Enhanced Predictive Models for Macular Hole Surgery Outcomes

Burak Kucukgoz

School of Computing
Newcastle University

This dissertation is submitted for the degree of
Doctor of Philosophy

May 2025

To my family

Declaration

"I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements."

Burak Kucukgoz
May 2025

Acknowledgements

I would like to express my deepest appreciation to all those who provided me with the possibility to complete this doctoral dissertation. I would not have been able to navigate this challenging yet rewarding journey without your support and guidance.

First and foremost, I would like to extend my sincere gratitude to my supervisor, Professor Boguslaw Obara, whose expertise, patience, and encouragement have been invaluable. Your insightful feedback and unwavering support have guided me through the complexities of my research and inspired me to push the boundaries of my work. Without this, it would not have been possible to complete the study.

I am also grateful to Professor David H. Steel and Dr. Huazhu Fu, my additional supervisors, for their dedicated guidance and invaluable insights throughout this research journey. Their helpfulness and encouragement made the process very rewarding.

In addition, I would like to express my sincere thanks to the Ministry of National Education of the Turkish Republic for providing me with the scholarship to cover the expenses and pursue my Ph.D. studies in the United Kingdom.

My deepest gratitude goes to my parents and my brother for their unconditional love and belief in my abilities, which have been my greatest source of strength.

A special thanks to my colleagues and friends at Newcastle University, A*STAR, and Durham University. Your intellectual exchange has been a source of motivation and joy. The collaborative environment and shared experiences have made this journey memorable.

Finally, I want to dedicate this doctoral dissertation to my honey, Aysen, for supporting me in everything I did and loving me unconditionally. Without her support, I would have never been able to accomplish this work. Thank you, my love!

Abstract

This thesis presents research work conducted in the field of retinal image analysis. More specifically, the work is directed at the employment of deep learning (DL) based image informatics for the analysis of diverse real world phenomena where features of interest are very difficult to distinguish. The evaluation of idiopathic full-thickness macular holes (MHs) holds critical clinical importance as MHs represent one of the strongest predictors of surgical success, influencing both anatomical closure and functional visual recovery – a key motivation for developing robust deep learning frameworks to quantify their characteristics and predict postoperative outcomes. In this context, three distinct parts to retinal image analysis are proposed. The first part addresses critical research questions on the quantitative assessment of MH, the role of DL in postoperative visual acuity (VA) prediction, the integration of automated optical coherence tomography (OCT) analysis for clinical decision-making, and the potential of DL models to improve diagnostic accuracy and support clinical practices. Hence, this part presents a comprehensive image informatics framework to create a high-quality spectral-domain OCT (SD-OCT) image dataset, providing a robust DL-based predictive model of VA in patients following surgery with MH and presenting an automated solution for non-standardised SD-OCT datasets. The imaging data undergoes preprocessing, quality assurance, and anomaly detection procedures. Seven state-of-the-art DL predictive models are then designed, implemented, trained, and tested with multiple two-dimensional (2D) input channels on the SD-OCT dataset. The models are quantitatively compared using four evaluation metrics. The method concludes the impact of the following surgery by predicting VA. Overall, the obtained results confirm that the fully automated approach with input from seven central SD-OCT images from each patient may robustly predict VA measurements using a high-quality SD-OCT image dataset. Following this, three-dimensional (3D) convolutional neural networks are integrated to train the model. 3D networks generally outperformed the 2D networks in some evaluation metrics; however, it came with the sacrifice of significantly more computational complexity. The second part identifies key research questions related to common sources of uncertainty in OCT images and proposes an effective method for representing and quantifying this uncertainty in DL-based predictive models. Furthermore, the study compares the proposed UQ method with existing approaches. In this context, the study highlights the significance of uncertainty, especially in dealing with the SD-OCT images. Predicting postoperative VA through DL models is crucial for decision-making and patient advisement, though their black-box behaviour is opaque to users and uncertainty associated with their predictions is not typically stated, leading to a lack of trust among clinicians and patients. To meet this need, an uncertainty-aware regression model is introduced for predicting postoperative VA using 3D SD-OCT images. The model not only

predicts VA post-surgery but also quantifies the associated uncertainty, enhancing reliability and trustworthiness. Qualitative evaluation shows that the proposed model outperforms commonly used methods in terms of prediction accuracy and reliability, demonstrating robust performance on out-of-sample data, including low-quality images and previously unseen instances. This makes the model a promising tool for clinical settings, improving the reliability of DL models in predicting VA. The third is the segmentation of the retinal external limiting membrane layer, where any disruptions in this layer are associated with worse visual outcomes in patients with idiopathic full-thickness MHs. Precise image-wise binary annotations are used to segment the retinal external limiting membrane (ELM) layer. Finally, qualitative and quantitative results are systematically compared with seven state-of-the-art DL-based segmentation methods to identify the ELM layer with an automated system. Additionally, it examines the feasibility of integrating automated ELM layer segmentation into clinical workflows while incorporating the latest advancements in DL-based ELM detection. The results confirm the efficacy of DL in retinal image analysis, providing a foundation for future enhancements in clinical applications. Future work will explore enhancing the models' performance and efficiency, and extending the approach to other retinal conditions.

Keywords: Image Analysis, Machine Learning, Deep Learning, Visual Acuity Measurement, Optical Coherence Tomography

Table of Contents

List of Figures

List of Algorithms

List of Tables

Chapter 1

Introduction

Contents

1.1 Introduction

Retinal imaging and retinal image analysis are pivotal in modern ophthalmology. These high-resolution imaging techniques allow for a detailed examination of the retina, which is instrumental in detecting and monitoring various eye conditions [62]. Examining these images presents a significant challenge for ophthalmologists, as it necessitates the precise interpretation of intricate retinal structures and subtle pathological changes (see Fig. 1.1).

The advancement in automated image analysis algorithms supports ophthalmologists in efficiently processing and interpreting the vast amounts of retinal imaging data generated, enhancing diagnostic accuracy and patient care. However, until the early 2000s, computer-assisted retinal imaging did not attract sufficient scientific attention to identify eye diseases accurately [119]. Despite the recent momentum gained by automated image analysis algorithms in improving the early detection and diagnosis of retinal diseases, significant research challenges remain.

This thesis aims to substantially impact this growing field by exploring state-of-the-art predictive modelling using image informatics, deep learning (DL), and uncertainty quantification (UQ). The potential benefits and excitement of this research are palpable.

This chapter will introduce the technology that enables the acquisition of three-dimensional (3D) retinal imagery, the importance of image analysis for clinical applications, an overview of computer-aided diagnosis (CAD) systems, including classical, machine learning, and DL-based image informatics techniques, and the UQ in retinal imaging analysis. It will also outline the research contributions and structure of this thesis.

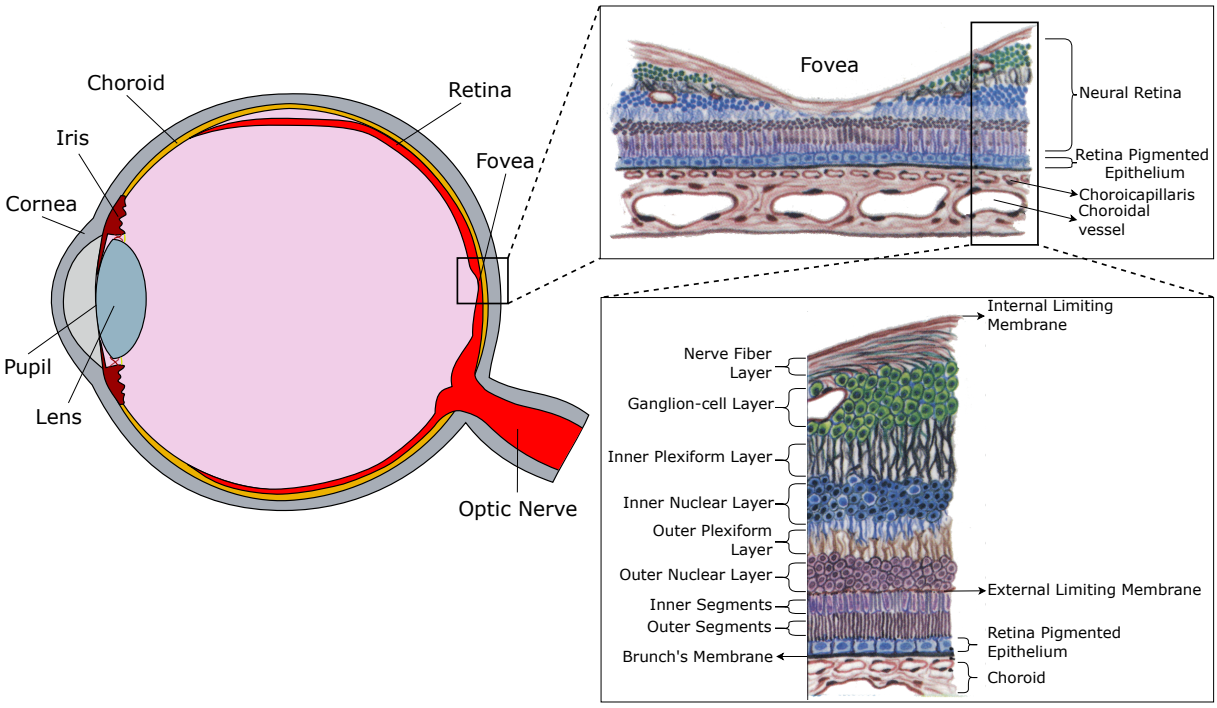


Fig. 1.1 The human eye. A cutaway view of the eye (left) demonstrates retina, fovea, optic nerve, lens, pupil, cornea, iris, and choroid. A cross section through the fovea (top right) indicates neural retina, retina pigmented epithelium, choroicapillaris and choroidal vessel. The enlarged cross section (bottom right) shows retinal layers [37, 38].

1.1.1 Retinal Imaging and Analysis

The anatomy of the human eye is presented in Fig. 1.1, illustrating a cross-sectional view of the human eye with various ocular structures. The human eye operates sequentially: Light enters through the cornea and pupil to reach the lens, which is surrounded by the iris. Subsequently, the lens focuses the light onto the retina at the back of the eye. The retina then converts the captured light into signals transmitted to the brain via the optic nerve. Given the critical role of the retina in vision, it is crucial to monitor its health and detect any abnormalities early on. Changes in the retinal structure can indicate various eye conditions that may lead to severe vision impairment or blindness if left untreated.

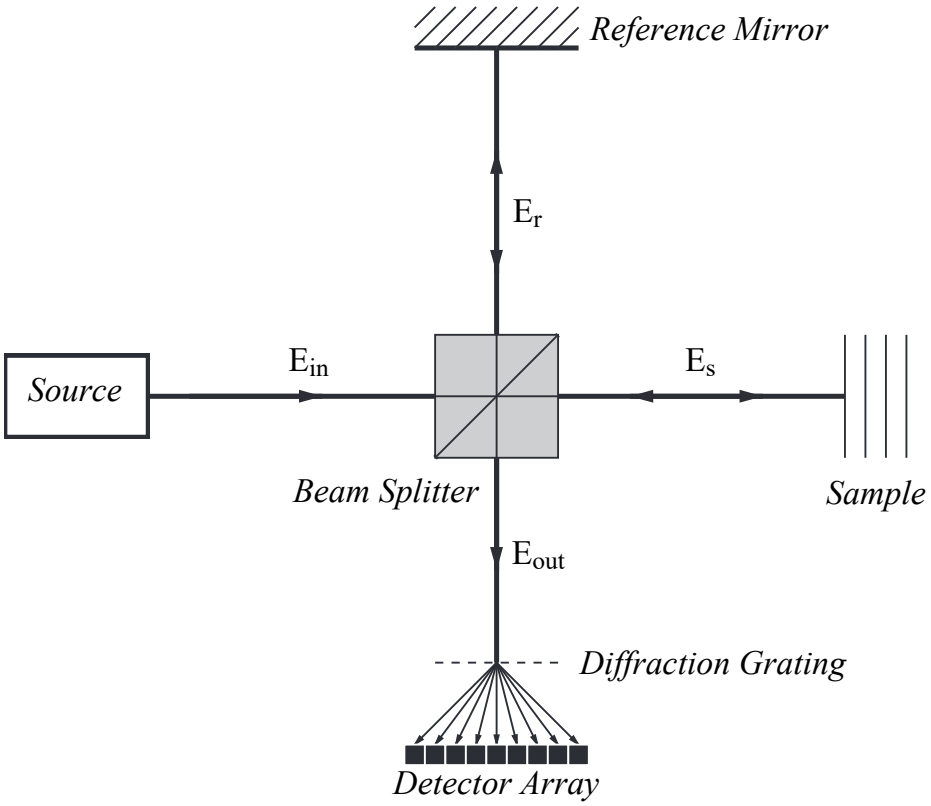


Fig. 1.2 Schematic diagram of spectral-domain optical coherence tomography system, based on a Michelson interferometer [161].

To meet this need, retinal imaging techniques have been developed to effectively capture the morphology of the retina and to model its changes over time. Fundus photography (FP), Optical coherence tomography (OCT) [67], and Fluorescein angiography (FA) are retinal imaging techniques that are widely used to identify any potential issues that may affect sight. Notably, OCT has revolutionised retinal imaging by providing non-invasive cross-sectional two-dimensional (2D) and 3D views of the retina and has proven effective in detecting and monitoring a variety of retinal diseases [62]. OCT resolution has advanced from 10 μm to 200 μm, and imaging speed has increased, reflecting the transition from time-domain OCT to spectral-domain OCT (SD-OCT). A typical SD-OCT schematic is shown in Fig. 1.2. The schematic illustrates key SD-OCT components (reference arm, sample arm, detector, light source). The light coming from source is split by a beam splitter into two paths; the reference mirror and sample-reflected

back and combined at the detector array. Interference occurs when the light travel times match, allowing measurement of depth within the retina. The horizontal SD-OCT image of a normal fovea illustrates retinal layers (Fig. 1.3).

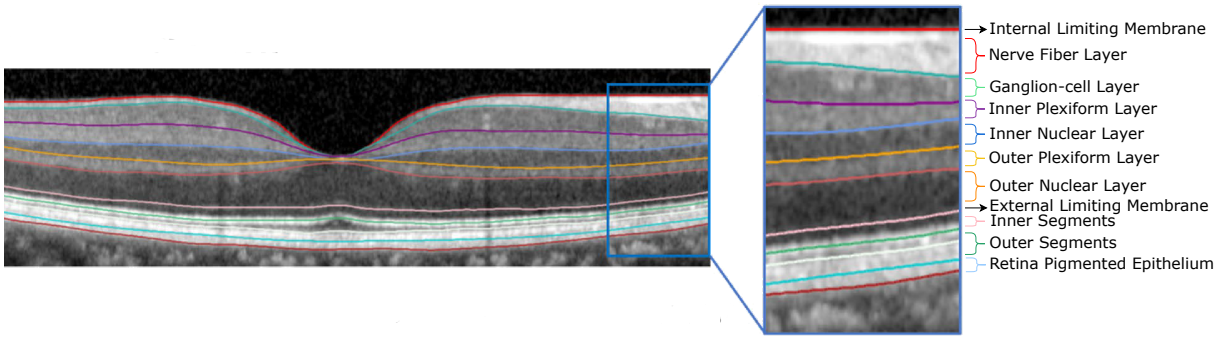
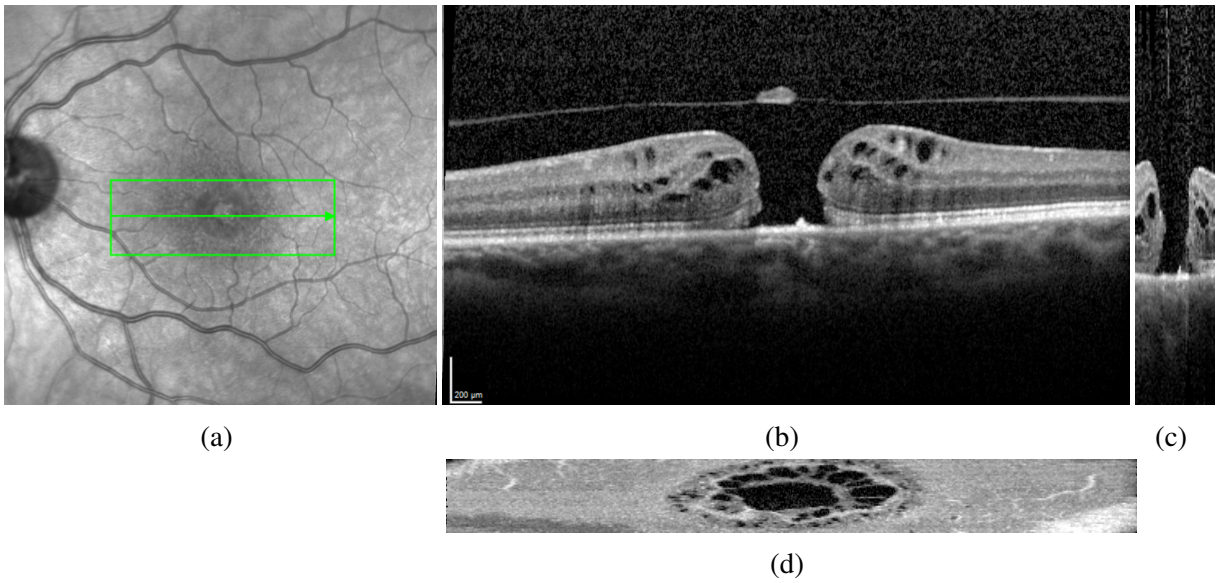


Fig. 1.3 Spectral-domain optical coherence tomography illustration of the various layers of retina [161]. Retina layers from inside are internal limiting membrane, nerve fiber layer, ganglion-cell layer, inner plexiform layer, inner nuclear layer, outer plexiform layer, outer nuclear layer, external limiting membranem, inner and outer segments and retina pigmented epithelium.

Image processing has played a crucial role in retinal imaging. It has been instrumental in analysing retinal layers and diagnosing retinal conditions over the past 15 years [4, 166]. Image processing is, therefore, of particular interest to ophthalmologists. In this way, ophthalmologists utilise a broad range of imaging-based techniques using SD-OCT scans to qualitatively and quantitatively characterise and monitor many retinal diseases, including diabetic retinopathy (DR), central serous chorioretinopathy, vitreomacular interface syndrome, age-related macular degeneration (AMD), diabetic macular edema (DMO), and idiopathic full-thickness macular holes (MH) [33, 117, 138]. Given the importance of accurate retinal imaging and analysis, this thesis focuses explicitly on MHs, a significant retinal condition that can cause severe vision loss (see Fig. 1.4).

MHs are a common vitreoretinal interface abnormality that affects approximately 1 in 200 people aged 60 years or older and can result in significant visual deterioration [6, 111]. MHs are small, typically 400 μm defects in the neurosensory retina at the centre of the fovea, which can be related to diabetes or ageing [144, 149]. Retinal imaging allows ophthalmologists to diagnose, measure and analyse MHs using SD-OCTs [33]. Surgery is the primary treatment for MHs, with pars plana vitrectomy and intraocular gas tamponade being the current gold standard intervention, which achieves successful postoperative MH closure in more than 90% of cases [107, 147]. After successful hole closure, visual acuity (VA) usually improves by a mean of three lines of vision. However, the final VA achieved can be variable, with, for instance, only 35-40% achieving a United Kingdom (UK) driving level of vision after otherwise successful surgery [73, 147] (see Figs. 1.5). This thesis initially seeks to evaluate the VA using preoperative SD-OCT images to ascertain if the potential benefits of surgical intervention outweigh the associated risks.

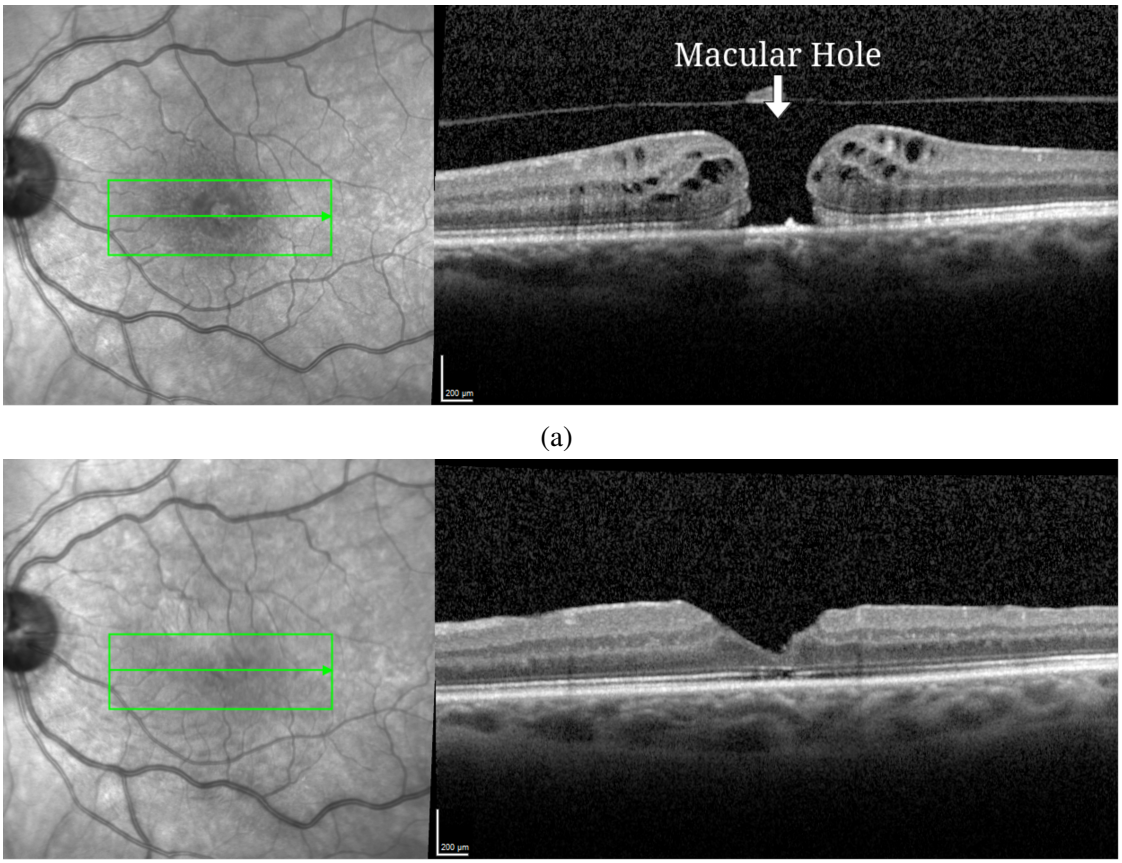
Anatomical and visual outcomes of MH closure surgery have recently improved with retinal imaging and analysis using SD-OCT devices. SD-OCT devices easily reveal distinct hyper-reflective layers corresponding to the external limiting membrane (ELM), the inner segments (IS), the outer segments (OS), and the retinal pigment epithelium (RPE) as seen in Fig. 1.3. The IS-OS



(a) (b) (c)

(d)

Fig. 1.4 A 2D fundus view of the eye (a), and coronal (b), sagittal (c), and axial (d) slice of the spectral-domain optical coherence tomography image showing a macular hole. The green box and arrow on the fundus view indicate the corresponding locations of 2D slice. OCT resolution is 200 μm.



(a)

(b)

Fig. 1.5 A preoperative 2D slice of a 3D SD-OCT image of a patient's eye with an idiopathic full-thickness macular hole and VA of 42 ETDRS letters (a), and the postoperative 2D slice of a 3D SD-OCT image after successful surgery with closure of the hole, restoration of the foveal depression and a VA of 71 ETDRS letters (b). The green box and arrow on the fundus view indicate the corresponding locations of 2D slice. OCT resolution is 200 μm.

junction hyper-reflective layer appears to coincide with the isthmus between the inner and outer segments of the photoreceptors [97]. The studies have noted a substantial correlation between the visual outcomes of MH closure surgery and the postoperative condition of the IS-OS junction layer, with disruptions in this layer being associated with poorer visual outcomes [97, 123]. Additionally, the ELM is another hyper-reflective layer in the outer retina, appearing just above the IS-OS junction layer. Recent studies have demonstrated that after retinal detachment (RD) repair, the combined disruptions of the ELM and IS-OS layers cause worse visual outcomes than eyes with isolated IS-OS layer disruptions [97, 123]. For these reasons, preoperative analysis of the ELM layer is crucial for assisting in MH closure surgery. The accurate identification of the ELM layer plays a key role in correct intervention, potentially improving VA following surgery. This thesis will also examine the analysis of the ELM layer (see Fig.1.6).

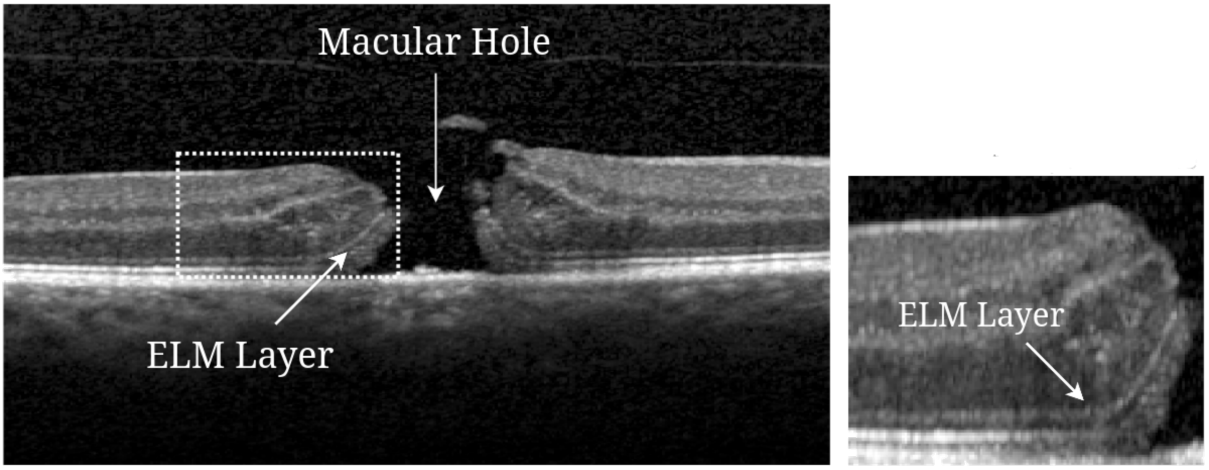


Fig. 1.6 Spectral-domain optical coherence tomography central slice through the macula demonstrating an idiopathic full-thickness macular hole. The external limiting membrane is clearly labelled.

1.1.2 Traditional Methods for Retinal Image Analysis

Retinal image analysis was traditionally performed using classical image processing and machine learning techniques, valued for their interpretability, lower data requirements, and applicability in low-resource environments. Classical image processing methods were applied to enhance image quality, reduce noise through techniques like histogram equalization, Gaussian filtering, and median filtering, and to perform tasks such as segmentation and detection of relevant structures using mathematical morphology.

