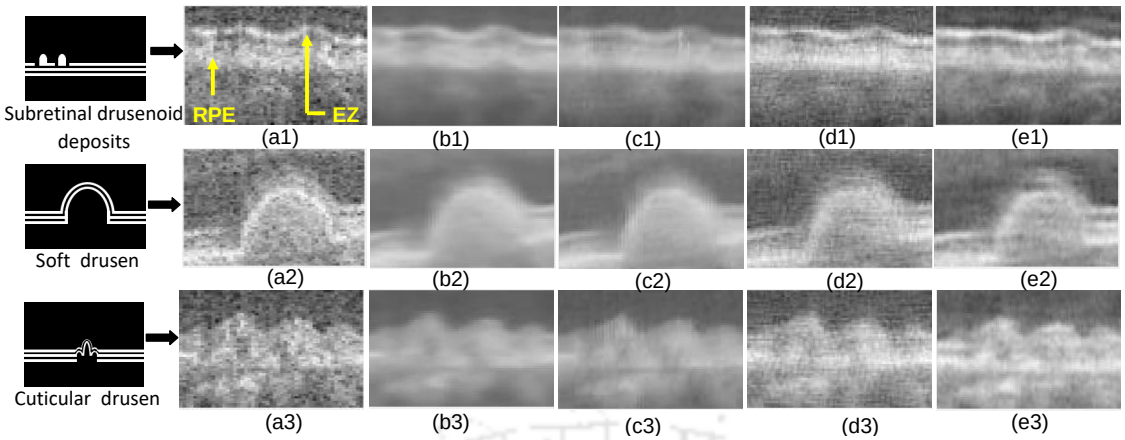


Figure 3.8: Visual comparison of the SR methods for drusen reconstruction for a magnification factor of 2. (a1-a3) LR images, images reconstructed by (b1-b3) SBSDI, (c1-c3) NWSR, (d1-d3) SRGAN and (e1-e3) proposed method.

The second highlighted region focuses on the reconstruction of the INL (see Figure 3.6 (h2)). Clinically, the INL thickening is the reason for atrophy of the RPE [132]. Hence, the measurement of INL thickness is an essential parameter for AMD diagnosis. The INL thickness can be accurately measured when it has sharp boundaries with the neighboring layers. It can be seen that the INL is hardly distinguishable from the neighboring layers in the images reconstructed by the SBSDI (Figure 3.6 (b2)), the SSR (Figure 3.6 (c2)), the NWSR (Figure 3.6 (d2)) and the LRSOTTV (Figure 3.6 (e2)) methods. The SRGAN (Figure 3.6 (f2)) reconstructs the INL better than the SBSDI and the NWSR methods. However, the granularity in the reconstructed image may lead to errors in thickness measurements. The reconstruction of the proposed method (Figure 3.6 (g2)) is much smoother than the SRGAN, while the boundaries are still preserved.

The third highlighted region aims at evaluating the reconstruction quality of the choroid. The choroid is the vascular layer that contains blood vessels and choriocapillaris for the nourishment of the retina [156]. Though the choroid has limited visualization in OCT images, it is an essential diagnostic factor for assessing the retinal health [156]. Figure 3.6 (h3) shows the zoomed version of a choroidal vessel (see dashed white arrow). It can be observed that the SBSDI (Figure 3.6 (b3)), the SSR (Figure 3.6 (c3)), the NWSR (Figure 3.6 (d3)), the LRSOTTV (Figure 3.6 (e3)) and the SRGAN (Figure 3.6 (f3)) have not reconstructed the vessel clearly. The proposed method (Figure 3.6 (g3)) has reconstructed the vessel with a clear demarcation of its boundaries. Thus, the visual results verify that the proposed method reconstructs the clinical details well. The reconstructed images for a magnification factor of 4 are provided for visualization in Figure 3.7. It can be observed that the reconstruction performance similar to that of the magnification factor of 2 is obtained.

We also analyze the reconstruction quality of the drusens for the proposed and the existing methods.

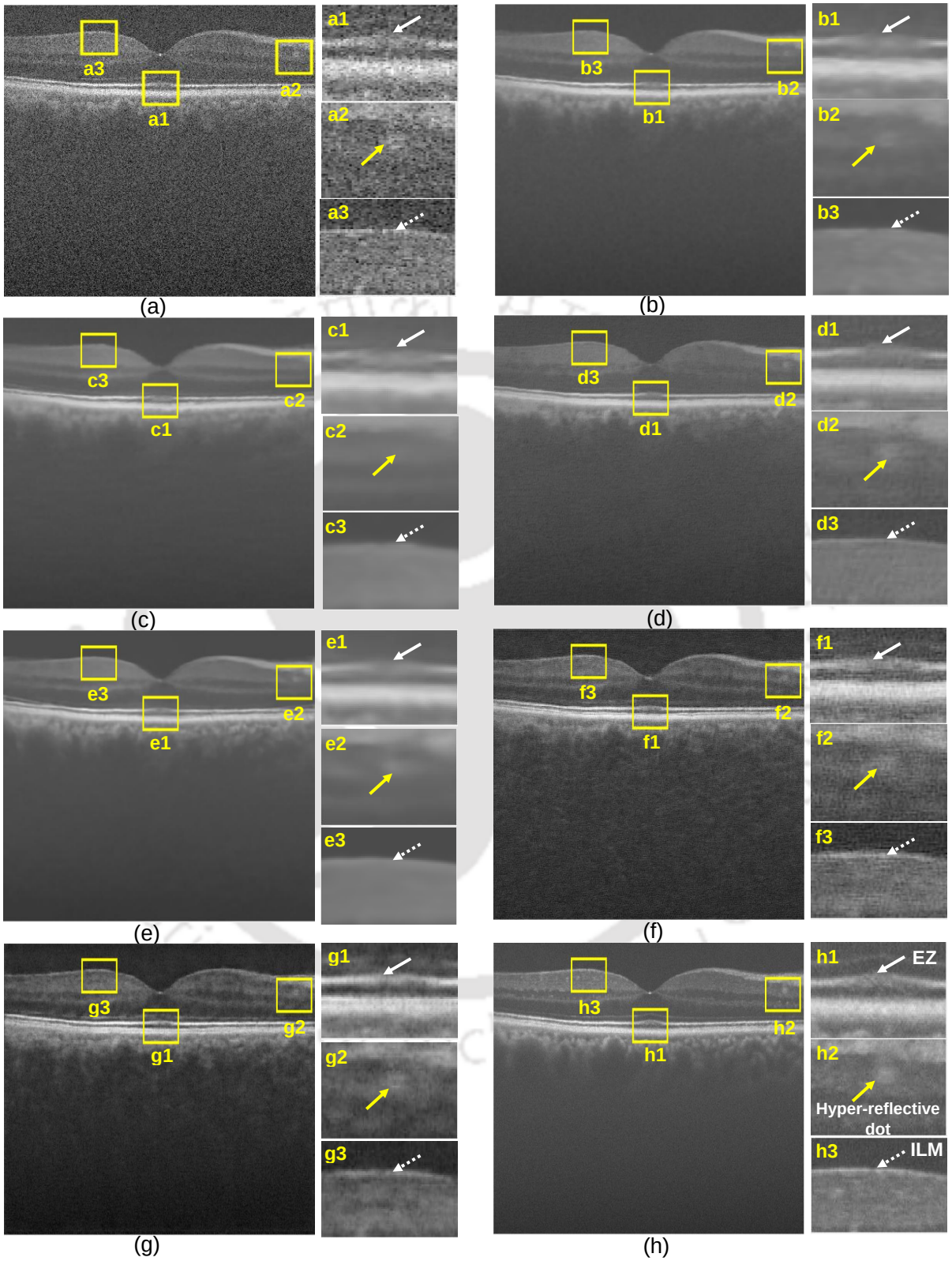


Figure 3.9: Visual comparison of the denoising and SR performance for a magnification factor of 2. (a) LR image, reconstructed images using (b) SBSDI, (c) SSR, (d) NWSR, (e) LRSOTTV, (f) SRGAN, (g) proposed method and (h) original HR image.

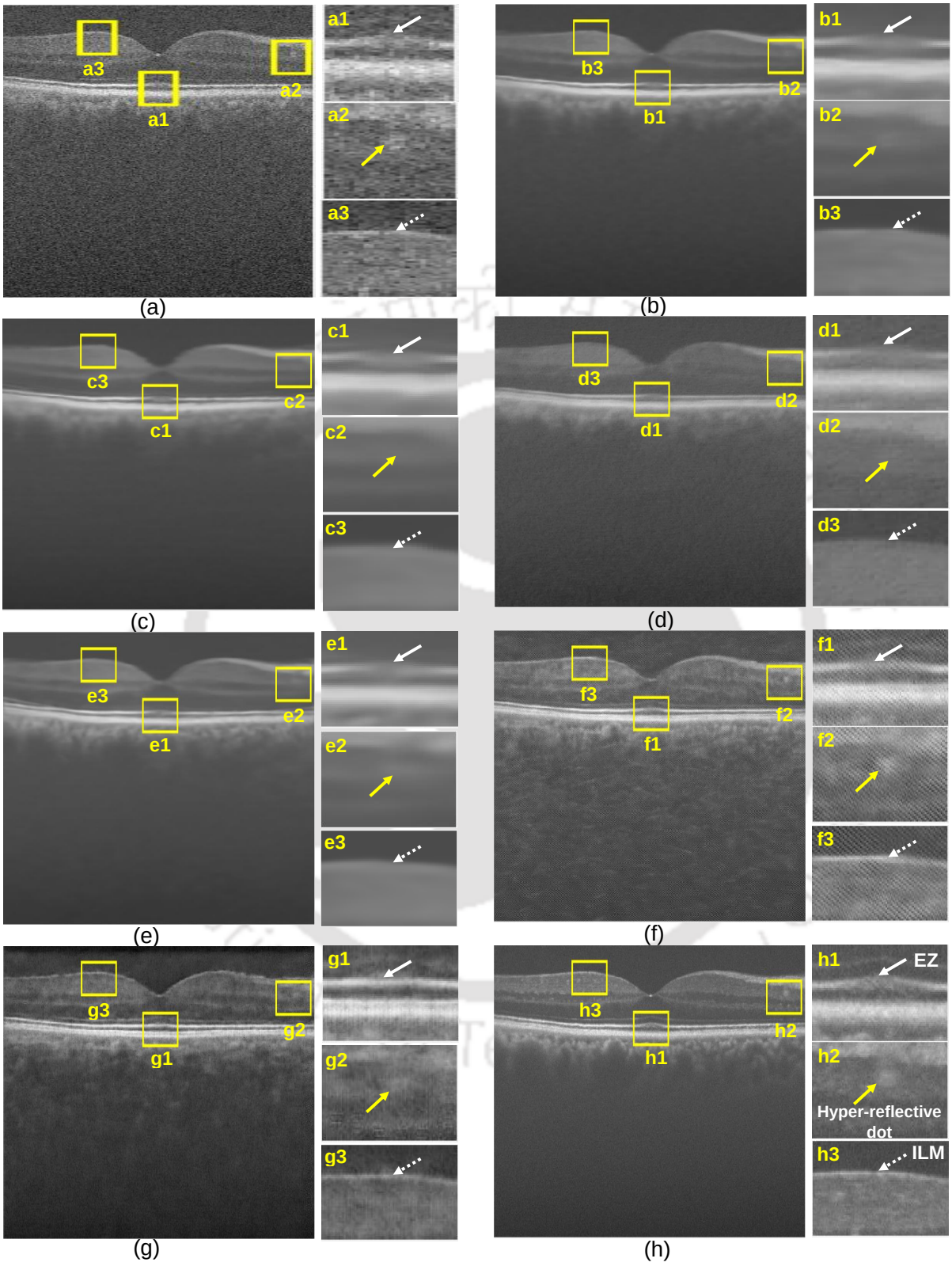


Figure 3.10: Visual comparison of the denoising and SR performance for a magnification factor of 4. (a) LR image, reconstructed images using (b) SBSDI, (c) SSR, (d) NWSR, (e) LRSOTTV, (f) SRGAN, (g) proposed method and (h) original HR image.

Figure 3.8 shows the visual reconstruction results for the different types of drusens, i.e., the subretinal drusenoid deposits, the soft drusen and the cuticular drusen [26]. Figure 3.8 (a1-a3) show the noisy LR drusen images. These image patches are taken from the AMD volumes of the DUIA database. As this dataset does not have clean HR counterparts, we only show the comparative reconstruction performance with the existing methods. Figure 3.8 (b1-b3) and (c1-c3) show the zoomed versions of reconstructed drusens using the SBSDI and the NWSR methods, respectively. It can be observed that the overall reconstruction is quite blurry compared to the SRGAN (Figure 3.8 (d1-d3)) and the proposed method (Figure 3.8 (e1-e3)). The reason for this is the excessive smoothing that takes place during the patch-based reconstruction and the denoising stages of the methods. This results in loss of the sharpness of the edges in the retinal layers, which can affect diagnostic measurements. For example, the analysis of the ellipsoid zone (EZ) and the RPE (see Figure 3.8 (a1)) are vital for the management of AMD [157]. Both manual and automated segmentation of these layers become challenging and erroneous from blurred images as generated from the SBSDI and the NWSR methods. Although the SRGAN generates sharp EZ and RPE boundaries (Figure 3.8 (d1-d3)), the granular patterns throughout indicate that the images are not denoised well. The drusens are well reconstructed by the proposed method (Figure 3.8 (e1-e3)). The proposed method does not rely on the neighboring slices for denoising and also does not involve a patch-based reconstruction strategy. The deep residual convolutional architectures of the G_{HR} in association with the adversarial learning help preserve the sharp edges of the retinal layers.

Figure 3.9 shows the visual reconstruction performance (for a magnification factor of 2) highlighting few more diagnostic details. The LR and the original HR images are shown in Figure 3.9 (a) and (h), respectively. The reconstructed images by the different methods are shown in Figure 3.9 (b)-(g). The zoomed versions of some of the important diagnostic regions like the EZ (shown by solid white arrow, Figure 3.9 (h1)), a hyperreflective dot (yellow arrow, Figure 3.9 (h2)) and the inner limiting membrane (ILM) (dotted white arrow, Figure 3.9 (h3)) are presented. It can be observed that the EZ is not well reconstructed by the SBSDI (Figure 3.9 (b1)), the SSR (Figure 3.9 (c1)), the NWSR (Figure 3.9 (d1)), the LRSOTTV (Figure 3.9 (e1)) and the SRGAN (Figure 3.9 (f1)) methods. It can also be seen that the SBSDI (Figure 3.9 (b2) and (b3)), the SSR (Figure 3.9 (c2) and (c3)), the NWSR (Figure 3.9 (d2) and (d3)) and the LRSOTTV (Figure 3.9 (e2) and (e3)) methods completely fail to reconstruct the hyperreflective dot and the ILM. Although the SRGAN (Figure 3.9 (f2) and (f3)) reconstructed these details better than the SBSDI and the NWSR, the noise present in the images affect the visual clarity. The proposed method, on the other hand, generates better quality images without losing the diagnostic information. The reconstructed images for a magnification factor of 4 are presented in Figure 3.10 for visualization.

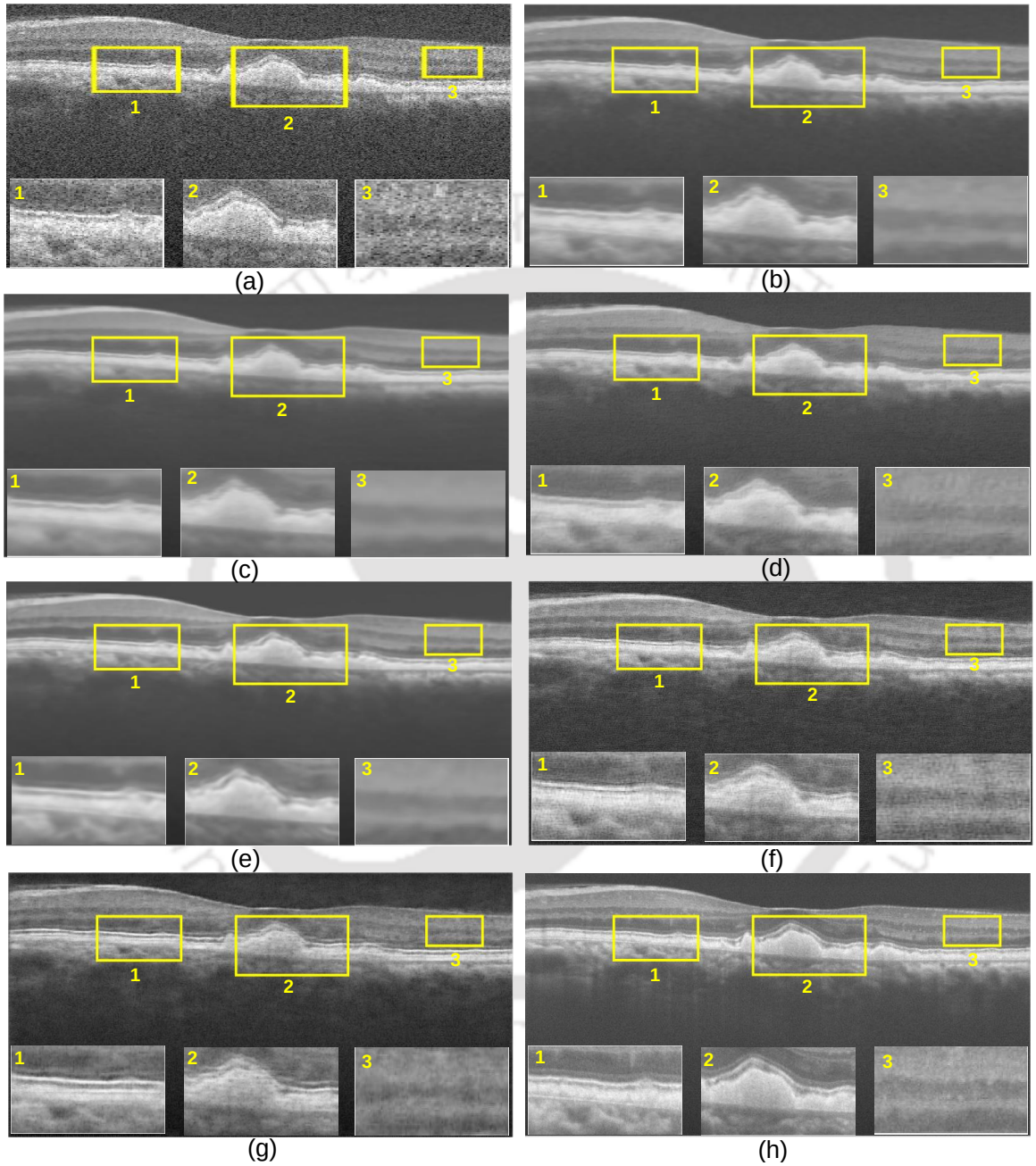


Figure 3.11: Visual comparison of the denoising and SR performance for a magnification factor of 2.(a) LR image, reconstructed images using (b) SBSDI, (c) SSR, (d) NWSR, (e) LRSOTTV, (f) SRGAN, (g) proposed method and (h) original HR image.