Matsui et al. [110] were among the first to propose a method for segmenting retinal blood vessels in 2D color fundus photographs using mathematical morphology. Similarly, other studies employed techniques such as the top-hat transform to detect pathological lesions associated with DR and AMD [78, 166]. Building on these early approaches, subsequent research introduced more sophisticated methods for vessel segmentation and lesion detection. Leandro et al. [98] combined mathematical morphology and wavelet transforms, employing top-hat filtering, Laplacian of Gaussian edge detection, and frequency-based noise suppression to improve vessel visibility. Budai et al. [24] used a multi-resolution Gaussian pyramid with Hessian-based analysis

to segment vessels of varying diameters, while Martinez-Perez et al. [109] proposed a multiscale Gaussian convolution method to isolate vessels at different scales. Despite these advances, classical fundus analysis methods often lacked robustness to illumination changes, artefacts, and anatomical variability, limiting their real-world performance.

As imaging technology evolved, the introduction of OCT fundamentally expanded the retinal image analysis from 2D color fundus images to 3D volumetric cross-sections, leading to more advanced spatial and volumetric analysis. Classical image processing techniques were adapted to handle the increased complexity of 3D volumetric data. Early approaches for retinal layer segmentation used edge detection, active contours, or graph-based shortest path algorithms to delineate intraretinal boundaries [36, 54]. Cabrera Fernández et al. [25] developed a method based on deformable models and statistical priors to segment fluid-filled regions and retinal layers in OCT images. Similarly, Baroni et al. [21] introduced dynamic programming for boundary detection. Structural abnormalities such as fluid accumulations or MH were also targeted using region growing, voxel intensity modeling, and morphological filtering [101, 165]. These handcrafted pipelines enabled interpretable analysis but often struggled with generalizability, sensitivity to image noise, and variability in anatomical presentation.

Traditional machine learning techniques were widely employed to enhance retinal image analysis by combining handcrafted feature extraction with statistical classification methods. In 2D fundus photography, features such as color histograms, texture descriptors (e.g., Local Binary Patterns), and morphological characteristics were commonly extracted to facilitate tasks including microaneurysm detection, drusen classification, and optic disc localization. These features were subsequently fed into classifiers (support vector machines (SVMs), k-nearest neighbors (k-NN), random forests, and decision trees) to distinguish between normal and pathological patterns [2, 120]. Following these studies, traditional machine learning methods were also applied to OCT images to classify macular diseases and detect fluid accumulations by leveraging volumetric intensity profiles, layer thickness measurements, and reflectivity-based features. For example, Venhuizen et al. [165] used a random forest model to detect intraretinal and subretinal fluid, while Srinivasan et al. [145] applied SVMs to differentiate between diabetic macular edema, AMD, and normal retinas. Similarly, Mujat et al. [116] used SVMs for segmentation and thickness mapping of retinal layers, and Draelos et al. [45] utilized logistic regression on extracted OCT features to detect retinal fluid in patients with exudative AMD. Although these models improved diagnostic accuracy and interpretability, they often required extensive feature engineering and remained sensitive to image quality and inter-patient variability.

1.1.3 Deep Learning Applications for Retinal Image Analysis

With the recent advancement in DL techniques applied to medical image processing problems, retinal image analysis has become increasingly influential and an indispensable area of research [13]. Early retinal image analysis relied on manual segmentation, which was labour-intensive and required significant expertise [119]. In contrast, computer-aided detection is cost-effective, objective, and feasible without needing highly trained ophthalmologists. Screening systems developed through these techniques enable early detection, diagnosis, and real-time

classification of retinal diseases [13, 117, 118]. However, these traditional methods used manual image feature descriptors, requiring domain knowledge, manual parameter tuning, and often lacking generalisation. Recent advancements in DL have transformed retinal image analysis by enabling automatic and generalised feature extractions from raw image data [13].

DL has significantly advanced retinal image analysis, providing powerful tools for various tasks aimed at diagnosing and classifying diseases (DR, AMD, glaucoma, and MH) [3, 59, 128], segmenting optic discs, retinal vessels, choroidal and retinal layers [114, 127, 136, 168], and monitoring disease progression (treatment efficacy) [94, 122, 176]. The essential DL technique leveraged in these tasks is the convolutional neural networks (CNNs), specifically designed for processing grid-like data such as images.

CNNs can accurately diagnose several retinal conditions. CNNs consistently demonstrate physician-level classifying and grading from SD-OCT images, which have been approved by the United States (US) Food and Drug (FDA) cleared system [3]. A study also used eyes containing patient records that CNNs may recognise [128]. Remarkably, it showed that CNNs could classify cardiovascular and diabetic risk factors from fundus images with patient records, including age, gender, smoking, haemoglobin level, body-mass index, systolic blood pressure, and diastolic blood pressure.

Besides, several different architectures of CNNs have been used for segmentation of retinal layers, retinal vessels, macular edema, and macular hole using SD-OCT images, including FCNN [108], U-Net [133], SegNet [14], and DeepLabV [27]. Additionally, the segmentation analysis of retinal layers was further extended to more severe retinal diseases, and the features from fundus and SD-OCT have been combined in multimodal DL, which is closer to one of the chapters in this thesis [114, 127, 136, 168].

Similarly, recent state-of-the-art DL architectures can predict the progression of retinal diseases, including DME, AMD, glaucoma, and MHs. More specifically, the study presented that in patients diagnosed with exudative wet AMD (exAMD) in one eye, CNNs perform better than five out of six experts in predicting progression to exAMD in the second eye [176]. Other studies also observed visual improvement after MH surgery using DL and clinical features, which is closer to one of the chapters in this thesis [94, 122]. One study predicted whether the VA improvement would be greater than or less than 15 letters using Early Treatment Diabetic Retinopathy Study (ETDRS) vision charts, while another study categorised outcomes into four groups. However, no study has considered predicting the exact number of letters that can be read on the ETDRS charts following surgery.

1.1.4 Uncertainty Quantifications

While state-of-the-art DL architectures have shown outstanding performances in retinal image analysis, more attention should be paid to uncertainty estimation/quantification in model outputs. UQ helps determine which cases need further inspection by ophthalmologists by providing a confidence evaluation [13]. This is particularly important in retinal disease diagnosis, classification, prediction, and segmentation, where clinical decisions may cause significant patient visual deterioration.

Despite the high performance of DL techniques, a study presented that doctors using DL make more errors in cases where the DL prediction is incorrect [81]. This indicates that the probabilistic outputs from DL models do not accurately reflect the true probability that their predictions are correct. Providing the uncertainty estimation/quantification of the DL techniques could help reduce errors in such settings [12]. However, this would only be effective if the UQ is well-calibrated, which means the uncertainty needs to be higher when more errors are made. In this way, Orlando et al. [125] proposed a Bayesian DL-based model to observe model uncertainty when segmenting a retinal layer. Similarly, Sedai et al. [139] introduced a Bayesian DL-based model for retinal layer segmentation, which also captured the model and data uncertainty. Another study presented a novel DL model for DR grading, providing UQs with each prediction [11]. Additionally, as collecting a dataset that covers all retinal abnormalities is practically impossible, the study developed an uncertainty-inspired open-set DL model capable of accurately classifying known retinal diseases while also identifying out-of-distribution samples without requiring further labelled data [167].

This thesis only considers commonly used UQ methods and our proposed method. The first is dropout sampling, proposed by Gal and Ghahramani [53], performing multiple predictions on the same input image during inference. Next are model ensembles, which provide multiple predictions for each input by training many DL models independently [95]. Lastly, we describe and perform an uncertainty-aware regression model (U-ARM), presenting prediction and its associated uncertainty with single-run implementation. The following section continues with the motivation of this work. It explains the main problems that are addressed by this thesis.

1.1.5 Summary

Retinal imaging has undergone significant advancements, particularly with the development of modalities like SD-OCT and automated analysis techniques. Classical image processing methods were instrumental in enhancing image quality and detecting structural features, but they were often limited in their robustness to variations in illumination, anatomical differences, and image artefacts. These limitations hindered their scalability and general application in clinical environments, particularly when deployed across diverse patient populations with varying imaging conditions.

In recent years, the rise of ML and, more prominently, DL has shifted the field towards more data-driven models capable of learning complex, rich image representations. DL-based methods have demonstrated state-of-the-art performance in several tasks, including disease classification, retinal layer segmentation, and disease progression prediction. However, while these advancements have had a significant impact on some retinal diseases, they have not been uniformly applied across all retinal conditions. Much of the focus of current DL research has been on more prevalent diseases, such as DR and AMD, while conditions like MH, which can lead to substantial vision loss, remain underrepresented. This underrepresentation of MH in DL applications is a notable gap in the current landscape of retinal image analysis.

Furthermore, while significant strides have been made in disease classification and segmentation, there is a notable gap in the prediction of postoperative VA, particularly after MH surgery.

Current models typically focus on binary classifications or coarse outcome categories, but fail to provide fine-grained predictions that are clinically meaningful. Specifically, models that predict continuous VA scores, such as those measured in exact ETDRS letters, are still underexplored. This is a critical shortcoming because, in clinical practice, exact visual acuity measurements are necessary to assess treatment efficacy and guide surgical decision-making accurately. In this regard, more advanced predictive models that estimate precise functional outcomes, based on comprehensive features from SD-OCT images, are needed.

Another significant gap in current research is the insufficient integration of anatomical features, particularly the ELM layer, in predictive models. The ELM is an important anatomical marker for postoperative visual recovery, as it reflects the integrity of photoreceptors that are critical for visual function. Despite its relevance, few models explicitly incorporate the ELM in their analysis pipelines, and even fewer consider its role in predicting functional recovery following MH surgery. A deeper integration of such anatomically relevant features would improve the predictive accuracy and clinical relevance of DL models.

Moreover, one of the most critical shortcomings in the field is the limited and often poorly calibrated incorporation of UQ in deep learning models. While recent advances in Bayesian deep learning and ensemble-based methods have introduced ways to estimate uncertainty, these approaches have not been adequately explored in the context of continuous outcome prediction, such as postoperative VA. In clinical settings, the confidence in a prediction is just as important as the prediction itself. For instance, accurately quantifying the uncertainty around VA improvement post-surgery can significantly impact clinical decision-making. Furthermore, few models provide mechanisms to effectively identify ambiguous or uncertain cases that may require additional review or more detailed clinical investigation. The lack of reliable uncertainty estimation prevents these models from being fully trustworthy and actionable in clinical practice.

In summary, while retinal image analysis has made considerable progress, current research still exhibits critical gaps: (1) the underrepresentation of complex retinal conditions, such as macular holes, in deep learning applications; (2) the lack of fine-grained, continuous predictive models for functional outcomes, such as exact ETDRS letter scores; (3) insufficient incorporation of anatomically relevant features like the ELM layer, which is key to understanding postoperative visual recovery; and (4) the underexplored use of well-calibrated uncertainty quantification in predictive models. Addressing these gaps is essential for advancing the development of clinically robust, interpretable AI systems that can significantly enhance decision-making in ophthalmology and ultimately improve patient outcomes.

1.2 Motivation

The majority of image analysis approaches in the past have used privately owned datasets, and there is currently no large publicly available benchmark dataset. Therefore, the most appropriate image analysis methodology is unclear in retinal-related tasks. Moreover, the impact of high variations in retinal image datasets requires an extensive imaging data assessment and quality assurance procedure to have a robust of the DL model. In particular, the published literature aiming to predict postoperative VA for patients with MHs is sparse and not sufficient

performance since a fully automated image informatics approach is needed in retinal images containing high variations. To address this gap, a robust, publicly available dataset and advanced models are essential to achieve consistent and accurate predictions. Also, UQ methods have not been thoroughly evaluated for VA prediction. The introduction of a thorough uncertainty quantification framework would not only improve the accuracy of the models but also help in assessing the reliability of predictions, which is crucial for clinical use. Similarly, ELM layer segmentation lacks a detailed analysis of image and annotation quality, which is essential for achieving precise ELM layer segmentation. Improving segmentation models and analyzing image annotation reliability will be key to advancing automated diagnostic tools in retinal care. Thus, there is a pressing need for more comprehensive studies that address these challenges and propose solutions for practical implementation in clinical workflows. There have only been a limited number of studies investigating these various tasks. To address the above challenges and limitations, this PhD research is attempting to fulfil the following research questions (RQ):

1. How have idiopathic full-thickness macular holes (MHs) been quantitatively assessed in the published literature, and what are the limitations of existing datasets?
2. How does the newly introduced 3D SD-OCT imaging benchmark dataset contribute to advancing DL models for MH analysis?
3. How do different state-of-the-art DL models perform in predicting postoperative VA from preoperative OCT images?
4. What relationships exist between preoperative OCT images, preoperative VA, and postoperative VA, and how can DL-based predictive models leverage these relationships?
5. How can an automated OCT image analysis framework for ophthalmologists be developed using DL algorithms trained on preoperative OCT images and postoperative VA outcomes?
6. What preprocessing, image quality assessment, and anomaly detection techniques enhance the robustness of DL-based OCT analysis?
7. What are the common sources of uncertainty in OCT images, and how can they be effectively represented and quantified in DL-based predictive models?
8. How does the proposed UQ method compare with commonly used UQ approaches in improving the reliability of 2D and 3D DL-based predictive models?
9. How do different state-of-the-art DL-based segmentation models perform in detecting the ELM layer in OCT images?
10. How can automated ELM layer segmentation be integrated into clinical workflows, and what are the latest advancements in DL-based ELM layer detection?

1.3 Thesis Contributions

This PhD research adds to the literature by:

(A) We present a novel 3D SD-OCT imaging benchmark dataset for 210 patients with idiopathic full-thickness MHs (10,339 2D slices).

(B) We propose a complete image informatics framework to generate a high-quality OCT image dataset for use in a robust DL-based predictive model of VA in patients with idiopathic full-thickness MHs. The framework involves image preprocessing, image quality assessment, and anomaly detection.

(C) We compare quantitatively seven 2D state-of-the-art DL-based predictive models of both preoperative and postoperative VA. To account for the 3D nature of the eye captured in 3D SD-OCT imaging data, we used multiple image slices during the training phase.

(D) We compare quantitatively four 3D state-of-the-art DL-based predictive models of both preoperative and postoperative VA using four evaluation metrics.

(E) We propose a novel method that present a quantitative evaluation of uncertainty in the 2D and 3D DL-based predictive models on the dataset.

(F) We compare the proposed method and most commonly used UQ methods on the dataset.

(G) We perform an extensive experiments including seven 2D state-of-the-art DL-based image segmentation models for the ELM layer using six evaluation metrics.

1.4 Thesis Structure

The motivation behind this work, the objectives of the thesis, and the contributions have been described so far. This section provides an outline of the thesis, alongside a brief overview of each chapter. The remainder of this thesis is structured as follows:

Chapter 1 presents the motivation for the work undertaken in this thesis, and highlights retinal imaging, traditional and advanced retinal image analysis approaches, DL-based predictive models, and UQ methods. The contributions and the peer-reviewed publications produced as a result of work undertaken in fulfilment of this thesis are also presented.

Chapter 2 provides the description and analysis of the benchmark OCT imaging dataset used in this study and existing datasets (Research questions 1 and 2 - Contribution (A)).

Chapter 3 outlines a comprehensive image informatics framework for predicting VA using DL-based predictive models. The implementation of the DL models is examined in detail, including model training, evaluation metrics, and hyperparameter tuning. In addition, the comparison of 2D and 3D CNN models are presented (Research questions 3, 4, 5, and 6 - Contributions (B), (C), and (D)).

Chapter 4 describes a novel uncertainty-aware regression model, presenting the prediction of VA and its associated uncertainty with single-run implementation (Research questions 7 and 8 - Contributions (E), and (F)).

Chapter 5 details the experimental design for segmenting ELM layer using state-of-the-art DL-based image semantic segmentation models (Research questions 9 and 10 - Contribution (G)).

Chapter 6 explains how the proposed framework (Chapter 3), uncertainty of the models (Chapter 4) and ELM segmentation (Chapter 5) are useful to understand overall MH assessment.

Chapter 7 provides a summary of the study findings. In addition, it summarises the thesis contributions and highlights future works and recommendations for how the research should proceed.

1.5 Publications

The following papers have been either submitted or published towards support of the research conducted in this thesis:

[143] Singh, V. K., Kucukgoz, B., Murphy, D. C., Xiong, X., Steel, D. H., and Obara, B. (2022). Benchmarking automated detection of the retinal external limiting membrane in a 3D spectral domain optical coherence tomography image dataset of full thickness macular holes. *Computers in Biology and Medicine*, 140:105070.

[89] Kucukgoz, B., Yapici, M. M., Steel, D. H., and Obara, B. (2023). Evaluation of 2D and 3D deep learning approaches for predicting visual acuity following surgery for idiopathic full-thickness macular holes in spectral domain optical coherence tomography images. In *2023 International Symposium on Image and Signal Processing and Analysis*, pages 1-6.

[88] Kucukgoz, B., Yapici, M. M., Murphy, D. C., Spowart, E., Steel, D. H., and Obara, B. (2024). Deep learning using preoperative optical coherence tomography images to predict visual acuity following surgery for idiopathic full-thickness macular holes. *IEEE Access*.

[90] Kucukgoz, B., Zou, K., Murphy, D. C., Steel, D. H.,Obara, B. and Fu, H.(2024). A Clinician's guide to predict postoperative visual acuity in patients with macular holes. *Computerized Medical Imaging and Graphics, Elsevier*.

Chapter 2

Macular Hole Data Collection

Contents

In this chapter, we will detail the datasets used in the thesis and the data collection process. The chapter will cover description of datasets, inclusion and exclusion criteria, and imaging protocols and data characteristics. It will provide a comprehensive overview of the datasets and data collection, forming the foundation for the subsequent analysis and evaluation in the thesis.

2.1 Data Collection

The proposed image informatics frameworks and DL models were designed, implemented and evaluated on two sets of SD-OCT imaging datasets, all captured using the Heidelberg Spectralis (Heidelberg, Germany) (Fig. 2.1a) using the same imaging protocol at Sunderland Eye Infirmary, UK and Rigshospitalet, Copenhagen, Denmark (see full SD-OCT image with fundus region Fig. 1.4 and without fundus region Fig. 1.4). Three different Spectralis cameras were used in the UK centre and one in Denmark.

(a) Heidelberg Spectralis OCT imaging system capturing retinal scans.

(b) Visual acuity assessment using ETDRS chart at 4 meters.

Fig. 2.1 Clinical assessment components: (a) OCT image acquisition procedure and (b) visual acuity measurement protocol used in the study. Both examinations were performed by trained clinicians following standardized operating procedures.

All machines were of the same model and manufacturer, but minor differences in calibration, operator technique, and local image acquisition conditions may exist, as is typical in real-world clinical environments. Operator involvement in SD-OCT acquisition—such as patient positioning, scan alignment, focus adjustments, and quality control—can introduce subtle variations in image quality and appearance. These factors may affect the distribution of features within the dataset, such as contrast, sharpness, or anatomical centering. Importantly, each imaging centre had a single designated operator responsible for scan acquisition: one in the UK and one in Denmark. Both operators were experienced clinicians with comparable training in retinal imaging. This consistency in operator expertise helped reduce systematic biases between the datasets.

Additionally, standardized imaging protocols and the Heidelberg "Automatic Real-Time" (ART) mode helped further mitigate variability, producing scans with consistent signal-to-noise characteristics across both sites. As a result, the noise characteristics were consistent in type across datasets; however, the degree of noise may still vary slightly due to differences in

image acquisition conditions. Visual inspection of representative samples confirmed generally comparable noise profiles between the UK and Danish cohorts.

The UK data were collected as part of routine clinical care between January 2017 and January 2021 and were fully anonymized prior to analysis. According to UK Health Research Authority (HRA) guidelines, using such anonymized retrospective data for service evaluation does not require ethical approval. The images from Denmark were obtained from a previously published randomized controlled trial, which received ethical approval from the Scientific Ethics Committee of the Capital Region of Denmark (Protocol Number: H-4-2013-091), and written informed consent was obtained from all participants.

This thesis included patients with a confirmed idiopathic full-thickness MH (on OCT) who had undergone vitrectomy and internal limiting membrane (ILM) peeling with gas tamponade surgery, successfully achieved primary hole closure (hole closure following a single surgery) evidenced by standardised OCT imaging two weeks after surgery, and had a best-corrected VA recording at three months ($\pm$ two weeks) postoperatively. All patients were pseudophakic postoperatively, which prevents the confounding influence of cataracts on the VA measurement. Patients who were phakic before surgery underwent combined phacovitrectomy if they were from the UK cohort, and phacoemulsification and intra-ocular lens insertion surgery one week before vitrectomy surgery if from the Denmark group.

This thesis excluded all secondary holes, non-full thickness holes, eyes with previous vitrectomy surgery and/or non-primary closure, and eyes with other co-existing causes for reduced vision, for example, AMD or amblyopia. This is because different medical treatments or operations might be needed. However, the techniques could be applied to other types of MH.

Image dataset and clinical data on 210 eyes from 210 patients meeting the inclusion and exclusion criteria were analysed: 67 from Denmark and 143 from the UK. The mean age was 70 years old (range 48–84), 172 (82%) were female, and 105 (50%) were right eyes. The mean minimum linear diameter of the holes was 383 μm, and the median duration of symptoms was six months. All scans used 16 automatic real-time settings, enabling multi-sampling and noise reduction over 16 images. While capturing the OCT images (see Fig. 2.1), the patients were asked to focus on a constant fixation object to minimise eye motion as a general procedure.

VA was measured in all patients using ETDRS vision charts (Fig. 2.1b), with testing at four metres [83]. These charts have a group of five letters per row, with multiple rows with a reducing letter size of 0.1 logMAR per line. The VA measurement is calculated by how many letters can be correctly read on the chart. A score of 70 letters equates to 0.3 logMAR (or 20/40 Snellen VA), whilst 35 letters equate to 1.0 logMAR (or a Snellen acuity of 20/200). The VAs were recorded by two experienced optometrists in a clinic unassociated with the study and were best-corrected VAs after refraction using a standardised protocol.

2.2 Data Analysis

For both image sets, the same standardised imaging protocol was used, namely a high-density central horizontal scanning protocol with 29–30 μm (microns) line spacing in the central 15 by 5 degrees. With 27–34 μm spacing between scans (Z-axis), there were typically 49 scans per

dataset. The captured OCT images, however, had variable pixel widths, heights, and depths (X is from 178 to 497 px (pixels), Y is from 321 to 776 px, and Z is from 49 to 96 px) relating to different captured image resolutions (see Fig. 2.3). The pixel resolutions of the OCT images were, therefore, between 5.04 and 12.66 μm per pixel in-width, but the same resolution of 3.87 μm per pixel in-height (see Fig. 2.2a and 2.2b). As illustrated in Fig. 2.2a, the in-width resolutions across UK and Danish datasets show considerable variation. This indicates inconsistencies in horizontal pixel sampling, likely due to differences in device settings or acquisition protocols across sites. In contrast, Fig. 2.2b demonstrates that Z-axis sampling (depth) is much more consistent between cohorts, supporting comparable volumetric data structures. Since Y-axis resolution is fixed at 3.87 μm for all images, vertical measurements are standardised. These figures collectively show that while vertical and depth resolutions are consistent, horizontal measurements vary and may need harmonisation in downstream analyses. The specific parameters of this standardized acquisition protocol are detailed in Table 2.1.

Table 2.1 Protocol Specifications for Optical Coherence Tomography Image Acquisition

Parameter	**Specification (Both Cohorts)**
Mean age	70 (Range from 48 to 84)
Sex	Female: 82 % and Male: 18 %
Eyes	Right: 50 % and Left: 50 %
Scan Type	High-density horizontal
Line Spacing	29–30 µm
Scan Area	Central 15×5° retina
X-axis Resolution	5.04–12.66 µm/px
Y-axis Resolution	Fixed 3.87 µm/px
Z-axis Spacing	27–34 µm
Scans per Volume	49–96
ART Frame Averaging	16 frames
MH Size	Mean 383 µm
VA Testing	ETDRS at 3 months

Also, the collected OCT images show differences in image quality related to patient movements, operator techniques, and the specific OCT camera used. Similarly, due to ocular anatomy and acquisition distortions, the OCT images could be scaled differently, shifted from the center, or randomly oriented. This can cause considerable variability (see Fig. 2.3). The VA is ranged from 5 to 83, outlined in Fig. 2.2c and 2.2d. Our image pixel resolutions and VA measurements have imbalanced distributions.

2.3 Existing Datasets

Table 3.1 summarizes a range of studies that utilize OCT imaging data to assess visual acuity or related clinical outcomes using both classical statistical methods and modern deep learning models. Across these studies, a key limitation is the lack of publicly available datasets tailored specifically for postoperative visual acuity prediction, particularly in MH cases.

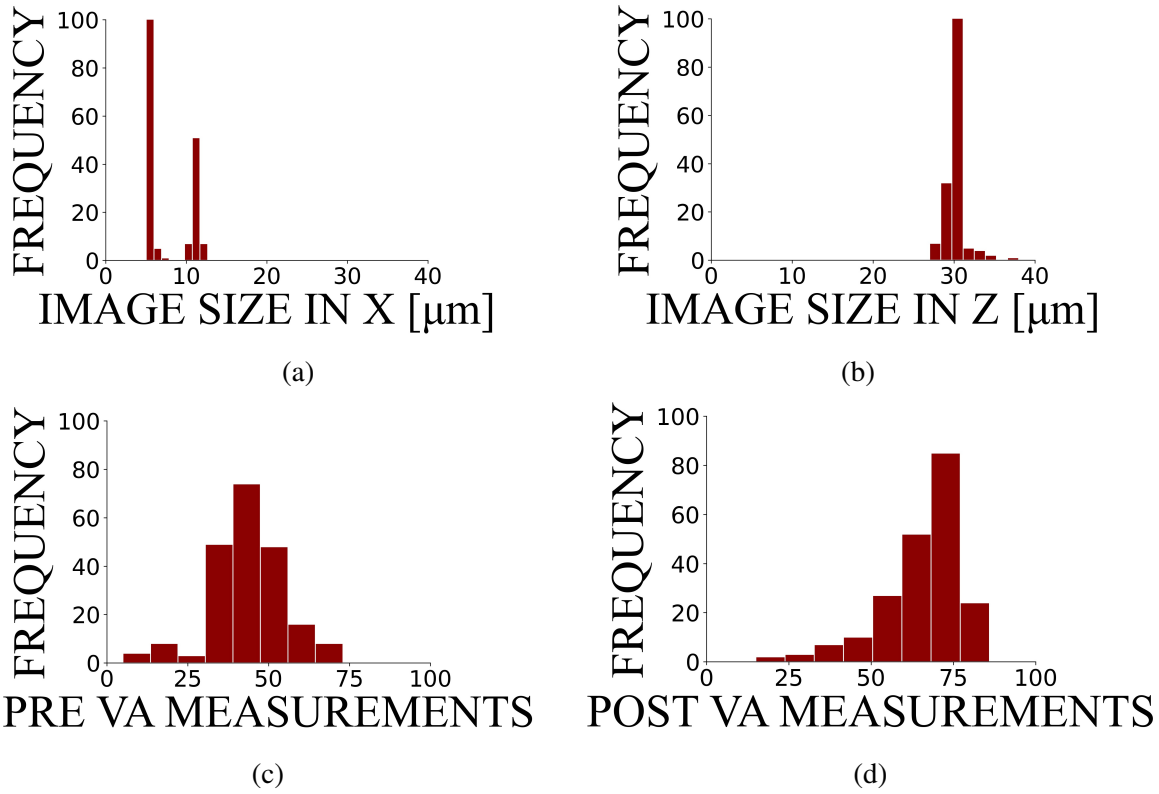


Fig. 2.2 The distribution of OCT image sizes in X (a), Z (b), distribution of preoperative (c) and postoperative (d) visual acuity measurements. Image size in Y is 3.87 μm for all images.