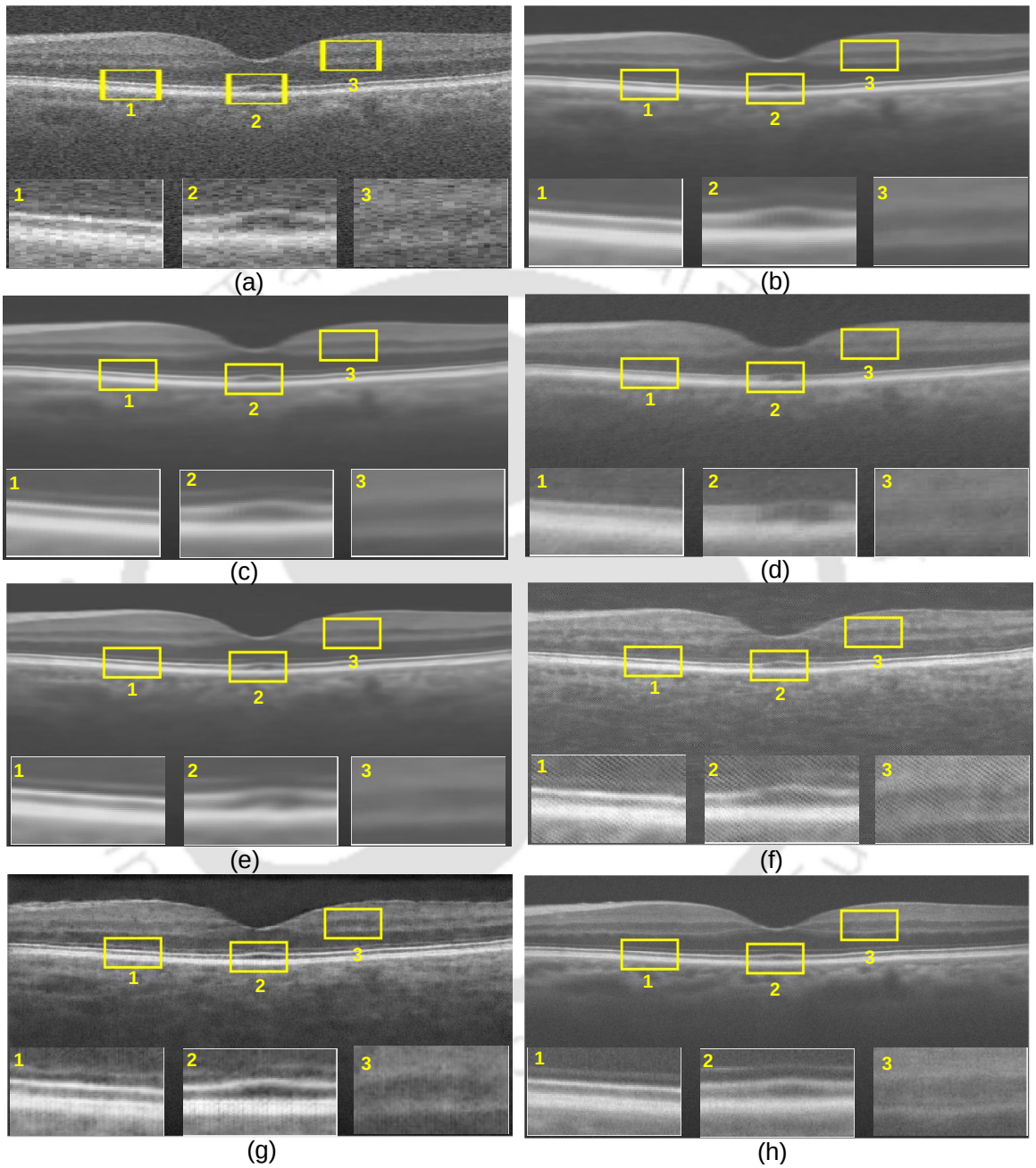


Figure 3.12: Visual comparison of the denoising and SR performance for a magnification factor of 4. (a) LR image, reconstructed images using (b) SBSDI, (c) SSR, (d) NWSR, (e) LRSOTTV, (f) SRGAN, (g) proposed method and (h) original HR image.

Figures 3.11 and 3.12 show more visual results of the reconstructed OCT images by the proposed approach and the existing methods for magnification factors of 2 and 4, respectively. It can be summarized from the visual results that the over-smoothness greatly influences the edge details of the OCT images reconstructed by the existing methods. These methods also lose a few key diagnostically significant features during reconstruction. The proposed method suppresses noise, efficiently reconstructs the edge details of the retinal layers and the drusen morphologies compared to the existing methods.

3.4.1.2 Quantitative Analysis

In this section, we quantitatively analyse the denoising and SR performance of the proposed and the existing methods for the magnification factors of 2 and 4. The details of the test dataset and the evaluation measures are presented in section 3.3.1 and 3.4.1, respectively. The mean and the standard deviation values of the performance measures are presented for the test dataset in Table 3.1. It can be seen from the table that the proposed method outperforms the existing methods in terms of CNR, achieving an average CNR of 4.69 and 4.64 for magnification factors of 2 and 4, respectively. This improvement in the CNR values validates that the foreground clinically significant regions (retinal layers) are well contrasted from the clinically non-significant background regions. Similarly, a minimum improvement of 4% is observed in terms of the PSNR values for the proposed method. The high PSNR values are indicative of the closeness in characteristics of the reconstructed images with the original HR images. It can be seen from Table 3.1 that the SBSDI, the SSR, the NWSR and the LRSOTTV methods have a significantly low value of IS. As discussed before, the low IS values can be attributed to the inefficient patch-based approach used for reconstruction. It is also observed that the SRGAN achieves the highest value for IS. These elevated levels are partially biased by the granularity present in the images rather than the well-preserved edges. Compared to the existing methods, the proposed method shows promising performance in terms of most of the measures.

Figure 3.13 compares the PSNR and the average test run times of the proposed and the existing methods for magnification factors of 2 and 4. These measurements are taken from a system with 16 GB RAM and Quadro K600 graphics processor. It can be observed that the sparse representation based methods, i.e., the SBSDI, the SSR and the NWSR have lower PSNR values compared to the proposed method. These methods also take an increasingly high amount of time to generate the HR OCT images. This is because of the serial processing of image patches during the reconstruction process. The large duration for the HR image generation limits the applicability of these methods for the reconstruction of the 3D-OCT volumes that require the SR of many B-scans for the analysis. The SRGAN has a low test run

Table 3.1: Performance comparison of the proposed and the existing simultaneous denoising and SR methods on the test dataset.

Method	Approach	Magnification	CNR	PSNR (dB)	IS
Shared sparse representation	SBSDI [14]	2×	4.52 ± 0.74	31.12 ± 2.06	10.16 ± 0.92
		4×	4.53 ± 0.74	34.73 ± 3.16	9.78 ± 0.83
	SSR [33]	2×	4.61 ± 0.75	31.07 ± 2.06	8.82 ± 0.82
		4×	4.61 ± 0.77	34.64 ± 3.09	8.8 ± 0.83
	NWSR [105]	2×	4.51 ± 0.73	32.31 ± 1.73	17.03 ± 0.83
		4×	4.52 ± 0.70	32.2 ± 1.68	19.18 ± 0.72
Low Rank Approximation	LRSOTTV [108]	2×	4.49 ± 0.77	33.99 ± 2.07	9.81 ± 0.12
		4×	4.49 ± 0.78	33.98 ± 2.08	9.57 ± 1.16
Deep learning	SRGAN [104]	2×	3.86 ± 0.71	35.10 ± 3.02	40.22 ± 1.33
		4×	3.57 ±0.62	26.51± 1.02	56.77 ± 3.54
	Proposed method	2×	**4.69 ± 0.68**	**39.15 ± 3.54**	**26.96 ± 1.17**
		4×	**4.64 ± 0.72**	**35.67 ± 2.68**	**27.83 ± 1.35**

The bold values show the performance of the proposed method.

time, but the inferior PSNR values provided by the method (especially for the magnification factor of 4), affect its reliability for clinical applications. On the contrary, the proposed method generates good quality images with high PSNR values in less than a second. Therefore, the proposed method can conveniently aid ophthalmologists for a quick and reliable diagnosis.

In clinical practice, the measurements obtained from the 3D OCT volumes have proven helpful for reliable diagnosis of progressive retinal diseases [158]. Considering the utility of these volumes in diagnostic ophthalmology, we have extended the proposed B-scan based denoising and SR method for the HR reconstruction of the OCT volumes. The proposed model is employed to generate the HR version for each of the B-scan images in the OCT volume to obtain the HR reconstruction for the entire volume. Table 3.2 shows the volume reconstruction results in terms of CNR, IS and average test time for the SRGAN and the proposed method for 169 LR AMD volumes from the DUIA database. The performance could not be verified in terms of the PSNR as there are no reference HR ground-truth images available for the dataset. The SRGAN is considered for the performance comparison considering its quick image reconstruction attribute compared to the other state-of-the-art methods. The SBSDI, the SSR and the NWSR methods have not been included in this study as they take a large amount of time to generate the reconstructed images for the entire volume. It can be observed from Table 3.2, that the proposed method exhibits better performance in terms of CNR and also average test time compared to the SRGAN, thereby making the proposed method suitable for reliable reconstruction of the OCT volumes for improved diagnosis.

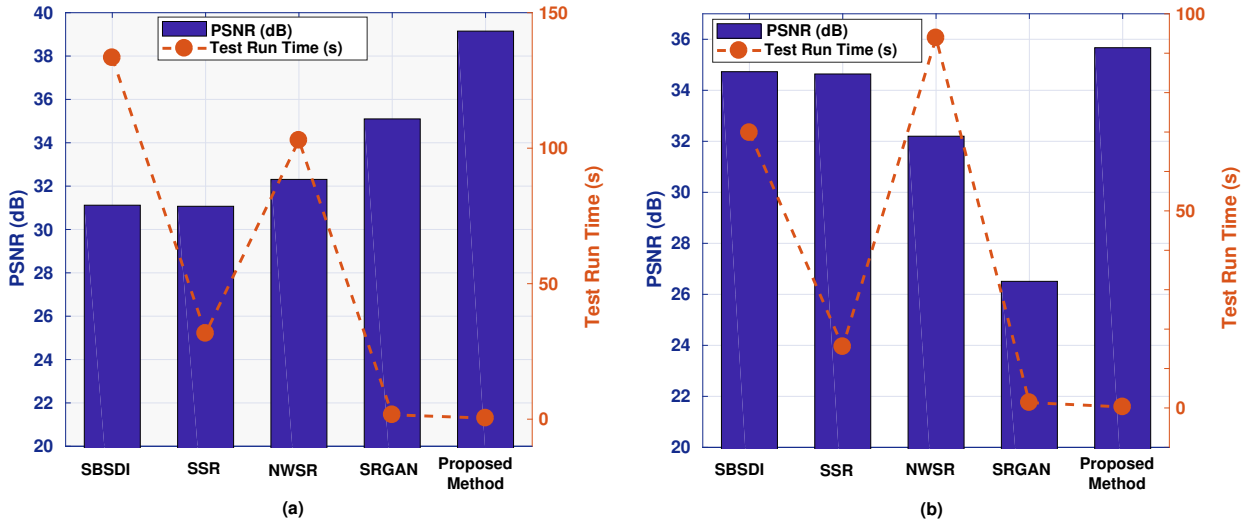


Figure 3.13: Comparison of the PSNR and the average run time of the methods for a magnification factor of (a) 2 and (b) 4.

Table 3.2: Quantitative performance analysis of the proposed and the existing denoising and SR methods for OCT volumes.

Approach	Magnification	CNR	IS	Test Run Time (s)
SRGAN [104]	2×	3.36 ± 0.72	42.66 ± 1.79	74.02 ± 0.02
	4×	2.71 ± 0.78	55.15 ± 2.10	50.83 ± 1.08
Proposed Method	2×	**4.02 ± 0.68**	**26.74 ± 0.88**	**12.89 ± 0.02**
	4×	**3.93 ± 0.76**	**26.68 ± 0.74**	**7.86 ± 0.03**

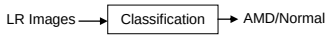
The bold values show the performance of the proposed method.

3.4.2 Evaluation of the Proposed Method for AMD Diagnosis

The usefulness of the proposed method is demonstrated by using the reconstructed images for the automated classification of intermediate AMD and healthy subjects. Specifically, the proposed denoising and SR method is used as a pre-processing block to generate clean HR images from noisy LR images. The generated images are then provided to an automated classification method to obtain the diagnosis (see the first column of Table 3.3). The classification performance with and without the pre-processing block is analyzed for the automated AMD diagnosis. Without the pre-processing block, the classification method takes the noisy LR images as input to generate the diagnosis decision (see the first column of Table 3.3).

A CNN based classification framework is adapted from the literature [35] for the task. The model has an architecture of $C_{(5,3)}BRP_{(2)} \rightarrow C_{(3,3)}BRP_{(2)} \rightarrow C_{(3,3)}BRP_{(2)} \rightarrow C_{(3,3)}BRP_{(2)} \rightarrow C_{(3,3)}BRP_{(2)} \rightarrow Flatten \rightarrow D_{(15)} \rightarrow D_{(2)} \rightarrow Softmax$, where $C_{(k,n)}BRP_{(p)}$ is a sequence of convolutional layer followed

Table 3.3: Performance comparison of automated AMD classification using the LR and the generated HR images.

Configuration	Denoising and SR Method	Image Size (Input to Classifier)	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUC	Kappa
	-	450×225 [LR ($\downarrow 4$)]	88.57	93.16	90.87	0.93	0.82
	-	450×450 [LR ($\downarrow 2$)]	93.78	94.48	94.13	0.95	0.88
	SRGAN [SR ($\uparrow 4$)]	450×900	92.15	93.01	92.58	0.94	0.85
	SRGAN [SR ($\uparrow 2$)]	450×900	96.19	94.41	95.30	0.96	0.90
	Proposed Method [SR ($\uparrow 4$)]	450×900	**94.71**	**94.17**	**94.44**	**0.96**	**0.88**
	Proposed Method [SR ($\uparrow 2$)]	450×900	**96.89**	**96.19**	**96.54**	**0.97**	**0.93**

The bold values show the performance of the proposed method.

by BN, ReLU activation and max-pooling layers. Here, k represents the kernel size. n and p represent the number of convolution filters and the pool size, respectively. $D_{(d)}$ is a dense layer with d output neurons. To train the classification model, a total of 12,645 B-scans are extracted from the 384 OCT volumes of AMD and control subjects from the DUIA database. The images from the dataset are down-sampled to create noisy LR images. Let us denote LR ($\downarrow$ 2) and LR ($\downarrow$ 4) as the LR OCT images obtained by downsampling the noisy images by factors of 2 and 4, respectively. Firstly, the noisy LR images are provided to the classifier. Training of the CNN model is performed on 80% of the images and testing is performed on the rest. The models are trained on mini-batches of size 64 for 50 epochs using the Adam [123] optimizer with a learning rate of 10^{-3}. The classification results are presented in Table 3.3. It can be seen that, when the LR images are directly used for classification, the performance degrades by nearly 5% as the resolution is reduced from 450×450 to 450×225. This is because the pathological symptoms of AMD cannot be captured well in the LR images. This verifies that the clean HR images are essential for the improved diagnosis of AMD.

In the second case, the images are first denoised and super-resolved using the pre-processing block and then applied to the classification framework. These images are denoted as SR ($\uparrow$ 2) and SR ($\uparrow$ 4) for magnification factors of 2 and 4, respectively. The classification results for the denoised and super-resolved images using the SRGAN and the proposed method are presented in Table 3.3. The methods like the SBSDI, the SSR, the NWSR and the LRSOTTV have not been considered in this study, as they require high computational time to generate the clean HR images. It is observed that the classification sensitivity improves by 3% over the LR images (LR ($\downarrow 2$)) when the images reconstructed by the proposed method (SR($\uparrow 2$)) are used for classification. The method also attains high AUC and kappa values of 0.97 and 0.93, respectively. Significant improvements in classification performance are observed when the reconstructed images by the proposed method (for a magnification factor of 4) are used for classification, achieving sensitivity, specificity and accuracy of 94.71%, 94.17% and 94.44%, respectively. These

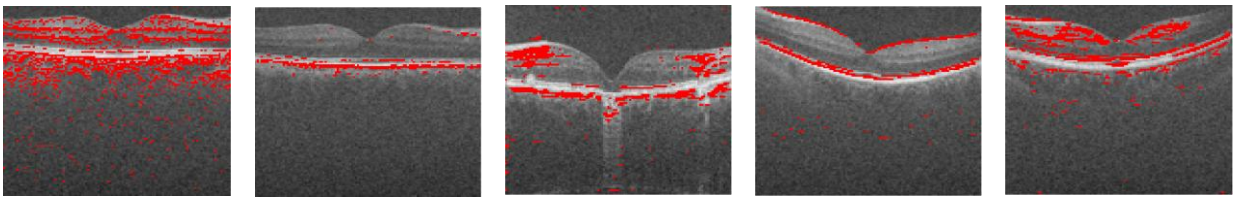


Figure 3.14: Examples of feature maps at different layers of the G_{HR} network.

improved classification results indicate that the proposed method efficiently reconstructed the LR images that aided in the correct diagnosis. It can also be observed from Table 3.3 that the classification performance for a magnification factor of 2 for the SRGAN and the proposed method are mostly similar. However, for the magnification factor of 4, the proposed method outperforms the SRGAN. This validates that the proposed method reconstructs the images better than the SRGAN for higher magnification factors. The suboptimal performance of SRGAN can be attributed to its inferior quality image reconstruction, as discussed in section 3.4.

3.4.3 Visualization of the Learned Features

To understand how the generator is approaching reconstruction, the feature maps of G_{HR} at different layers of the network are observed. These feature maps are obtained by passing the LR OCT images through G_{HR} and extracting the neural activations at the output of the desired network layers. Figure 3.14 shows some of these neural activations (shown in red) at different layers of the network overlaid on the LR OCT images. Due to the convolutional nature of the network, the input LR images and the feature maps may not have the same spatial dimensions. Therefore, the input LR OCT images are rescaled to the size of the feature maps to obtain the images in Figure 3.14. The highlighted areas in the figure are the regions that have higher activations. It can be observed that the feature maps highlight the different layers of the retina during the SR process. This shows that the model focuses on the clinically significant portions of the image during the reconstruction. The model transparency provided by the feature visualization can be viewed as an additional advantage of the proposed method as such observations cannot be made for the shared sparse representation or the low-rank approximation models.

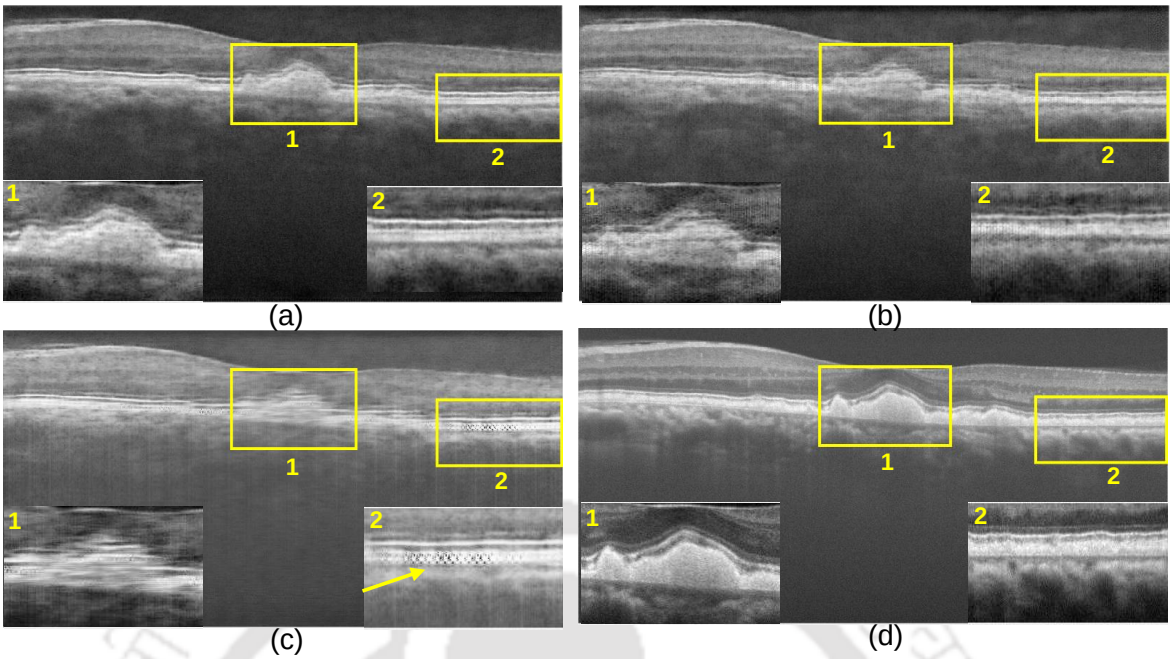


Figure 3.15: Reconstructed images using the proposed method: (a) at 50% sampling (b) at 25% sampling, (c) at 13% sampling and (d) true HR image.

3.4.4 Marginal Sampling Rate for the Proposed Method

This chapter presents and discusses reconstruction results for a magnification factor of 2 (50% missing data) and 4 (25% missing data). In order to identify the marginal sampling rate at which the method fails to reconstruct the clinical details well, we learned different models with training noisy LR images obtained by uniform sub-sampling with 50%, 25% and 13% data. Figure 3.15 (a)-(c) show the visual reconstruction result of OCT images reconstructed with 50%, 75% and 87% missing data. Figure 3.15 (d) shows the true HR image for reference. The key regions used in this analysis are highlighted in yellow and the zoomed versions of these regions are presented at the bottom of each reconstructed image. It can be seen that the reconstruction of the drusen (see the zoomed region marked as 1 in Fig. 3.15) has been compromised when the noisy LR image with 13% sampling is used. It is also observed that artefacts (shown by the yellow arrow in Figure 3.15) are introduced in the RPE layer at this sampling rate which is not present on the reconstructed images with 50% and 25% sampling. Similar behavior was observed for other images of the test set as well. Hence, based on the above observation, it can be said that the marginal sampling rate for the method is 25% below which diagnostic distortions can be introduced during reconstruction.

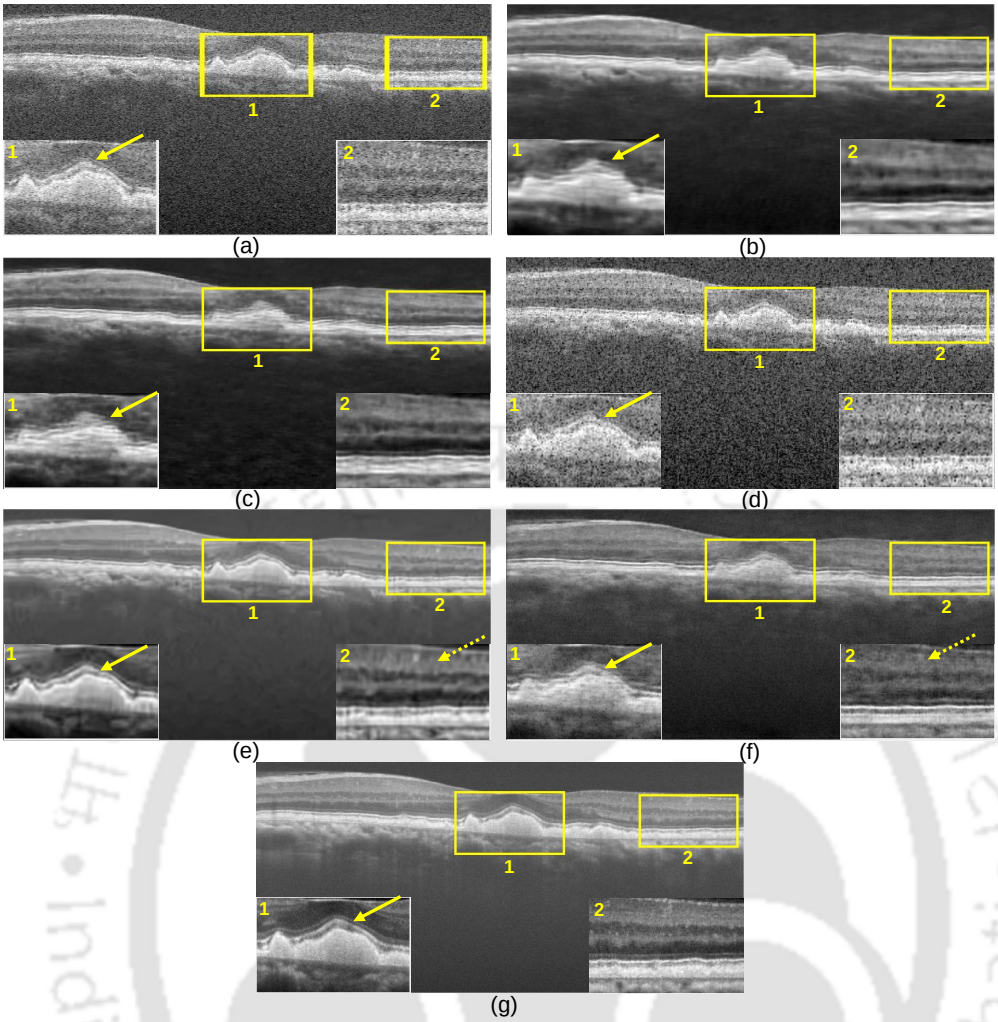


Figure 3.16: Visual comparison of the denoising and SR performance for a magnification factor of 2. (a) LR image, reconstructed images using (b) EDSR, (c) WDSR, (d) ZSSR, (e) BM3D+ZSSR, (f) proposed method and (g) original HR image.

3.4.5 Comparison of the Proposed Method with Computer Vision Models for SR

The proposed simultaneous denoising and SR method is compared with three computer vision SR models, namely, the enhanced deep residual networks for single image super-resolution (EDSR) [159], the wide activation for efficient and accurate image SR (WDSR) [160] and the zero-shot super-resolution using deep internal learning (ZSSR) [161]. The EDSR and the WDSR are supervised in nature and the ZSSR is unsupervised.

The supervised models are trained on paired LR and HR image patches of size 150×300 and 300×600, respectively. The image patches were extracted from the DUSR [14] and the DUD databases [150]. The models are trained using 12,000 images patches with a batch size of 8 for 50 epochs. The unsupervised ZSSR model does not rely on prior training. It exploits the internal recurrence of information in a single

Table 3.4: Performance comparison of the proposed and the existing computer-vision SR methods.

Method	Approach	CNR	PSNR (dB)	IS
EDSR [159]	Supervised	4.13	36.0	18.40
WDSR [160]	Supervised	4.31	37.0	20.14
ZSSR [161]	Unsupervised	3.28	28.29	41.02
BM3D [40] + ZSSR [161]	Unsupervised	4.32	33.72	12.30
Proposed Method	Unsupervised	4.69	39.15	26.96

image and trains a small image-specific CNN at test time on examples extracted solely from the input image itself. The details of the image-specific CNN and the training details can be found in [161]. As the ZSSR is entirely unsupervised and learns from the input test image, it may not be sufficient to denoise the OCT image during SR. Therefore, we have also performed an experiment where we first denoise the OCT image using BM3D [40] and then super-resolve the denoised image using ZSSR.

Fig. 3.16 shows the visual results of the reconstruction performance of the methods for a test OCT image. Fig. 3.16 (a) shows the noisy LR image. The zoomed versions of two highlighted regions are presented for the OCT images for better visualization of the minute details. Fig. 3.16 (b)-(f) show the reconstructed images using the EDSR, the WDSR, the ZSSR, the BM3D+ZSSR and the proposed unsupervised GAN based methods for a magnification factor of 2. Fig. 3.16 (g) shows the true HR image for reference. As can be seen, the EDSR and the WDSR methods have failed to efficiently reconstruct the drusen (shown by solid yellow arrow in (1) in Fig. 3.16 (b) and (c)). The ZSSR method has reconstructed the drusen well, but the image is not denoised well (Fig. 3.16 (d)). The reconstruction performance of BM3D+ZSSR is quite encouraging compared to the other existing approaches. However, close observation reveals the presence of some smearing artefacts (see yellow dotted arrow in (2) in Fig.3.16 (g)), which may have been introduced by the BM3D denoising. The proposed method, on the other hand, has reconstructed the drusen well and has not introduced any artefacts.

The methods are also quantitatively evaluated on the 17 test images and the performance comparison in terms of PSNR, CNR and IS is presented in Table 3.4. As can be seen from the table, the proposed method achieves better performance compared to the existing methods, which corroborates well with the visual results.

3.5 Summary

This chapter explored the use of denoising and SR to enhance the visualization of the early AMD manifestations. An unsupervised GAN based SR method is proposed for the task. The significant advantage of the proposed method is its unpaired training strategy, which improves generalizability and indeed eliminates the burden of large-scale paired image acquisition. The method has been extensively compared with the well-known state-of-the-art SR methods. The experimental results show that the proposed method can efficiently denoise and super-resolve the SD-OCT images while maintaining the retinal layers' edge information and the drusen details. Additionally, the method has a significantly low test time compared to the existing methods. The encouraging SR performance and the low testing time of the method make it highly suitable for manual and automated OCT analysis in clinical settings. The method can be reliably used by ophthalmologists to obtain clean HR images for better manual quantification of the pathological manifestations. The method can also be used as a pre-processing block in automated diagnosis systems to improve the diagnosis performance.

4

B-scan Attentive CNN for OCT Volume Classification

Contents

In the previous two chapters, we discussed the B-scan based OCT classification methods for diagnosing retinal diseases. The methods employed only one B-scan to obtain the diagnosis decision. Such approaches are appropriate for preliminary diagnosis and mass-screening applications where quick diagnosis results are desired. However, it is challenging to identify the retinal region for scanning to capture the pathological manifestations for diagnosis. Such selection requires clinical expertise and may be subjected to inter- and intra-observer variations. Moreover, the clinicians mostly prefer the OCT volumes over single B-scans for a comprehensive analysis of the spread and severity of the diseases. The volumetric OCT data constructed from a collection of densely sampled B-scans are more suitable as they scan over a larger retinal region. The automated classification of OCT volumes can result in a comprehensive and reliable diagnostic tool for medical practitioners.

Most of the existing automated methods in literature classify the individual B-scans in the volume and aggregate the diagnosis decision using manual threshold-based inference strategies [35–37, 64]. However, all B-scans in an OCT volume do not manifest disease symptoms. Summarizing the classification decision using all the B-scans irrespective of their diagnostic relevance can provide misleading results. Such methods also require fine-grained expert annotations of the B-scans in the volumes to learn the model. Fine-grained annotation procedures are very expensive and time-consuming as they involve a complex grading system with many trained graders for image labeling and verification [98].

In clinical practice, the ophthalmologists identify the salient B-scans with disease manifestations and fuse the clinical information from these scans to diagnose diseases and their severity stages. Therefore, this chapter presents a classification method that automatically encodes discriminative information from the clinically relevant B-scans to make a diagnosis decision for the volume. Recently, the development of deep attention networks has enabled the neural network to pay more attention to the useful features for efficient classification. The attention mechanism has been widely used in various fields. For example, Bahdanau *et al.* [162] introduced attention into the neural machine translation, Rush et al. [163] proposed an attention based network for sentence summarization, Chorowski *et al.* [164] employed an attention mechanism for speech recognition and Xu *et al.* [165] proposed an attention based framework for image caption generation. In the classification tasks, the attention mechanism has been used to provide selective focus on relevant regions of the visual space to acquire discriminative feature representations, thereby boosting the recognition performance [166]. Wang *et al.* [167] proposed a residual attention network that incorporates soft-attention for generating attention-aware features for classification. The authors in [97] proposed a SE-block which is a lightweight gating mechanism, to model inter-channel dependencies. Woo *et al.* [168] proposed the convolutional block attention module (CBAM) that can be seamlessly integrated

into any CNN architecture to emphasize meaningful features in spatial and channel-wise dimensions.

Inspired by the success of the attention networks, we propose the B-scan attentive CNN (BACNN), which uses a self-attention mechanism to provide weights of importance to the salient B-scans of the volume. The features from the B-scans of the volume are then aggregated based on the attention weights to obtain a high-level attentive representation for classification. The self-attention also eliminates the requirement of fine-grained annotation of the B-scans and enables the utilization of only the volume level labels for model learning.

The rest of the chapter is organized as follows. The proposed BACNN method is discussed in section 4.1. The experimental results are presented in section 4.2. The chapter is summarized in section 4.3.

4.1 B-scan Attentive CNN for OCT Volume Classification

The schematic representation of the proposed framework is presented in Figure 4.1. The method contains three modules: the feature extraction module, the attention module, and the classification module. The feature extraction module extracts spatial features from each of the B-scans in the volume and encodes them as representative vectors. The attention module generates weights for the B-scan features, which are aggregated through weighted summation. The aggregated feature vector is finally applied to the classification module to generate the class label for the input OCT volume. The following sub-sections provide the details of each of the modules.

4.1.1 Feature Extraction Module

This module focuses on extracting features from the B-scans of the OCT volumes. Let us define an OCT volume as $\boldsymbol{\mathcal{V}} = \{\boldsymbol{V_1}, \boldsymbol{V_2}, ..., \boldsymbol{V_{\hat{N}}}\}$ containing B-scans $\boldsymbol{V_i} \in \mathbb{R}^{d_m \times d_n}$, $i \in \{1, 2, ..., \hat{N}\}$. Here, $\hat{N}$ represents the number of B-scans in the volume and d_m and d_n are the dimensions of the B-scans. Spatial feature representations are extracted from each of the B-scans of the volume $\boldsymbol{\mathcal{V}}$ using CNNs. The CNNs are used for the task as they can capture local spatial changes in the OCT B-scans through the learnable convolution kernels. As can be seen from Figure 4.1, each of the B-scans are fed to different CNNs for feature extraction. The output, $\boldsymbol{f_{v_i}}$ of the *CNN* for the input B-scan $\boldsymbol{V_i}$ is given as

$$\boldsymbol{f_{v_i}} = f(\boldsymbol{V_i}; \boldsymbol{\theta_e}) \tag{4.1}$$

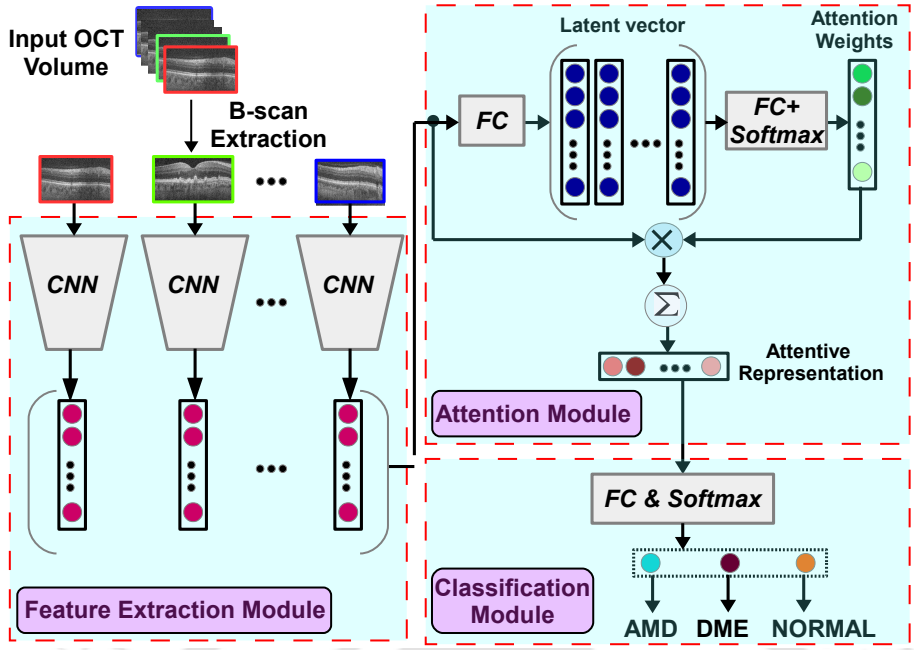


Figure 4.1: Block diagram of the BACNN classification method.

where $f(.)$ represents the composite function for the multiple linear and non-linear operations of the *CNN*, θ_e represents the parameters of the *CNN*. A parameter sharing strategy is employed for the convolutional layers of the different *CNNs*. This reduces the number of trainable parameters and also constrains the CNNs to learn generalizable features across the B-scans. Thus $\boldsymbol{f}_{v_i} \in \mathbb{R}^{d_e \times 1}$, $i \in \{1, 2, ..., \hat{N}\}$ is the encoded CNN feature vector for the B-scan $\boldsymbol{V}_i$ and d_e represents the dimension of $\boldsymbol{f}_{v_i}$.

Recently, Rasti et al. [35] presented the single scale CNN for the classification of AMD, DME and healthy control B-scans. The method employed a series of convolution, BN and pooling layers followed by an FC layer for classification and achieved promising classification performance. Considering the success of the network for OCT classification, we have employed a similar architecture for extracting spatial information from the OCT B- scans. The network architecture of the *CNNs* are given as

$$C_{(5,3)}BRP_{(2)} \rightarrow C_{(3,3)}BRP_{(2)} \rightarrow C_{(3,3)}BRP_{(2)} \rightarrow C_{(3,3)}BRP_{(2)} \rightarrow C_{(3,3)}BRP_{(2)} \rightarrow Flatten \rightarrow D_{(32)} \tag{4.2}$$

where $C_{(k,n)}BRP_{(p)}$ is a sequence of convolutional layer followed by BN, non-linear ReLU activation function and max pooling layers. Here, k represents the kernel size. n and p represent the number of convolution filters and the pool size respectively. $D_{(\hat{d})}$ is a dense layer with $\hat{d}$ output neurons.

4.1.2 Attention Module

As discussed previously, the disease symptoms do not manifest in all the B-scans of the volume. Hence, identifying the salient B-scan features is crucial for discriminative feature extraction and reliable classification. The attention mechanism has proven useful in locating the salient features and guiding the network to focus on these features during classification [58]. In this study, a self-attention mechanism is employed to assign appropriate weights of importance to the B-scan features based on the diagnostic content manifested in the B-scans. The features are then weighted by the attention weights and aggregated to obtain an attentive feature representation. The attentive vector can represent the vital discriminative disease features well and hence, guarantee improved classification performance. The design of the attention module is given as follows.