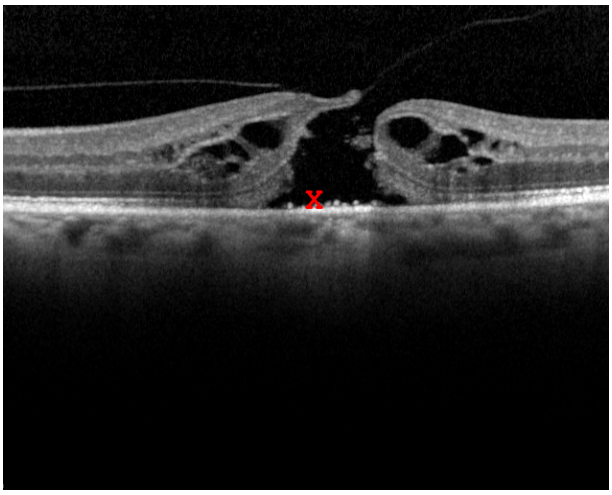
Steel et al.[148], Murphy et al.[117], and Obata et al. [122] conducted their work using private 3D OCT datasets focused on MH patients, with sample sizes ranging from 67 to 1527 scans. These datasets are not publicly accessible, which limits reproducibility and comparative evaluation.

Srinivasan et al.[146] employed a publicly available 3D OCT dataset, but their work focused on classifying retinal diseases such as AMD and DME rather than MH or visual acuity prediction. Romo-Bucheli et al.[132] also performed regression and classification tasks, though the dataset used in their study was private. Kawczynski et al.[79] and Xu et al.[173] explored visual acuity prediction (e.g., BCVA) using private datasets with 1071 and 56 3D OCT volumes, respectively.

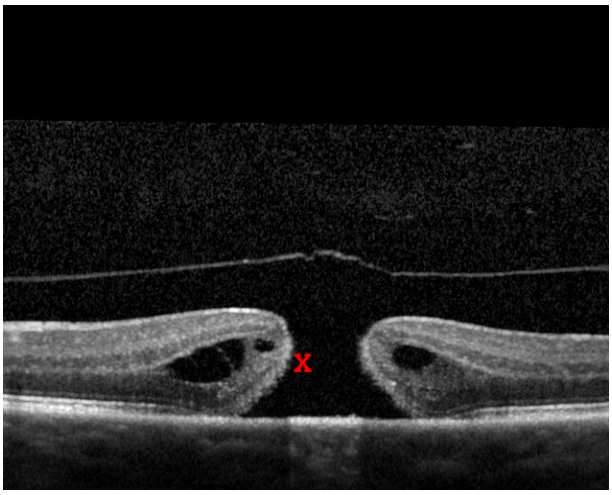
Alqudah et al.[8] worked with a large publicly available 2D OCT dataset (over 137,000 images); however, this dataset is primarily intended for disease classification tasks (e.g., AMD, CNV, Drusen) and lacks MH-specific information or postoperative visual acuity. While other public OCT datasets exist (e.g., Duke, RETOUCH, UCSD), they are not designed for or applied to visual acuity prediction or postoperative outcomes in MH cases.

Lachance et al. [94] presents a closely related publicly available dataset, which includes 121 cases comprising HD-OCT B-scans (2D) and corresponding clinical data for patients undergoing MH surgery. While this dataset enables binary classification of postoperative visual improvement ($\geq$ 15 ETDRS letters at 6 months), it is limited to 2D single-slice B-scans rather than full volumetric (3D) OCT data, which offers richer spatial information.

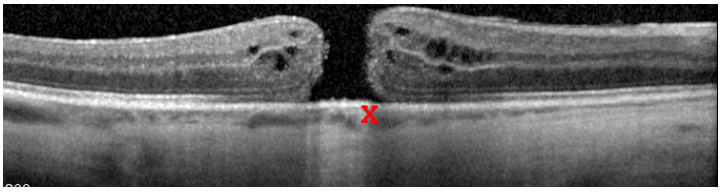
To address the above gaps, we introduce a new publicly available dataset of 210 preoperative 3D OCT volumes from MH patients, paired with corresponding postoperative visual acuity



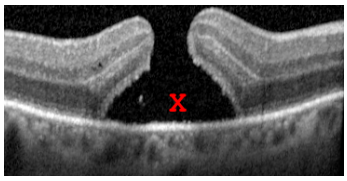
(a) 632×490 *px* -5.43×30 μm



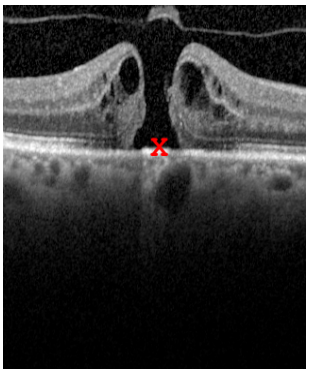
(b) 632×490 *px* -5.59×30 μm



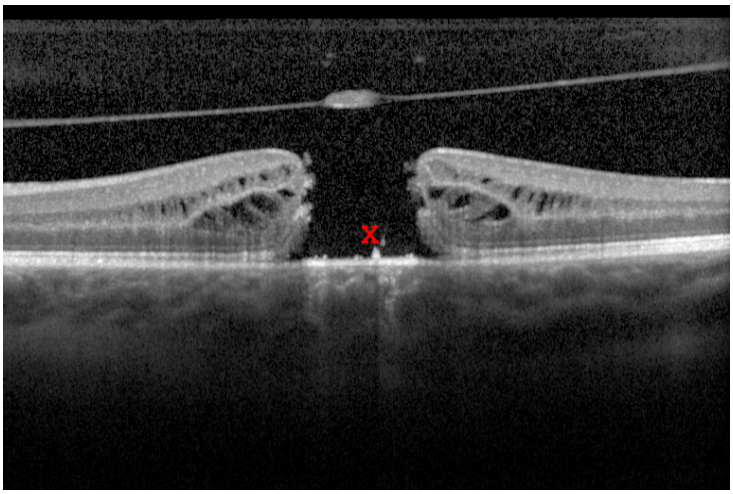
(c) 750×192 *px* -7.05×38 μm



(d) 365×178 *px* -10.75×29 μm



(e) 321×376 *px* -11.12×30 μm



(f) 768×496 *px* -6.47×35 μm

Fig. 2.3 2D mid-slices of six randomly selected 3D images, with their calculated centres of mass (red crosses), and with their corresponding image sizes in *px* and μm.

scores. To our knowledge, it is among the few datasets designed explicitly for fine-grained postoperative visual acuity prediction in MH cases, providing a more detailed and regression-focused alternative to existing classification-based approaches.

2.4 Summary

Within the chapter, we introduced the dataset. Chapter 3 and 4 utilize the entire dataset, incorporating all the collected imaging data for comprehensive analysis and discussion. These chapters also use clinical data, including preoperative and postoperative VA measurements as a label to conduct the analysis. Chapter 5 focuses on a subset of the datasets. The selection criteria for this subset are detailed within the section, ensuring the relevance and specificity of the analysis. Additionally, Chapter 5 makes use of annotations provided by doctors for this subset of datasets.

Chapter 3

Predicticition of Visual Acuity using Deep Learning Models

Contents

This chapter introduces a comprehensive image informatics framework for predicting VA outcomes in patients undergoing surgery for idiopathic full-thickness MHs, using SD-OCT images. The framework addresses real-world challenges in OCT image analysis by incorporating advanced image preprocessing, quality assurance, and anomaly detection methods.

We explore both 2D and 3D CNNs based deep learning models to predict preoperative and postoperative VA measurements. In this chapter, we specifically:

- Introduce a novel 3D SD-OCT imaging benchmark dataset comprising 10,339 2D slices from 210 patients, providing a valuable resource for future OCT-based deep learning research.
- Propose an automated image informatics framework that enables the development of a robust deep learning-based predictive model for VA outcomes, addressing the challenges posed by non-standardized OCT datasets.
- Conduct a comprehensive quantitative comparison of nine state-of-the-art 2D deep learning models, optimizing them with a custom loss function and evaluating their predictive performance using multiple metrics.
- Perform a direct comparison between 2D and 3D CNN models, assessing their effectiveness in predicting postoperative VA and analyzing the trade-offs in terms of predictive accuracy, computational complexity, and spatial information capture.

By systematically analyzing both 2D and 3D CNN-based predictive models, this chapter provides critical insights into the strengths and limitations of each approach. The findings contribute to advancing automated VA prediction using OCT imaging and highlight the trade-offs between 2D and 3D deep learning approaches for OCT image analysis.

3.1 2D Convolutional Neural Network based Deep Learning Models

This section presents a fully automated image informatics framework. The framework is combined with a DL approach to automatically predict VA outcomes for people undergoing surgery for idiopathic full-thickness MHs using 3D SD-OCT images. To overcome the impact of high variation in real-world image quality on the robustness of DL models, comprehensive imaging data preprocessing, quality assurance, and anomaly detection procedures were utilised. We then implemented, trained, and tested nine state-of-the-art DL predictive models through our designed loss function with multiple 2D input channels on the imaging dataset. Finally, we quantitatively compared the models using four evaluation metrics. Overall, the predictive model achieved a MAE of 6.47 ETDRS letters score, demonstrating high predictability. This confirms that our fully automated approach with input from seven central SD-OCT images from each patient can robustly predict visual acuity measurements.

3.1.1 Introduction

Idiopathic full-thickness MHs form secondary to age-related abnormalities of the vitreoretinal interface with a prevalence of up to 3 in 1000 people over the age of 55 [50]. They appear as a small dehiscence in the neurosensory retina at the centre of the fovea, a highly specialised part of the human retina responsible for fine acuity and colour vision [49]. Today's technology allows ophthalmologists to diagnose, classify and measure MHs using SD-OCT scans. OCT is a non-invasive, high-resolution imaging technique that uses infrared light to provide 3D imaging of the retina [62] (see Fig. 1.4).

Macular holes can be effectively treated by closing the hole using vitrectomy surgery. Predicting the visual outcome after surgery is important to guide the decision to operate and manage patients' expectations. Several studies [84, 93, 148] have shown that postoperative VA is highly correlated with preoperative VA, as well as a variety of measures of macular hole size that can be measured on SD-OCT. Various studies have attempted to precisely predict postoperative VA using manual 2D measurements of MHs and preoperative VA, although their predictive ability has been limited [163]. 3D automated image reconstruction has improved this ability [74, 119], but there are no current standards for shape, size, and resolution of OCT imaging data captured by different OCT devices for this task [174]. There are also many qualitative features and subtle alterations in retinal anatomy, for example, associated with chronicity, which may be predictive of acuity outcomes and that are difficult to measure [89, 143]. Additionally, image artefacts related to a patient's eye movement and media opacity pose a further challenge in developing image informatics methods [15]. Recently, some researchers have highlighted the low signal strength of OCT devices which results in issues such as image noises, blurriness and contrast reduction [77, 91]. Similarly, another study expanded the analysis to include scan centring and retinal region checks [160]. These challenges constitute our primary motivation.

To overcome those challenges, most machine learning (ML) and DL approaches have focused on the automated classification of macular diseases, such as age-related macular degeneration (AMD), diabetic macular oedema (DME), and MHs from OCT images data [8, 10, 115, 117, 162, 177]. More recently, some DL approaches have improved the prediction of VA outcomes [79] using OCT data [66, 122, 131]. In particular, convolutional neural network (CNN) models have achieved high performance in OCT image analysis studies; however, there have only been a limited number of studies investigating VA measurements [79, 94, 122]. Considering the success of prominent CNN-based networks in medicine [70, 71, 126], they used a ResNet [61] in the [79], VGG [141] in the [175], and CBR-Tiny models [129] in the [94] as a backbone. These studies also presented that CNN-based networks excel in extracting spatial features from OCT images. Consequently, the implementation of CNN-based models for predicting VA measurements has gained significant importance for the next motivation. Subsequently, vision transformers (ViTs) [44] have recently demonstrated great potential in assisting clinicians with clinical diagnosis [69], particularly in OCT image analysis [169]. However, to the best of our knowledge, ViTs have not yet been thoroughly applied for predicting VA measurements. Since ViTs consider global context and dynamic attention, it revealed the need for comparing standard

CNN-based state-of-art models (VGG, ResNet, Inception v3 and DenseNet, EfficientNetV2) and ViTs in this study, acting as another motivation.

This research presents a comprehensive image informatics framework for predicting both the preoperative and postoperative VA measurements for patients with idiopathic full-thickness macular holes using an SD-OCT image dataset based on image preprocessing, image quality assessment, image anomaly detection, and deep learning models-based prediction.

The remainder of the section is organised as follows: Section 3.1.2 describes related works and our contributions, including regression models, classification models and predictive models of VA measurements based on image analysis approaches. Section 3.1.3 summarises image preprocessing, quality measurement, anomaly detection, and the DL models-based visual acuity prediction methodology and experimental design. Section 3.1.4 details the study results, and Section 3.1.6 concludes the study findings and provides recommendations for future research.

3.1.2 Related Works

Numerous image informatics approaches to assess macular diseases using OCT imaging data have been proposed in the published literature (see the summary in Table 3.1). As shown in Table 3.1, there has been only limited research investigating how to predict VA outcomes for specific retina diseases using OCT imaging data. In particular, the published literature aiming to predict postoperative vision for patients with idiopathic full-thickness macular holes is particularly sparse [122, 131]. Although the datasets used in different studies vary in size, imaging modality, and disease type, Table 3.1 remains valuable as it provides a comprehensive overview of existing methods, their applications, and performance metrics. This comparison highlights the gaps in current research and underscores the need for more robust predictive models, particularly for postoperative VA prediction in specific retinal conditions. By analyzing these diverse studies, we discuss the approaches that have been used to date.

Image Based Classification Approaches for Different Retinal Diseases

Zhang et al. [177] proposed a binary classification of OCT image data based on kernel principal component analysis (PCA) model ensembles to predict patients with AMD-affected eyes from normal eyes. Also, a Bayesian network classifier was introduced by [5] and then tested on the same image dataset. Another study implemented the bag-of-words (BoW) model by keeping the most salient points corresponding to the top vertical gradient values calculated in the OCT images [165]. However, this approach was limited by relying on key points and predicting only two classes: DME and normal eyes. Anantrasirichai et al. [10] proposed a support vector machine (SVM)-based approach to differentiate between normal eyes and eyes with glaucoma. Similarly, [146] used SVM to predict the presence of AMD and DME from normal eyes using a small image data set with fewer outliers. However, the obtained accuracy was extensively impacted by retinal layer discontinuities caused by the disease pathology and motion artefacts.

In addition, Lemaitre et al. [99] developed a local binary pattern (LBP) classifier to identify AMD and DME from normal eyes. Liu et al. [106] proposed an approach using image gradient information, LBP, and SVM with a radial basis function (RBF) kernel as a classifier. The

approach first classified eyes as either normal or abnormal, and then was further sub-classified into either DME, AMD, MH or normal eyes.

Motozawa et al. [115] proposed two typical CNN-based models for classifying OCT data into AMD and normal, and they were also able to differentiate between wet and dry AMD. Likewise, Alqudah et al. [8] showed that by using image denoising and resizing and tuning a CNN model with an ADAM optimiser, higher accuracy and lower time cost could be achieved when classifying OCT images into five classes: choroidal neovascularisation (CNV, a feature of wet AMD), DME, dry AMD, drusen only (a feature of early AMD), and normal.

Another OCT image classification approach using a deep multi-scale CNN model was proposed by Rasti et al. [130]. The proposed model employed a prior decomposition and new cost function to discriminate and fast-learn representative image features. The authors used the modified versions of VGG, ResNet, and Inception models to detect normal, AMD, and DME features. Li et al. [102] suggested a novel DL model for predicting CNV, DME, drusen, and normal eyes, called OCTD_Net and based on modified DenseNet and ReLayNet models. Tsuji et al. [162] proposed a method using a capsule neural network (CapsNet) model to classify the same eye disorders by learning spatial information from the OCT images.

Table 3.1 The summary of the image informatics approaches focused on assessing visual acuity using OCT imaging data. The best results are highlighted in bold for each metric in our proposed methods. Results for the proposed 2D networks are obtained using ResNet-50, while results for the proposed 3D networks are obtained using ResNet-18.

AUTHORS	TASK	DATASET	METHODS	DISEASE TYPE	OUTPUT	EVALUATION METHODS	RESULTS (Avg.)
Steel et al.[148]	Regression	3D OCT (1527 images)	Classical Statistics	MH	Postoperative VA	AUC	71.72%
Murphy et al.[117]	Regression	3D OCT (67 images)	Generalized Linear Model	MH	Postoperative VA	R-squared	0.45
Srinivasan et al.[146]	Classification	3D OCT (45 images)	SVM	Others	AMD, Normal, DME	Accuracy	95.55%
Alqudah et al.[8]	Classification	2D OCT (137437 images)	Multi-task CNN	Others	AMD, Normal, DME, CNV, Drusen	Accuracy	97.10%
Kawczynski et al.[79]	Regression Classification	3D OCT (1071 images)	ResNet-50	Others	BCVA 2 classes	RMSE AUC	9.01 0.91%
Obata et al.[122]	Regression Classification	3D OCT (259 images)	CNN	MH	BCVA 4 classes	R-squared Precision	0.46 75%,43%, 38%,50%
Romo-Bucheli et al.[132]	Regression Classification	3D OCT (350 images)	DenseNet, RNN	Others	Low, Intermediate, High treatment requirement	R-squared Accuracy	0.22 75.00%
Rizzo et al.[131]	Classification	3D OCT-A (35 images)	Inception v3, VGG16, VGG19, SqueezeNet	MH	Postoperative VA	P value	0.005
Lachance et al.[94]	Classification	2D OCT (121 images)	CBR-Tiny	MH	Postoperative VA 2 classes	Accuracy	78.7±2.9%
Xu et al.[173]	Regression	3D OCT (56 images)	Segmentation Algorithm	MH	Geometric measurements and Postoperative VA	P value	0.0028
The proposed 2D networks	Regression	3D OCT (210 images)	ResNet-50	MH	Postoperative VA	MAE R-squared RMSE	**6.84** **0.46** **9.01**
The proposed 3D networks	Regression	3D OCT (210 images)	ResNet-18	MH	Postoperative VA	MAE R-squared Pearson	**6.11** **0.46** **0.70**

These image informatics approaches have several limitations, including (1) a limited ability to identify different pathologies affecting the macula, such as MHs, (2) they are typically time-consuming due to high computation costs, which means they are inappropriate for use in clinical practice where issues may need to be resolved in real-time, and (3) by only relying on a limited number of key points, other essential ocular characteristics may not be noticed during the classification tasks.

Image Based Prediction Approaches for Visual Acuity After Macular Hole Surgery

Classical Methods: Several authors have used regression to predict postoperative VA using routinely collected clinical data. For example, Steel et al. [148], using logistic regression and the univariate level using $\tilde{\chi}^2$ tests, achieved a model area under the receiver operator curve of 71.72% for predicting a visual acuity of 0.3 logMAR or better after surgery. Generalised linear modelling has been used to predict actual acuity using an automated multi-scale 3D image analyser of OCT scans for MHs [117]. The study shows preoperative VA and MH height were important predictors of postoperative VA, achieving an R-squared value of 0.45. When preoperative vision was not included in the model and only OCT parameters were included, the most predictive model was 0.39. Interestingly, using only manual clinician-measured values, R-squared was only 0.20.

Other research teams have also investigated the 3D parameters of MH using different methodologies, such as automatically calculating three dimensions based on the sum of 2D images [173]. The 3D macular hole size parameters, such as MH volume, base area, base diameter, and MH height, were significantly correlated to postoperative VA (P value from 0.0003–0.011). [173].

Deep Learning Based Methods: Some recent studies were not only able to classify eye disorders on OCT image datasets, but were also able to predict associated VA measurements and recommend potential treatment requirements. In particular, the study by [132] presented an end-to-end DenseNet-based model for recommending treatment options in patients with wet AMD, where the model's output range was low, intermediate and high treatment requirement scores. In a further study, Kawczynski et al. [79] proposed a ResNet-50 v2 model-based approach that predicted the best-corrected VA (BCVA) measurement for patients with wet AMD eyes following treatment. BCVA measurement was obtained from the regression model, and considering the regression model results, they classified higher than 69 letters and lower than 69 letters into two classes.

The research presented in [131] assessed the ability to predict VA in two groups of 35 people with surgically treated MH using unsupervised DL models, including Inception v3, VGG16, VGG19 and SqueezeNet. Similarly, Lachance et al. [94] proposed a hybrid model classifying VA as higher than 15 letters and lower than 15 letters. Another study proposed a model for predicting postoperative VA using a typical CNN-based model [122] and four classes: class A is higher than 85 letters, class B is between 75 and 80 letters, class C is between 60 and 75 letters, and class D is lower than 50 letters. The prediction of postoperative VA using DL was compared

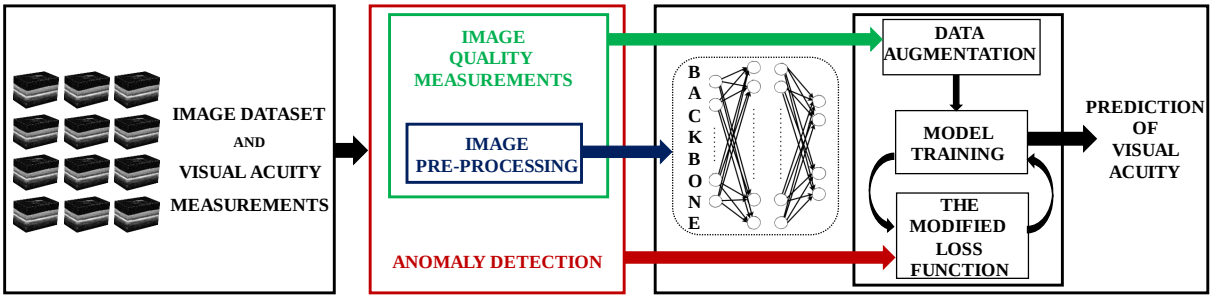


Fig. 3.1 Workflow of the proposed image informatics framework. The first stage corresponds to the input OCT image dataset and VA measurements obtained by ophthalmologists, the second stage incorporates OCT data preparation (i.e. scaling, the centre of mass detection, and cropping), OCT image quality analysis (i.e. noise score, blurriness score, contrast score, motion score, and brightness-darkness score) and anomaly detection. With the obtained high-quality image dataset and labels, multiple state-of-the-art DL models are trained and optimised by our designed loss function to predict VA measurements in the final stages.

to 3 typical regression models using preoperative clinical data. The DL model had a superior precision value of 46% compared to approximately 40% for the regression models.

These DL-based image informatics approaches have limitations related to (1) image and label data preparation, (2) data volume, (3) data quality, and (4) low-level model robustness and generalisation when using a wide range of OCT machines at different hospitals. To address some of these limitations, we recently presented a more comprehensive image informatics framework utilising robust data preparation and anomaly detection approaches combined with state-of-art DL models on a closely allied OCT analysis problem of external limiting membrane detection [143].

3.1.3 Methods

In this section, we present contributions and a comprehensive description of the imaging data preprocessing steps, data quality assessment and anomaly detection methods to create a high-quality standardised 3D OCT image dataset for DL-based prediction of VA. This comprehensive approach significantly improved our proposed model's results.

Contributions: This section adds to the literature by:

- Introducing a new 3D SD-OCT imaging benchmark dataset for 210 patients with idiopathic full-thickness macular holes (10,339 2D slices).
- Proposing a comprehensive image informatics framework to create a high-quality OCT image dataset used for a robust deep learning-based predictive model of visual acuity in patients following surgery with idiopathic full-thickness macular holes and presenting an automated solution for non-standardised OCT datasets (see Fig. 3.1). The method concludes the impact of the following surgery by predicting visual acuity.
- Quantitatively comparing nine 2D state-of-the-art DL-based predictive models of both preoperative and postoperative visual acuity using four evaluation metrics by optimising the models with our designed loss function. To account for the 3D nature of the eye

captured in 3D OCT imaging data, we used multiple image slices during the training phase.

Image Preprocessing

In 3D OCT images, due to ocular anatomy and acquisition distortions, the MHs may be scaled, shifted, and oriented randomly. Consequently, this causes high variability in the MH location and resolution, as shown in Fig. 2.3. To deal with those image acquisition issues, we used the following image preprocessing steps (see Fig. 3.2):

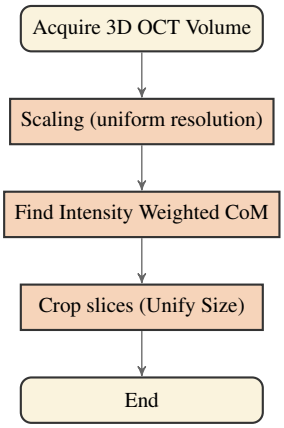
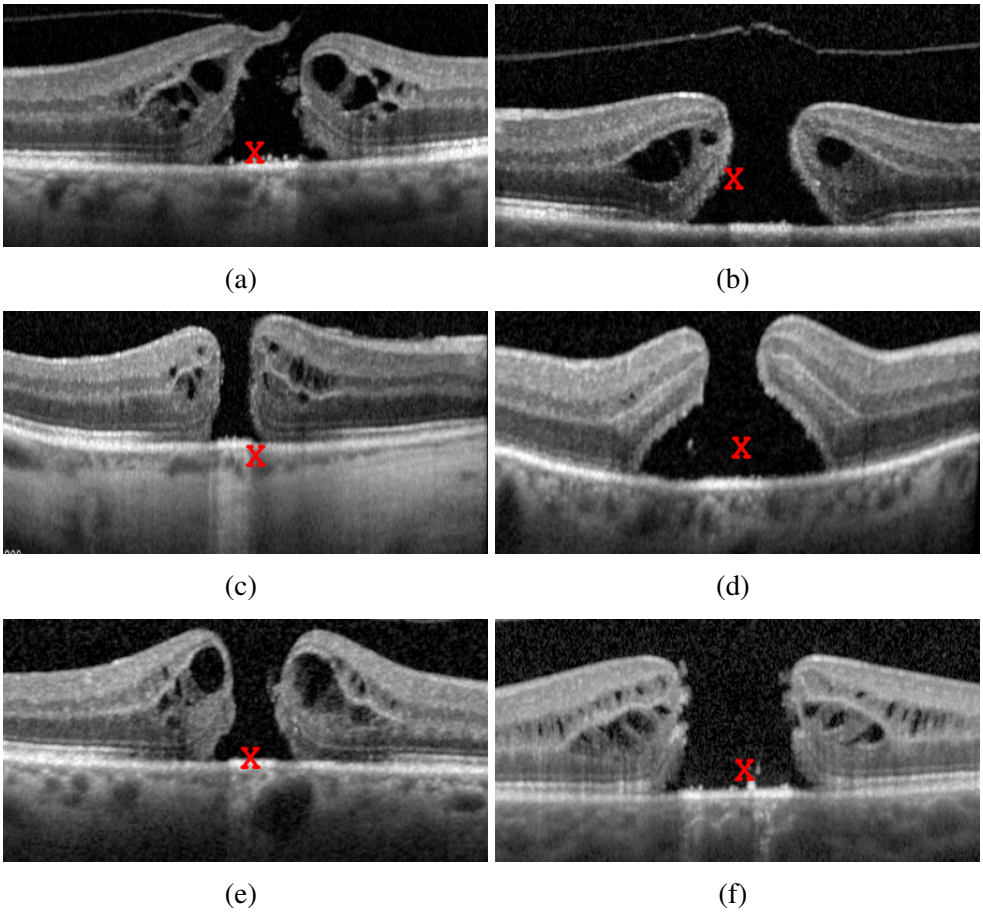


Fig. 3.2 Image preprocessing workflow

- Scaling (uniform resolution): All acquired OCT images (Fig. 2.3) were re-scaled across X, Y, and Z dimensions using the following sizes ($7.41 \times 3.87 \times 30.1\ \mu m$), and the resulting images are shown in Fig. 3.3. It ensured consistency in scale values for all OCT images.
- Intensity Weighted Centre of Mass: The MHs were located in a range of different positions in the 3D OCT image slices, as presented in Fig. 2.3. To centre the images around the positions of the MHs, we used pixel intensity weighted centres of mass calculated for each dimension as a focal point of the image, as shown in Fig. 3.3. It was used to ensure consistency in image acquisition position for all OCT images. The determining centres of mass were also considered when selecting the parameter for the data augmentation stage during the training of the DL model.
- Cropping: The scaled OCT images were then centred around intensity-weighted centres of mass (red cross) and cropped across the X, Y, and Z dimensions to the 3D size $452 \times 204 \times 49\ px$, as shown in Fig. 3.3. It ensured consistency in sizes for all OCT images.