The module accepts the encoded feature vectors from the feature extraction module and generates soft attention weights for each of the encoded vectors. It is expected that the attention weights are high for the features that represent B-scan features manifesting disease characteristics and low for the features of B-scans having no pathological manifestations. This is achieved by first non-linearly transforming the feature vectors into a latent space using an FC layer containing d_l output neurons followed by a $tanh$ activation function. The output of the FC layer is given as

$$\boldsymbol{l_i} = tanh(\boldsymbol{W_l}\boldsymbol{f_{v_i}} + \boldsymbol{b_l}) \tag{4.3}$$

where $\boldsymbol{W_l} \in \mathbb{R}^{d_l \times d_e}$ and $\boldsymbol{b_l} \in \mathbb{R}^{d_l \times 1}$ are the weight matrix and the bias vector for the FC layer, $\boldsymbol{l_i} \in \mathbb{R}^{d_l \times 1}$, $i = \{1, 2, ..., \hat{N}\}$ is the obtained latent vector and d_l is the dimension of the latent vector $\boldsymbol{l_i}$. The vectors $\boldsymbol{l_i}$, $i = \{1, 2, ..., \hat{N}\}$ are provided to another FC layer with softmax activation to obtain the attention weights. The self attention weights α_i, $i = \{1, 2, ..., \hat{N}\}$ for the encoded CNN vectors are then obtained as [166, 169, 170]

$$\alpha_i = \frac{exp(\boldsymbol{w_{att}}\boldsymbol{l_i} + b_{att})}{\sum_{j=1}^{\hat{N}} exp(\boldsymbol{w_{att}}\boldsymbol{l_j} + b_{att})} \tag{4.4}$$

where $\boldsymbol{w_{att}} \in \mathbb{R}^{1 \times d_l}$ and $b_{att} \in \mathbb{R}^1$ are the learnable parameters of the network. The softmax operation in Eq. 4.4 ensures that the computed attention weights are positive and sum up to one. These weights act as a feature selector to highlight the discriminative features from the B-scans with pathological manifestations and suppress the features from the trivial B-scans. In this way, the weights guide the network to focus on clinically relevant features thereby assuring reliable classification. Finally, the *CNN* feature vectors of the B-scans, $\boldsymbol{f_{v_i}}$s, $i \in \{1, 2, ..., \hat{N}\}$, are aggregated through weighted summation of the $\boldsymbol{f_{v_i}}$s and α_is ,

$i \in \{1, 2, ..., \hat{N}\}$ to obtain the attentive feature vector $\boldsymbol{h} \in \mathbb{R}^{d_e \times 1}$ as

$$\boldsymbol{h} = \sum_{i=1}^{\hat{N}} \alpha_i \boldsymbol{f}_{\boldsymbol{v_i}}. \tag{4.5}$$

4.1.3 Classification Module

In the classification stage, $\boldsymbol{h}$ is fed to an output FC layer with softmax activation. This layer computes the probability distribution of the output categories for the volume $\boldsymbol{\mathcal{V}}$ as

$$p_v(c|\boldsymbol{\mathcal{V}}) = Softmax(\boldsymbol{W_v h} + \boldsymbol{b_v}) \tag{4.6}$$

where $p_v(c|\boldsymbol{\mathcal{V}})$ is the probability of the volume $\boldsymbol{\mathcal{V}}$ belonging to class c, $c \in \{1, 2, ..., C\}$, C is the number of output classes, $\boldsymbol{W_v} \in \mathbb{R}^{C \times d_e}$ and $\boldsymbol{b_v} \in \mathbb{R}^{C \times 1}$ are the weights and biases of the output layer. The final predicted output label for volume $\boldsymbol{\mathcal{V}}$ is obtained as

$$class(\boldsymbol{\mathcal{V}}) = \underset{c=1,2,...,C}{\arg\max}\, p_v(c|\boldsymbol{\mathcal{V}}). \tag{4.7}$$

Training Loss Optimization: In the training phase, the network parameters are trained using backpropagation by minimizing the multi-class categorical cross entropy loss (Eq. 4.8) between the probabilistic output and one hot encoded labels for a set of training samples $(\boldsymbol{\mathcal{V}_m}, y_m), m \in \{1, 2, ..., M\}$.

$$L = -\frac{1}{M} \sum_{m=1}^{M} \sum_{c=1}^{C} \mathcal{I}(y_m = c)\, log\, p_v(c|\boldsymbol{\mathcal{V}_m}) \tag{4.8}$$

where $\mathcal{I}(.)$ is an indicator function which is equal to one if y_m equals to c.

Once the proposed model is trained, it can be used to obtain the class predictions for unseen OCT volumes. In the following section, we analyse the performance of the proposed method and compare it with the state-of-the-art OCT classification models.

4.2 Experimental Results

In this section, we present the performance evaluation of the proposed BACNN method. The clinical database used for evaluation, the network details, the ablation study and the results are also discussed.

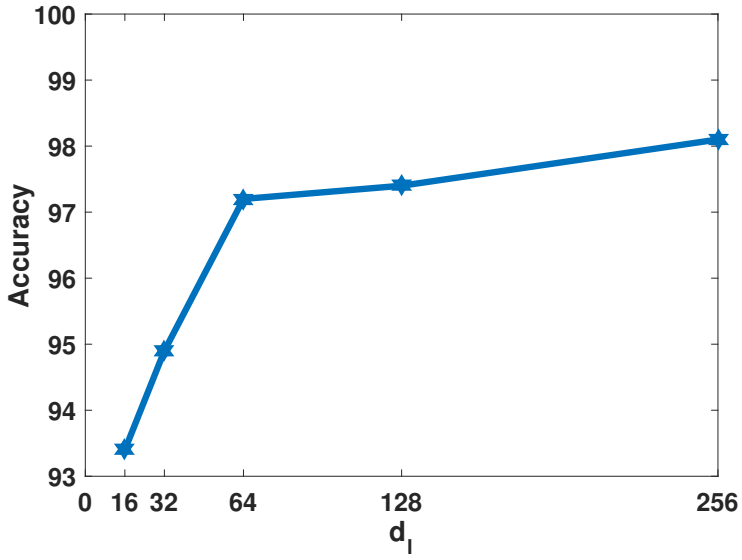


Figure 4.2: Variation in the accuracy for different sizes of d_l.

4.2.1 Clinical Database

The proposed method is evaluated on two publicly available OCT volume databases: the DUIA [46] and the NEH [35]. The DUIA dataset is the world's largest SD-OCT database for intermediate AMD and control subjects. It contains OCT volumes from 269 AMD and 115 control subjects. All the subjects in this database have 100 B-scan images per volume. The NEH dataset contains 3D SD-OCT volumes from 48 AMD, 50 DME and 50 normal subjects. The number of B-scans per volume varies from 19 to 61. We have down-sampled all the volumes in this dataset to 19 B-scans for the NEH database for ease of processing.

4.2.2 Evaluation Scheme and Performance Measures

In this work, a 5-fold cross-validation protocol is considered for both the databases. For efficient training, data augmentation is performed on the training folds by flipping the B-scans of the OCT volumes. However, no augmentation process is carried out for the testing folds. The classification performance for the DUIA database containing two classes (AMD and normal) is evaluated using the sensitivity, precision, F1 score, accuracy values from the binary confusion matrix. The performance of the NEH database containing three classes (AMD, DME and normal) is evaluated using the class-wise and the overall measures. The class-wise measures include the sensitivity (SE), the precision (PR) and the F1 score (F1). The overall measures used are the overall sensitivity (OS), the overall precision (OP), the overall F1 score (OF1)

Table 4.1: The effect of the attention module on the classification performance for the DUIA dataset.

Method	Sensitivity(%)	Precision(%)	F1 score(%)	Accuracy(%)	AUC	Kappa
CNN- w/o Attention	96.64 ± 2.47	96.37 ± 2.79	96.47 ± 1.79	95.02 ± 2.52	0.93 ± 0.03	0.88 ± 0.06
BACNN	**97.76 ± 3.07**	**98.14 ± 1.84**	**97.97 ± 2.76**	**97.12 ± 2.98**	**0.97 ± 0.03**	**0.93 ± 0.07**

The bold values show the performance of the proposed method.

and the overall accuracy (OA). The AUC and kappa values are also analyzed for both the databases.

4.2.3 Network Parameters and Ablation Study

The details of the different hyper-parameters for obtaining the desired results are discussed here. The B-scan images of the volume are flattened, cropped and resized to 128×256 before the training process [35]. The number of B-scans per volume, i.e., the data-dependent parameter $\hat{N}$ are 100 and 19 for the DUIA and the NEH databases, respectively. The parameters of the CNN used for feature extraction, i.e., the kernel size, the number of filters, the pool size and the neurons in the dense layers, are provided in Eq. 4.2. In the attention module, the size d_l of the non-linearly transformed CNN vectors is obtained by performing a grid search. The values of d_l in the set {16, 32, 64, 128, 256} are considered and the corresponding classification accuracies are observed for the DUIA database. Figure 4.2 shows the variation of the accuracy with respect to d_l. It can be seen that the accuracy increases with the increase in d_l until $d_l = 64$. No significant increase in accuracy is observed beyond that point. To build a light-weight network with fewer parameters, d_l is set to 64 for the analysis of the proposed method. A drop-out factor of 20% is applied to the output layer to prevent over-fitting. In the classification stage, the parameter C represents the number of output classes. The value of C for the DUIA and the NEH datasets are 2 and 3, respectively. These hyper-parameters are set to the above discussed values for obtaining the classification result of the proposed method and for the comparison with the existing methods. For efficient training, data augmentation is performed on the training folds by horizontally flipping the B-scans of the OCT volumes. However, no augmentation is carried out for the testing folds. The model is trained with an Adam [123] optimizer with a learning rate of 10^{-3} on mini-batches of size 9 for 3800 iterations. An NVIDIA Tesla V100 GPU is used to perform the experiments.

We present an ablation study to highlight the importance of the attention module in the proposed framework.

Significance of the attention module: To verify the usefulness of the attention module, the results for the proposed framework without the attention module are reported. This configuration is named 'CNN

Table 4.2: The effect of the attention module on the classification performance for the NEH dataset.

Method	Class	SE(%)	PR(%)	F1 (%)	OS/OP/OF1(%)	OA(%)	AUC/Kappa
CNN- w/o Attention	AMD	79.50 ± 2.88	94.09 ± 6.84	86.25 ± 6.95	89.71 ± 3.74	89.94 ± 3.77	0.92±0.03 0.85 ± 0.06
	DME	100.0 ± 0.00	95.46 ± 5.25	97.62 ± 2.75	90.53 ± 4.06		
	Normal	89.94 ± 3.97	82.08 ± 6.29	85.68 ± 5.57	89.85 ± 3.86		
BACNN	AMD	**92.00 ± 4.41**	**89.78 ± 0.49**	**90.74 ± 4.35**	**93.23 ± 2.23**	**93.24 ± 2.28**	**0.95 ± 0.01** **0.90 ± 0.03**
	DME	**100.0 ± 0.00**	**98.18 ± 4.07**	**99.05 ± 2.13**	**93.44 ± 2.28**		
	Normal	**87.78 ± 4.37**	**92.36 ± 7.73**	**89.75 ± 3.01**	**93.17 ± 2.24**		

The bold values show the performance of the proposed method.

- w/o Attention'. To realize this architecture, we employed the GAP in place of the attention module. Specifically, the feature vectors obtained from each of the input B-scan images are pooled using GAP and concatenated to form a representative vector which is then applied to the classifier for classification. The classification performances for the CNN - w/o Attention and the BACNN models for the DUIA and the NEH datasets are given in Tables 4.1 and 4.2, respectively. It can be observed from both the tables that the proposed method outperforms the CNN - w/o Attention model. Notably, the BACNN method achieves an improvement of 1.8% in precision, 2.1% in accuracy and 4.3% in AUC over the CNN - w/o Attention for the DUIA database. Similarly, enhancements of 3.7% in OF1 and OA are observed for the proposed method on the NEH database. This verifies that the attention module aids in efficient classification by highlighting the discriminative features from the salient B-scans of the volume.

4.2.4 Performance Comparison with Existing Methods

The proposed method is compared with the B-scan based and the 3D volume based classification models. The B-scan based volume classification methods classify each of the B-scans in the volume, and the volume level classification scores are obtained by the following rule: in a given volume, if the number of B-scans predicted as pathological is higher than 30%, then the volume is classified as pathological else normal [35]. Recently proposed OCT B-scan volume classification methods like the SC-CNN [35] and the fine-tuned InceptionV3 [34] are considered in this study. Popularly used feature engineering HOG+SVM [36] method and transfer learning based models like the VGG16 [77] and the ResNet [75] are also used for comparison. For the 3D volume based method, the 3D-CNN (conv3D) architecture presented by Maetschke *et al.* [171] is adapted for the purpose of comparison. The conv3D method utilizes 3D convolution operations with GAP and an output FC layer to classify the OCT volumes.

The training of the B-scan based volume classification methods for the DUIA dataset is performed by extracting 20 pathological B-scans per AMD volume and 50 normal B-scans per healthy control volume

Table 4.3: Performance evaluation of the proposed BACNN and the existing methods on the DUIA database using 5-fold cross-validation.

Method	Sensitivity(%)	Precision(%)	F1 score(%)	Accuracy(%)	AUC	Kappa
HOG+SVM [36]	88.47 ± 6.06	98.34 ± 1.80	93.07 ± 3.59	90.83 ± 4.60	0.93 ± 0.04	0.79 ± 0.09
SC-CNN [35]	93.66 ± 5.01	97.05 ± 2.74	95.23 ± 2.34	93.45 ± 3.06	0.93 ± 0.03	0.85 ± 0.07
VGG16 [77]	94.80 ± 3.03	96.32 ± 2.89	95.51 ± 1.78	93.73 ± 2.47	0.93 ± 0.03	0.84 ± 0.06
InceptionV3 [34]	91.46 ± 4.24	99.23 ± 1.05	95.13 ± 1.95	93.46 ± 2.41	0.95 ± 0.01	0.85 ± 0.05
ResNet [75]	89.59 ± 4.05	99.19 ± 1.10	94.10 ± 2.20	92.14 ± 2.76	0.94 ± 0.02	0.82 ± 0.06
Conv3D [171]	94.22 ± 4.72	88.39 ± 5.35	91.23 ± 4.05	87.20 ± 5.86	0.83 ± 0.08	0.68 ± 0.15
BACNN	**97.76 ± 3.07**	**98.14 ± 1.84**	**97.97 ± 2.76**	**97.12 ± 2.98**	**0.97 ± 0.03**	**0.93 ± 0.07**

The bold values show the performance of the proposed method.

from the training set. Similarly, for the NEH dataset, all the B-scans from the training OCT volumes are acquired. The fine-tuning of the transfer learning models (VGG16, ResNet and InceptionV3) is carried out by first replacing the output layer neurons with the number of classes (two and three for the DUIA and NEH dataset, respectively). All the network parameters except the output layer are frozen. The Adam [123] optimizer is used for fine-tuning on mini-batches of size 64 for 11,000 iterations. The training of the conv3D [171] is performed directly on the training OCT volumes with similar settings as the proposed method.

4.2.5 Results on the DUIA database

Table 4.3 shows the 5-fold cross-validation classification performance of the proposed and the existing methods on the DUIA database. It can be observed that the conv3D method shows inferior performance in terms of precision, F1 score, accuracy, AUC and kappa values compared to the B-scan based classification methods (HOG+SVM, SC-CNN and transfer learning). This is because the minute clinical details which are essential for efficient classification are lost in the voxel based convolution process. The HOG+SVM shows very limited improvements in the classification performance over the Conv3D. As discussed previously, the simple gradient-based features captured by the HOG may not be sufficient for encoding the complex diagnostic features of the OCT B-scans, thereby providing limited performance. The SC-CNN and the pre-trained networks (VGG16, InceptionV3 and ResNet) perform quite similarly and achieve an F1 score and accuracy of nearly 95% and 93%, respectively. The proposed BACNN method performs favorably well, exhibiting a minimum improvement of nearly 3% in sensitivity, 3.5% in accuracy, 2% in AUC and 8% in kappa values over the existing methods. Specifically, the method achieves sensi-

Table 4.4: Performance evaluation of the proposed BACNN and the existing methods on the NEH database using 5-fold cross-validation.

Method	Class	SE(%)	PR(%)	F1 (%)	OS/OP/OF1(%)	OA(%)	AUC/Kappa
HOG+SVM [36]	AMD	90.00 ± 14.14	86.03 ± 11.76	86.83 ± 7.50	89.70 ± 4.43	89.74 ± 4.29	0.92 ± 0.06
	DME	100 ± 0	94.85 ± 7.55	97.23 ± 4.09	91.35 ± 3.79		0.85 ± 0.06
	Normal	79.11 ± 14.31	93.16 ± 10.27	83.94 ± 8.44	89.33 ± 4.64		
SC-CNN [35]	AMD	90.00 ± 12.25	88.21 ± 13.06	87.78 ± 5.50	88.37 ± 3.29	88.38 ± 3.24	0.91 ± 0.03
	DME	98.00 ± 4.47	94.67 ± 7.67	96.18 ± 5.24	90.81 ± 4.01		0.83 ± 0.05
	Normal	77.11 ± 12.08	89.57 ± 10.85	81.73 ± 9.54	88.56 ± 4.96		
VGG16 [77]	AMD	93.56 ± 5.97	72.88 ± 13.17	81.39 ± 9.05	80.59 ± 9.29	80.74 ± 8.90	0.86 ± 0.06
	DME	100.0 ± 0	85.80 ± 10.56	92.08 ± 6.17	85.39 ± 5.31		0.72 ± 0.13
	Normal	48.22 ± 24.47	97.50 ± 5.59	60.79 ± 24.82	78.09 ± 12.19		
InceptionV3 [34]	AMD	86.00 ± 12.91	87.18 ± 9.83	84.47 ± 11.83	89.78 ± 5.29	89.81 ± 5.24	0.92 ± 0.03
	DME	98.00 ± 4.47	98.18 ± 4.07	97.99 ± 2.75	91.57 ± 4.02		0.85 ± 0.07
	Normal	85.33±14.45	89.33 ± 15.35	93.78 ± 9.08	89.41 ± 5.59		
ResNet [75]	AMD	94.00 ± 13.42	79.43 ± 14.93	84.40 ± 6.14	86.30 ± 3.86	86.36 ± 3.57	0.89 ± 0.03
	DME	98.00 ± 4.47	93.03 ± 7.07	95.23 ± 3.23	89.70 ±1.82		0.79 ± 0.05
	Normal	66.88 ± 21.82	96.67 ± 7.46	76.65 ± 12.33	85.46 ± 4.40		
Conv3D [171]	AMD	66.36 ± 11.73	65.83 ± 14.24	64.42 ± 6.54	67.41 ± 3.18	67.60 ± 3.45	0.75±0.02
	DME	85.00 ± 12.91	76.97 ± 6.59	80.70 ± 9.38	67.50 ± 4.27		0.51 ± 0.05
	Normal	54.80 ± 12.41	59.69 ± 12.40	53.28 ± 19.11	66.13 ± 4.78		
BACNN	AMD	**92.00 ± 4.41**	**89.78 ± 0.49**	**90.74 ± 4.35**	**93.23 ± 2.23**	**93.24 ± 2.28**	**0.95 ± 0.01**
	DME	**100.0 ± 0.00**	**98.18 ± 4.07**	**99.05 ± 2.13**	**93.44 ± 2.28**		**0.90 ± 0.03**
	Normal	**87.78 ± 4.37**	**92.36 ± 7.73**	**89.75 ± 3.01**	**93.17 ± 2.24**		

The bold values show the performance of the proposed method.

tivity, precision and accuracy of 97.76%, 98.14% and 97.12%, respectively. An AUC of 0.97 and kappa of 0.93 are also obtained. The improved classification performance demonstrates the effectiveness of the attention module that provides more weight to the features vital for efficient classification.

4.2.6 Results on the NEH database

The experimental results on the NEH dataset are shown in Table 4.4. As can be observed, a performance trend similar to the DUIA database is seen for this dataset as well. The proposed BACNN method outperforms the existing methods and attains an impressive OS of 93.23%, OP of 93.44%, OA of 93.24%, AUC of 0.95 and kappa value of 0.9. It can be seen from Table 4.4 that the proposed method outperforms the existing methods without compromising on the class-wise detection performance. It is also observed that the proposed method shows less variance in performance across the 5-folds, thereby rendering better generalization.

Figure 4.3 compares the average test run times for the different methods. The measurements for the

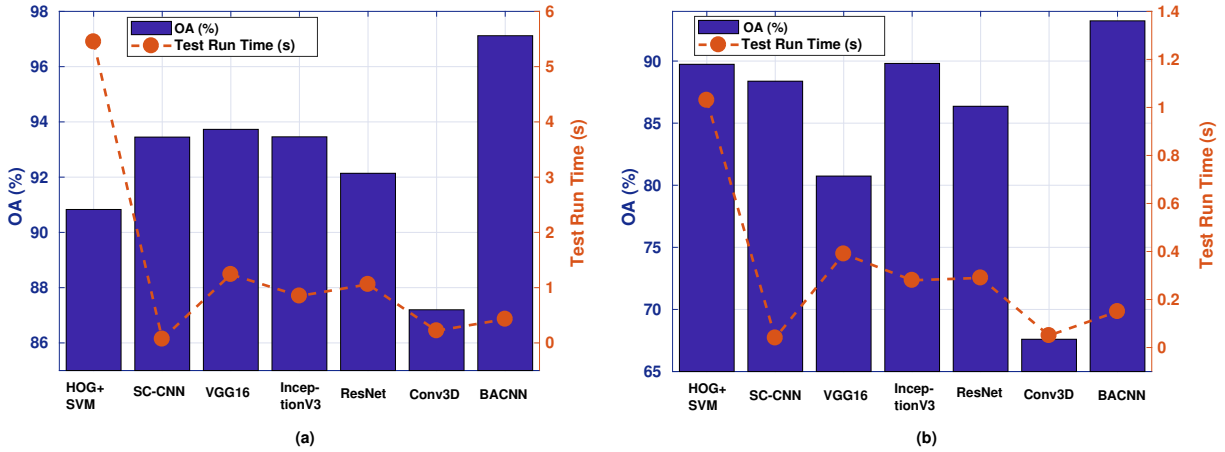


Figure 4.3: Comparison of the OA and the average run time of the methods on (a) the DUIA and (b) the NEH databases.

testing times are taken from a system with an i5 processor with 16 GB RAM and Quadro K600 graphics processor. It can be observed that the Conv3D method has the lowest test run time. This is because it processes the entire OCT volume together as opposed to the B-scan based methods. However, the low OA of the method makes it unsuitable for clinical applications. The HOG+SVM takes the highest run time as it serially processes the B-scans in the volume to obtain the classification result. It can be observed that the average test run time for the proposed method is lower than the VGG16, the InceptionV3 and the ResNet. This is because of the shared CNN network used in the feature extraction module and the light-weight nature of the proposed method with fewer parameters compared to the bulky networks like the VGG16, the InceptionV3 and the ResNet. The proposed method takes an average time of 0.43 seconds and 0.15 seconds to generate the classification outputs for the DUIA and the NEH databases, respectively.

4.2.7 Visualization of the Learned Attention Weights

The self-attention module of the proposed method is designed to provide weights of importance to the B-scans based on their diagnostic relevance. To verify the effectiveness of the module, the attention weights are visualized to identify the salient B-scans that are given more priority in the classification process. Figure 4.4 (a) shows the B-scans from a DME volume along with the attention weights in the form of a heat map. The white arrows on the B-scans highlight the pathological manifestations. It can be observed that the attention weights are very low for the B-scans 1, 8 and 19 that have no pathological manifestations. As the DME manifestations become prominent in the B-scans, the attention weights also increase. It can be seen that the highest attention weights are obtained for the B-scans (see Figure 4.4

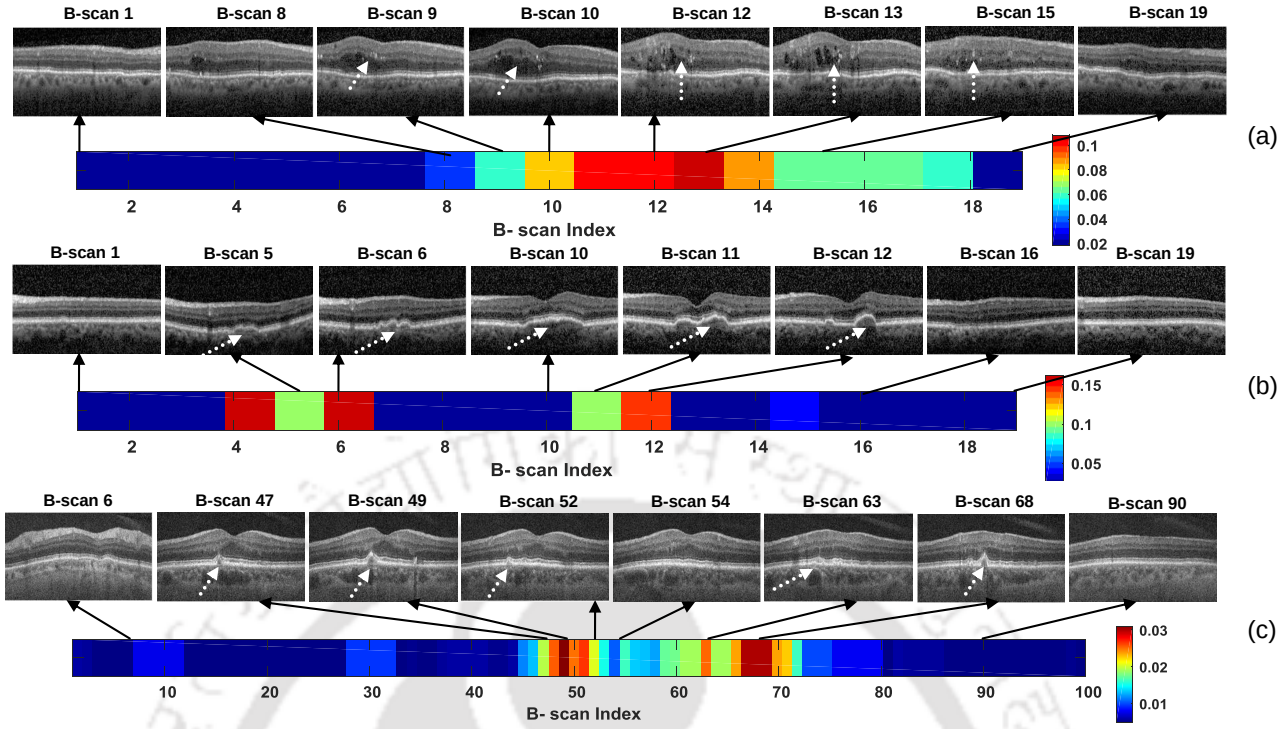


Figure 4.4: Examples of OCT volumes and the corresponding attention maps. (Best viewed in color)

(a): B-scan 12 and 13) that have severe and prominent DME manifestations. Similar observations are seen for other volumes of the database. Hence, it is clear that the method automatically identifies the affected B-scans and provides appropriate weightage to these scans during the classification based on the severity of the manifestations.

Figure 4.4 (b) and (c) show attention weights for AMD volumes from the NEH and DUIA datasets, respectively. It can be seen that the pathological symptoms in the B-scans of Figure 4.4 (c) (shown by white arrows) are smaller in size compared to Figure 4.4 (b). However, the attention weights (Figure 4.4 (b)) have successfully attended to the B-scans that manifest these minute symptoms. Therefore, the BACNN method is also suitable for diagnosing the early stages of the diseases. It is evident from the attention weights that the proposed method correlates with the ophthalmologists' way of diagnosing retinal diseases. The method also eliminates the need for manual tuning of thresholds, which is obligatory for the B-scan based classification approaches.

To provide a more transparent diagnostic basis, we also visualize the regions of the B-scan images that are important for correct classification using Grad-CAM [129]. Figure 4.5 shows the Grad-CAM outputs for the B-scans from AMD volumes that have the highest attention weights. The highlighted portions in the images are the regions that the network focuses on during classification. It can be seen that the proposed method attends to the pathological manifestations in the B-scans to make a diagnosis decision.

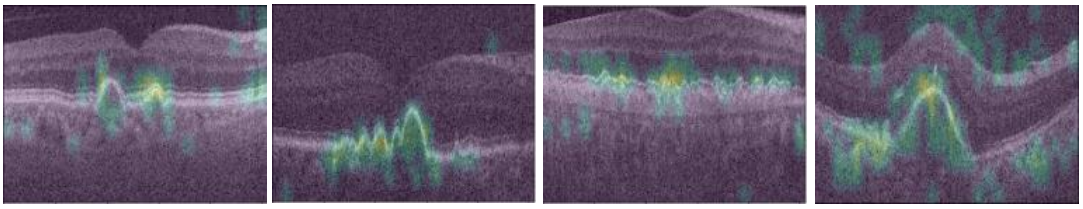


Figure 4.5: Grad-CAM outputs for the B-scan images with the highest attention.

This makes the method reliable and establishes trust in the predictions for the users.

4.3 Summary

In this chapter, a new B-scan attentive CNN is presented for the automated classification of the retinal OCT volumes. The method extracts local spatial features from the B-scans in the volume and inherently provides weights of importance to the clinically relevant B-scan features for efficient classification. The experimental results verify that the proposed method outperforms the existing methods with a large margin. A notable advantage of the method is the diagnostic transparency provided by the attention weights. The visualization of the attention weights highlights that the method provides higher weight values to the pathological B-scans while making the classification decision. This process correlates well with the ophthalmologists' way of diagnosing retinal diseases. Thus the method mimics the ophthalmologists' ways of clinical diagnosis and can be reliably used in eye clinics and hospitals for the automated diagnosis of retinal diseases. The method also has a generalizable architecture, i.e., the attention module can be integrated with any deep learning based feature extractor and can be easily adapted to classify other diseases.

5

Conclusions

Contents

5.1 Summary of the Work

According to the National Eye Institute, United States of America (U.S.A), the prevalence of AMD and DME are estimated to increase by the end of 2030 [172]. The high prevalence will result in the insufficiency of medical experts per patient [173]. It would be troublesome for any eye care centre to provide screening services at a large scale with a limited number of skilled personnel. Thus, computer-aided diagnosis can be more suitable as it can accomplish fast and reliable screening, enable the semi-skilled medical personnel at remote locations to perform preliminary diagnosis and reduce the overall cost of care. In this regard, this thesis presented various classification frameworks for the automated diagnosis of retinal diseases using OCT images and volumes. The methods considered the pathological evidence of the retinal diseases and handled the imaging bottlenecks (speckle noise and poor resolution) of OCT to develop robust DL based classifiers. A summary of the thesis is given as follows.

In Chapter 1, an introduction to OCT imaging of the retina and the challenges encountered during diagnosis were discussed. Along with these, the literature review of the existing classification methods for OCT images and volumes was presented. A brief survey of the different simultaneous denoising and SR methods for improving the diagnostic details in the images was also reported. Although many methods were proposed in the literature for the OCT classification, a few research challenges existed that were not effectively addressed. One such challenge was the efficient representation of the diverse multi-scaled disease characteristics of the retinal lesions for improved classification of the B-scans. Most of the existing DL based methods had employed single-scale CNNs for feature extraction and classification of the OCT B-scans [34,37,58]. However, it was challenging to encode the variabilities in the pathological morphologies of different diseases efficiently by ignoring potentially useful information on different scales. Therefore, we proposed two multi-scale feature fusion based methods in Chapter 2 to maximally encode the disease characteristics manifested in the B-scans for improving the diagnosis performance.

The first proposed method, namely, the MDFF, explored the use of spatial pyramid based image representations to obtain the multi-scale views of the images. Multiple CNN based feature extractors were then employed to encode the scale-specific features. These features were fused through concatenation to obtain high-level discriminative feature representation for classification. The MDFF method attained an OA of 94.52% with an improvement of over more than 5% on the UCSD database compared to the recent state-of-the-art OCT classification methods like the IFCNN [56] and LGCNN [89]. To further improve the classification performance, the LM-DFF method was proposed. The method employed multiple CNNs with different receptive fields to automatically extract scale-specific features from the OCT images.

The receptive fields were varied using dilated convolutions with different dilation factors for each of the CNNs. The scale-specific features were fused through concatenation to obtain powerful multi-scale features for classification. A joint multi-loss optimization strategy was also designed to effectively learn the scale-specific and cross-scale features for classification. The LM-DFF method outperformed the MDFF method by a margin of 1.5% and 2% in terms of OA on the UCSD and the NEH databases, respectively. Specifically, the method achieved an impressive OA of 96.03% and 99.6% on the UCSD and the NEH databases, respectively. The improved performance of the LM-DFF method can be attributed to the learnable multi-scale feature extraction obtained using the dilated convolutions and the joint loss optimization that can effectively capture the within and cross-scale discriminative feature representations for efficient classification.

The second challenge we addressed was the simultaneous denoising and SR of the OCT images for improved diagnosis of intermediate AMD. Recent methods have relied on example-based strategies that require paired LR and HR images during model training [14,33,108]. However, the large-scale acquisition of paired LR-HR images for efficient supervised learning is challenging due to various reasons [136]. It is also well known that the DL based models require a large amount of data to perform reasonably well [98]. To mitigate the data scarcity issue and still make use of the advantages of DL, an unsupervised GAN based denoising and SR framework was proposed in Chapter 3. The method did not rely on one-to-one alignment between the LR and the HR images during training. The option of using unpaired images provided flexibility to leverage the already available OCT data for better generalization of the denoising and SR performance. The adversarial learning of the GANs and the sophisticated optimization objectives helped in building a model that can reliably reconstruct the clinical details even in the absence of paired training images. The experimental results on clinical-grade OCT images confirmed that the proposed method outperformed the existing methods by a large margin both in terms of SR performance and computational time. The method achieves a PSNR of 39.15 dB and 35.67 dB for a magnification factor of 2 and 4, respectively. The method takes nearly 4 seconds to generate the clean HR images from the noisy LR images. The usefulness of the method was also demonstrated by comparing the classification performance of intermediate AMD and normal subjects using the noisy LR and the reconstructed clean HR images. An improvement of 3.8% in classification accuracy was observed when the reconstructed clean HR images were used for classification compared to the noisy LR counterparts.The results verified that the generated clean HR images effectively reconstructed the essential clinical details which led to the improvements in the intermediate AMD classification performance.

In Chapter 4, the task of OCT volume classification was addressed. The OCT volumes span a larger

retinal region and are useful for assessing the severity stages of the retinal diseases. The state-of-the-art methods in literature classified the individual B-scans within the volume and aggregated the diagnosis decision for the OCT volume using manual threshold-based rules [35–37]. However, the retinal lesions may not manifest in all the B-scans of the volumes [99]. Hence, summarizing the classification decision using all the B-scans irrespective of their diagnostic relevance could provide misleading results. To tackle this issue, we presented the BACNN method that selectively focused on the features of the diagnostically relevant B-scans to make a classification decision. The method included three modules: the feature extraction module, the attention module and the classification module. The feature extraction module used a CNN framework to extract local spatial feature representations from the B-scans of the volume. The attention module provided weights of importance to the obtained feature vectors based on their diagnostic relevance. The features were then weighted by the attention weights and aggregated to obtain an attentive feature representation. The discriminative attentive representation was fed to the classification module to obtain the class labels for the input volume. The proposed method improved the diagnostic accuracy of AMD volume classification by nearly 4% compared to the B-scan based approaches, achieving an OA of 97.12%. A notable advantage of the proposed method was the diagnostic transparency that was achieved by the visualization of the attention weights. The weights provided information about the B-scans that were given higher importance during the classification process. Visual analysis of the weights and the corresponding B-scan images verified that the network was correctly identifying the salient B-scans for classification. This analysis of the weights imparted interpretability to the model and made it more reliable for clinical use.

5.2 Future Directions

Despite the automated solutions for diagnosis presented in the thesis, some limitations still exist that need to be addressed in future works. Some of them are given below.

- It is well known that the DL based methods require large amounts of data to generate reasonably good performance. However, considering the asymptomatic nature of the retinal diseases, large-scale databases for the early stages of the diseases are less abundantly available. The limited availability affects the classification results of the automated methods for these stages. For example, Table 2.5 in Chapter 2, shows a drop in the class-wise performance for the drusen class compared to the other classes. This is because the number of samples of the drusen class is quite less compared to the other classes. Even though a cost-sensitive learning strategy has been adopted, there is still

room for improvement. Recently, the data-efficient DL based methods have shown to provide generalizable results in limited training data conditions. Odena [174] presented a semi-supervised GAN architecture, where the discriminator was altered to perform multi-class classification. It was found that the changed architecture created a more data-efficient classifier. The data-efficient attribute of the modified GAN can be suitably used for the classification of clinical OCT images as well [112]. A detailed study on such data-efficient approaches can be performed in the future.

- In Chapter 3, we proposed an unsupervised GAN based method for simultaneous denoising and SR method of OCT B-scans. The method is trained and evaluated on the images acquired from the Bioptigen SD-OCT system. It would also be interesting to study the behaviour and generalizability of the proposed method on images acquired from different OCT systems such as Heidelberg, Zeiss and Topcon devices. Also, at present, the proposed method is extended to 3D OCT data by denoising and super-resolving individual B-scans of the volume. Recently, it has been proven that 3D SR for medical volumetric data delivers better visual results than conventional two-dimensional (2D) processing [175]. Hence, it would be interesting to explore the 3D denoising and SR of the OCT volumes for superior reconstruction performance.

- The proposed SR method in Chapter 3 (as well as the existing methods [14, 33, 105, 108]) uses the PSNR to quantify the reconstruction performance. The PSNR is computed as the ratio of the maximum power of the image to the mean squared error between the original and the reconstructed images. Such an approach is highly suitable for natural images as they have high texture variations throughout. However, in the OCT images, the clinical information is confined only to a small region (retinal layers), and the rest can be treated as non-diagnostic background regions. Global non-diagnostic measures like the PSNR provide equal priority to all pixels in the image irrespective of their diagnostic content. As the non-diagnostic pixels are much higher in number than the diagnostic pixels in the OCT images, the diagnostic distortions occurring during the reconstruction process may get diluted in the error quantification process. Therefore, specialized quality assessment metrics need to be developed that focus more on the diagnostic features in the OCT images to quantify the distortion.

- In Chapter 4, the BACNN method is proposed for the classification of the OCT volumes. The method provides attention weights to the B-scans in the volume based on their diagnostic relevance during classification. The attention weights can be further analysed to predict the spread and the severity of the diseases as well. For example, if most of the B-scans have high attention weights, it can be

inferred that the disease has spread to a large retinal region and vice versa. Further, the attention weights can be used as cues to identify the salient B-scans on which detailed analysis such as the quantification of the disease lesions can be performed for grading the severity of the diseases.