Image Quality Assessment

The OCT images collected as part of routine clinical care inevitably differed in image quality related to patient movements, operator controls, and the OCT camera used. These resulted in several image imperfections, including speckle noise, contrast changes, and motion artefacts, which we measured using a variety of image quality measurement methods. Since these quality



(a) (b)

(c) (d)

(e) (f)

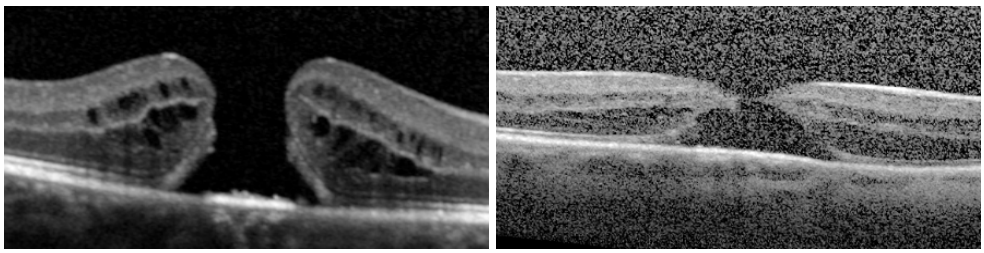
Fig. 3.3 Results of image preprocessing steps applied to images from Fig. 2.3. Final image size: $(452 \times 204 \times 49\ px\ -7.41 \times 3.87 \times 30.1\ \mu m)$.

metrics generally have no inherent upper limit, we performed min-max normalization by scaling each feature separately to a fixed range between 0 and 1. Specifically, for each metric, the minimum observed value across the dataset was mapped to 0, and the maximum observed value was mapped to 1, ensuring a consistent scale across different quality measures. High scores mainly denote a heightened presence of the measured imperfection. To enable a high-quality dataset to be selected and to optimise the DL model, these methods served as a guide to detect and remove abnormal images. We measured image quality using the following evaluation metrics.

- Noise score: Noise can be a significant problem in OCT images. Many researchers have proposed wavelet transformations to assess the lower, average, and upper bound of noise in these images [43, 60]. We performed a wavelet-based estimator of the Gaussian noise standard deviation, which revealed significant noise variances. Specifically, we computed the noise score using the median absolute deviation (MAD) method in the wavelet domain. The noise standard deviation σ was estimated as follows:

$$\sigma = \frac{\text{median}(|Y_d|)}{0.6745} \tag{3.1}$$

  where Y_d represents the wavelet coefficients at the finest scale (high-frequency subbands). The constant 0.6745 is derived from the assumption that noise follows a zero-mean Gaussian distribution and scales the MAD estimator to match the standard deviation. The resulting noise score was then normalized to a range of 0 to 1 using min-max normalization, where the minimum and maximum values were obtained from the dataset, see Fig. 3.4.



(a) Low (b) High

Fig. 3.4 Spectral-domain optical coherence tomography images demonstrating a small and large noise score.

- Blurriness score:

  Another important issue was the blur and sharpness of the OCT images. Recent studies have suggested the use of a Laplacian operator with Gaussian filters in measuring the blurriness and sharpness of images [20, 51, 85]. The Gaussian filter is defined as in Equation 3.2.

$$G(x,y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(x^2+y^2\right)}{2\sigma^2}}, \tag{3.2}$$

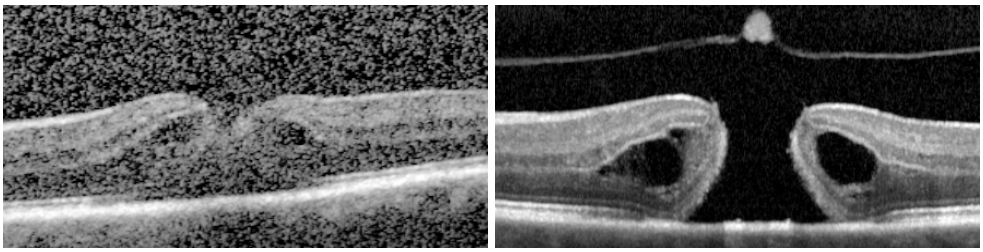
where x and y are coordinates of an image $I(x,y)$. σ is the Gaussian distribution standard deviation. The Gaussian scale-space representation L of an image $I(x,y)$ is defined as in the Equation 3.4.

$$L(x,y) = I(x,y) * G(x,y) \tag{3.3}$$

where $*$ is the convolution operator. The Laplacian operator is then applied to this smoothed imag $L(x,y)$, rather than directly to $I(x,y)$, to enhance edge structures and measure sharpness:

$$\nabla^2 L = \frac{\partial^2 L}{\partial x^2} + \frac{\partial^2 L}{\partial y^2}, \tag{3.4}$$

where $\nabla^2 L$ represents the Laplacian of the smoothed image. This operation emphasizes rapid intensity changes, helping distinguish *blurry* (low variance) and *sharp* (high variance) images (see Fig. 3.5).



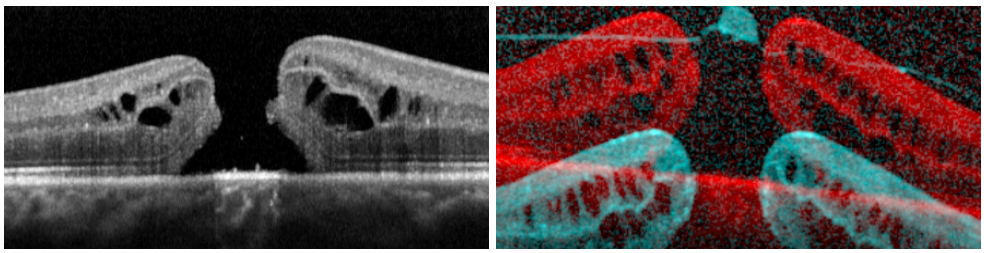
(a) Low (b) High

Fig. 3.5 Spectral-domain optical coherence tomography images demonstrating blurriness and sharpness.

- Contrast score: The differences in the chromaticity and brightness of any pixel and any other pixels within the same scene represent image contrast [140]. We therefore measured the gradients of each image pixel, including their standard deviation, with marked differences in black and white luminance.

- Motion score: Due to the movement of the eye during OCT scanning, consideration has to be given to motion artefacts. Many researchers have proposed the Horn-Schunck optical flow motion estimation method [57, 64, 112]. This method implements first-order derivatives, allowing the velocity in the flow between sequential images to be measured with high accuracy and resolution. We measured the observed motion and perceived distortions in the smooth flow of information on the Z-axis of every 3D OCT image, as seen in Fig. 3.6. An RGB colour map was used for motion visualisation, with grey representing low motion and red or blue representing high motion between any two neighbouring 2D slices in a 3D image.

- Brightness-Darkness score: The brightness and darkness of images are associated with perceived luminance. Therefore, we calculated a luminance measurement [22]. Darkness



(a) Low (b) High

Fig. 3.6 An example of optical flow velocity vector magnitudes between two neighbouring 2D slices in a spectral-domain optical coherence tomography image with ((a) - grey colour) small and ((b) - red/blue colour) large motion.

is perceived if the luminance level is low, whereas brightness is perceived if the luminance is high.

Anomaly Detection

To define a high-quality image dataset for training the DL-based VA predictive models, an anomaly detection method was used to eliminate low-quality images. This led to an improvement in our proposed model's results.

Algorithm 1: 2D Image Anomaly Candidates and Anomaly Scores Calculation.

Require: $\{I\}$ //Set of 3D images
Ensure: $\{a\},\{d\}$ //Sets of anomaly candidates and anomaly scores for all 2D image slices of 3D images
$p \leftarrow |\{I\}|$ //Number of 3D images
for $i \leftarrow 1$ **to** p **do**
  $s_{x,y,z} \leftarrow |I^i|$ //Image size for x,y,z
  for $z \leftarrow 1$ **to** s_z **do**
    $\{f_z^i\} \leftarrow quality(I_z^i)$ //2D image quality scores
  end for
end for
$\{f_z^i\} \leftarrow normalise(\{f_z^i\})$
$\{a_z^i\},\{d_z^i\} \leftarrow anomaly(\{f_z^i\})$

According to [56, 143], although several anomaly detection methods have been developed, unsupervised anomaly detection methods are preferred. This is because they have the most flexible setup and do not require any labels or prior knowledge about the dataset [48]. Methods used for unsupervised anomaly detection include nearest-neighbour, clustering, statistical, subspace, and classifier-based methods [56].

This section used the nearest-neighbour method based on the local outlier factor (LOF) method. The LOF method computes the local density deviation of the entire dataset, showing how much a data point's local density differs from its neighbours. If the data has a significantly lower density compared to its neighbours, it has a high-density deviation, suggesting it may be abnormal. Here, we applied the elbow method to iteratively determine the optimal number of neighbours in the dataset, which was found to be 10. As discussed, we employed quality

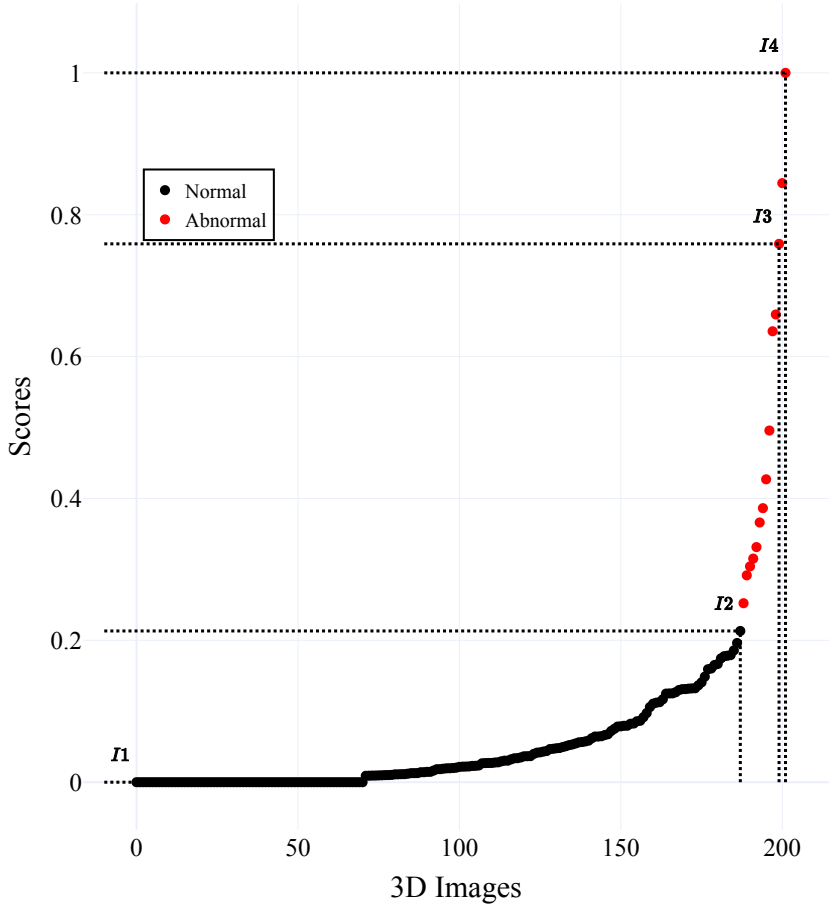


Fig. 3.7 A graph depicting the 3D OCT image anomaly detection results: black and red points represent normal and abnormal images, respectively. I_1, I_2, I_3, and I_4 demonstrate randomly selected 3D images to be presented in Fig. 3.8.

assessment measurements for each 2D image slice in every 3D image. Then, the quality scores were used as input to detect abnormal and normal images by the LOF method.

Algorithm 1 shows how the LOF-based approach determines normal and abnormal image candidates using anomaly prediction scores d and stored in a. Where f_z corresponds to the image quality assessment measurement for each 2D image slice I_z in a 3D image $\{I^i\}$. $\{a\}$ corresponds to 2D image-based anomaly candidates and $\{d\}$ corresponds to 3D image-based anomaly scores calculated across the 3D image dataset $\{I\}$.

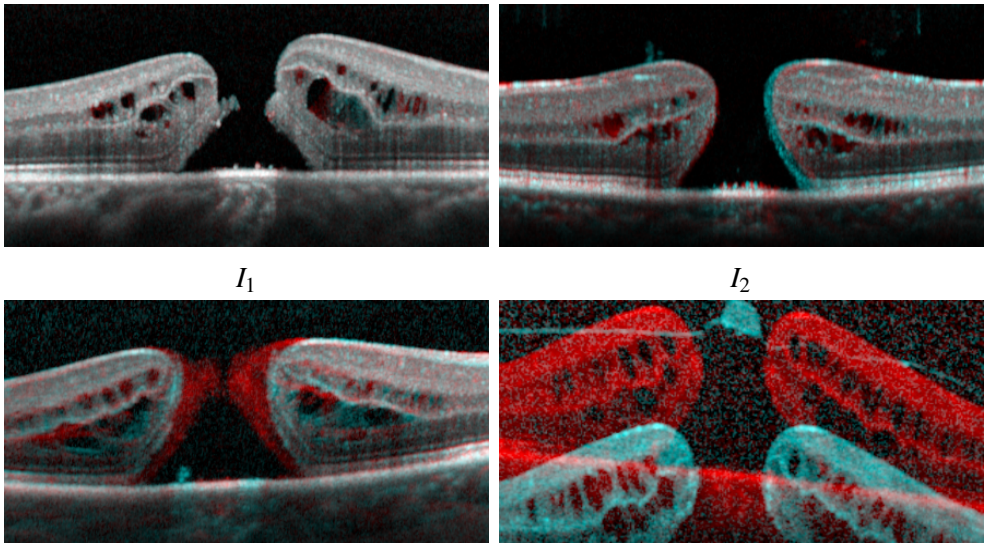


Fig. 3.8 The best to worst 2D slices from 3D spectral domain optical coherence tomography images indicated as I_1, I_2, I_3, and I_4 in Fig. 3.7.

The distribution of normalised anomaly scores among 3D images was visualised in Fig. 3.7, where the red points represent the abnormal OCT image, and the black points represent normal OCT images. These scores were calculated following the procedure described in Algorithm1, using the LOF method applied to the quality scores of 2D image slices. Images with significantly lower local density relative to their neighbours receive higher anomaly scores. To confirm the accuracy of the proposed anomaly detection procedure, 2D slices of the selected 3D image samples in Fig. 3.7, from the best to the worst (I_1, I_2, I_3, and I_4), are displayed in Fig. 3.8 - where, red/blue colours represent a high anomaly score, and a grey colour represents a low anomaly score.

Deep Learning Models

Deep learning models, as an established but still continuously advancing technology, have significantly improved disease screening through medical imaging [16, 71, 94]. Numerous deep neural networks have been developed to classify, segment and predict various diseases since they have achieved comparable performance to human experts [69, 70]. To provide a comprehensive overview, nine well-known and state-of-the-art 2D DL-based predictive models were implemented, trained and tested in this study. The input to typical 2D convolutions is C * H * W, where C is the number of input channels, H is the height, and W is the width. The

kernels move in two dimensions. In this study, we changed the first convolutional layers to feed the following DL-based predictive models with the input channels of single (1) and multiple OCT image slices (3, 5, 7, 9, 11, 13, 17, 19, and 21) since it led to better results. To adapt the classification networks for regression tasks, the output layer was changed to have a single neuron with a linear activation function, replacing the softmax or sigmoid used for classification. The loss function was also modified, switching from categorical cross-entropy to mean squared error (MSE) or mean absolute error (MAE) or Huber loss (HL). Additionally, evaluation metrics were adjusted to focus on continuous predictions, using R-squared (R^2), mean squared error (MSE), or root mean squared error (RMSE) instead of classification-specific metrics. These changes allowed the models to predict continuous values instead of discrete class labels.

Visual Geometry Group (VGG): The VGG model [141], consisting of 5 groups of convolution layers and 1 group of a fully connected layer, provided the feature extraction process in the OCT images. During training, the input OCT image is propagated through convolution layers using 3×3 kernel-sized filters and a stride of 2. Although excessive depth can be computationally expensive and time-consuming, the most relevant salient features are extracted, including edges, corners, and interest points. In this study, this model was selected for its relatively small filter size compared with other models. VGG-based models with 11 and 19 layers were used.

Residual Neural Network (ResNet) The ResNet is composed mainly of 5 groups of residual blocks and one fully connected layer [61]. These residual blocks were used to combine the information from the OCT images across multiple time points of the initiation phase. These residual blocks also link each other with skip connections. With this cross-layer connectivity, the convergence of deep networks is sped up. Thus, it prevents the problem of a diminishing gradient and provides a robust model against over-fitting in this study. The model used a 7×7 kernel-sized filter and a stride of 2 in the first convolution layer. ResNet-based models with 18, 34, and 50 layers were considered.

Inception v3 To suppress the high computational complexity problems encountered in VGG, ResNet and other mentioned models, the Inception v3 model with different kernel sizes, a max-pooling layer, and a stride, called Inception blocks, has been proposed by [153]. Due to this, it is remarkably useful for processing data in multiple resolutions and multilevel features, which makes this model suitable for OCT images. The model also overcomes high computation times by factorising the convolution (3×3) into asymmetric convolutions (3×1 and 1×3) [153].

Densely Connected Convolutional Networks (DenseNet) DenseNet consists of sets of convolutional layers and direct connections from any convolutional layer to all subsequent layers [68]. Each layer receives a piece of collective information from all preceding layers with kernel sizes 1×1 and 3×3 in a dense and uses its feature maps as input. Then, 1×1 convolutional layer followed by 2×2 average pooling as the transition layers between sequential dense blocks. Thus, our OCT image dataset has not been exposed to a vanishing-gradient problem by showing a strengthened feature propagation during training.

EfficientNetV2 EfficientNetV2 has recently been further introduced with training faster and relatively smaller parameters, a new version of the well-known EfficientNet [156, 157]. Different from EfficientNet, EfficientNetV2 uses FusedMBConv in the earlier stages of the network, which replaces the depthwise convolutions (3×3) and expansion convolutions (1×1) in MBConv with single regular convolutions (3×3) [156, 157]. EfficientNetV2 also comes in different sizes. The large-sized (EfficientNetV2-L) model was selected as our OCT images' size is relatively large.

Vision Transformers (ViTs) Vision Transformers (ViTs) have revolutionized computer vision using a transformer architecture [44]. They first split images into fixed-size patches. Each patch is linearly embedded into high-dimensional vectors. Then, to retain spatial information, positional embeddings are added to the patch embeddings. The embedded patch vectors, along with positional embeddings, are processed through a stack of Transformer encoder layers. Each layer includes self-attention mechanisms, allowing the model to capture relationships between different patches. The output of the Transformer encoder is typically a sequence of vectors. A classification head, often a linear layer, is added to obtain the final output for tasks. These steps allowed ViTs to discern global contextual information on OCT images. Therefore, we selected ViT-Base with an image patch size of 16 among several variations (ViT-B/16).

Experimental Setup

This section introduces the DL models' training and evaluation details, including the k-fold cross-validation, the DL framework, data augmentation methods, parameter selections, and the evaluation criteria used.

Table 3.2 OCT imaging data used and splitting.

	Number of 3D Images	
	Initial	After data preparation
Train	168	152
Test	42	38

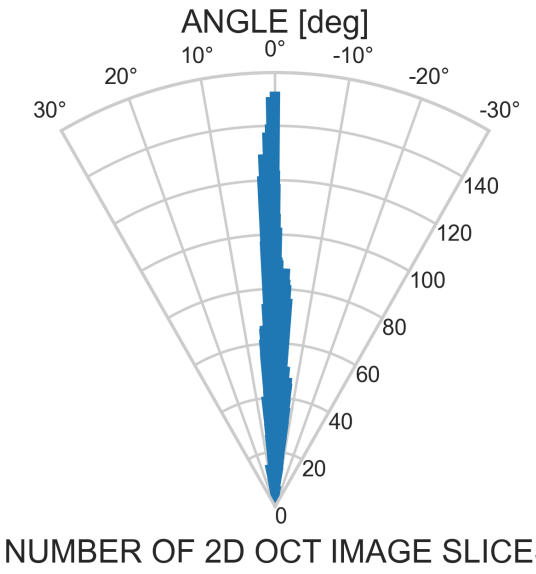
Training Following image preprocessing, image quality assessment and anomaly detection procedures, twenty images were excluded from the initial OCT image dataset (see Table 3.2) (fifteen images - image quality assessment and anomaly detection, five images - image preprocessing). The final 3D OCT images used for the training were of the size of $452 \times 204 \times 49$ px. The dataset was split uniformly into training and test sets using random five-fold cross-validation with a ratio of 80% and 20%. Specifically, each folded cross-validation consisted of 152 and 38 for training and testing, respectively.

Furthermore, DL models were trained using the same image size, with the use of single and multiple OCT image slices (1, 3, 5, 7, 9, 11, 13, 17, 19, and 21) at the first convolutional layer, centred around the mid-slice defined by the 3D intensity weighted centre of mass. To train the

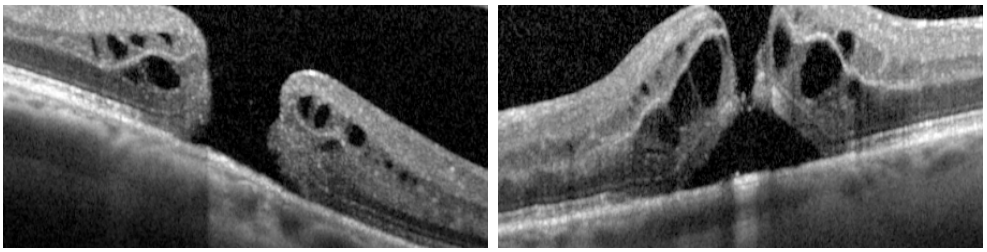
nine DL models used, we utilised Python 3.8.10, CUDA 11.4, cuDNN 8, PyTorch 1.9.0+cu102 running on a 64-bit Ubuntu operating system using a 3.4 GHz Intel Core-i9 with 32 GB of RAM and NVIDIA GTX 1080 Ti GPU with a frame buffer of 11 GB GDDR5X.

Data Augmentation To enlarge the variety and amount of data artificially, we employed a range of image data augmentation techniques, helping us overcome over-fitting issues while maintaining the data properties that existed initially in the data. The data augmentation techniques used were:

- Rotation: To estimate the eye orientation during the scanning procedure and to define the data augmentation range for the rotation, we measured the orientation distribution across the image dataset and presented it in Fig. 3.9. The dominant orientation is around ± 0 degrees, with the rotation augmentation range between -22 and 15 degrees.



(a)



(b) $-22.46°$ (c) $15.33°$

Fig. 3.9 Eye orientation distribution in the OCT imaging dataset (a) and two sample images corresponding to the eye orientations, $-22.46°$ (b) and $15.33°$ (c), respectively.

- Vertical and Horizontal Translation: Based on the calculated intensity weighed centres of mass for all OCT images (see Fig. 3.3), we defined a range of vertical and horizontal translation augmentations, $\pm 5px$ for the vertical translation (up-down) and $\pm 8px$ the horizontal translation (right-left) respectively.

- Horizontal Flip: Reversing all rows and columns of an OCT image's pixels allowed a mirror image OCT to be obtained, representing a fellow eye (right for left and left for right) and allowing the model to learn this unpredictable variability.
- Gaussian Blur: The calculated noise scores in Section 3.1.3 were used to guide a Gaussian blur data augmentation technique. A Gaussian filter with a mean of $\mu = 0$ and variance $\sigma^2 = 1.0$ was applied using a 5×5 kernel.

Loss Function The mean absolute error (MAE), the mean squared error (MSE), and the Huber loss (HL) functions were separately used. These functions are commonly used by the optimiser to minimise training errors. While MAE presents the average of the absolute differences between the actual and predicted visual acuity, scored as ETDRS letters, MSE loss is calculated as the average of the squared differences. We utilised HL as a combination of the MAE and the MSE, meaning that HL represents a quadratic behaviour for minor errors and a linear behaviour for significant errors. Moreover, HL and MAE are more robust to data with outliers. Minor error values calculated by these loss functions reflect a better model. In addition, to further improve the performance of DL models, we also modified the HL function, considering image quality assessment to guide the optimisation. In our formulation, the parameter δ defines the threshold at which the loss function transitions from the quadratic to the linear regime (see Eq. 3.5). Importantly, δ is adaptive. It is updated during training based on the anomaly scores derived from the image quality assessments (see Section 3.1.3 and Algorithm 1). A smaller δ made the loss function more sensitive to small errors, while a larger δ made it more robust to anomalies. This adaptive adjustment enables the optimizer to focus on minimizing prediction errors at the first epochs and gradually the model focuses on the model's robustness

$$L_\delta = \begin{cases} \frac{1}{2}(y-\hat{y})^2 & if\,|(y-\hat{y})| < \delta \\ \delta((y-\hat{y}) - \frac{1}{2}\delta) & otherwise \end{cases} \tag{3.5}$$

where $\delta = \{d\}$ is the adaptively adjusted parameter of the HL function, which could provide more robustness to anomalies.

Parameter Selection The stated loss functions were optimised via the ADAM algorithm with a fixed number of epochs ($n_{epoch} = 1000$), resulting in sufficient learning within fewer epochs in our experiments. The learning rate was set to 1×10^{-5} with weight decay ($w = 1 \times 10^{-5}$) and momentums ($\beta_1 = 0.5, \beta_2 = 0.9$), and an automatic learning schedule were added. The DL models were trained by dividing the dataset into batches (batch size=38). Lastly, we saved the parameters of each DL model when the model's performance started to decrease since this reduction is a strong indication of over-fitting.

Evaluation Metrics To evaluate the performance of the models to predict both pre and postoperative visual acuity, we used the following metrics: R-squared, root-mean-square error (RMSE), mean absolute error (MAE), and the Pearson correlation coefficient. The R-squared value ranges from 0 to 1, with a higher value indicating a better fit between the predicted and actual values.