- This thesis focused on grossly identifying AMD and DME B-scans/volumes from the healthy controls. However, retinal diseases such as AMD, can manifest in several stages, mainly described as early, intermediate and advanced [45]. Patients with early symptoms of AMD can suddenly progress to advanced stages (without any noticeable visual changes) with abnormal vessel growth and protein leakage under the macula, leading to irreversible damage and rapid vision loss [176]. Thus, identifying patients with the impending risk of progression is an interesting and challenging research problem. The SD-OCT has enabled the accurate representation of the AMD manifestations such as drusens, pigmentary changes in the RPE and intra-retinal fluid deposits [177]. The characterization of these features from the SD-OCT volumes are useful biomarkers for the staging of AMD [47]. Advanced studies demonstrate the interaction of genetic, demographic (age, gender and medical history) and environmental risk factors (diet and smoking) in accelerating the disease progression [178]. Hence, combined analysis of imaging and clinical features (genetic, demographic and environmental factors) can play a key tool in the prognostic evaluation of AMD, which needs to be addressed in the future.

- The SD-OCT provides information about the anatomical structures of the retinal layers. However, it is limited in terms of its ability to provide the retinal microvasculature details [179]. The clinical investigation of the retinal vasculature provides useful information about vascular leakage, occlusions, vessel density and branching patterns [180,181]. Such clinical information is useful in the early diagnosis of retinovascular diseases, such as diabetic retinopathy, DME, CNV and retinal vein occlusion. For many years, the FA has been considered as the gold standard for imaging the retinal vasculature [182]. However, FA is time-consuming, requires intravenous access, and can have adverse effects, including nausea and more serious allergic reactions [183]. The OCT- angiography (OCT-A) is a recently emerging non-invasive imaging technique that can generate volumetric angiography images in a few seconds [184]. Recent findings demonstrate the effectiveness of the concomitant use of the en-face OCT-A with the SD-OCT for the diagnosis of retinal diseases [185]. Hence, multimodal automated analysis of OCT-A and SD-OCT can provide better insights for staging and risk prediction of retinal diseases.

- Lastly, we perceive that the lack of adequate databases has mostly limited the research in this field.

Large-scale databases featuring different diseases and multi-modal imaging outcomes from subjects need to be created to forward the research in the field.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

A

Retina Physiology

Retina and its Micro-structures

The retina is the light-sensitive layer of the eye that converts light from the object to electrical impulses that are communicated to the brain for visual perception [3]. Figure A.1 shows the sagittal view of the human eye and the hierarchical arrangement of the different cells of the retina. As can be seen from Figure A.1 (a), the retina is present in the posterior part of the eye and is surrounded by the tough external fibrous layer called the sclera and the vascular choroid layers. The retinal micro-structure contains the photoreceptor cells, also known as rods and cones cells (see Figure A.1 (b)) that convert the electro-magnetic waves of light from the object to electro-chemical gradient potentials. These potentials further excite the bipolar and the ganglion cells (see Figure A.1 (b)) to generate action potentials that are transmitted to the central nervous system through the nerve fibers. The retina also contains supporting cells like the Müller cells (see Figure A.1 (b)) to keep the retinal structure intact by fixing the photo-receptor and the neuronal cells at their respective spatial locations. Similarly, the melanin-rich retinal pigment epithelium (RPE) cells (see Figure A.1 (b)) present posterior to the photoreceptor cells absorb the excess light falling onto the retina in order to prevent reflections and degradation of the optical image [186]. The horizontal and the amacrine cells (see Figure A.1 (b)) are the laterally interconnecting neurons for integrating and regulating the inputs from the multiple cells. The hierarchical arrangement of the different cells forms the ten layers of the retina and are given as follows.

(i) **inner limiting membrane (INL)** formed by the foot-plates of the Müller cells.

(ii) **retinal nerve fiber layer (RNFL)** created by the axons of the ganglion cells.

(iii) **ganglion cell layer (GCL)** comprised of the cell bodies of the ganglion cells.

(iv) **inner plexiform layer (IPL)** made up of the fibers and synapses of the ganglion cells and the bipolar neurons.

(v) **inner nuclear layer (INL)** constituted by the cell bodies of the bipolar neurons, the amacrine and the horizontal cells.

(vi) **outer plexiform layer (OPL)** comprised of fibers and synapses of the bipolar neurons and rod and cone cells.

(vii) **outer nuclear layer (ONL)** composed of the cell bodies of the photoreceptors.

(viii) **external limiting membrane (ELM)** formed by the thin membrane formed by the Müller cell processes.

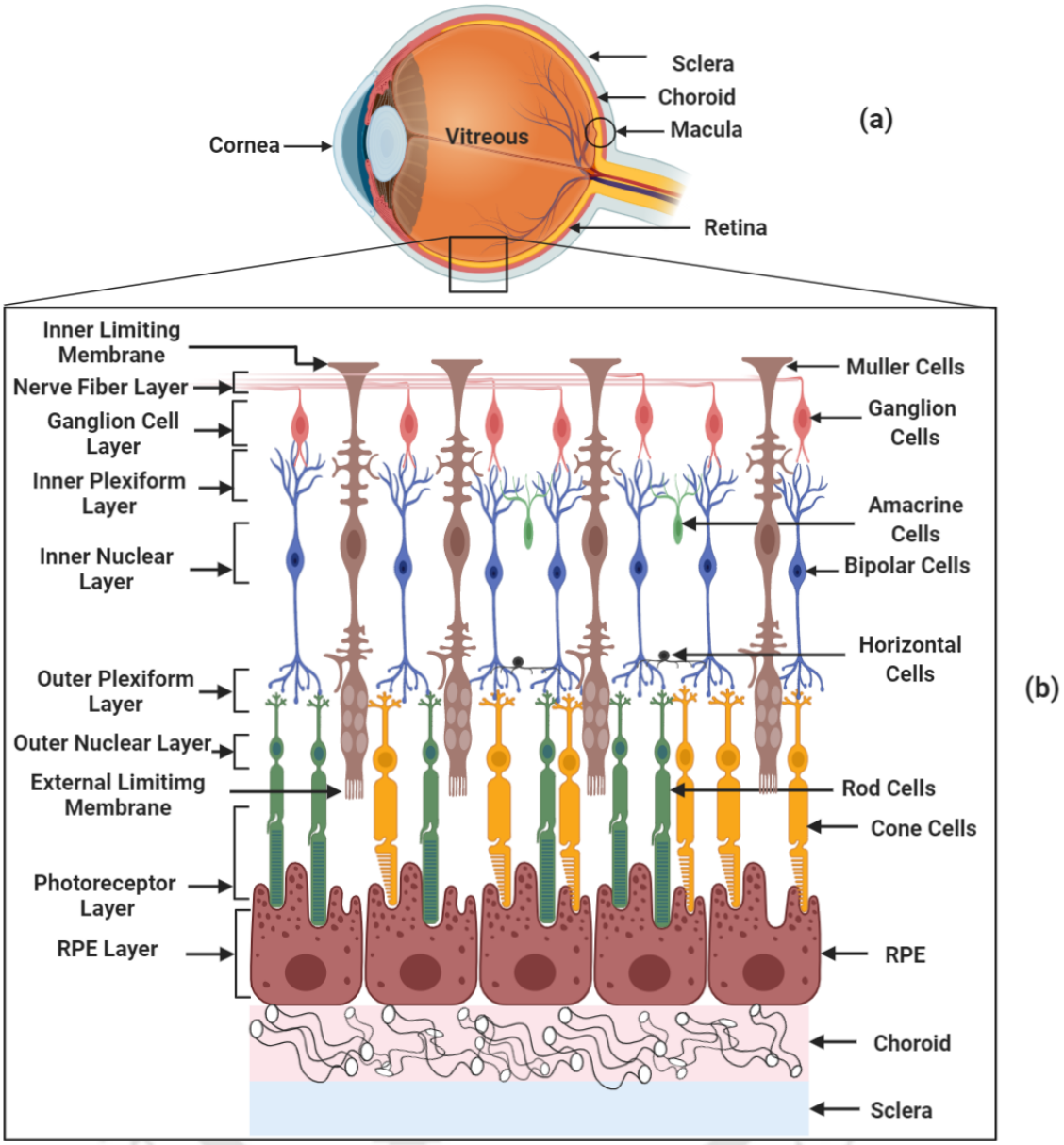


Figure A.1: Visual illustration of the different layers of the retina.

(ix) **photoreceptor layer (PR)** made up of the outer segments of the rods and cones and lies anterior to the RPE layer.

(x) **RPE layer** created by the multitude of the RPE cells of the retina.

The different layers of the retina presented above are highlighted in Figure A.1. The OCT B-scan images provide the visualization of these layers for the diagnosis of various retinal conditions. Figure A.2 shows an OCT B-scan highlighting the different layers of the retina.

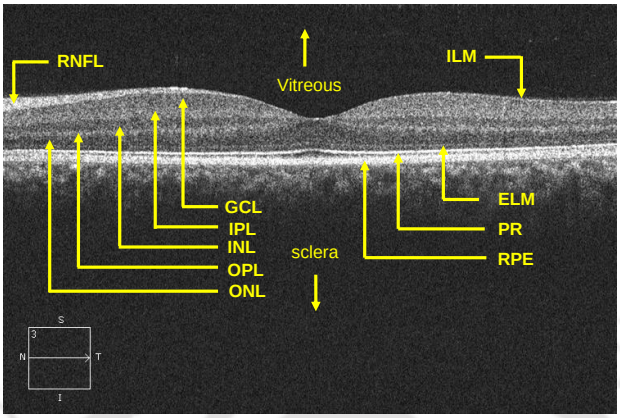


Figure A.2: OCT B-scan highlighting the ten different layers of the retina.

B

CNN Components

The CNN is a type of DL model designed to automatically and adaptively learn hierarchical spatial features from the input image data [187]. The CNN architecture includes several building blocks: convolution layer, activation layer, pooling layer and FC layer. Typically, the CNN architecture consists of a stack of several convolutions, activations and pooling layers followed by FC layers in the end [188]. The details of the different components of the CNN are given below.

- **Convolution layer**: This layer is the core building block of the CNN that performs feature extraction using the convolution operation. The layer is parametrized by learnable kernels (filters) that convolve over the height and width of the input to generate feature maps [152]. The convolution with multiple learnable filters results in feature maps that represent different characteristics (features) of the input tensors. Conventionally, multiple such convolution layers are cascaded in a CNN to extract discriminative hierarchical features. Mathematically, for an input volume $\boldsymbol{I} \in \mathbb{R}^{H_i \times W_i \times C_i}$ and a convolution kernel $\boldsymbol{k} \in \mathbb{R}^{h \times w \times c}$, the output feature map $\boldsymbol{F_o} \in \mathbb{R}^{H_o \times W_o}$ for a convolution operation is given as

$$\boldsymbol{F_o}(x, y) = \sum_{n_h=1}^{H_i} \sum_{n_w=1}^{W_i} \sum_{n_c=1}^{C_i} \boldsymbol{k}(n_h, n_w, n_c) \boldsymbol{I}(x + n_h - 1, y + n_w - 1, n_c). \tag{B.1}$$

  The height and width of F_o can be given as $H_o = H_i - h + 1$ and $W_o = W_i - w + 1$, respectively. Here, H_i, W_i and C_i are the height and width and channels of the input volume. h, w and c are the dimensions of k. It can be observed that the convolution operation reduces the size of the output feature map. With successive convolution operations, the feature maps would get smaller. Therefore, to maintain the same size of the inputs and outputs, zero padding is generally used. The padding adds rows and columns of zeros on each side of the input tensor to retain the same dimensions of the input and output feature maps. Another key parameter of the convolution operation is the stride. The stride determines the distance between two consecutive positions of the kernel during the convolution. The size of the output feature map $\boldsymbol{F_o}$ with padding (P) and stride (S) is given as, $H_o = \frac{H_i - h + 2P}{S} + 1$ and $W_o = \frac{W_i - w + 2P}{S} + 1$ [71, 189].

- **Activation layer**: The activation functions are non-linear transformations applied to the outputs of the linear operations like the convolution. Without the non-linear activations, the neural networks would perform as linear functions. The non-linear elements allow for greater flexibility and the creation of complex functions during the learning process. Figure B.1 shows some of the commonly used activation functions. The rectified linear unit (ReLU) is the widely used activation for the CNNs and is defined as

$$\sigma(x) = max(0, x). \tag{B.2}$$

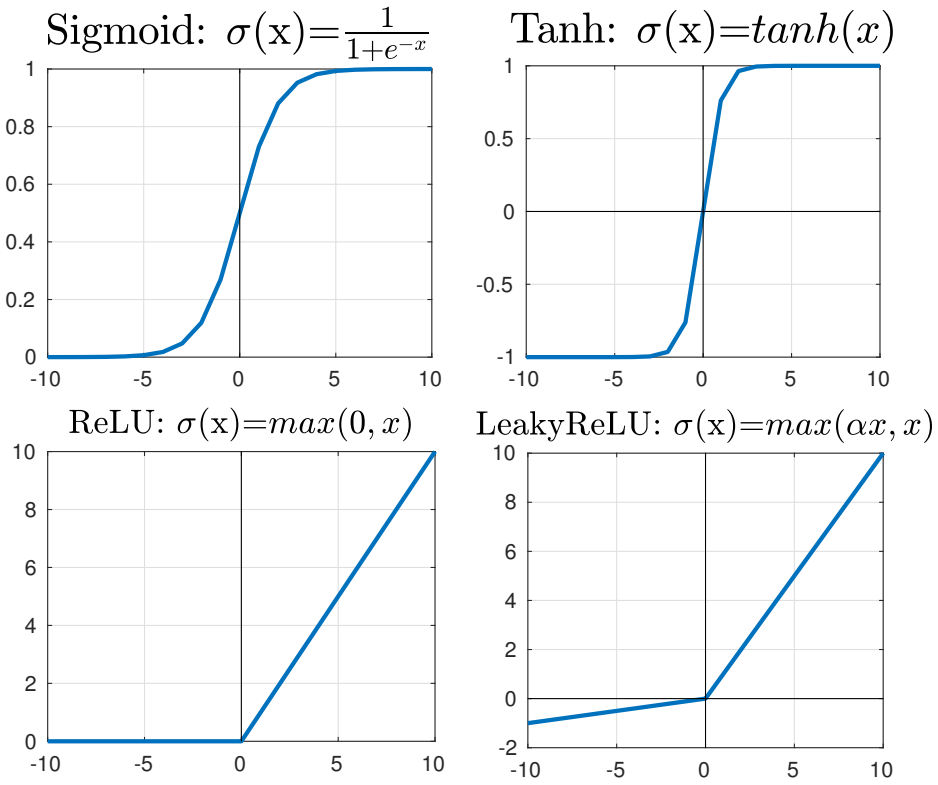


Figure B.1: Activation functions used in the CNNs.

Here $\sigma(x)$ represents the output of the activation function for input x. As can be seen from the above equation, the ReLU suppresses the negative inputs to zero, thereby preventing the activation of all neurons at a time. This reduces the computational complexity of the system, thereby leading to faster convergence [116].

- **Pooling layer**: The pooling layer fuses the neighborhood spatial information from the feature maps using average or max operations. Pooling reduces the spatial resolution of the feature maps, thereby reducing the number of parameters and improving translation invariance [190]. It is worth mentioning that the pooling layers do not have any learnable parameters, whereas the hyper-parameters (pool size, stride and padding) have similar functionalities as the convolution layer. The max-pooling operation is usually preferred in the CNNs and is performed by selecting the highest activation within the $p \times p$ pooling window size. This window is shifted across the feature map to obtain the pooled feature representation.

- **Fully connected (FC) layer**: The output feature maps of the final convolution or pooling layers are

transformed into a one-dimensional vector and connected to one or more FC layers. These layers are the dense layers where each node is connected to all the nodes in the previous layer by learnable weights. The final FC layer, also known as the output layer, contains the same number of neurons as the number of output categories with a softmax activation to obtain the class probabilities.

- **Global average pooling (GAP)**: The conventional CNNs use convolutional layers for feature extraction. The last convolutional layer features are transformed into a one-dimensional vector and fed to the FC layers for classification. However, the FC layers are prone to overfitting and may hamper the generalization ability of the network [127]. Therefore, the GAP can be used in place of the FC layer to transform the convolutional feature maps into vectors. The GAP reduces the spatial dimensions (by averaging) of the convolutional feature maps resulting in a vector that can be fed to the output layer for classification. The advantages of applying GAP are as follows: (a) reduces the number of trainable parameters, hence reducing overfitting and (b) enables the CNN to accept inputs of variable size.

For better training of the CNN, few regularization techniques like the dropout, batch normalization (BN) and instance normalization are also incorporated into the CNN framework. The details of which are given below.

- **Dropout**: It is a regularization technique that deactivates the output of a set of neurons chosen at random with a specific predefined probability during training [191]. These deactivated units do not take part in the forward and the backward propagation. The dropout prevents the network from being highly dependent on only a few neurons of the network as it may be randomly eliminated. Therefore, the dropout encourages the network to spread out and assign a bit of weight to all the neurons. This results in shrinking the squared norm of the weights, similar to L2 regularization, thereby preventing overfitting.

- **Batch normalization (BN)**: Training neural networks can be complicated as the distribution of each layer inputs changes during training [115]. This problem is also known as the internal covariance shift. For proper training, low learning rates and careful parameter initialization are required, which makes the networks very difficult to train. The problem is addressed by the BN that normalizes the output of the previous layer by subtracting the batch mean and dividing by the batch standard deviation. The batch mean and standard deviation are calculated for each channel across all samples and both spatial dimensions. This allows the use of much higher learning rates and less careful

parameter initialization. It helps in the fast and stable training of the network.

- **Instance normalization** [154]: Unlike BN, the mean and variance are calculated for each channel for each sample across both spatial dimensions. The normalization technique is generally used for style transfer and SR applications. It aims to normalize the contrast of the feature map to reconstruct the images at the target domain better.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

Bibliography

[1] G. Panozzo, E. Gusson, B. Parolini, and A. Mercanti, "Role of OCT in the diagnosis and follow up of diabetic macular edema," in *Seminars in Ophthalmology*, vol. 18, no. 2. Taylor & Francis, 2003, pp. 74–81.

[2] A. García-Layana, F. Cabrera-López, J. García-Arumí, L. Arias-Barquet, and J. M. Ruiz-Moreno, "Early and intermediate age-related macular degeneration: update and clinical review," *Clinical Interventions in Aging*, vol. 12, p. 1579, 2017.

[3] M. D. Abràmoff, M. K. Garvin, and M. Sonka, "Retinal imaging and image analysis," *IEEE Reviews in Biomedical Engineering*, vol. 3, pp. 169–208, 2010.

[4] S. Chiu, "Graph theory and dynamic programming framework for automated segmentation of ophthalmic imaging biomarkers," Ph.D. dissertation, Duke University, 2014.

[5] W. Drexler and J. G. Fujimoto, *Optical coherence tomography: technology and applications*. Springer Science & Business Media, 2008.

[6] J. G. Fujimoto, C. Pitris, S. A. Boppart, and M. E. Brezinski, "Optical coherence tomography: an emerging technology for biomedical imaging and optical biopsy," *Neoplasia (New York, NY)*, vol. 2, no. 1-2, p. 9, 2000.

[7] R. J. Zawadzki, A. R. Fuller, M. Zhao, D. F. Wiley, S. S. Choi, B. A. Bower, B. Hamann, J. A. Izatt, and J. S. Werner, "3D OCT imaging in clinical settings: toward quantitative measurements of retinal structures," in *Ophthalmic Technologies XVI*, vol. 6138. International Society for Optics and Photonics, 2006, p. 613803.

[8] A. Maalej, W. Cheima, K. Asma, R. Riadh, and G. Salem, "Optical coherence tomography for diabetic macular edema: early diagnosis, classification and quantitative assessment," *Journal of Clinical and Experimental Ophthalmology*, vol. 2012, p. 2, 2012.

[9] A. Al-Mujaini, U. K. Wali, and S. Azeem, "Optical coherence tomography: clinical applications in medical practice," *Oman medical journal*, vol. 28, no. 2, p. 86, 2013.

[10] C. Volz, F. Grassmann, R. Greslechner, D. A. Maerker, P. Peters, H. Helbig, and M.-A. Gamulescu, "Impact of optical coherence tomography (OCT) on decision to continue treatment for neovascuar age-related macular degeneration," *Investigative Ophthalmology & Visual Science*, vol. 57, no. 12, pp. 3707–3707, 2016.

[11] J. M. Schmitt, S. Xiang, and K. M. Yung, "Speckle in optical coherence tomography," *Journal of Biomedical Optics*, vol. 4, no. 1, pp. 95–106, 1999.

[12] M. Li, R. Idoughi, B. Choudhury, and W. Heidrich, "Statistical model for OCT image denoising," *Biomedical Optics Express*, vol. 8, no. 9, pp. 3903–3917, 2017.

[13] A. Baghaie, Z. Yu, and R. M. D'Souza, "Involuntary eye motion correction in retinal optical coherence tomography: Hardware or software solution?" *Medical Image Analysis*, vol. 37, pp. 129–145, 2017.

[14] L. Fang, S. Li, R. P. McNabb, Q. Nie, A. N. Kuo, C. A. Toth, J. A. Izatt, and S. Farsiu, "Fast acquisition and reconstruction of optical coherence tomography images via sparse representation," *IEEE Transactions on Medical Imaging*, vol. 32, no. 11, pp. 2034–2049, 2013.

[15] Y.-Y. Liu, M. Chen, H. Ishikawa, G. Wollstein, J. S. Schuman, and J. M. Rehg, "Automated macular pathology diagnosis in retinal OCT images using multi-scale spatial pyramid and local binary patterns in texture and shape encoding," *Medical Image Analysis*, vol. 15, no. 5, pp. 748–759, 2011.

[16] R. R. Bourne, G. A. Stevens, R. A. White, J. L. Smith, S. R. Flaxman, H. Price, J. B. Jonas, J. Keeffe, J. Leasher, K. Naidoo *et al.*, "Causes of vision loss worldwide, 1990–2010: a systematic analysis," *The Lancet Global Health*, vol. 1, no. 6, pp. e339–e349, 2013.

[17] M. Wojtkowski, T. Bajraszewski, I. Gorczyńska, P. Targowski, A. Kowalczyk, W. Wasilewski, and C. Radzewicz, "Ophthalmic imaging by spectral optical coherence tomography," *American Journal of Ophthalmology*, vol. 138, no. 3, pp. 412–419, 2004.

[18] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, C. A. Puliafito *et al.*, "Optical coherence tomography," *Science*, vol. 254, no. 5035, pp. 1178–1181, 1991.

[19] J. Pluciński, R. Hypszer, P. Wierzba, M. Strąkowski, M. Jędrzejewska-Szczerska, M. Maciejewski, and B. Kosmowski, "Optical low-coherence interferometry for selected technical applications," *Bulletin of the Polish Academy of Sciences: Technical Sciences*, pp. 155–172, 2008.

[20] R. A. Costa, M. Skaf, L. A. Melo Jr, D. Calucci, J. A. Cardillo, J. C. Castro, D. Huang, and M. Wojtkowski, "Retinal assessment using optical coherence tomography," *Progress in Retinal and Eye Research*, vol. 25, no. 3, pp. 325–353, 2006.

[21] A. G. Podoleanu, "Optical coherence tomography," *Journal of Microscopy*, vol. 247, no. 3, pp. 209–219, 2012.

[22] Z. Yaqoob, J. Wu, and C. Yang, "Spectral domain optical coherence tomography: a better OCT imaging strategy," *Biotechniques*, vol. 39, no. 6, pp. S6–S13, 2005.

[23] R. J. Ross, V. Verma, K. I. Rosenberg, C.-C. Chan, and J. Tuo, "Genetic markers and biomarkers for age-related macular degeneration," *Expert Review of Ophthalmology*, vol. 2, no. 3, pp. 443–457, 2007.

[24] M. Fleckenstein, P. Mitchell, K. B. Freund, S. Sadda, F. G. Holz, C. Brittain, E. C. Henry, and D. Ferrara, "The progression of geographic atrophy secondary to age-related macular degeneration," *Ophthalmology*, vol. 125, no. 3, pp. 369–390, 2018.

[25] M. P. S. Group *et al.*, "Subfoveal neovascular lesions in age-related macular degeneration. guidelines for evaluation and treatment in the macular photocoagulation study," *Archives of Ophthalmology*, vol. 109, pp. 1242–1257, 1991.

[26] R. Zhao, A. Camino, J. Wang, A. M. Hagag, Y. Lu, S. T. Bailey, C. J. Flaxel, T. S. Hwang, D. Huang, D. Li *et al.*, "Automated drusen detection in dry age-related macular degeneration by multiple-depth, en face optical coherence tomography," *Biomedical Optics Express*, vol. 8, no. 11, pp. 5049–5064, 2017.

[27] Y. Kanagasingam, A. Bhuiyan, M. D. Abramoff, R. T. Smith, L. Goldschmidt, and T. Y. Wong, "Progress on retinal image analysis for age related macular degeneration," *Progress in Retinal and Eye Research*, vol. 38, pp. 20–42, 2014.

[28] S. R. Cohen and T. W. Gardner, "Diabetic retinopathy and diabetic macular edema," in *Retinal Pharmacotherapeutics*. Karger Publishers, 2016, vol. 55, pp. 137–146.

[29] P. Romero-Aroca, M. Baget-Bernaldiz, A. Pareja-Rios, M. Lopez-Galvez, R. Navarro-Gil, and R. Verges, "Diabetic macular edema pathophysiology: vasogenic versus inflammatory," *Journal of Diabetes Research*, vol. 2016, 2016.

[30] P. Sudeep, S. I. Niwas, P. Palanisamy, J. Rajan, Y. Xiaojun, X. Wang, Y. Luo, and L. Liu, "Enhancement and bias removal of optical coherence tomography images: an iterative approach with adaptive bilateral filtering," *Computers in Biology and Medicine*, vol. 71, pp. 97–107, 2016.

[31] Y. Ma, X. Chen, W. Zhu, X. Cheng, D. Xiang, and F. Shi, "Speckle noise reduction in optical coherence tomography images based on edge-sensitive cGAN," *Biomedical Optics Express*, vol. 9, no. 11, pp. 5129–5146, 2018.

[32] S. Yun, G. Tearney, J. De Boer, and B. Bouma, "Motion artifacts in optical coherence tomography with frequency-domain ranging," *Optics Express*, vol. 12, no. 13, pp. 2977–2998, 2004.

[33] L. Fang, S. Li, D. Cunefare, and S. Farsiu, "Segmentation based sparse reconstruction of optical coherence tomography images," *IEEE Transactions on Medical Imaging*, vol. 36, no. 2, pp. 407–421, 2017.

[34] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[35] R. Rasti, H. Rabbani, A. Mehridehnavi, and F. Hajizadeh, "Macular OCT classification using a multi-scale convolutional neural network ensemble," *IEEE Transactions on Medical Imaging*, vol. 37, no. 4, pp. 1024–1034, 2018.

[36] P. P. Srinivasan, L. A. Kim, P. S. Mettu, S. W. Cousins, G. M. Comer, J. A. Izatt, and S. Farsiu, "Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images," *Biomedical Optics Express*, vol. 5, no. 10, pp. 3568–3577, 2014.

[37] S. P. K. Karri, D. Chakraborty, and J. Chatterjee, "Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration," *Biomedical Optics Express*, vol. 8, no. 2, pp. 579–592, 2017.

[38] J. J. Gómez-Valverde, C. Sinz, E. A. Rank, Z. Chen, A. Santos, W. Drexler, and M. J. Ledesma-Carbayo, "Adaptive compounding speckle-noise-reduction filter for optical coherence tomography images," *Journal of biomedical optics*, vol. 26, no. 6, p. 065001, 2021.

[39] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2. IEEE, 2005, pp. 60–65.

[40] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Transactions on Image Processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[41] M. Tajmirriahi, R. Kafieh, Z. Amini, and H. Rabbani, "A lightweight mimic convolutional auto-encoder for denoising retinal optical coherence tomography images," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–8, 2021.

[42] N. A. Kande, R. Dakhane, A. Dukkipati, and P. K. Yalavarthy, "Siamesegan: a generative model for denoising of spectral domain optical coherence tomography images," *IEEE Transactions on Medical Imaging*, vol. 40, no. 1, pp. 180–192, 2020.

[43] S. J. Chiu, X. T. Li, P. Nicholas, C. A. Toth, J. A. Izatt, and S. Farsiu, "Automatic segmentation of seven retinal layers in SDOCT images congruent with expert manual segmentation," *Optics Express*, vol. 18, no. 18, pp. 19 413–19 428, 2010.

[44] R. Kafieh, H. Rabbani, M. D. Abramoff, and M. Sonka, "Curvature correction of retinal OCTs using graph-based geometry detection," *Physics in Medicine & Biology*, vol. 58, no. 9, p. 2925, 2013.

[45] L. de Sisternes, N. Simon, R. Tibshirani, T. Leng, and D. L. Rubin, "Quantitative SD-OCT imaging biomarkers as indicators of age-related macular degeneration progression," *Investigative Ophthalmology & Visual Science*, vol. 55, no. 11, pp. 7093–7103, 2014.

[46] S. Farsiu, S. J. Chiu, R. V. O'Connell, F. A. Folgar, E. Yuan, J. A. Izatt, C. A. Toth, A.-R. E. D. S. . A. S. D. O. C. T. S. Group *et al.*, "Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography," *Ophthalmology*, vol. 121, no. 1, pp. 162–172, 2014.

[47] F. G. Venhuizen, B. van Ginneken, F. van Asten, M. J. van Grinsven, S. Fauser, C. B. Hoyng, T. Theelen, and C. I. Sánchez, "Automated staging of age-related macular degeneration using optical coherence tomography," *Investigative Ophthalmology & Visual Science*, vol. 58, no. 4, pp. 2318–2328, 2017.

[48] S. J. Chiu, J. A. Izatt, R. V. O'Connell, K. P. Winter, C. A. Toth, and S. Farsiu, "Validated automatic segmentation of amd pathology including drusen and geographic atrophy in sd-oct images," *Investigative ophthalmology & visual science*, vol. 53, no. 1, pp. 53–61, 2012.

[49] J. Y. Lee, S. J. Chiu, P. P. Srinivasan, J. A. Izatt, C. A. Toth, S. Farsiu, and G. J. Jaffe, "Fully automatic software for retinal thickness in eyes with diabetic macular edema from images acquired by cirrus and spectralis systems," *Investigative ophthalmology & visual science*, vol. 54, no. 12, pp. 7595–7602, 2013.

[50] P. A. Dufour, L. Ceklic, H. Abdillahi, S. Schroder, S. De Dzanet, U. Wolf-Schnurrbusch, and J. Kowal, "Graph-based multi-surface segmentation of oct data using trained hard and soft constraints," *IEEE transactions on medical imaging*, vol. 32, no. 3, pp. 531–543, 2012.

[51] S. J. Chiu, M. J. Allingham, P. S. Mettu, S. W. Cousins, J. A. Izatt, and S. Farsiu, "Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema," *Biomedical optics express*, vol. 6, no. 4, pp. 1172–1194, 2015.

[52] L. Fang, C. Wang, S. Li, J. Yan, X. Chen, and H. Rabbani, "Automatic classification of retinal three-dimensional optical coherence tomography images using principal component analysis network with composite kernels," *Journal of Biomedical Optics*, vol. 22, no. 11, p. 116011, 2017.

[53] A. G. Roy, S. Conjeti, S. P. K. Karri, D. Sheet, A. Katouzian, C. Wachinger, and N. Navab, "Relaynet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks," *Biomedical Optics Express*, vol. 8, no. 8, pp. 3627–3642, 2017.

[54] Q. Li, S. Li, Z. He, H. Guan, R. Chen, Y. Xu, T. Wang, S. Qi, J. Mei, and W. Wang, "Deepretina: Layer segmentation of retina in oct images using deep learning," *Translational Vision Science & Technology*, vol. 9, no. 2, pp. 61–61, 2020.

[55] S. Apostolopoulos, S. De Zanet, C. Ciller, S. Wolf, and R. Sznitman, "Pathological oct retinal layer segmentation using branch residual u-shape networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 294–301.

[56] L. Fang, Y. Jin, L. Huang, S. Guo, G. Zhao, and X. Chen, "Iterative fusion convolutional neural networks for classification of optical coherence tomography images," *Journal of Visual Communication and Image Representation*, vol. 59, pp. 327–333, 2019.

[57] S. S. Mishra, B. Mandal, and N. Puhan, "Multi-level dual-attention based CNN for macular optical coherence tomography classification," *IEEE Signal Processing Letters*, vol. 26, no. 12, pp. 1793–1797, 2019.

[58] L. Fang, C. Wang, S. Li, H. Rabbani, X. Chen, and Z. Liu, "Attention to lesion: Lesion-aware convolutional neural network for retinal optical coherence tomography image classification," *IEEE Transactions on Medical Imaging*, 2019.

[59] G. Lemaître, M. Rastgoo, J. Massich, C. Y. Cheung, T. Y. Wong, E. Lamoureux, D. Milea, F. Mériaudeau, and D. Sidibé, "Classification of SD-OCT volumes using local binary patterns: experimental validation for dme detection," *Journal of ophthalmology*, vol. 2016, 2016.

[60] A. Albarrak, F. Coenen, Y. Zheng *et al.*, "Age-related macular degeneration identification in volumetric optical coherence tomography using decomposition and local feature extraction," in *Proceedings of International Conference on Medical Image, Understanding and Analysis*, 2013, pp. 59–64.

[61] Y. Sun, S. Li, and Z. Sun, "Fully automated macular pathology detection in retina optical coherence tomography images using sparse coding and dictionary learning," *Journal of Biomedical Optics*, vol. 22, no. 1, p. 016012, 2017.

[62] P. Dash and A. Sigappi, "Detection and classification of retinal diseases in spectral domain optical coherence tomography images based on SURF descriptors," in *IEEE International Conference on System, Computation, Automation and Networking (ICSCA)*. IEEE, 2018, pp. 1–6.

[63] F. G. Venhuizen, B. van Ginneken, B. Bloemen, M. J. van Grinsven, R. Philipsen, C. Hoyng, T. Theelen, and C. I. Sánchez, "Automated age-related macular degeneration classification in OCT using unsupervised feature learning," in *Medical Imaging 2015: Computer-Aided Diagnosis*, vol. 9414. International Society for Optics and Photonics, 2015, p. 94141I.

[64] D. Sidibe, S. Sankar, G. Lemaitre, M. Rastgoo, J. Massich, C. Y. Cheung, G. S. Tan, D. Milea, E. Lamoureux, T. Y. Wong *et al.*, "An anomaly detection approach for the identification of DME patients using spectral domain optical coherence tomography images," *Computer Methods and Programs in Biomedicine*, vol. 139, pp. 109–117, 2017.

[65] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[66] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The Handbook of Brain Theory and Neural Networks*, vol. 3361, no. 10, p. 1995, 1995.

[67] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE Transactions on Neural Networks*, vol. 8, no. 1, pp. 98–113, 1997.

[68] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[69] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[70] N. O'Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. V. Hernandez, L. Krpalkova, D. Riordan, and J. Walsh, "Deep learning vs. traditional computer vision," in *Science and Information Conference*. Springer, 2019, pp. 128–144.

[71] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT Press Cambridge, 2016, vol. 1.

[72] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Cision and Pattern Recognition*, 2015, pp. 1–9.

[73] F. Li, H. Chen, Z. Liu, X.-d. Zhang, M.-s. Jiang, Z.-z. Wu, and K.-q. Zhou, "Deep learning-based automated detection of retinal diseases using optical coherence tomography images," *Biomedical Optics Express*, vol. 10, no. 12, pp. 6204–6226, 2019.

[74] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

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

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

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

[78] S. Kaymak and A. Serener, "Automated age-related macular degeneration and diabetic macular edema detection on OCT images using deep learning," in *IEEE 14th International Conference on Intelligent Computer Communication and Processing (ICCP)*. IEEE, 2018, pp. 265–269.

[79] G. C. Chan, S. A. Shah, T. Tang, C.-K. Lu, H. Muller, and F. Meriaudeau, "Deep features and data reduction for classification of SD-OCT images: Application to diabetic macular edema," in *International Conference on Intelligent and Advanced System (ICIAS)*. IEEE, 2018, pp. 1–4.

[80] G. C. Chan, A. Muhammad, S. A. Shah, T. B. Tang, C.-K. Lu, and F. Meriaudeau, "Transfer learning for diabetic macular edema (DME) detection on optical coherence tomography (OCT) images," in *IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. IEEE, 2017, pp. 493–496.

[81] G. An, H. Yokota, N. Motozawa, S. Takagi, M. Mandai, S. Kitahata, Y. Hirami, M. Takahashi, Y. Kurimoto, and M. Akiba, "Deep learning classification models built with two-step transfer learning for age related macular degeneration diagnosis," in *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 2049–2052.

[82] M. Awais, H. Müller, T. B. Tang, and F. Meriaudeau, "Classification of SD-OCT images using a deep learning approach," in *IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. IEEE, 2017, pp. 489–492.

[83] J. de Moura, J. Novo, and M. Ortega, "Deep feature analysis in a transfer learning-based approach for the automatic identification of diabetic macular edema," in *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[84] W. Lu, Y. Tong, Y. Yu, Y. Xing, C. Chen, and Y. Shen, "Deep learning-based automated classification of multi-categorical abnormalities from optical coherence tomography images," *Translational Vision Science & Technology*, vol. 7, no. 6, pp. 41–41, 2018.

[85] R. M. Kamble, G. C. Chan, O. Perdomo, M. Kokare, F. A. Gonzalez, H. Müller, and F. Mériaudeau, "Automated diabetic macular edema (DME) analysis using fine tuning with inception-resnet-v2 on OCT images," in *IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*. IEEE, 2018, pp. 442–446.

[86] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[87] N. Motozawa, G. An, S. Takagi, S. Kitahata, M. Mandai, Y. Hirami, H. Yokota, M. Akiba, A. Tsujikawa, M. Takahashi *et al.*, "Optical coherence tomography-based deep-learning models for classifying normal and age-related macular degeneration and exudative and non-exudative age-related macular degeneration changes," *Ophthalmology and Therapy*, vol. 8, no. 4, pp. 527–539, 2019.

[88] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.

[89] L. Huang, X. He, L. Fang, H. Rabbani, and X. Chen, "Automatic classification of retinal optical coherence tomography images with layer guided convolutional neural network," *IEEE Signal Processing Letters*, vol. 26, no. 7, pp. 1026–1030, 2019.

[90] G. Lemaitre, M. Rastgoo, J. Massich, S. Sankar, F. Mériaudeau, and D. Sidibé, "Classification of SD-OCT volumes with LBP: application toDME detection," 2015.

[91] O. Perdomo, S. Otálora, F. A. González, F. Meriaudeau, and H. Müller, "OCT-NET: A convolutional network for automatic classification of normal and diabetic macular edema using SD-OCT volumes," in *IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 1423–1426.

[92] X. He, L. Fang, H. Rabbani, X. Chen, and Z. Liu, "Retinal optical coherence tomography image classification with label smoothing generative adversarial network," *Neurocomputing*, 2020.