RMSE and MAE values show better performance when they are closer to zero. Pearson correlation coefficients range between -1 and +1. A value of -1.0 shows a perfect negative correlation, whilst + 1.0 shows a perfect positive correlation. A zero correlation shows no relationship between the predicted and actual values. We used these four metrics to allow a comprehensive comparison with previous studies in the literature.

3.1.4 Experimental Design and Results

We present results for both preoperative and postoperative visual acuity. Preoperative visual acuity is known before surgery, but we were interested in assessing the model's performances for both measurements. To ensure the preoperative VA models were trained using the same data set as the postoperative VA models, we kept the same data augmentation methods and model parameters for both tasks. However, the cross-validation setup for the preoperative and postoperative VA predictions was adjusted due to the different distributions of these measurements. We utilized stratified five-fold cross-validation, where the folds were created by preserving the distribution of the VA measurements across all folds. This ensured that each fold contained a similar range and distribution of VA measurements, providing balanced training and testing sets for both preoperative and postoperative data. Specifically, for each fold, 152 OCT images were used for training, and 38 OCT images were used for testing, with the stratification ensuring that both preoperative and postoperative measurements were well-represented in each fold. The modified HL function performed better than the other mentioned loss functions; therefore, we represented the results of optimising the modified HL function in this section.

Results based on preoperative VA

Table 3.3 shows the quantitative comparison between our trained DL models using a different number of OCT image slices, using MAE as the evaluation metric. All evaluation metric values are the means obtained using five-fold cross-validation. Statistically significant results are highlighted in bold. The obtained results clearly show that the majority of tested DL predictive models performed best with seven OCT image slices.

Table 3.4 illustrates the performance of nine DL models using the seven central OCT image slices, with all four evaluation metrics given as means obtained using five-fold cross-validation. The experimental results demonstrated that ResNet-18 was the most predictive in all scores, achieving 0.47 for R-squared, 7.34 for RMSE, 0.65 for the Pearson correlation coefficient, and 5.96 for MAE.

Results based on postoperative VA

Table 3.5 shows the quantitative comparison between our trained DL models using a different number of OCT image slices, with MAE as the evaluation metric (means obtained using five-fold cross-validation). Statistically significant results are highlighted in bold. The obtained results clearly show that the majority of tested DL predictive models performed best with the central seven OCT image slices.

Table 3.3 The mean absolute error values, based on preoperative VA measurements, were obtained for nine state-of-the-art DL predictive models using a different number of OCT image slices through our designed loss function (the best results are highlighted in bold). All evaluation metric values are given as the means and standard deviations obtained using five-fold cross-validation.

MODELS	OCT IMAGE SLICES									
	1	**3**	**5**	**7**	**9**	**11**	**13**	**17**	**19**	**21**
VGG-11	**6.72±1.07**	6.88±1.11	7.09±0.98	7.21±0.89	7.48±0.82	7.21±0.83	7.67±0.69	7.56±0.87	7.42±0.62	7.46±0.77
VGG-19	6.56±1.12	**6.51±0.99**	6.74±0.97	6.88±0.82	7.07±0.55	6.94±0.91	7.20±0.66	7.41±0.89	7.05±0.78	7.11±0.77
ResNet-18	6.32±1.3	6.38±0.94	6.79±0.93	**5.96±0.72**	7.04±0.67	6.90±0.72	6.77±0.87	7.05±0.92	7.29±1.12	7.07±1.19
ResNet-34	6.27±1.20	6.60±1.23	6.78±1.17	**6.01±0.78**	6.96±0.79	7.03±0.68	7.26±0.65	6.72±0.89	6.94±0.91	6.75±0.97
ResNet-50	6.94±0.97	7.59±0.82	7.20±0.62	**6.75±0.89**	7.40±0.86	7.53±0.80	7.42±0.49	7.26±0.49	7.08±0.30	7.22±0.26
Inception v3	6.95±1.14	6.66±1.12	7.48±0.94	**6.42±0.83**	7.41±0.58	8.61±0.30	8.78±0.26	8.47±0.41	8.20±0.48	8.58±0.73
DenseNet-121	**6.08±0.91**	7.24±1.29	7.06±1.15	6.97±0.77	7.18±0.65	7.17±0.93	6.75±0.78	6.88±0.86	6.82±0.81	6.92±0.83
EfficientNetV2-L	6.58±1.01	6.83±0.97	6.82±1.01	**6.57±0.89**	6.90±1.15	6.98±1.23	6.95±1.18	7.00±0.93	6.97±0.95	7.06±1.11
ViT-B/16	7.05±0.70	6.88±0.62	6.93±0.62	6.90±0.63	6.98±0.60	6.82±0.64	**6.70±0.58**	6.71±0.51	6.87±0.51	6.95±0.52

Table 3.4 Quantitative comparison of nine state-of-the-art DL predictive models with seven OCT image slices, using four different evaluation metrics through our designed loss function, as the means and standard deviations obtained with five-fold cross-validation, based on preoperative VA measurements (the best results are highlighted in bold).

MODELS	**R2**	**RMSE**	**Pearson**	**MAE**
VGG-11	0.27	9.70	0.52	7.21±0.89
VGG-19	0.34	9.227	0.58	6.88±0.82
ResNet-18	**0.47**	**7.34**	**0.65**	**5.96±0.72**
ResNet-34	0.40	7.75	0.62	6.01±0.78
ResNet-50	0.25	9.04	0.39	6.75±0.89
Inception v3	0.38	8.27	0.61	6.42±0.83
DenseNet-121	0.30	9.37	0.55	6.97±0.77
EfficientNetV2-L	0.37	8.49	0.59	6.57±0.89
ViT-B/16	0.39	9.48	0.52	6.90±0.63

Table 3.5 The mean absolute error values, based on postoperative VA measurements, were obtained for nine state-of-the-art DL predictive models using a different number of OCT image slices through our designed loss function (the best results are highlighted in bold). All evaluation metric values are the means and standard deviations obtained using the five-fold cross-validation.

MODELS	OCT IMAGE SLICES									
	1	**3**	**5**	**7**	**9**	**11**	**13**	**17**	**19**	**21**
VGG-11	8.19±1.43	8.74±1.38	8.20±1.21	**7.94±1.11**	8.10±1.36	8.57±1.20	8.82±1.22	8.25±1.12	8.13±1.09	8.62±1.01
VGG-19	7.94±1.66	**7.76±1.12**	8.47±1.04	7.87±1.11	8.28±1.21	8.02±1.28	8.46±1.56	8.85±1.32	8.95±1.22	8.77±1.23
ResNet-18	7.29±1.76	7.31±1.25	7.67±1.01	**6.81±0.81**	7.52±1.24	8.09±1.47	7.45±1.50	7.45±0.95	7.57±1.12	7.35±0.80
ResNet-34	7.06±1.10	7.58±1.84	7.29±1.25	7.15±1.31	7.16±1.55	**6.96±1.59**	7.55±1.73	7.48±1.36	7.46±1.22	7.63±1.55
ResNet-50	7.37±1.61	7.70±1.56	7.37±1.51	**6.84±0.37**	7.26±0.54	7.58±0.90	7.59±1.23	7.64±1.30	7.41±1.14	7.75±1.62
Inception v3	**7.38±1.41**	7.57±1.28	8.01±1.43	7.57±1.51	7.60±1.12	7.78±1.01	8.01±1.51	7.66±1.60	8.07±1.35	7.91±1.23
DenseNet-121	7.87±1.19	7.75±1.27	7.68±1.01	**7.64±0.88**	8.02±1.22	7.79±1.37	7.60±1.41	8.01±1.29	7.67±1.41	7.90±1.39
EfficientNetV2-L	7.98±2.11	7.76±1.29	7.73±1.13	7.57±1.01	7.61±0.98	7.47±0.98	7.35±1.08	**6.97±1.16**	6.99±1.13	7.2±0.98
ViT-B/16	8.12±0.80	8.04±0.72	7.97±0.64	**7.89±0.61**	7.89±0.66	7.90±0.75	7.93±0.73	7.88±0.74	7.92±0.70	8.02±0.69

Table 3.6 illustrates the performance of the nine DL models using seven OCT image slices, using all four outcome metrics: R-squared, RMSE, Pearson correlation coefficient, and MAE. The experimental results demonstrated that ResNet-50 was the most effective and superior in most evaluation scores, achieving 0.46 for R-squared, 9.01 for RMSE, 0.69 for the Pearson correlation coefficient, and 6.84 for MAE.

In Fig.3.10 and3.11, we show descriptive results for the relationship between ground truth and predicted VA values and corresponding confidence intervals to demonstrate how closely our best-trained model predicts VA values on the test dataset. In Fig.3.10, predictions that fall near the red dotted line, representing perfect correlation, indicate well-performing cases due to high-quality training data and clear OCT image features. Conversely, significant deviations from this line suggest underperformance, resulting from cases with high anomaly scores, insufficient training examples in certain VA ranges, or complex disease characteristics that the model struggles to capture. Similarly, Fig.3.11 shows differences between the ground truth and predicted postoperative VA values. The red dotted lines denote the 95% confidence interval obtained by the ResNet-50 model on the test dataset (as highlighted in Table3.6), while the solid red line represents the gold standard. This figure not only quantifies the prediction errors but also emphasizes that the well-performing results generally stem from cases where the input data is clean and well-represented in the training set. In contrast, larger prediction errors are associated with more challenging cases—where image anomalies or inherent variability lead to higher uncertainty in the model's output. Overall, these results confirm that our proposed fully automated image informatics framework can robustly predict both preoperative and postoperative visual acuity measurements for patients with idiopathic full-thickness macular holes using a high-quality SD-OCT image dataset.

Table 3.6 Quantitative comparison of nine DL predictive models using seven OCT image slices, showing four different evaluation metrics through our designed loss function (with means obtained by five-fold cross-validation), as the means and standard deviations obtained with five-fold cross-validation, based on postoperative VA measurements (the best results are highlighted in bold).

MODELS	R2	RMSE	Pearson	MAE
VGG-11	0.27	10.77	0.52	7.94±1.11
VGG-19	0.25	11.52	0.45	7.87±1.11
ResNet-18	0.43	9.31	0.65	**6.81±0.81**
ResNet-34	0.36	9.95	0.60	7.15±1.31
ResNet-50	**0.46**	**9.01**	**0.69**	6.84±0.37
Inception v3	0.26	10.12	0.51	7.57±1.51
DenseNet-121	0.23	10.35	0.47	7.64±0.88
EfficientNetV2-L	0.36	10.05	0.53	7.57±1.01
ViT-B/16	0.38	10.79	0.60	7.89±0.61

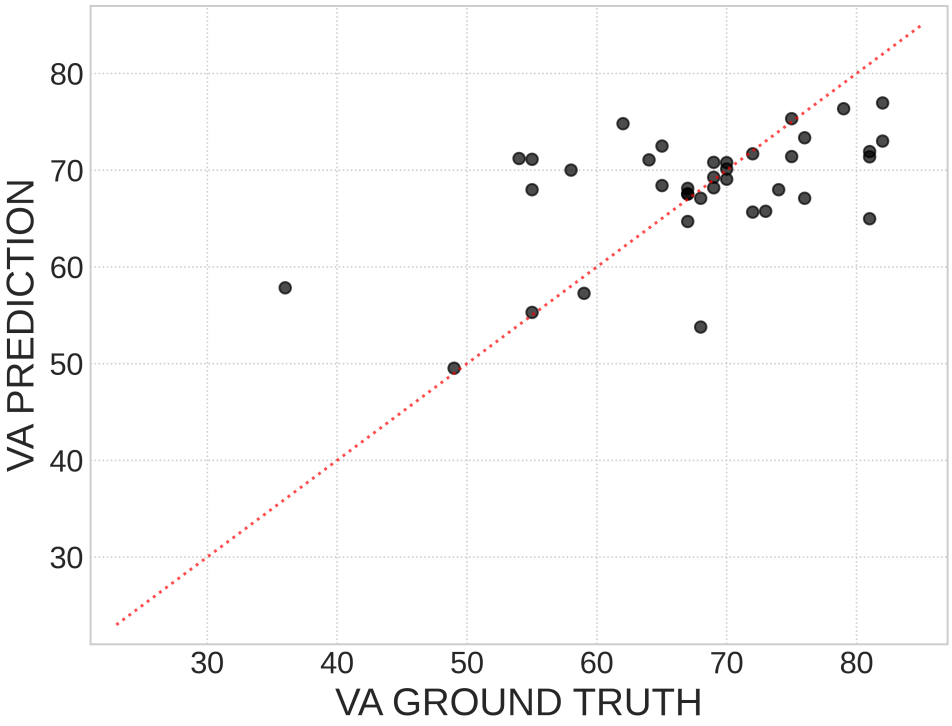


Fig. 3.10 The scatter plot visualises the relationship between the ground truth and predicted postoperative VA measurements obtained by the ResNet-50 model on the test dataset (the highlighted result in Table 3.6). The red dotted line depicts the gold standard.

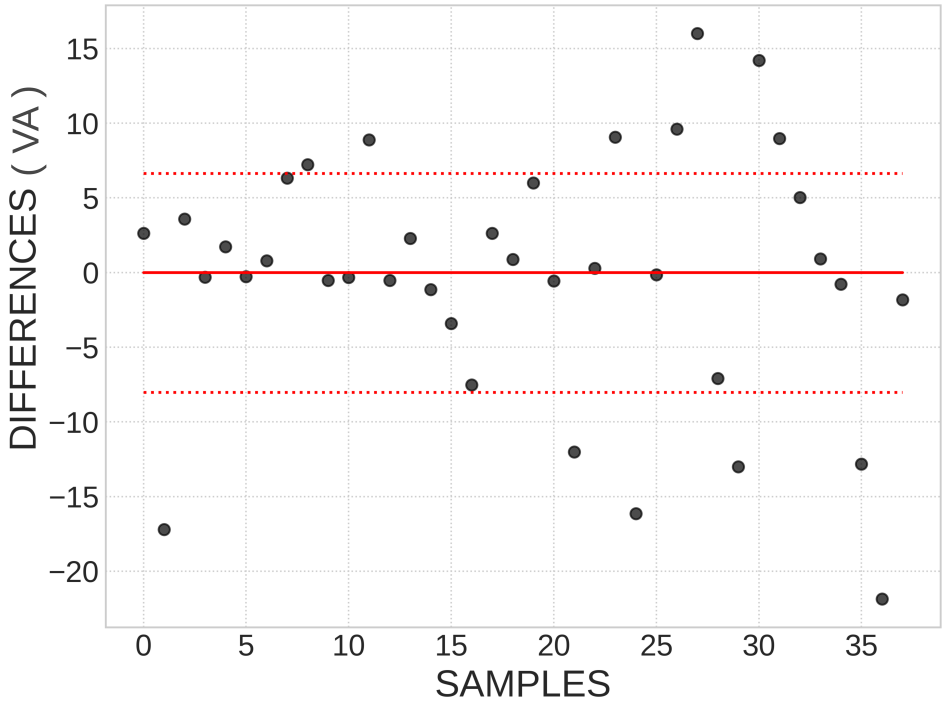


Fig. 3.11 The 95% confidence interval between the ground truth and predicted postoperative VA values is shown with the red dotted lines (-8.02, 6.46) obtained by the ResNet-50 model on the test dataset (the highlighted result in Table 3.6). The solid red line depicts the gold standard.

3.1.5 Clinician Evaluation

Feedback was obtained from three retinal specialists who reviewed both the methodology and example outputs to assess the potential clinical utility of the proposed predictive framework. They expressed strong interest in the method's ability to provide quantitative predictions of visual outcomes, which could support clinical decision-making and patient counselling before MH surgery. Clinicians particularly valued the model's robustness to variability in OCT image quality and noted that automated image quality screening reduced their concerns about interpretability. However, they also noted that the integration of such tools into clinical workflows would require further validation on larger, more diverse datasets and real-time implementation capabilities. Overall, they considered the proposed approach promising, especially in its ability to complement expert judgment with objective, reproducible outputs.

3.1.6 Discussion and Conclusion

We present a full image informatics approach to predict visual acuity outcomes in people undergoing surgery to treat MHs using preoperative SD-OCT images and deep learning-based predictive models.

To overcome the impact of high variations in real-world image quality on the robustness of the deep learning model, an extensive imaging data assessment and quality assurance procedure was implemented. Data preparation steps, including scaling, centre of mass detection, and cropping, were used to unify the imaging dataset's scale, size and centration. Further, data

quality assessment measurements, including noise, blurriness, contrast, motion, and brightness-darkness scores, were calculated to identify and exclude abnormalities in the imaging dataset.

The resultant high-quality imaging dataset was then used to train nine state-of-the-art 2D deep learning-based predictive models for both pre and postoperative VA using multiple channels (2D+), followed by a quantitative performance comparison with our designed loss function.

All tested models were able to predict preoperative visual acuity with less than an 8.78 MAE letter score, with the best predictive model achieving a 5.96 MAE score with 0.47 for R-squared, 7.34 for RMSE, 0.65 for the Pearson correlation coefficient. Similarly, all tested models were able to predict postoperative visual acuity with less than an 8.95 letter MAE score, with the best predictive model achieving a 6.84 MAE score with 0.46 for R-squared, 9.01 for RMSE, 0.69 for the Pearson correlation coefficient.

The CNN-based backbone networks mostly demonstrated high predictive performance, as evidenced by their competitive results in predicting preoperative and postoperative VA measurements. The reason might be that they leverage their inherent ability to analyse hierarchical features for complex structure analysis on OCT images [39]. Another might be that they likely recognize patterns regardless of the location of relevant structures in the input OCT image [150]. On the contrary, a certain standard has not been achieved regarding robustness when the number of input channels is altered.

In addition, the ViTs model did not perform as effectively as the CNN-based state-of-the-art models. This is likely because the ViT-based models improve training efficiency on large-scale datasets [44]. However, our dataset is relatively small. Furthermore, the ViTs come with expensive overhead, causing large parameter sizes [44, 178]. Hence, it is a computationally expensive and time-consuming process. On the contrary, the ViT models presented more robust performance. This is also because the ViT-based models may incorporate attention mechanisms to focus on the most relevant part of images and disregard irrelevant noise [69]. However, the introduction of redundant information and insufficient sparsity can impede the improvement of robustness in ViT-based models, leading to performance degradation [19].

As a result, ResNet architectures show slightly better results among nine state-of-the-art DL models. One of the reasons might be that it has residual blocks, leading to fewer vanishing problems. Additionally, ResNet-50 obtained the best results, which may be due to a deeper architecture and having a bottleneck. Overall, the prediction of preoperative visual acuity had better performance than postoperative VA in most of the metrics. As other studies have shown, however, preoperative VA is strongly correlated with postoperative VA. While the current study did not explicitly model the correlation between pre- and postoperative VA, it is clinically recognised and briefly noted that preoperative VA is a significant, but not always a reliable indicator of postoperative outcomes. Some patients experience meaningful vision gains after surgery, while others may not improve significantly, depending on individual retinal condition, chronicity, or other factors not fully captured in the OCT scan alone [117, 148].

Although other studies have reported similar results for the important and informative prediction of postoperative VA [79, 117, 122], our study provides more robust results, as we validated our results using an independent data set. Indeed, our best model achieved the highest metrics in all evaluation scores, achieving 0.52 for R-squared, 9.23 for RMSE, 0.71 for the

Pearson correlation coefficient, and 6.47 for MAE, as shown in Table 3.1. These compare very favourably to previous traditional regression modelling methodologies. Furthermore, while we trained the predictive model on a limited dataset acquired by only one type of OCT imaging system, albeit in two hospitals and four different devices, the proposed 2D DL-based predictive approach contains a comprehensive image informatics framework with our designed loss function that can be applied across a breadth of many 3D medical image datasets.

To overcome those limitations, our future research work will focus on adapting full 3D deep learning-based predictive models, the uncertainty of 2D and 3D DL-based predictive models, and a substantive scale-up of the OCT imaging data size and types.

3.2 3D Convolutional Neural Network based Deep Learning Models

In this section, we compared the performance of 2D and 3D versions of five state-of-the-art DL neural networks on predicting VA following surgery for idiopathic full-thickness MHs using an image dataset of SD-OCT scans. To make this study more comparable, using the same dataset revealed the differences between 2D and 3D versions of DL neural networks. Based on our results, 3D networks generally outperformed the 2D networks in R-squared and Pearson correlation coefficient; however, they fell behind in mean absolute error. 3D networks also come with the sacrifice of significantly more computational complexity.

3.2.1 Introduction

CNNs have achieved high performance in OCT image analysis studies; however, there have only been a limited number of studies investigating postoperative VA measurements and mostly performing 2D networks [79, 122]. With the recent development of CNN structures, graphical processing power, and the increasing accessibility of high resolution OCT imaging data, 3D CNN's applications in OCT image analysis have been growing rapidly [142].

Compared to 2D, the advantage of 3D networks is that they can capture spatial information in one more axis [52, 105]. Evaluating the performance of a 3D network as compared to a 2D approach to predicting postoperative VA in our dataset would offer a significant test of the technique's potential benefits, with high applicability to other domains [172]. Additionally, even though CNNs have been applied to retinal layers in OCT images with significant effort [159], the majority of techniques are still 2D [8, 122, 132], and the researchers consider a single 2D mid-plane image since it may contain the essential information needed for the research or clinical diagnosis. More specifically, the study [122] suggested a model for predicting postoperative VA, utilising a standard CNN-based model from preoperative clinical data. However, the predictions were performed in the 2D mid-planes of OCT images and the study was limited by the absence of publicly available visual acuity labelled data for the intended analysis.

Similarly, we were unable to access publicly available data with their visual acuity labels in this study. However, we select several best-performing 2D CNNs based on our OCT dataset

and apply their 3D counterparts on the same dataset to compare their performance directly. This comparison of the same dataset reveals the differences between 2D and 3D versions of deep learning neural networks. We believe this section provides a comprehensive understanding of how 2D and 3D CNNs evaluate and predict postoperative visual acuity from preoperative OCT images (see Figure 1.4). For future reference, a summary of the advantages and disadvantages of using 2D or 3D networks is provided based on the results.

3.2.2 Methods

This section presents a comprehensive image informatics framework to create a high-quality OCT image dataset, including image preprocessing, image quality assessment, and anomaly detection as explained in Section 3.1.3. It then summarises training and experimental setup for 2D and 3D SD-OCT images where the kernels are 2D and 3D convolutions, respectively (see Figure 3.12).

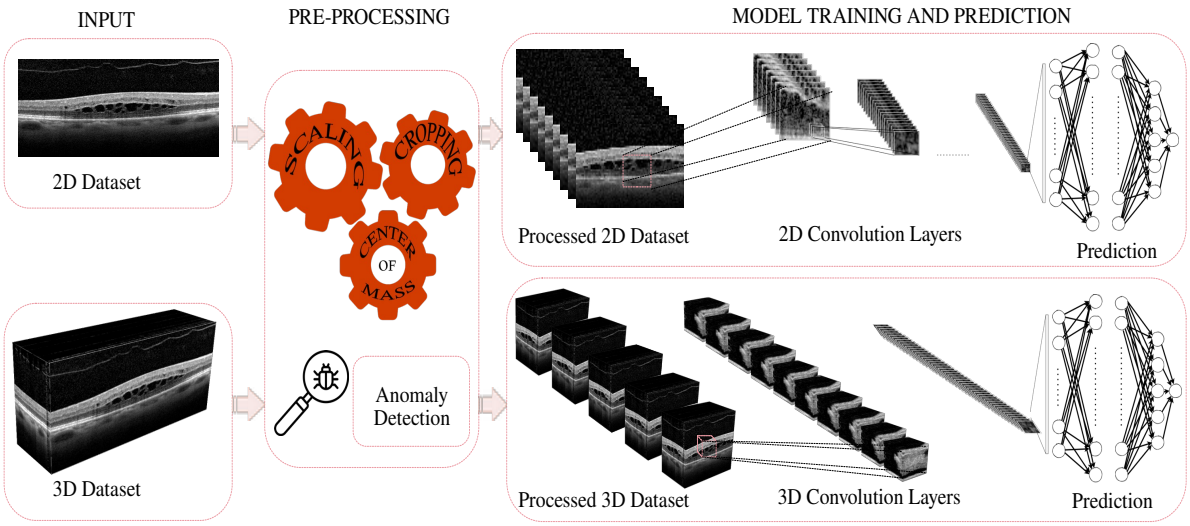


Fig. 3.12 Flowchart of the proposed methods. The first stage corresponds to the input 2D and 3D SD-OCT image dataset and VA measurements obtained by ophthalmologists, the second stage incorporates OCT data preprocessing (i.e. scaling, the centre of mass detection, cropping, and anomaly detection based on image quality measurements). With the obtained high-quality image dataset and labels, multiple state-of-the-art DL models are trained to predict postoperative VA measurements in the final stages.

Data Preparation

The preparation of 3D SD-OCT images involves several critical steps to manage the high variability in the MH location and resolution due to ocular anatomy and acquisition distortions. These steps include scaling, intensity weighted centre of mass, cropping, and anomaly detection, based on image quality measurements. These steps, as elaborated in Section 3.1.3, ensure that the dataset used for training the models is of high quality, ultimately leading to more reliable predictions.

2D CNNs and Their 3D Counterparts

The top-five best-performing networks were chosen to be applied in this work: DenseNet-121 [68], ResNet-18 [61], ResNet-34 [61], and MobileNet v2 [65]. The 3D versions of these networks share the same general structure, with all 2D convolutional kernels that move in 2 directions—axial and coronal switched to 3D kernels that move in 3 directions—axial, coronal, and sagittal.

Training and Experimental Setup

All networks were trained using Python 3.8.10, CUDA 11.4, cuDNN 8, PyTorch 1.9.0+cu102 running on a 64-bit Ubuntu operating system using a 3.4 GHz Intel Core-i9 with 32 GB of RAM and NVIDIA GTX 1080 Ti GPU with a frame buffer of 11 GB GDDR5X. Following Section 3.2.2, the remaining data was split uniformly into only training and test sets using random five-fold cross-validation with a ratio of 80% and 20% since the dataset size is small. To ensure the 3D networks are trained using the same data set as the 2D networks, we kept the same five-fold cross-validation set-up, where each fold contains 152 and 38 patients as train and test data sets. For training the 2D networks, we kept the training set up identical to the 3D networks by using an input image size of 452×204 *px* is the mid-slice of each 3D volume. For training the 3D networks, the size of all 3D volumes is $452 \times 204 \times 49$ *px*. Note that to maintain a balanced choice of data augmentation methods, we kept the same augmentations: rotation ± 10 degree, vertical and horizontal translation $\pm 3px$, horizontal flip, and Gaussian blur.

Loss Function and Parameter Selection

The mean absolute error (MAE) function was used for both networks since MAE is more robust to data with outliers. We used ADAM [82] optimiser with a fixed number of epochs($n_{epoch} = 1000$), and an initial learning rate was set to 1×10^{-5}, which was reduced during training following the "PolynomialLR" as described in [27] with weight decay ($w = 1 \times 10^{-4}$) and momentums ($\beta_1 = 0.5, \beta_2 = 0.9$). Note that to maintain a balanced choice of batches, we divided the dataset into 38 batches for all tests.

Evaluation Metrics

To evaluate the performance of the models to predict postoperative visual acuity, we used the following metrics: R-squared, mean absolute error (MAE), and Pearson correlation coefficient. The R-squared value ranges from 0 to 1, and a greater value denotes a better fit between the predicted and actual values. Note that when MAE values are close to zero, the performance of the model improves. Pearson correlation coefficients range from +1 to -1, with +1 representing a positive correlation, -1 representing a negative correlation, and 0 representing no relationship between the predicted and actual values. We selected these three metrics to provide a good comparison by considering previous studies in the literature.

3.2.3 Results

We present results for predicting postoperative visual acuity. With the results, we evaluate the performances and complexities of the five state-of-the-art DL neural networks where the kernels are 2D convolutions for the input 2D OCT images and 3D convolutions for 3D SD-OCT images.

Prediction Performance Comparison

Figure 3.13 shows the quantitative comparison between our trained 2D networks DL algorithms considering each fold separately, with MAE as the evaluation metric. The box plot displays statistically the predictions of the minimum, maximum and mean values of five-fold cross-validation along with the position of the lower and upper quartiles.

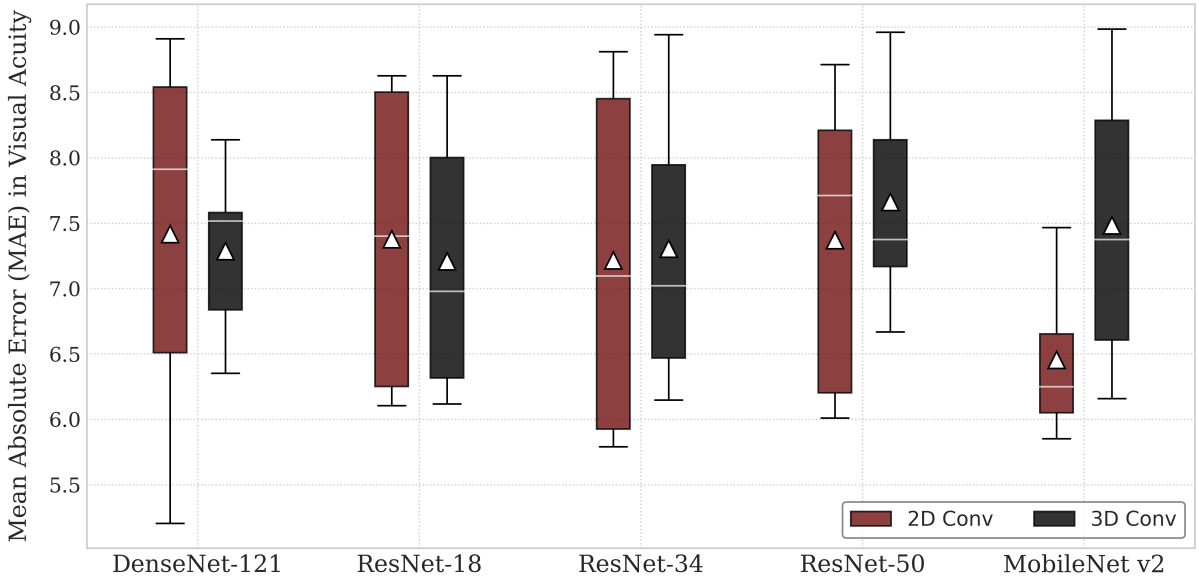


Fig. 3.13 Box plot for 2D and 3D networks - A five-fold cross-validation MAE results in VA score with five different deep learning algorithms. The median marks the mid-point of each model prediction result and is shown by the white line. The mean of the model prediction results is marked as a white triangle.

Similarly, Figure 3.13 displays the quantitative comparison between our trained 3D networks DL algorithms considering each fold separately, with MAE as the evaluation metric. The box plot illustrates the predictions of the minimum, maximum and mean values of five-fold cross-validation along with the position of the lower and upper quartiles statistically.

The obtained results show that while the 3D networks have better performance than 2D networks in DenseNet-121, ResNet-34, ResNet-18, and ResNet-50, the 3D networks in MobileNet v2 lagged behind the 2D networks, shown by the white line in Figure 3.13. The performance advantages of 3D networks in MAE come from the fact that 3D networks have access to the whole 3D volume, while 2D networks only have access to individual B-scans (axial slices in 2D).

Considering the minimum MAE values of five-fold cross-validation, DenseNet-121 in 2D networks shows the best performance among 2D and 3D versions of all DL neural networks. Considering the maximum MAE values of five-fold cross-validation, MobileNet v2 and ResNet-50 in 3D performs worse among all DL algorithms (see Figure 3.13).

Table 3.7 Quantitative comparison of five state-of-the-art DL models on a uniform test dataset with 2D and 3D networks, using three different evaluation metrics (the best results are highlighted in bold for each evaluation metric).

MODELS	2D Networks			3D Networks		
	R2	**Pearson**	**MAE**	**R2**	**Pearson**	**MAE**
DenseNet-121	0.35	0.61	6.51	0.15	0.45	6.83
ResNet-18	0.36	0.62	6.10	**0.46**	**0.70**	6.11
ResNet-34	0.41	0.65	**5.79**	0.43	0.69	6.46
ResNet-50	0.30	0.54	6.32	0.20	0.43	6.54
MobileNet v2	0.21	0.50	6.38	0.10	0.35	6.60

In addition, to make a more straightforward comparison, we also examined R-squared and Pearson correlation coefficient evaluation metrics and the quantitative prediction performance of 2D and 3D versions of the five DL algorithms used on the same uniform dataset. As seen in Table 3.7, the outcomes of the experiments in 2D networks showed that ResNet-34 had consistently good performance in most evaluation metrics, achieving 0.41 for R-squared, 0.65 for the Pearson correlation coefficient, and 5.79 for MAE. And the outcomes of the experiments in 3D networks illustrated that ResNet-18 consistently performed well in most evaluation scores, achieving 0.46 for R-squared, 0.70 for the Pearson correlation coefficient, and 6.11 for MAE. These results conclude that ResNet-18 is more suited to 3D adaptation than the other three networks. In the meantime, Resnet-34 is more appropriate for 2D networks. Also, MobileNet v2 in 3D performs worse among all DL algorithms in R-squared and Pearson correlation coefficient evaluation metrics (see Table 3.7).

Additionally, to demonstrate how closely our best-trained by 2D and 3D Resnet-34 models could predict VA values on the test dataset, we visualised descriptive results for the relationship between the ground truth and the predicted VA values and associated confidence intervals in Figure 3.14.

Computational Complexity

Despite applying the same network structure, the computational complexity varied significantly between the 2D and 3D versions of the DL neural networks. Table 3.8 presents the computational complexity of the 2D and 3D versions of the five DL neural networks utilised. Compared to 2D networks, 3D networks have significantly more parameters and higher computational complexity (MACs–a number of multiply-and-accumulate operations). This is because 3D convolutional kernels have more parameters than 2D convolutional kernels, and 3D volume inputs have more pixels than 2D inputs.

3.2.4 Conclusion

In this work, we directly compared the performance of 2D and 3D versions of five state-of-the-art DL neural networks to predict postoperative VA measurements in people undergoing surgery to

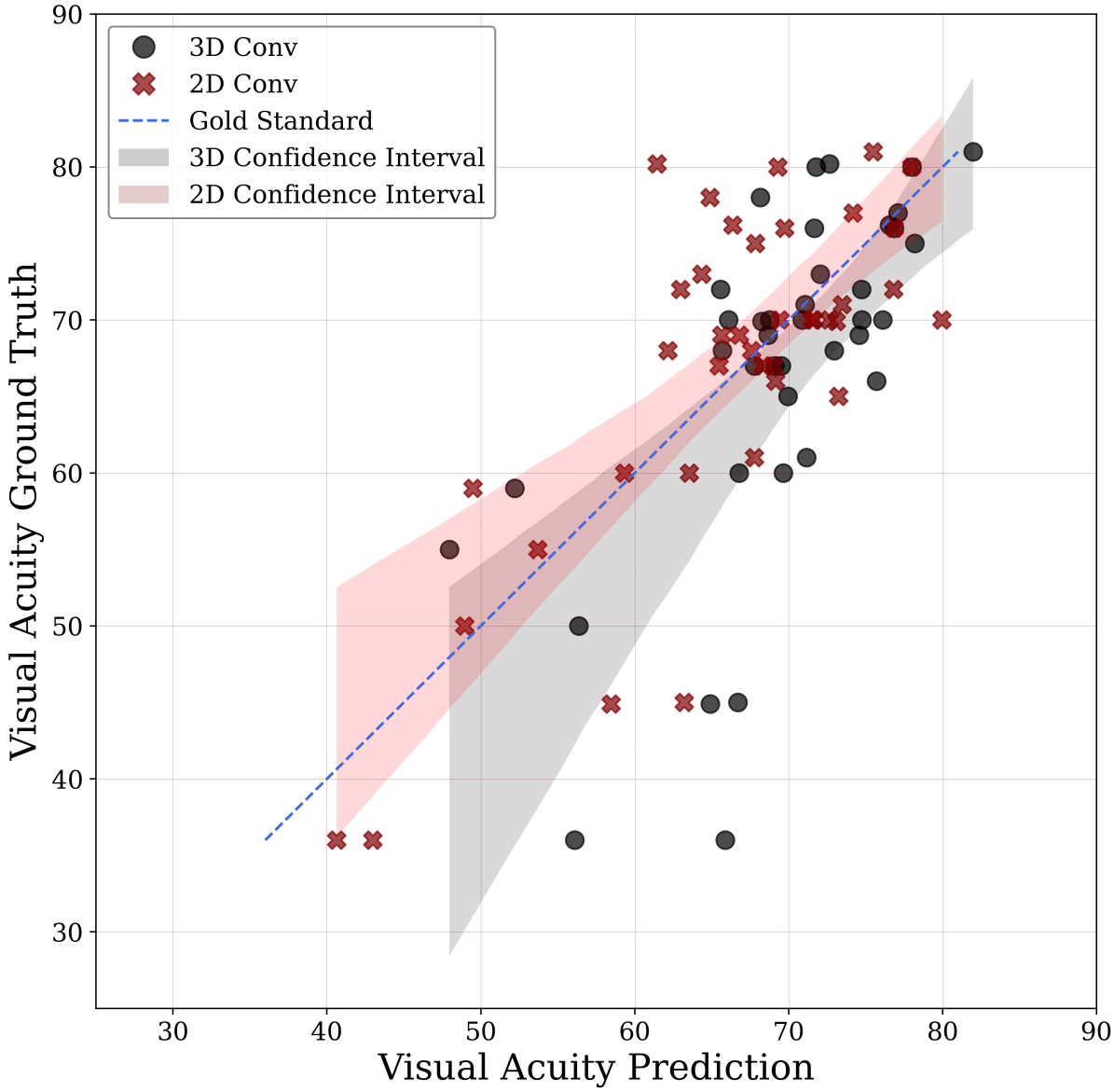


Fig. 3.14 The scatter plot for 2D and 3D networks visualises the relationship between the ground truth and predicted postoperative VA measurements obtained by the ResNet-34 model on the test dataset (the highlighted result in Table 3.7). The blue dotted line depicts the gold standard.

Table 3.8 The computational comparison (the best results are highlighted in bold for each network model).

MODELS	2D Networks		3D Networks	
	Parameters	**MACs**	**Parameters**	**MACs**
DenseNet-121	6.94	2.80	11.24	102.19
ResNet-18	11.17	1.74	**33.16**	**77.70**
ResNet-34	**21.27**	**3.60**	63.47	135.33
ResNet-50	41.12	4.8	121.98	248.36
MobileNet v2	2.24	0.31	2.35	5.96

treat idiopathic full-thickness macular holes using real-world preoperative 3D SD-OCT images. The prediction results from 3D networks have better performance than their 2D counterparts in most performance metrics, except for MAE. However, 3D networks require significantly more computational power compared to 2D networks. In conclusion, if there is enough computing power, choosing 3D networks should be prioritised when predicting postoperative VA in 3D SD-OCT scans.

3.3 Summary

In this chapter, we presented a robust image informatics framework for predicting visual acuity outcomes in patients undergoing macular hole surgery, leveraging spectral-domain optical coherence tomography (SD-OCT) images. The framework integrates advanced techniques in image preprocessing, quality assurance, and anomaly detection to enhance the reliability of predictive models.

We conducted a thorough evaluation of both 2D and 3D convolutional neural network (CNN) based deep learning models. The 2D CNN models demonstrated considerable efficacy in predicting visual acuity outcomes, achieving a mean absolute error (MAE) of 6.47 ETDRS letters score. These models showed high predictive accuracy while handling various image quality issues.

In contrast, the 3D CNN models, which utilize additional spatial dimensions, generally provided superior performance in terms of R-squared and Pearson correlation coefficient. However, they also presented increased computational complexity and, in some instances, a higher MAE compared to their 2D counterparts. This comparison underscores the trade-offs between the enhanced spatial resolution of 3D models and their computational demands.

Overall, the findings of this chapter offer valuable insights into the strengths and limitations of both 2D and 3D CNN approaches in the context of OCT image analysis. The choice between 2D and 3D models should be guided by the specific requirements of the application, including the need for computational efficiency versus the potential benefits of additional spatial information. Future work will focus on optimising these models and exploring hybrid approaches to further improve predictive accuracy and practical applicability.

Chapter 4

Uncertainity of Deep Learning Models in Predicting Visual Acuity

Contents

The surgery of idiopathic full-thickness macular holes (iFTMHs) has received considerable interest in retinal diseases, particularly in the era of high-street spectral-domain optical coherence tomography (SD-OCT). Current outcomes of the following interventions rely heavily on predicting postoperative visual acuity (VA) through artificial intelligence models to enable decision-making and optimally advising patients, delivering promising performance. However, their black-box behaviour is opaque to users and uncertainty associated with their predictions is not typically stated, leading to a lack of trust among clinicians and patients. In this chapter, we describe an uncertainty-aware regression model (U-ARM) for predicting VA for people undergoing macular hole surgery using preoperative SD-OCT images. In addition to predicting VA following surgery, U-ARM also displays its associated uncertainty. We then qualitatively evaluated the performance of U-ARM. Lastly, we demonstrate out-of-sample data performance, generalising well to data outside the training distribution, low-quality images, and unseen instances not encountered during training. The results show the capability of U-ARM compared to commonly used methods in terms of prediction and reliability.

4.1 Introduction

Macular holes (MH) can be closed by vitreoretinal surgery. Successful closure of iFTMHs improves VA and, ultimately, the quality of life for patients [155]. However, the degree of VA improvement following surgery is variable and influenced by factors including the duration of symptoms, iFTMHs size, preoperative VA and the type of interventions [148]. Also, the surgery itself is not without risks and side effects [148]. Hence, patients and retinal surgeons have to make important decisions on whether surgery should proceed. Recent studies have proposed that postoperative VA outcomes are closely related to preoperative VA and MH size measured by SD-OCT machines [84, 93, 148]. They have also focused on the prediction of postoperative VA using manual 2D measurement (base diameter (BD) and minimum linear diameter (MLD)) of MHs and preoperative VA despite their limited predictive abilities [32, 80, 117, 148, 163].

Previous studies also considered medical records variables, including age, gender, axial length (AL), symptom duration, randomization group, clinician-measured BD and MLD using a 2D slice of a horizontal OCT image in predicting postoperative VA [32, 117]. Moreover, [118] used 3D measurements of MHs on OCT images (minima and maxima of BD, and minima and maxima of MLD). Although using 3D (rather than 2D) measurement has improved the ability to predict VA [118, 163, 174], further improvements in predictive ability are still needed [89, 117]. There were also other morphologic variables that they did not include, such as external limiting membrane height and outer retinal integrity. These variables are important for the prediction of postoperative VA and reliability in practice [32].

Presently, research on DL-based models for the diagnosis [8, 115, 162, 171, 177], classification [66, 117, 131], and prediction [79, 94, 122] of VA has attracted considerable interest after vitrectomy surgery using 2D or 3D preoperative OCT images, with both studies reporting promising results. [89] achieved better results which not only predicted the postoperative VA but also evaluated using 2D and 3D images as inputs for DL models. However, the failure of these studies to consider uncertainty related to model prediction leads to limitations in areas such

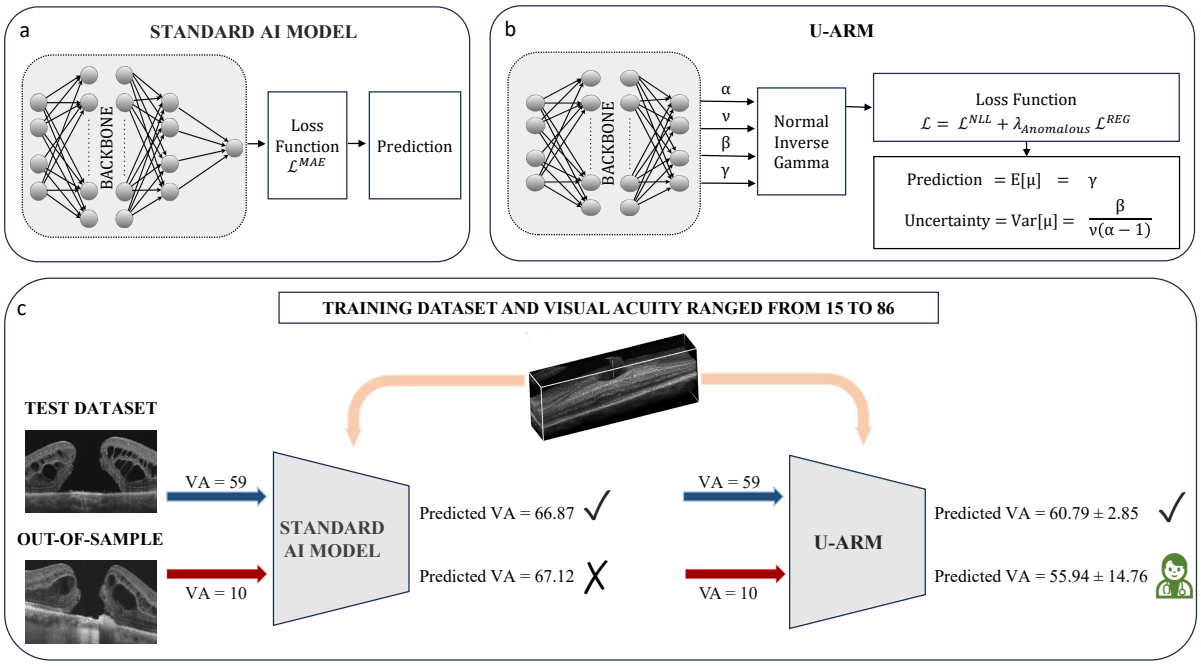


Fig. 4.1 (**a**) The structure of the standard AI model. (**b**) The structure of our proposed uncertainty-aware regression model (U-ARM). (**c**) The overall framework of DL models.

as transparency, trustworthiness, and reliability, which need to be overcome before DL models can be applied in practice. A potential solution to this gap is using uncertainty quantification (UQ) methods that mitigate the inherent black-box behaviour of DL models [1, 180]. UQ methods play a pivotal role in representing model predictions with uncertainties in a trustworthy manner and can increase the confidence of model predictions [79]. To the best of our knowledge, UQ methods have not been thoroughly evaluated for VA prediction, leading to this study's main motivation. Whilst DL methods are a key factor in the development of advanced tools capable of surpassing human experts for this particular task [32, 170], the use of DL models on out-of-sample (OOS) data (including data outside the training distribution, low-quality images and atypical instances) can be difficult in clinical applications, even with UQ methods [31, 42, 100, 164]. These challenges reveal the clinical need for a robust and reliable DL model that can be generalised on OOS data, providing additional necessity for this study.

In this section, we tackle the problem of (i) predicting postoperative VA and its associated evidence to learn the model uncertainty, and of (ii) objectively and rigorously validating in-sample (IS) and OOS data performance. Our contributions can be summarized as follows:

- We introduce an uncertainty-aware regression model (U-ARM) based on deep evidential regression to estimate postoperative VA and its associated evidence to learn model uncertainty.
- We adopt abnormalities in the data to prevent over-confidence in predictions and over-inflation of uncertainty during training.
- We test and compare three typical UQ methods to our proposed U-ARM.

- We conduct internal and external evaluation tests using publicly available datasets (OCT-Newcastle [89] and HD-OCT of MH [94]),
- We analyze the performance of U-ARM and a standard AI model on IS and OOS data to assess their potential generalizability. If the test data is within the range of the training dataset label and with lower abnormality, U-ARM will predict the label with a low uncertainty score to indicate a reliable prediction. On the contrary, if the test data is OOS, such as falling outside the range of the training dataset labels, being a low-quality image, or one with high abnormality, U-ARM will predict the label with a high uncertainty score to indicate an unreliable prediction. In such cases, a manual check by an experienced ophthalmologist becomes necessary. Hence, the estimated uncertainty allows U-ARM to provide reliable predictions for OCT images and labels involved in training data and to avoid confusion from OOS data.

4.2 Methods

Here, we describe the imaging data collection, the data analysis and preparation, the framework of our proposed U-ARM and other baseline models, the experimental setup and the evaluation metrics used.

4.2.1 Data collection and description

We used two sets of SD-OCT imaging datasets for this study. The first set is the OCT-Newcastle dataset described in Section 2 as 3D images [88, 89]. We excluded 25 images due to variations in image size and quality. This dataset was subsequently split into training (148) and test sets (37) with a ratio of 80% and 20%, respectively.

To further test the performance of U-ARM and the other baselines, we used a second set of OCT images (HD-OCT of MH [94]) that is a publicly available dataset and published as 2D images. This dataset consists of 2658 confirmed iFTHMs images from 493 patients, along with their clinical data. However, in this study, we selected only 320 OCT images from patients who had undergone successful hole closure surgery and had the best-corrected VA at three months post-operatively. There were mostly two scans per image (horizontal and vertical OCT scans) of $750 \times 500px$ in size. The VA of the HD-OCT of the MH dataset ranged from 10 to 85, showing an imbalanced distribution. We excluded 35 images due to variations in quality and VA scores falling outside the defined range. The VA range was determined using the upper and lower limits of the OCT-Newcastle dataset's VA scores, with any VA scores below 29 or exceeding 86 considered outside the range. The HD-OCT of the MH dataset then was split into training (228) and test sets (57) with a ratio of 80% and 20%, respectively.

We also used the excluded images (in both the OCT-Newcastle and the HD-OCT of MH dataset) to evaluate the performance of all models in detecting OOS data. The excluded dataset consists of 25 excluded images from the OCT-Newcastle dataset and 35 excluded images from the HD-OCT of MH dataset.

4.2.2 Data analysis and preparation

We first standardised the image size to reduce overfitting and to improve the models. For all images, the centre position of images was found using the intensity-weighted centre of mass, calculated for each dimension separately. We first cropped the images in the OCT-Newcastle dataset with the size of the smallest image ($452 \times 204 \times 49px$) as a reference and used the centre position of each image. We subsequently cropped the images in the HD-OCT of the MH dataset ($554 \times 250px$), maintaining a consistent ratio along both x and y axis. To augment the variety and amount of data artificially during the training of the models, the intensity-weighted centre of mass was also used for vertical ($\pm 5px$) and horizontal ($\pm 8px$) translation since differences in the centre position of images for both datasets remain. We then calculated the eye orientation to define the data augmentation range for the rotation (-22° and $+15^\circ$).

Since DL models depend highly on the supplied data, the quality of data must be considered. As would be intuitively expected, the acquired OCT images vary in quality because of lens opacities, ocular saccades, blink artefacts, and OCT operator skills. We considered this problem to be unavoidable, and measured speckle noise [43], contrast [140], blurriness [51], brightness and darkness [22], and motion [64, 112] of all images. We thus demonstrated the effect of data quality during the model training. This served as a guide for anomaly detection and proved how effective the data quality is through UQ methods.

DL models define high-level features from data, and any nonlinear relationships within data impact the learning. The images show notable non-linearity and a high variance in quality assessment measurements. Therefore, we separately implemented the LOF [23] anomaly detection method for each dataset and then excluded them. The LOF also produced anomaly scores for 2D scans per image, assisting in observing the uncertainty in predictions made by the models.

4.2.3 Framework of U-ARM and other baseline models

Uncertainty caused by variability in data, errors in DL model architecture, and unknown data can be estimated with a few different approaches [1, 180]. These methods largely provide a probabilistic representation of single-point estimates with uncertainty quantification methods in model predictions, with sampling during inference or training with out-of-distribution data. Hence, we considered the standard AI model, dropout sampling, model ensembling and U-ARM.

The standard AI model consisted of a backbone network to conduct the experiment, predicting only a single value (Fig. 4.1). ResNet-50 is the most popular DL architecture used for medical image analysis; hence, we employed it as a backbone network. Considering the uncertainty quantification and to enable reasonable comparisons with the other baseline models, the model was trained over ten times, and the final prediction was generated using the predictive mean over ten experiments.

The other common approach (dropout sampling) involves randomly removing a proportion of nodes inside the model architecture, mostly after the last pooling layer; instead of making a single deterministic prediction, it runs multiple forward passes with multiple dropout rates enabled. The outputs of these multiple runs are then averaged to produce a final prediction, and the variance among these predictions is used as a measure of uncertainty. In this work, we

employed pre-trained ResNet-50 as a backbone network and incorporated dropout sampling only during model training. Our experiments were conducted using different dropout rates, specifically $= \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.75, 0.8, 0.9\}$.

The other methodology for understanding UQ is the use of model ensembling. Model ensembling provides multiple predictions for each input by training many DL models independently, where the ensemble members are trained in parallel with the same input but without any interaction. This results in a distribution of predictions for input data. The variance and mean of the range of predictions is then used for UQ. In this work, we quantified the uncertainty using ten state-of-the-art DL models with the same input data and parameters: ResNet-18, ResNet-34, ResNet-50, VGG-11, VGG-19, DenseNet-121, DenseNet-169, Inception-v3, Res2Net50-14w-8s, Res2Net50-26w-6s.

Based on these commonly used deterministic-based methods for regression problems, we proposed U-ARM based on deep evidential regression [9]. Besides the other baseline methods, U-ARM does not require sampling during inference or training with out-of-distribution data. Furthermore, it tunes the regularisation coefficient when calibrating uncertainty, which is a primary limitation of deep evidential regression [9]. It trains a backbone network model to learn the hyperparameters of evidential distribution (Fig. 4.1). U-ARM is simply composed of three distinct parts:

(1) Maximising the DL model fit. To accomplish this, [9] proposed the observations, y_i, from a Gaussian distribution with unknown means (μ) and variance (σ^2) place a Gaussian prior on the unknown means and an Inverse-Gamma prior on the unknown variance:

$$\mu \sim \mathcal{N}(\gamma, \sigma^2 \nu^{-1}), \qquad \sigma^2 \sim \Gamma^{-1}(\alpha, \beta), \tag{4.1}$$

where $\Gamma(\cdot)$ is the gamma function, $m = (\gamma,\ \nu,\ \alpha,\ \beta)$, and $\gamma \in \mathbb{R}$, $\nu > 0$, $\alpha > 1$, $\beta > 0$ are the hyperparameters. Together, the distributions of μ and σ^2 form the Normal Inverse-Gamma (NIG) evidential prior:

$$\begin{aligned} p(\underbrace{\mu, \sigma^2}_{\theta} | \underbrace{\gamma, \nu, \alpha, \beta}_{m}) &= \mathcal{N}(\gamma, \sigma^2 \nu^{-1})\, \Gamma^{-1}(\alpha, \beta), \\ &= \frac{\beta^{\alpha}\sqrt{\nu}}{\Gamma(\alpha)\sqrt{2\pi\sigma^2}} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \\ &\quad \exp\left\{-\frac{2\beta + \nu(\gamma-\mu)^2}{2\sigma^2}\right\}, \end{aligned} \tag{4.2}$$

The interpretation of parameters in the conjugate prior distribution involves "virtual-observations," where the mean of a NIG distribution is estimated from ν virtual-observations with a sample mean γ, and its variance is estimated from α virtual-observations with sample mean γ and a sum of squared deviations 2ν. Total evidence, denoted as Φ, is defined as the sum of all inferred virtual-observations counts: $\Phi = 2\nu + \alpha$ [75]. Given a NIG distribution, [9] used the first-order moments of $p(\mu, \sigma^2 | \gamma, \nu, \alpha, \beta)$ representing the maximum likelihood prediction, $\mathbb{E}[\mu]$, aleatoric

uncertainty, $\mathbb{E}[\sigma^2]$, and epistemic uncertainty, $Var[\mu]$:

$$\underbrace{\mathbb{E}[\mu] = \gamma}_{\text{prediction}}, \underbrace{\mathbb{E}[\sigma^2] = \frac{\beta}{\alpha - 1}}_{\text{aleatoric}}, \underbrace{Var[\mu] = \frac{\beta}{\nu(\alpha - 1)}}_{\text{epistemic}}. \tag{4.3}$$

To maximise the DL model fit, [9] derived the model evidence of the NIG distribution of the i-th prediction, y_i, over the likelihood parameters θ:

$$\begin{aligned} p(y_i|m) &= \int_\theta p(y_i|\theta)p(\theta|m)d\theta \\ &= St\left(y_i; \gamma, \frac{\beta(1+\nu)}{\nu\sigma}, 2\alpha\right). \end{aligned} \tag{4.4}$$

$St(y_i;\ \mu_{St},\ \sigma^2_{St},\ \nu_{St})$ presents the Student-t distribution evaluated at y with location μ_{St}, scale parameter σ^2_{St}, and ν_{St} degrees of freedom. Using this result, [9] computed the negative log-likelihood loss, $\mathscr{L}_i^{NLL}$, for sample i as:

$$\begin{aligned} \mathscr{L}_i^{NLL} &= -\log p(y_i|m), \\ &= -\log\left(St\left(y_i; \gamma, \frac{\beta(1+\nu)}{\nu\sigma}, 2\alpha\right)\right), \\ &= -\frac{1}{2}\log\left(\frac{\pi}{\nu}\right) - \alpha\log(\Omega) \\ &+ (\alpha + \frac{1}{2})\log((y-\gamma)^2\nu + \Omega) \\ &+ \log\left(\frac{\Gamma(\alpha)}{\Gamma(\alpha + \frac{1}{2})}\right). \end{aligned} \tag{4.5}$$

where $\Omega = 2\beta(1+\nu)$.

(2) Minimizing evidence on errors. [9] also proposed a regularisation term, $\mathscr{L}_i^R$, using incorrect evidence penalty to minimise model evidence in instances of high predictive error:

$$\mathscr{L}_i^R(x) = |y_i - \mathbb{E}[\mu_i]| \cdot \Phi = |y_i - \gamma| \cdot (2\nu + \alpha). \tag{4.6}$$

(3) Rescaling the total loss based on data anomaly scores. We subsequently normalised the data anomaly scores obtained through image data quality in the previous section to tune the regularisation coefficient λ. Therefore, the total loss can be written as:

$$\mathscr{L}_i(x) = \mathscr{L}_i^{NLL}(x) + \lambda_{ANOMALY\_SCORE}\mathscr{L}_i^R(x), \tag{4.7}$$

$\lambda_{ANOMALY\_SCORE}$ denotes trades off between uncertainty inflation and model fit. $\lambda_{ANOMALY\_SCORE}$ approaching zero leads to an over-confident estimate, which if too high, results in over-inflation [9]. In this work, we thus utilized the total loss $\mathscr{L}_i(x)$, consisting $\mathscr{L}_i^{NLL}$ and $\mathscr{L}_i^R$, scaled by the regularisation coefficient over data anomaly scores. Following this, we used a pre-trained ResNet-50 as a backbone network to predict the hyperparameters, μ, ν, α, β. During training, while the pre-

dictions - $\mathbb{E}[\mu]$ were obtained by γ, the uncertainty was computed by the other hyperparameters (ν, α, β) using Eq. 4.3.

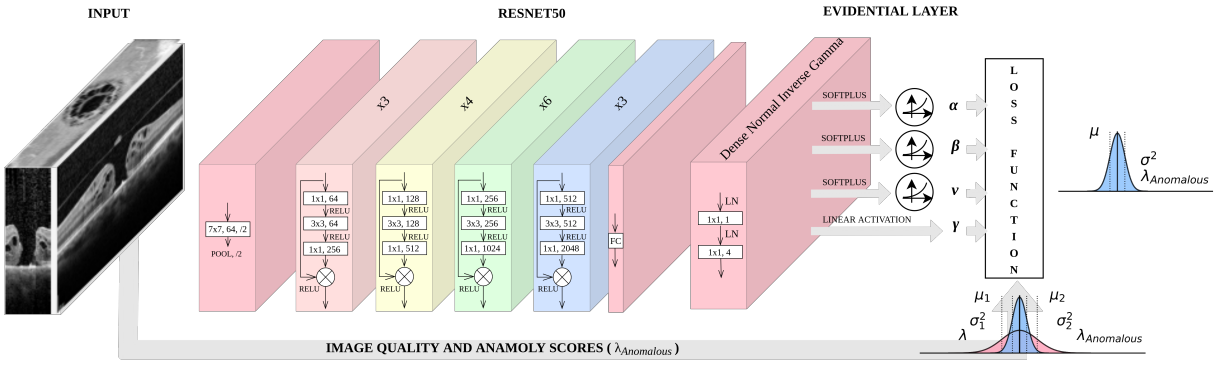


Fig. 4.2 U-ARM with Resnet50 backbone network and layer settings

In conclusion, while U-ARM builds upon the evidential model proposed by Amini et al. [9], it introduces key advancements that distinguish it from the original approach. The motivation for using deep evidential regression in U-ARM is to provide robust uncertainty estimation, particularly for challenging regression problems where quantifying uncertainty is crucial for decision-making. Deep evidential regression enables the model to output predictions and an associated uncertainty measure, which is vital in domains that require confidence in model predictions. However, while deep evidential regression offers a solid starting point, it faces two key challenges: handling heterogeneous data quality (such as noisy or outlier-prone inputs) and ensuring well-calibrated uncertainty estimates.

- **Dynamic Regularisation Scaling for Uncertainty Calibration:** The evidential model in [9] uses a fixed regularisation coefficient (λ), which can lead to suboptimal uncertainty estimation across varying data distributions. In contrast, U-ARM dynamically scales λ based on data anomaly scores, ensuring better uncertainty calibration that adapts to input quality.

- **Adaptive Uncertainty Estimation via Data Quality Scores:** The evidential model assumes a single λ value is sufficient for all input samples, which may result in over-inflation or overconfidence in uncertainty estimates. U-ARM addresses this by normalising anomaly scores and adjusting λ accordingly, providing a more flexible and data-driven uncertainty estimation strategy.

These modifications make U-ARM more robust, adaptable, and reliable for uncertainty quantification compared to the original evidential model [9].

4.2.4 Experimental setup

We trained the standard AI model, dropout sampling, ensemble models, and U-ARM on the OCT-Newcastle and the HD-OCT of MH datasets. All models in this chapter were trained from random initialization, without using pre-trained weights from Chapter 3 and with the following

hyperparameters: batch size of 37 in the OCT-Newcastle dataset, 57 in the HD-OCT of MH dataset, Adam optimisation with learning rate (1×10^{-4}), weight decay (1×10^{-3}), momentums ($\beta_1 = 0.4, \beta_2 = 0.8$), a fixed number of epochs (1000) with random initialisation to produce all presented results. According to the training set's MAE, the best model is saved and used for testing. U-ARM used the loss function defined in Eq. 4.7, and the other baseline models used the Huber loss function. Each model was trained 10 times except for U-ARM. We lastly preserved the same augmentations in all experiments to maintain a balanced selection of data augmentations: rotation, vertical and horizontal translation. All experiments were conducted on an AMD EPYC 7763 CPU and NVIDIA® GeForce Tesla A100. All models were implemented in Python 3.9 with the PyTorch 1.11.

In addition, the models were also trained and tested using the OCT-Newcastle dataset with the use of single and multiple OCT image scans (1, 3, 5, 7, 11, and 21), centred around the mid-scan defined by the intensity-weighted centre of mass - calculated in data analysis and preparation section. The standard 2D convolutions have an input of C * H * W, where C is the number of input channels, H is the height, and W is the width. The kernel shifts along two dimensions on the image. Considering this, the first convolutional layers of the models were modified. We also trained and tested the models with the input of a whole OCT image. The standard 3D convolutions have an input of C * D * H * W, where it additionally had the depth, D, meaning that 3D convolutions capture spatial information in one more axis. Therefore, the kernel shifted along three dimensions on the image.

4.2.5 Evaluation metrics

We utilised the following metrics to evaluate the model's uncertainty in the prediction of post-operative VA: MAE, R-squared, and RMSE. MAE refers to the average of absolute errors, which is the magnitude of the difference between the prediction and actual value. Similarly, the RMSE is a quadratic scoring method that quantifies the average magnitude of the difference. Therefore, lower MAE and RMSE values indicate better performance. R-squared evaluates a regression model's goodness of fit, and its values vary from 0 to 1, with a higher value indicating a better fit between actual values and predictions. In addition, we used the p-value to assess the statistical significance of the model's predictions and uncertainty prediction. The p-value represents the probability of observing the given results, or more extreme ones, assuming that the null hypothesis (no effect or no relationship) is true. It helps determine whether the relationship between the model's predicted and actual post-operative VA values is statistically significant. A low p-value (typically less than 0.05) indicates strong evidence against the null hypothesis, suggesting that the model's predictions are not due to random chance. Conversely, a higher p-value suggests that the model's predictions may not be significantly different from random variations in the data. In this study, the significance level is set at 0.05.

Table 4.1 Evaluation of model performance in the internal testing dataset across channel configurations. RMSE, MAE, and R2 for a standard AI model, dropout sampling, model ensembling, and U-ARM. Top scores for each metric are highlighted in bold (within statistical significance), n = 10 for sampling baselines. In both cases, the variance of model output was used for uncertainty estimation. All models were trained and tested on the OCT-Newcastle dataset with their VA scores.

Number of Channels	Standard AI Model				Dropout				Ensemble				U-ARM			
	MAE	**RMSE**	**R2**	**p-value**	**MAE**	**RMSE**	**R2**	**p-value**	**MAE**	**RMSE**	**R2**	**p-value**	**MAE**	**RMSE**	**R2**	**p-value**
1 Channels (2D Convolution)	6.77±0.34	9.50 ± 0.42	0.43	0.281	6.79 ± 0.40	9.66 ± 0.55	0.28	0.331	7.32 ± 0.34	10.00 ± 0.40	0.221	0.20	**6.34 ± 0.25**	**9.58 ± 0.20**	**0.45**	**0.075**
3 Channels (2D Convolution)	6.93±0.44	9.90 ± 0.43	0.26	0.310	6.88 ± 0.36	9.58 ± 0.59	0.32	0.307	7.03 ± 0.34	10.04 ± 0.44	0.20	0.227	**6.56 ± 0.20**	**9.97 ± 0.32**	**0.38**	**0.045**
5 Channels (2D Convolution)	6.89±0.41	9.84 ± 0.43	0.36	0.241	7.06 ± 0.37	10.13 ± 0.45	0.12	0.401	6.89 ± 0.31	9.58 ± 0.48	0.24	0.198	**6.39 ± 0.22**	**9.74 ± 0.23**	**0.46**	**0.017**
7 Channels (2D Convolution)	6.81±0.26	10.15 ± 0.43	0.27	0.198	6.89 ± 0.36	10.09 ± 0.68	0.35	0.299	6.87 ± 0.32	9.26 ± 0.40	0.30	0.207	**6.27 ± 0.17**	**9.25 ± 0.25**	**0.46**	**<0.005**
11 Channels (2D Convolution)	7.14±0.32	10.24 ± 0.42	0.28	0.207	7.14 ± 0.34	10.05 ± 0.38	0.10	0.417	7.22 ± 0.38	9.78 ± 0.34	0.18	0.213	**6.57 ± 0.30**	**9.88 ± 0.25**	**0.34**	**<0.005**
21 Channels (2D Convolution)	6.87±0.35	10.01 ± 0.60	0.36	0.271	7.07 ± 0.33	9.92 ± 0.41	0.13	0.408	7.26 ± 0.38	9.69 ± 0.45	0.21	0.245	**6.72 ± 0.21**	**9.78 ± 0.19**	**0.37**	**0.011**
49 Channels (3D Convolution)	6.90±0.28	10.00 ± 0.32	0.33	0.225	6.99 ± 0.34	9.59 ± 0.48	0.24	0.324	7.04 ± 0.57	9.01 ± 0.87	0.45	0.192	**6.57 ± 0.33**	**9.70 ± 0.32**	**0.49**	**<0.005**

4.3 Experimental Results

We first qualitatively tested and compared the performance of U-ARM trained separately with the OCT-Newcastle dataset across channel configurations and with the HD-OCT of MH dataset against the standard AI model, dropout sampling [53] and model ensembling [96] on internal testing datasets in the prediction of postoperative VA. We then demonstrated the performance of U-ARM against results presented for the set of baselines based on mean absolute error (MAE), root mean squared error (RMSE), R-squared (R2), and p-value. Following this, we evaluated the performance of U-ARM trained with the OCT-Newcastle dataset against the set of baselines and tested on the HD-OCT of the MH dataset, showing the performance of the external testing dataset. Additionally, we presented the performance with OOS test data. Training and additional details of the method to build the UQ models for this study are available in the Methods section.

4.3.1 Performance in the internal testing dataset

In the internal testing set with 37 images randomly selected from the OCT-Newcastle dataset, we tested the standard AI model, dropout sampling, model ensembling and U-ARM using different channel configurations trained with the OCT-Newcastle dataset (Table 4.1, statistically significant results in bold text) to quantify the performance of the model predictions and its associated uncertainties. U-ARM consistently outperformed the standard AI model, dropout, and ensemble across different channel configurations in predicting VA following surgery. U-ARM was the most effective prediction model and displayed the best evaluation metrics, achieving lower MAE, RMSE, and p-value indicating better results, and higher R2, suggesting improved model fit. With regards to uncertainty, U-ARM showed notably lower MAE variance and p-value, suggesting better uncertainty awareness and more reliable predictions than other models. The standard AI model and dropout sampling both performed favourably, while model ensembling indicated a slightly higher MAE in prediction. However, both the standard AI model and dropout sampling exhibited variability in MAE and RMSE in terms of uncertainty, rarely performing well as model ensembling.

In addition, we also monitored the performances of model prediction in relation to the number of channels of the inputs feeding the models (Table 4.1). The input with one and seven channels

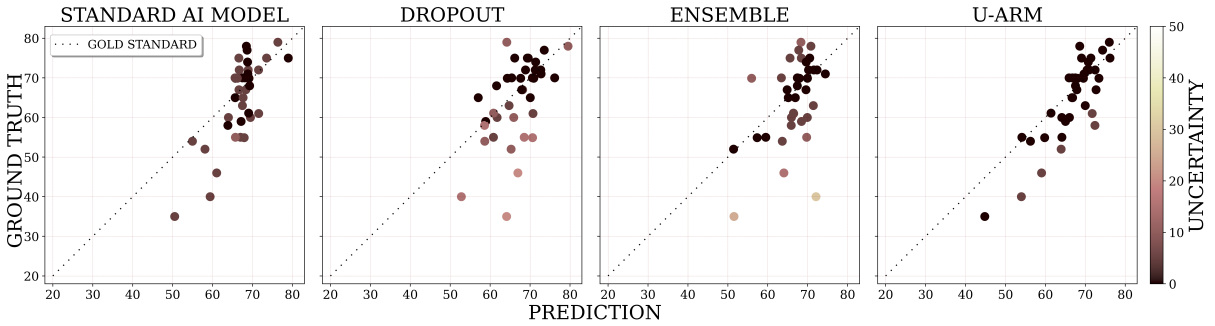


Fig. 4.3 The relationship between the ground truth, predictions and uncertainties obtained by all models on the internal testing dataset using seven OCT image scans (the highlighted result in Table 4.1); n = 10 for sampling baselines, and in both cases, the variance of model output was used for uncertainty estimation. The dashed red lines depict the gold standard. The colour bar visualises the uncertainty of the predictions, with dark colour representing low uncertainty and light colour indicating high uncertainty.

notably tended to have lower MAE values, uncertainties, and p-value compared to those of other configurations. This suggests that a moderate number of channels in the 2D convolution models may strike a balance between capturing closer predictions and minimising uncertainty in predictions. While the input with one channel performed competitively in all evaluation metrics, it was not as good and as reliable as the input with seven channels.

In the relationship between the ground truth, predictions, and uncertainties obtained by all models fed with seven OCT image scans (Fig. 4.3), U-ARM performed the best for prediction and uncertainty (6.27$\pm$0.17 for MAE, 9.25$\pm$0.25 for RMSE, 0.46 for R2 and p-value of <0.005). Using the standard AI model and U-ARM with one input channel resulted in values of 6.34$\pm$0.25 for MAE, 9.58$\pm$0.20 for RMSE, 0.45 for R2 and p-value of 0.075 (Table 4.1 and Fig. 4.4). Examining the effect of data abnormality revealed U-ARM to be more robust to anomalies than the standard AI model (Fig. 4.4). U-ARM was also found to have greater awareness of the uncertainty of the predictions, as image samples of a high uncertainty tended to have a high anomaly score.

As with the experiments with the OCT-Newcastle dataset, we also trained and tested the set of the models with HD-OCT of the MH dataset for predicting VA following surgery (Table 4.2, Fig. 4.5). U-ARM again performed better than other models in all evaluation metrics (6.88 for MAE, 9.58 for RMSE, 0.42 for R2, and p-value of <0.005). For uncertainty, U-ARM again demonstrated lower variance for both MAE, RMSE, and p-value indicating a better awareness of uncertainty and greater reliability of its predictions compared to other models. Although the standard AI model and ensemble showed comparable performance, the dropout sampling displayed variability in both prediction and uncertainty.

We also examined the ability of both the standard AI model and U-ARM to display the effect of data abnormality in their outcomes (Fig. 4.6). To further support the superiority of U-ARM and its underlying principle, we present a set of image samples from the worst to the best anomaly score (i, ii, iii, iv). From these, it is evident that U-ARM performed more robustly in response to an anomaly than the standard AI model, with samples having higher anomaly scores demonstrating higher uncertainty.

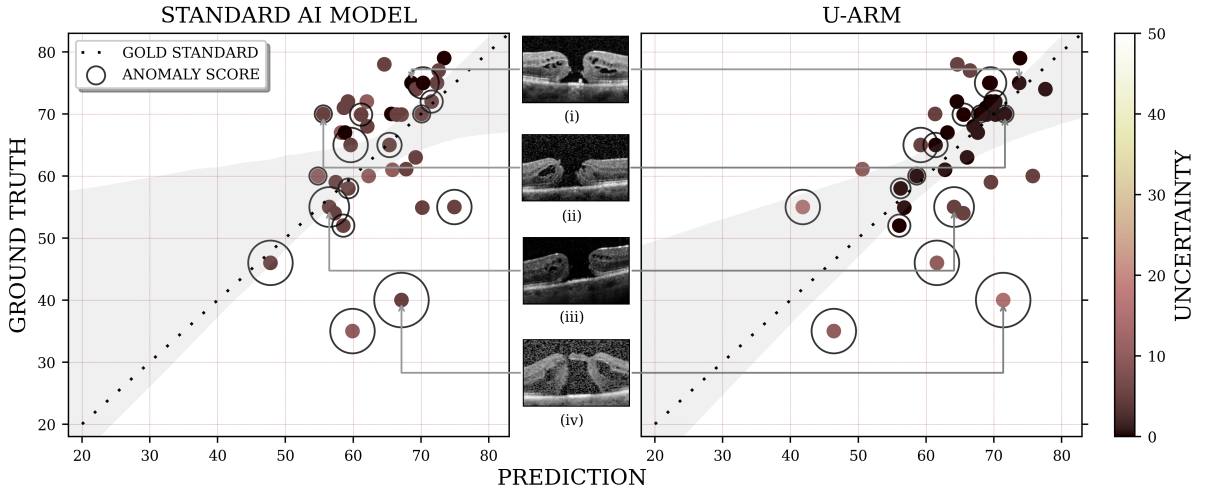


Fig. 4.4 The relationship between predictions, uncertainties, and data abnormality for standard AI model and U-ARM on the internal testing dataset using one OCT image scan (the highlighted result in Table 4.1); n = 10 for sampling baselines, and in both cases, the variance of model output was used for uncertainty estimation. The dashed red lines depict the gold standard. The grey area displays the regression fit. The black circle shows anomaly scores based on image quality. The colour bar visualises the uncertainty of the predictions with dark colour representing low uncertainty and light colour indicating high uncertainty.

Table 4.2 Evaluation of model performances in the testing datasets. RMSE, MAE, and R2 for a standard AI model, dropout sampling, model ensembling, and U-ARM. Top scores for each metric are in bold (within statistical significance), n = 10 for sampling baselines, and in both cases, the variance of model output was used for uncertainty estimation. While the first dataset was trained and tested on the HD-OCT of the MH dataset with their VA scores, the second dataset was trained with the OCT-Newcastle dataset and tested on the HD-OCT of the MH dataset with their VA scores for all models. The excluded and artificial dataset consists of both the OCT-Newcastle dataset and the HD-OCT of MH dataset.

Name of Test Dataset	Standard AI Model				Dropout				Ensemble				U-ARM			
	MAE	**RMSE**	**R2**	**p-value**	**MAE**	**RMSE**	**R2**	**p-value**	**MAE**	**RMSE**	**R2**	**p-value**	**MAE**	**RMSE**	**R2**	**p-value**
HD-OCT of MH-Internal	7.03 ± 0.44	10.34 ± 0.43	0.32	0.391	7.22 ± 0.42	11.39 ± 0.32	0.28	0.302	7.07 ± 0.36	10.40 ± 0.42	0.35	0.298	**6.88 ± 0.26**	**9.58 ± 0.33**	**0.41**	**<0.005**
HD-OCT of MH-External	7.47 ±0.45	10.34 ± 0.43	0.32	0.402	7.43 ± 0.45	11.39 ± 0.49	0.28	0.297	7.34 ± 0.39	10.68 ± 0.44	0.34	0.294	**7.15 ± 0.29**	**9.68 ± 0.39**	**0.39**	**0.039**
Excluded Dataset	10.25 ± 1.13	11.67 ± 1.20	0.21	0.420	10.17 ± 0.99	11.67 ± 1.01	0.27	0.346	9.97 ± 0.93	11.01 ± 0.95	0.27	0.301	**9.40 ± 0.79**	**10.45 ± 0.88**	**0.26**	**0.113**
Artificial Dataset	10.48 ± 1.17	12.01 ± 1.51	0.17	0.433	10.41 ± 0.99	11.71 ± 1.07	0.27	0.360	10.40 ± 1.00	11.63 ± 1.03	0.22	0.338	**10.36 ± 1.06**	**11.46 ± 1.02**	**0.13**	**0.117**

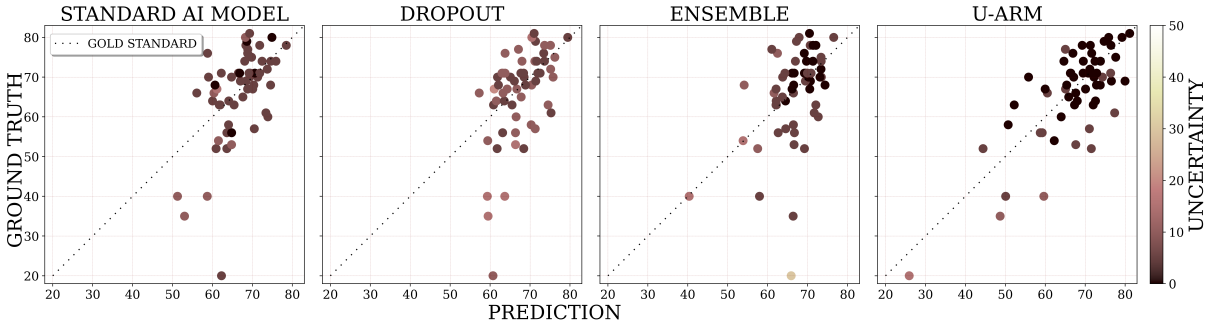


Fig. 4.5 The relationship between the ground truth, predictions and uncertainties obtained by all models on the internal testing HD-OCT of MH dataset (the highlighted result in Table 4.2); n = 10 for sampling baselines, and in both cases, the variance of model output was used for uncertainty estimation. The dashed red lines depict the gold standard. The colour bar visualises the uncertainty of the predictions with dark colour representing low uncertainty and light colour indicating high uncertainty.

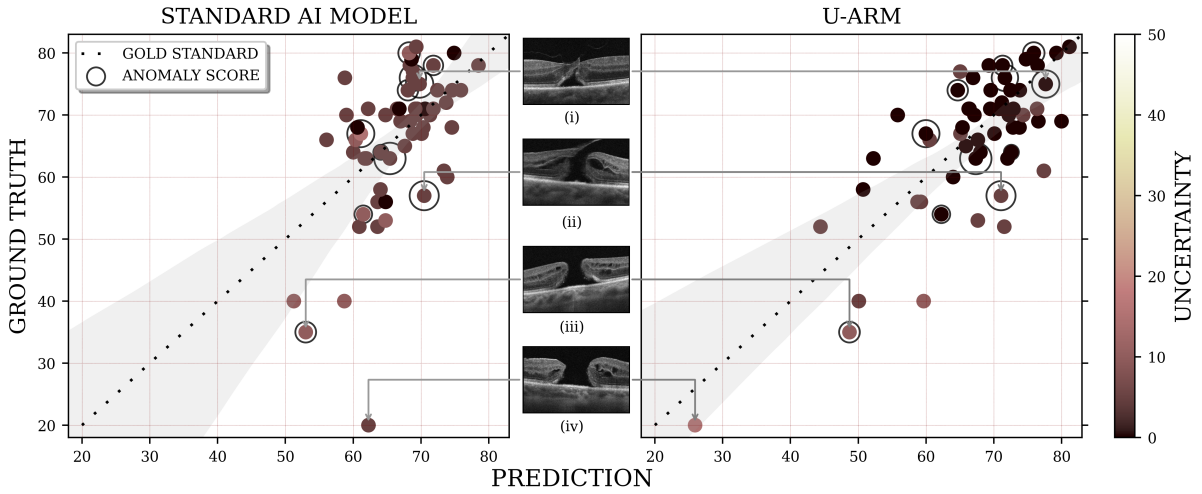


Fig. 4.6 The relationship between predictions, uncertainties, and data abnormality for both standard AI model and U-ARM on the internal testing HD-OCT of MH dataset (the highlighted result in Table 4.2); n = 10 for sampling baselines, and in both cases, the variance of model output was used for uncertainty estimation. The dashed red lines depict the gold standard. The grey area displays the regression fit. The black circle shows data anomaly scores based on image quality. The colour bar visualises the uncertainty of the predictions with dark colour representing low uncertainty and light colour indicating high uncertainty.

4.3.2 Performance in the external dataset

We also trained the models with the OCT-Newcastle dataset and then tested the models on the HD-OCT of the MH dataset to demonstrate the effectiveness of using the external dataset for further evaluating the potential generalisability of U-ARM and the other models. U-ARM again performed exceptionally well across all evaluation metrics, achieving 7.15 for MAE, 9.68 for RMSE, 0.39 for R2, and p-value of 0.039 (Table 4.2). As for uncertainty, U-ARM also showed notably lower variance for both MAE and RMSE, delivering more reliable predictions than other models.

In the relationship between the ground truth, predictions and uncertainties obtained by all models trained on the OCT-Newcastle dataset and tested on the HD-OCT of the MH dataset (Fig. 4.7), U-ARM performed slightly better, but the baseline models performed poorly in generalisation and were not able to learn meaningful patterns, likely due to the lack of data diversity.

In six sample images of the HD-OCT from the MH dataset with the predictions and uncertainties of the standard AI model and U-ARM trained on the OCT-Newcastle dataset (Fig. 4.8), the standard AI model yields a single output value as the final prediction for a given image, whereas U-ARM generates a prediction with an uncertainty score that reflects the reliability of the prediction. In the first two samples with lower anomaly scores, both the standard AI model and U-ARM provided similar predictions, even where distribution shifts were uncommon. This suggests that U-ARM is more conservative in its predictions, as indicated by close predictions and lower uncertainties. In the next two samples with moderate anomaly scores, the standard AI model and U-ARM generated good predictions, while U-ARM showed moderate uncertainties. In the last two image samples, both the standard AI model and U-ARM generated poor predic-

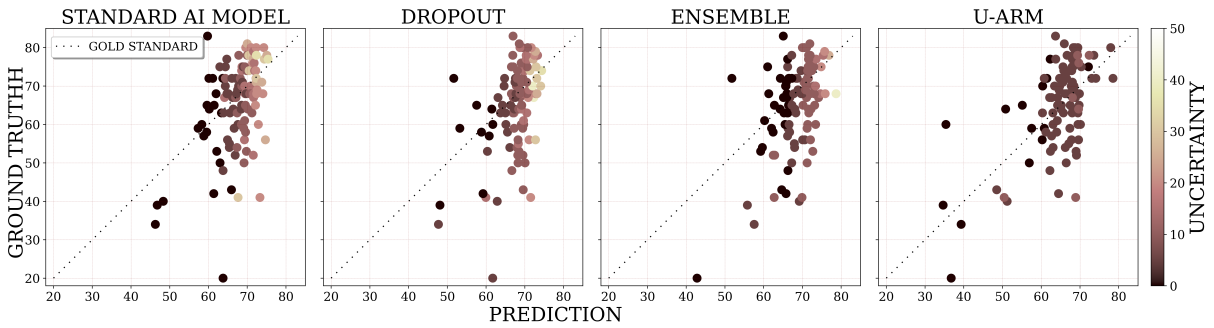


Fig. 4.7 The relationship between the ground truth, predictions and uncertainties obtained by all models trained on the OCT-Newcastle dataset and tested on the HD-OCT of MH dataset (the highlighted result in Table 4.2); n = 10 for sampling baselines, and in both cases, the variance of model output was used for uncertainty estimation. The dashed red lines depict the gold standard. The colour bar visualises the uncertainty of the predictions with dark colour representing low uncertainty and light colour indicating high uncertainty.

tions, with U-ARM presenting high uncertainties to indicate the low confidence the model had in its predictions. Overall, the results for U-ARM indicate that images with low and moderate anomaly scores generated lower uncertainties, resulting in good predictions. Conversely, images with high anomaly scores tended to generate higher uncertainties along with the prediction. What cannot be refuted is the fact that the standard AI model generates a single value as output regardless of its uncertainties.

4.3.3 Out-of-sample generalisation performance of the U-ARM and the other baseline models

We investigated the ability of U-ARM to capture uncertainties in OOS datasets compared with the standard AI model. When faced with abnormal samples outside the training dataset, the U-ARM demonstrated robust performance across diverse datasets (Fig. 4.8). The OOS datasets used in this section originate from the dataset described in Chapter 2 and include both the excluded samples and artificially generated data. Poor-quality images, identified by high anomaly scores, were classified as OOS data, while high-quality samples with low anomaly scores remained in-sample. Additionally, artificial datasets were created by systematically modifying image properties—including introducing noise and degrading image resolution to simulate challenging real-world conditions. Considering the excluded and artificial datasets contained out-of-range images, artificially noised images, and low-quality images (60 samples for both datasets), U-ARM largely indicated high uncertainties with a prediction that required manual verification to avoid possibly inaccurate predictions (Table 4.2). In the uncertainty density distribution of these datasets (Fig. 4.9), U-ARM exhibited high uncertainties for the samples from two groups of OOS datasets.

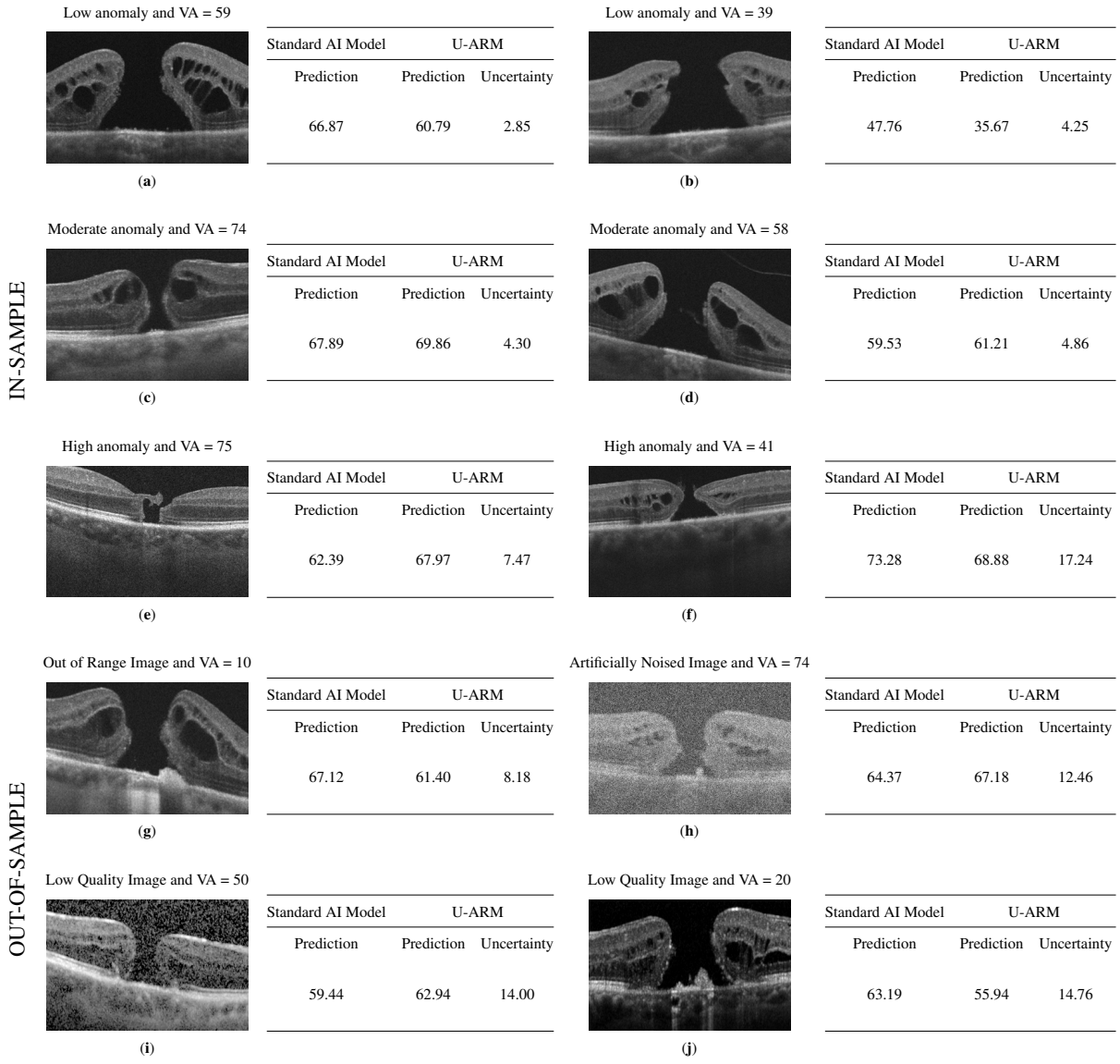


Fig. 4.8 Ten sample images of IS data from the HD-OCT of MH dataset and OOS data that were not included in the training. **a**, **b** Two samples with close predictions to their VA are from both the standard AI model and U-ARM, also indicating low anomaly and low uncertainty. **c**, **d** Two samples with roughly close predictions to their VA are from both the standard AI model and U-ARM, also indicating moderate anomaly with low uncertainty. **e**, **f** with far predictions to their VA are from both the standard AI model and U-ARM, also indicating high anomaly and high uncertainty, **g** An OOS sample is from HD-OCT of MH dataset due to being out of range, **h** An OOS sample is from OCT-Newcastle dataset and artificially added noise, **i** and **j** OOS samples are also from OCT-Newcastle dataset but are in low-quality. Samples predicted with the standard AI model and U-ARM. While the standard AI model only predicts VA scores as the final result, U-ARM will not only give the prediction but also provide the corresponding uncertainty score to reflect the reliability of the prediction result.

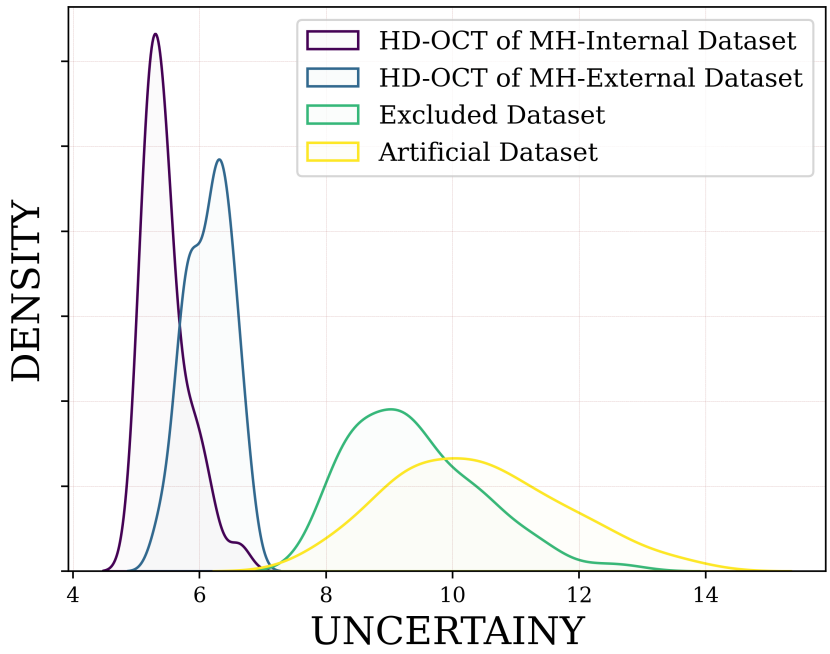


Fig. 4.9 Uncertainty density distribution for different datasets. Different coloured solid lines indicate different test datasets for predicting VA scores, presented in Table 4.2.

4.4 Discussion and Conclusion

In this study, we have developed a novel uncertainty-aware DL model that uses preoperative OCT images to predict postoperative VA following macular hole surgery at three months. Despite the limited number of OCT images, U-ARM demonstrated its ability to predict VA with good discriminative performances when testing the input of single and multiple internal OCT scans on two different datasets (MAE of 6.34±0.25, R2 of 0.45, and p-value of 0.075 for one channel on the OCT-Newcastle dataset; MAE of 6.27±0.17, R2 of 0.46, and p-value of <0.005 for seven channels on the OCT-Newcastle dataset; MAE of 6.88±0.26, R2 of 0.41, and p-value of <0.005 for one channel on the HD-OCT of MH dataset). These results were consistent when cross-validation was applied. Since the strongest predictor of visual improvement was MH size, [32] and [117] proposed fully automated 3D image analysis methods. However, these methods come with limitations that may lead to inaccuracies and variability. For example, OCT-based MH size and index ratios do not account perfectly for the hole asymmetry and overall hole shape, and their OCT images were restricted to patients undergoing surgery in one centre. In addition, the measurement of MH dimensions from OCT is time-consuming and requires expertise. Also, [94], [122], and [88, 89] proposed standard AI models using preoperative 2D and 3D OCT images, reporting similar results to U-ARM. However, U-ARM is a simple and promising alternative in predicting VA to the studies based on medical records or 2D and 3D measurements of MHs on preoperative OCT images (BD and MLD) or standard AI models using preoperative OCT images for prediction of outcomes following surgery.

In addition to predicting VA, another important contribution of this work was determining the associated uncertainty. Estimations of uncertainty must be employed to ensure the safe and ethical use of DL models in practice. To the best of our knowledge, this is the first clinical

trial to prove that a DL model can predict VA and its associated uncertainty using preoperative OCT images, even when internally testing the OCT scans. Moreover, along with predictions, U-ARM thoroughly assessed the uncertainty of the DL model and its generalisability on an external dataset; using U-ARM thus avoids inaccurate predictions. Our results showed that U-ARM trained on HD-OCT of MH dataset achieved MAE of 6.88$\pm$0.26, R2 of 0.41 , and p-value of <0.005 on the internal dataset and trained on OCT-Newcastle dataset achieved MAE of 7.15$\pm$0.29, R2 of 0.39, and p-value of 0.039 on the external dataset - HD-OCT of MH dataset. This study, therefore, can lead to new applications in ophthalmology in the future.

Different techniques have been proposed to assess uncertainty in AI models. For further comparison, we evaluated several commonly used uncertainty techniques. Dropout sampling has been proposed to approximate Bayesian inference in neural networks by randomly removing some of the nodes [53]. We calculated the uncertainty of the predictions over multiple runs. The model ensembling technique is another uncertainty quantification method that involves training many different DL models with different architectures [96]. Similarly, we used the variance of the distribution over the predictions obtained through this method as the uncertainty. However, dropout sampling and model ensembling are computationally expensive and time-consuming processes since they need to be run several times. Another uncertainty quantification method is deep evidential regression, which forms the basis of U-ARM [9]. This places evidential priors over the Gaussian likelihood output and trains the neural networks to infer the hyperparameters of the evidential distribution. Deep evidential regression is thus a single model that can be trained in a single run. Although this is a promising approach in terms of efficiency, it does not strike a balance between prediction and uncertainty. We introduced U-ARM, which considers image data quality and tunes a regularisation coefficient to balance prediction and uncertainty. Thus, U-ARM ensures that optimisation initially prioritises prediction over uncertainty and then gradually increases the weight assigned to uncertainty in later epochs to enhance overall reliability.

Generally speaking, AI models tend to perform better with IS datasets than with OOS datasets [41]. Considering this, we initially evaluated the models for IS testing by using the internal dataset. Next, we tested the models on the external dataset, which showed domain generalisation. U-ARM achieved the best performance on the IS datasets even if test samples from the internal and external datasets were uncommon cases. Following this, we compared the models on the OOS dataset. Similar to the IS testing, U-ARM indicated its ability to generalise well to data falling OOS. Overall, these tests indicate that U-ARM exhibits well-calibrated uncertainties in addition to good predictions compared to the other models. Furthermore, U-ARM is widely applicable for predicting VA due to its efficiency, single-run implementation, and ease of implementation. Lastly, U-ARM demonstrated remarkable robustness against the OOS dataset.

We also note additional factors that limit this study. First, even though the internal tests were successfully implemented using the input of single (2D) and multiple (3D) OCT scans, the external tests were only performed using single OCT scans. We believe that the performance of external tests in prediction and uncertainty can be improved if U-ARM is fed with the input of multiple OCT scans. In addition to this, there remain very limited OCT datasets with a confirmed

iFTMH, most of which are not publicly available as 3D datasets. This limits our testing of U-ARM to only two OCT datasets. In future studies, we will collect more preoperative 3D OCT datasets with their outcome VA and examine the uncertainty. The final limitation relates to OCT image quality, directly affecting the prediction and its associated uncertainty as stated in Fig. 4.4 and 4.6-(i, ii, iii, iv). Our results are a useful starting point in improving the reliability of DL models for predicting postoperative VA from preoperative OCT images.

4.5 Summary

In this chapter, we introduced an uncertainty-aware regression model (U-ARM) for predicting postoperative visual acuity (VA) in patients undergoing macular hole surgery using preoperative spectral-domain optical coherence tomography (SD-OCT) images. U-ARM addresses the limitations of existing AI approaches by incorporating uncertainty quantification, enhancing the transparency and reliability of predictions. The model effectively prevents over-confidence in predictions and over-inflation of uncertainty by accounting for abnormalities in the data. Through extensive internal and external evaluations using publicly available datasets (OCT-Newcastle and HD-OCT of MH), U-ARM demonstrated superior performance in predicting VA and estimating associated uncertainties compared to traditional methods. The model's ability to generalize well to out-of-sample data, including low-quality images and unseen instances, underscores its potential for practical clinical use. U-ARM is thus a promising approach for clinical settings and can improve the reliability of artificial intelligence models in predicting VA. Overall, U-ARM represents a significant advancement in applying AI to predict VA outcomes post-surgery, offering a more trustworthy and reliable tool for clinicians.

Chapter 5

Detection of the Retinal External Limiting Membrane

Contents

In this section, we display precise image-wise binary annotations to segment the retinal external limiting membrane (ELM) in a patient population with idiopathic full-thickness MHs. Specifically, we consider a subset of the entire dataset described in Section 2, containing 107 OCT images. The OCT images present high variations in image contrast, motion, brightness, and speckle noise which can affect the robustness of applied algorithms so we performed an extensive OCT imaging and annotation data quality analysis. Imaging data quality control included noise, blurriness and contrast scores, motion estimation, darkness and average pixel scores, and anomaly detection, using the methods in Section 3. Annotation quality was measured using gradient mapping of ELM line annotation confidence, and idiopathic full-thickness macular hole detection. Finally, we compared qualitative and quantitative results with seven state-of-the-art DL-based segmentation methods to identify the ELM line with an automated system.

5.1 Introduction

The ocular fundus, which comprises the retina, macula, optic disc, fovea and blood vessels, can be visualised in detail using OCT which allows more than ten retinal laminations to be identified. Figure 1.6 shows the appearance of an MH when imaged using spectral-domain OCT, and specifically identifies the ELM. Accurate visualization of retinal anatomy allows ophthalmologists to determine which retinal layers are affected in specific disease processes. The retina can be anatomically divided into inner and outer zones. The inner retina defines all structures from the internal limiting membrane (ILM) to the ELM, and the outer retina includes all structures from the photoreceptors to the choroid. The ELM is formed by junctional complexes between photoreceptor and Müller cells, located between the photoreceptor nuclei and their inner segments. The integrity of the ELM is important in many disease states. In this section, we specifically investigate the ELM in MHs.

Idiopathic full-thickness MHs form due to age-related changes at the vitreoretinal interface, with antero-posterior vitreous traction on the inner retinal surface being transmitted by Müller cells to the outer retina, which leads to outer retinal traction and dehiscence. Subsequently, a full-thickness MH forms with the associated movement of the outer retinal layers towards the inner retinal surface [154]. The integrity of the ELM after surgery is associated with postoperative vision. The height of the ELM on the sides of the MH also appears to provide prognostic information, and it is thought that postoperative vision is partly related to the extent of tractional changes in the outer retina when the MH forms, as well as its chronicity [55, 103]. Detecting the ELM in MHs is not only clinically useful, but also a particularly challenging image segmentation task with high applicability to other diseases and biomedical imaging modalities.

There is significant potential for computer-aided diagnosis (CAD) systems to aid OCT image analysis of MHs and assist ophthalmologists in the detection and characterisation of the ELM. Identification and quantification of the ELM can be processed manually, but the many challenges make this a slow and error-prone process. Accurately identifying the ELM is also difficult because the detection or inhibition of partially coherent optical beams causes speckle noise in OCT images, causing difficulty in identifying the ELM line and differentiating it from neighboring retinal layers.

There are two main types of automated image analysis approaches that have been utilized for CAD system development, which includes classical image analysis techniques and machine learning-based image informatics techniques [35, 46, 104, 127, 135, 158]. Classical image informatics methodologies have several limitations such as threshold methods which struggle with discontinuities and intensity variations across different retinal layers, and they also cannot combine previously obtained information such as retinal layer thicknesses from previous OCT scans. These limitations fail to provide a fully automated solution [34, 72, 86].

The majority of image analysis approaches in the past have used privately owned datasets, and there is currently no large publicly available benchmark dataset to use for ELM line segmentation from OCT images. Therefore, the most appropriate image analysis methodology for automated ELM line segmentation is unclear.

To address the challenges, we present an automated detection of the retinal ELM using a 3D SD-OCT publicly available image dataset of idiopathic full-thickness MHs. This study provides the baseline characteristics of current deep learning-based image segmentation methods and serves as the first building block to foster rapid and objective progress in tackling the research problem at hand as a research community. The main contributions of this work include:

1. We introduce a general framework for constructing a dataset for retinal ELM line segmentation as shown in Fig. 5.1. We used 3D SD-OCT (5243 2D planes) images from 107 patients.

2. To provide a high-quality benchmark dataset, robust and rigorous quality checks performed on OCT images, including noise score, blurriness score, contrast score, motion score, brightness-darkness score, and average pixel width scores. Moreover, annotation quality analysis includes gradient-based ELM line detection and MH detection.

3. We perform extensive experiments including a broad range of ablation studies using seven state-of-the-art biomedical and semantic image segmentation methods. Quantitative and qualitative outcomes are compared with six evaluation metrics.

5.2 Related Works

A number of retinal layer segmentation approaches have been described in published literature. Below, we present and discuss the most common segmentation methods, ranging from classical to machine learning-based image informatics approaches.

5.2.1 Classical Image Informatics Approaches

Automatic segmentation of retinal layers based on classical approaches can be categorised as follows: (1) threshold-based methods, (2) level set-based techniques, and (3) dynamic programming/graph cut-based methods. Mostly, these techniques have extracted hand-crafted features that utilize the pixel values, texture, colour, and shape for the segmentation process. To accurately determine the border of the retinal layers, Ishikawa et al. [72] utilized an adaptive

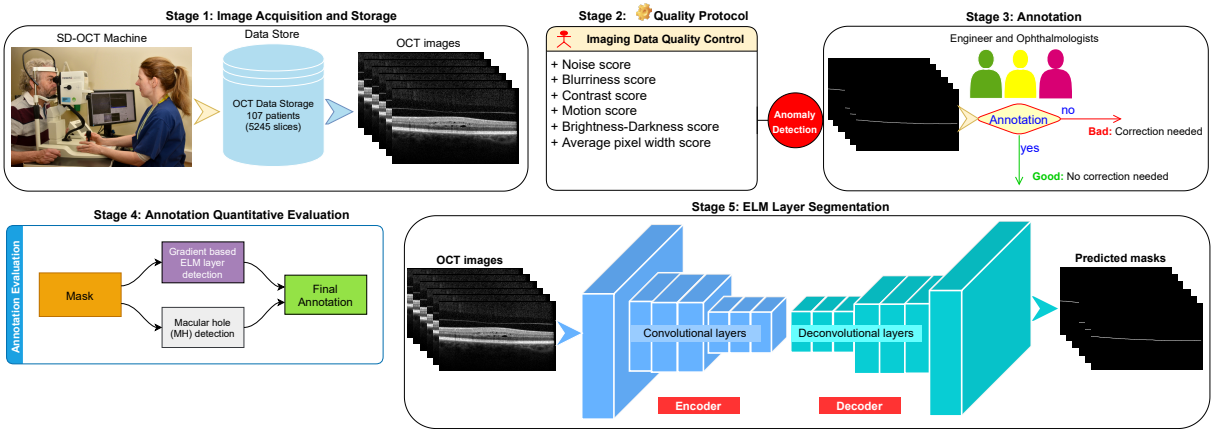


Fig. 5.1 Illustration of the proposed workflow. Stage 1 is data acquisition; stage 2 involves OCT imaging data quality check using several methods (i.e. noise score, blurriness score, contrast score, motion score, brightness and darkness score, and average pixel width); stage 3 incorporates ELM line annotation; stage 4 is the assessment of the quality of the annotated data (i.e. gradient-based ELM line detection, and idiopathic full-thickness macular hole (MH) detection); and stage 5 describes ELM line segmentation using multiple state-of-the-art segmentation methods. OCT refers to optical coherence tomography; ELM relates to an external limiting membrane; 2D applies to two-dimensional (2D) imaging; and MH refers to idiopathic full-thickness macular holes.

thresholding technique. Previous work has also applied intensity-based Markov boundary models [86] and a texture and shape analysis [76] approach to segment different retinal layers from OCT data. Furthermore, Novosel et al. [121] adapted a level-set technique based on Bayesian inference and anatomical information of the retina to delineate surfaces between layers. In the study [26], the author employed the deformable model method to segment eight retinal layers across the complete macular cube. This suggested mechanism maintains object relationships and topology whilst limiting overlaps.

A fully automated graph theory and dynamic programming-based method was applied to segment three retinal layers from SD-OCT images of an eye with drusen and geographic atrophy (GA) [35]. Another group [46] used a small dataset of OCT images for statistical shape modelling. Specifically, the Mumford-Shah functional method was employed which allows the development of a parametric illustration of open contours. Later, they formed small bands around the open contours which enables segmentation by incorporating local information. Furthermore, the kernel regression (KR)-based classification approach [34] was adopted to accurately measure abnormal fluid within and under the retina, and the position of the retinal layers. This classification strategy identifies a pattern to delineate retinal layers based on graph theory and dynamic programming techniques. However, this method also has limitations due to its parameter inference on images with noise. To overcome this challenge, Tian et al. [158] recommended the shortest path-based graph search method which identifies retinal layers by seeking the shortest path among two end nodes using Dijkstra's algorithm. Moreover, Bai et al. [18] proposed an adaptive-curve detection technique that explores the retinal area with boundary growth. The recommended approach used the simple linear iterative clustering (SLIC) superpixels, and the adjusted active contour, to

sequentially delineate the remaining boundaries. The method was tested using 3D OCT images captured from two different OCT systems.

Classical image informatics methods have several limitations, for example (1) the thresholding methods can cause discontinuities and variances of intensity within the same retinal layer, (2) they lack the ability to consider prior information about the retinal layer thicknesses obtained from previous imaging, and (3) their typically slow performance (especially graph-based techniques) means they are inappropriate for real-time clinical practices.

5.2.2 Machine Learning-Based Image Informatics Approaches

Convolutional neural networks (CNNs) are a deep learning-based technique used in many image-based segmentation applications [87]. Most CNNs are applied in areas with large quantities of imaging data with annotations available to be trained on. In the case of medical imaging, there are numerous unqiue challenges, such as the availability of data, dimensionality (3D or 4D) of images, annotation complexity, and quality. In addition, personal patient data and information must be anonymized, which is a complex and regulated process [47]. As deep learning methods improve, there are new available models for medical image segmentation tasks such as FCN [108], SegNet [14], and U-Net [134].

To overcome the aforementioned challenges, an end-to-end multi-scale nested U-Net [104] based method has been employed to segment seven retinal layers and three types of retinal fluid respectively. The important feature of this model is the utilization of multi