[93] S. Apostolopoulos, C. Ciller, S. De Zanet, S. Wolf, and R. Sznitman, "Retinet: Automatic amd identification in OCT volumetric data," *Investigative Ophthalmology & Visual Science*, vol. 58, no. 8, pp. 387–387, 2017.

[94] X. Wang, F. Tang, H. Chen, L. Luo, Z. Tang, A.-R. Ran, C. Y. Cheung, and P. A. Heng, "UD-MIL: uncertainty-driven deep multiple instance learning for OCT image classification," *IEEE Journal of Biomedical and Health Informatics*, 2020.

[95] D. E. Romo-Bucheli, U. Schmidt-Erfurth, and H. Bogunovic, "End-to-end deep learning model for predicting treatment requirements in neovascular AMD from longitudinal retinal OCT imaging," *IEEE Journal of Biomedical and Health Informatics*, 2020.

[96] R. Rasti, M. J. Allingham, P. S. Mettu, S. Kavusi, K. Govind, S. W. Cousins, and S. Farsiu, "Deep learning-based single-shot prediction of differential effects of anti-VEGF treatment in patients with diabetic macular edema," *Biomedical Optics Express*, vol. 11, no. 2, pp. 1139–1152, 2020.

[97] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[98] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature Medicine*, vol. 24, no. 9, pp. 1342–1350, 2018.

[99] J. Qiu and Y. Sun, "Self-supervised iterative refinement learning for macular OCT volumetric data classification," *Computers in Biology and Medicine*, vol. 111, p. 103327, 2019.

[100] M. D. Robinson, S. J. Chiu, J. Lo, C. Toth, J. Izatt, and S. Farsiu, "New applications of super-resolution in medical imaging," *Super-Resolution Imaging*, vol. 2010, pp. 384–412, 2010.

[101] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[102] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[103] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *IEEE 12th International Conference on Computer Vision*. IEEE, 2009, pp. 349–356.

[104] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Vattern Recognition*, 2017, pp. 4681–4690.

[105] A. Abbasi, A. Monadjemi, L. Fang, and H. Rabbani, "Optical coherence tomography retinal image reconstruction via nonlocal weighted sparse representation," *Journal of Biomedical Optics*, vol. 23, no. 3, p. 036011, 2018.

[106] M. Asif, M. U. Akram, T. Hassan, A. Shaukat, and R. Waqar, "High resolution OCT image generation using super resolution via sparse representation," in *Eighth International Conference on Graphic and Image Processing (ICGIP 2016)*, vol. 10225. International Society for Optics and Photonics, 2017, p. 1022512.

[107] E. Bousi and C. Pitris, "Lateral resolution improvement in optical coherence tomography (OCT) images," in *IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE)*. IEEE, 2012, pp. 598–601.

[108] P. G. Daneshmand, A. Mehridehnavi, and H. Rabbani, "Reconstruction of optical coherence tomography images using mixed low rank approximation and second order tensor based total variation method," *IEEE Transactions on Medical Imaging*, 2020.

[109] Y. Huang, Z. Lu, Z. Shao, M. Ran, J. Zhou, L. Fang, and Y. Zhang, "Simultaneous denoising and super-resolution of optical coherence tomography images based on generative adversarial network," *Optics Express*, vol. 27, no. 9, pp. 12 289–12 307, 2019.

[110] M. I. Razzak, S. Naz, and A. Zaib, "Deep learning for medical image processing: Overview, challenges and the future," *Classification in Biomedical Image Applications*, pp. 323–350, 2018.

[111] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, pp. 2672–2680, 2014.

[112] V. Das, S. Dandapat, and P. K. Bora, "A data-efficient approach for automated classification of OCT images using generative adversarial network," *IEEE Sensors Letters*, vol. 4, no. 1, pp. 1–4, 2020.

[113] B. Antony, L. Tang, M. Abràmoff, K. Lee, M. Sonka, and M. Garvin, "Automated method for the flattening of optical coherence tomography images," *Investigative Ophthalmology & Visual Science*, vol. 51, no. 13, pp. 1781–1781, 2010.

[114] P. Burt and E. Adelson, "The laplacian pyramid as a compact image code," *IEEE Transactions on Communications*, vol. 31, no. 4, pp. 532–540, 1983.

[115] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[116] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.

[117] A. F. Agarap, "Deep learning using rectified linear units (ReLU)," *arXiv preprint arXiv:1803.08375*, 2018.

[118] K. Janocha and W. M. Czarnecki, "On loss functions for deep neural networks in classification," *arXiv preprint arXiv:1702.05659*, 2017.

[119] Y. Ho and S. Wookey, "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," *IEEE Access*, vol. 8, pp. 4806–4813, 2019.

[120] S. Panchapagesan, M. Sun, A. Khare, S. Matsoukas, A. Mandal, B. Hoffmeister, and S. Vitaladevuni, "Multi-task learning and weighted cross-entropy for DNN-based keyword spotting." in *Interspeech*, vol. 9, 2016, pp. 760–764.

[121] H. Phan, M. Krawczyk-Becker, T. Gerkmann, and A. Mertins, "DNN and CNN with weighted and multi-task loss functions for audio event detection," *arXiv preprint arXiv:1708.03211*, 2017.

[122] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[123] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[124] M. Kubat, R. Holte, and S. Matwin, "Learning when negative examples abound," in *European Conference on Machine Learning*. Springer, 1997, pp. 146–153.

[125] J. Du, C.-M. Vong, C.-M. Pun, P.-K. Wong, and W.-F. Ip, "Post-boosting of classification boundary for imbalanced data using geometric mean," *Neural Networks*, vol. 96, pp. 101–114, 2017.

[126] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[127] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[128] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[129] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[130] S. G. Schuman, A. F. Koreishi, S. Farsiu, S.-h. Jung, J. A. Izatt, and C. A. Toth, "Photoreceptor layer thinning over drusen in eyes with age-related macular degeneration imaged in vivo with spectral-domain optical coherence tomography," *Ophthalmology*, vol. 116, no. 3, pp. 488–496, 2009.

[131] A. Oishi, M. Hata, M. Shimozono, M. Mandai, A. Nishida, and Y. Kurimoto, "The significance of external limiting membrane status for visual acuity in age-related macular degeneration," *American Journal of Ophthalmology*, vol. 150, no. 1, pp. 27–32, 2010.

[132] A. Ebneter, D. Jaggi, M. Abegg, S. Wolf, and M. S. Zinkernagel, "Relationship between presumptive inner nuclear layer thickness and geographic atrophy progression in age-related macular degeneration," *Investigative ophthalmology & visual science*, vol. 57, no. 9, pp. OCT299–OCT306, 2016.

[133] K. Shen, H. Lu, S. Baig, and M. R. Wang, "Improving lateral resolution and image quality of optical coherence tomography by the multi-frame superresolution technique for 3d tissue imaging," *Biomedical Optics Express*, vol. 8, no. 11, pp. 4887–4918, 2017.

[134] R. de Kinkelder, J. Kalkman, D. J. Faber, O. Schraa, P. H. Kok, F. D. Verbraak, and T. G. van Leeuwen, "Heartbeat-induced axial motion artifacts in optical coherence tomography measurements of the retina," *Investigative Ophthalmology & Visual Science*, vol. 52, no. 6, pp. 3908–3913, 2011.

[135] S. Adiga and J. Sivaswamy, "Shared encoder based denoising of optical coherence tomography images." 2018.

[136] A. Madani, J. R. Ong, A. Tibrewal, and M. R. Mofrad, "Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease," *NPJ Digital Medicine*, vol. 1, no. 1, pp. 1–11, 2018.

[137] J. Ambati and B. J. Fowler, "Mechanisms of age-related macular degeneration," *Neuron*, vol. 75, no. 1, pp. 26–39, 2012.

[138] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[139] S. Bell-Kligler, A. Shocher, and M. Irani, "Blind super-resolution kernel estimation using an internal-gan," *Advances in Neural Information Processing Systems*, vol. 32, pp. 284–293, 2019.

[140] A. Bulat, J. Yang, and G. Tzimiropoulos, "To learn image super-resolution, use a gan to learn how to do image degradation first," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 185–200.

[141] N. C. Rakotonirina and A. Rasoanaivo, "ESRGAN+: Further improving enhanced super-resolution generative adversarial network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3637–3641.

[142] H. N. Pathak, X. Li, S. Minaee, and B. Cowan, "Efficient super resolution for large-scale images using attentional gan," in *IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 1777–1786.

[143] Y. Xie, E. Franz, M. Chu, and N. Thuerey, "tempoGAN: A temporally coherent, volumetric GAN for super-resolution fluid flow," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–15, 2018.

[144] Y. Yan, C. Liu, C. Chen, X. Sun, L. Jin, and X. Zhou, "Fine-grained attention and feature-sharing generative adversarial networks for single image super-resolution," *arXiv preprint arXiv:1911.10773*, 2019.

[145] W. Liu, X. Liu, H. Ma, and P. Cheng, "Beyond human-level license plate super-resolution with progressive vehicle search and domain priori GAN," in *Proceedings of the 25th ACM International Conference on Multimedia*, 2017, pp. 1618–1626.

[146] J. Cai, H. Hu, S. Shan, and X. Chen, "FCSR-GAN: End-to-end learning for joint face completion and super-resolution," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–8.

[147] M. M. Majdabadi and S.-B. Ko, "Capsule GAN for robust face super resolution," *Multimedia Tools and Applications*, vol. 79, no. 41, pp. 31 205–31 218, 2020.

[148] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[149] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin, "Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 701–710.

[150] L. Fang, S. Li, Q. Nie, J. A. Izatt, C. A. Toth, and S. Farsiu, "Sparsity based denoising of spectral domain optical coherence tomography images," *Biomedical Optics Express*, vol. 3, no. 5, pp. 927–942, 2012.

[151] M. Lee, J. Izatt, E. A. Swanson, D. Huang, J. Schumun, C. Lin, C. Puliafito, and J. Fujimoto, "Optical coherence tomography for ophthalmic imaging: new technique delivers micron-scale resolution," *IEEE Engineering in Medicine and Biology Magazine*, vol. 14, no. 1, pp. 67–76, 1995.

[152] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," *arXiv preprint arXiv:1603.07285*, 2016.

[153] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[154] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[155] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015.

[156] M. K. Farazdaghi and K. B. Ebrahimi, "Role of the choroid in age-related macular degeneration: a current review," *Journal of ophthalmic & vision research*, vol. 14, no. 1, p. 78, 2019.

[157] T. J. Gin, Z. Wu, S. K. Chew, R. H. Guymer, and C. D. Luu, "Quantitative analysis of the ellipsoid zone intensity in phenotypic variations of intermediate age-related macular degeneration," *Investigative Ophthalmology & Visual Science*, vol. 58, no. 4, pp. 2079–2086, 2017.

[158] C. Cukras, Y. D. Wang, C. B. Meyerle, F. Forooghian, E. Y. Chew, and W. T. Wong, "Optical coherence tomography-based decision making in exudative age-related macular degeneration: comparison of time-vs spectral-domain devices," *Eye*, vol. 24, no. 5, pp. 775–783, 2010.

[159] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136–144.

[160] J. Yu, Y. Fan, J. Yang, N. Xu, Z. Wang, X. Wang, and T. Huang, "Wide activation for efficient and accurate image super-resolution," *arXiv preprint arXiv:1808.08718*, 2018.

[161] A. Shocher, N. Cohen, and M. Irani, ""zero-shot" super-resolution using deep internal learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3118–3126.

[162] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[163] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," *arXiv preprint arXiv:1509.00685*, 2015.

[164] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *Advances in Neural Information Processing Systems*, vol. 28, pp. 577–585, 2015.

[165] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, 2015, pp. 2048–2057.

[166] L. Meng, B. Zhao, B. Chang, G. Huang, F. Tung, and L. Sigal, "Where and when to look? spatial-temporal attention for action recognition in videos," 2018.

[167] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156–3164.

[168] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[169] D. Zhang, L. Yao, K. Chen, and J. Monaghan, "A convolutional recurrent attention model for subject-independent EEG signal analysis," *IEEE Signal Processing Letters*, vol. 26, no. 5, pp. 715–719, 2019.

[170] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[171] S. Maetschke, B. Antony, H. Ishikawa, G. Wollstein, J. Schuman, and R. Garnavi, "A feature agnostic approach for glaucoma detection in OCT volumes," *PloS One*, vol. 14, no. 7, p. e0219126, 2019.

[172] *Eye Conditions and diseases*, (accessed January 11, 2021). [Online]. Available: https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases

[173] N. De Souza, Y. Cui, S. Looi, P. Paudel, L. Shinde, K. Kumar, R. Berwal, R. Wadhwa, V. Daniel, J. Flanagan *et al.*, "The role of optometrists in india: An integral part of an eye health team," *Indian Journal of Ophthalmology*, vol. 60, no. 5, p. 401, 2012.

[174] A. Odena, "Semi-supervised learning with generative adversarial networks," *arXiv preprint arXiv:1606.01583*, 2016.

[175] Y. Li, Y. Iwamoto, L. Lin, R. Xu, and Y.-W. Chen, "Volumenet: A lightweight parallel network for super-resolution of medical volumetric data," *arXiv preprint arXiv:2010.08357*, 2020.

[176] N. A. Nathoo, C. Or, M. Young, L. Chui, N. Fallah, A. W. Kirker, D. A. Albiani, A. B. Merkur, and F. Forooghian, "Optical coherence tomography–based measurement of drusen load predicts development of advanced age-related macular degeneration," *American Journal of Ophthalmology*, vol. 158, no. 4, pp. 757–761, 2014.

[177] D. Ferrara, R. E. Silver, R. N. Louzada, E. A. Novais, G. K. Collins, and J. M. Seddon, "Optical coherence tomography features preceding the onset of advanced age-related macular degeneration," *Investigative Ophthalmology & Visual Science*, vol. 58, no. 9, pp. 3519–3529, 2017.

[178] H. Cook, P. Patel, and A. Tufail, "Age-related macular degeneration: diagnosis and management," *British Medical Bulletin*, vol. 85, no. 1, pp. 127–149, 2008.

[179] C. Shen, S. Yan, M. Du, H. Zhao, L. Shao, and Y. Hu, "Assessment of capillary dropout in the superficial retinal capillary plexus by optical coherence tomography angiography in the early stage of diabetic retinopathy," *BMC ophthalmology*, vol. 18, no. 1, p. 113, 2018.

[180] B. H. Najeeb, C. Simader, G. Deak, C. Vass, J. Gamper, A. Montuoro, B. S. Gerendas, and U. Schmidt-Erfurth, "The distribution of leakage on fluorescein angiography in diabetic macular edema: a new approach to its etiology," *Investigative Ophthalmology & Visual Science*, vol. 58, no. 10, pp. 3986–3990, 2017.

[181] J. Y. Kim, M. Y. Choi, E. J. Seo, S. Lee, J. S. Kim, J. B. Chae, D. Y. Kim, and J.-G. Kim, "Preliminary study of ultra-widefield peripheral retinal angiographic patterns in children and its association to the perinatal condition," *Scientific Reports*, vol. 10, no. 1, pp. 1–6, 2020.

[182] L. Laatikainen, "The fluorescein angiography revolution: a breakthrough with sustained impact," *Acta Ophthalmologica Scandinavica*, vol. 82, no. 4, pp. 381–392, 2004.

[183] H. Al-Khersan, J. F. Russell, T. A. Lazzarini, N. L. Scott, J. W. Hinkle, N. A. Patel, N. A. Yannuzzi, B. J. Fowler, R. M. Hussain, A. Barikian *et al.*, "Comparison between graders in detection of diabetic neovascularization with swept source OCT angiography and fluorescein angiography," *American Journal of Ophthalmology*, 2020.

[184] T. E. De Carlo, A. Romano, N. K. Waheed, and J. S. Duker, "A review of optical coherence tomography angiography (OCTA)," *International Journal of Retina and Vitreous*, vol. 1, no. 1, p. 5, 2015.

[185] M. Inoue, J. J. Jung, C. Balaratnasingam, K. K. Dansingani, E. Dhrami-Gavazi, M. Suzuki, E. Talisa, A. Shahlaee, M. A. Klufas, A. El Maftouhi *et al.*, "A comparison between optical coherence tomography angiography and fluorescein angiography for the imaging of type 1 neovascularization," *Investigative Ophthalmology & Visual Science*, vol. 57, no. 9, pp. OCT314–OCT323, 2016.

[186] H. Davson, *Physiology of the Eye*. Macmillan International Higher Education, 1990.

[187] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into imaging*, vol. 9, no. 4, pp. 611–629, 2018.

[188] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[189] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational Intelligence and Neuroscience*, vol. 2018, 2018.

[190] D. Scherer, A. Müller, and S. Behnke, "Evaluation of pooling operations in convolutional architectures for object recognition," in *International conference on Artificial Neural Networks*. Springer, 2010, pp. 92–101.

[191] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

List of Publications Related to Thesis

Journal Publications

1. **Vineeta Das**, S. Dandapat and P. K. Bora, "Automated Classification of Retinal OCT Images using a Deep Multi-Scale Fusion CNN," *IEEE Sensors Journal*, 2021.

2. **Vineeta Das**, E. Prabhakararao, S. Dandapat and P. K. Bora, "B-Scan Attentive CNN for the Classification of Retinal Optical Coherence Tomography Volumes," *IEEE Signal Processing Letters*, vol. 27, pp. 1025 - 1029, 2020.

3. **Vineeta Das**, S. Dandapat and P. K. Bora, "Unsupervised Super-Resolution of OCT Images Using Generative Adversarial Network for Improved Age-Related Macular Degeneration Diagnosis," *IEEE Sensors Journal*, vol. 20, pp. 8746 - 8756, 2020.

4. **Vineeta Das**, S. Dandapat and P. K. Bora, "Multi-scale deep feature fusion for automated classification of macular pathologies from OCT images," *Biomedical Signal Processing and Control*, vol. 110, pp. 76-89, 2019.

Conference Publications

1. **Vineeta Das**, S. Dandapat and P. K. Bora, "Diagnostic Information based Super-Resolution of Retinal Optical Coherence Tomography Images," in *Proceedings of the International Conference on Signal Processing and Communications (SPCOM)*, pp. 182-186, 2020.

2. Simran Barnwal, **Vineeta Das**, and P. K. Bora, "Deep Learning Based Fully Automated Decision Making for Intravitreal Anti-VEGF Therapy," in *Proceedings of the International Conference on Pattern Recognition and Machine Intelligence*, pp. 147-155, 2019.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

List of Other Publications

1. **Vineeta Das**, S. Dandapat and P. K. Bora, "A diagnostic information based framework for super-resolution and quality assessment of retinal OCT images," in *Computerized Medical Imaging and Graphics*, 2021.

2. **Vineeta Das**, S. Dandapat and P. K. Bora, "A Data-Efficient Approach for Automated Classification of OCT Images Using Generative Adversarial Network," in *IEEE Sensors Letters*, vol. 4, pp. 1 - 4, 2020.

3. **Vineeta Das**, S. Dandapat and P. K. Bora, "A novel diagnostic information based framework for super-resolution of retinal fundus images," in *Computerized Medical Imaging and Graphics*, vol. 72, pp. 22 - 33, 2019.

4. **Vineeta Das**, S. Dandapat and P. K. Bora, "Region Selective Information Augmentation for Retinal Images," in *Proceedings of the National Conference on Communications (NCC)* , pp. 1-5, 2018.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati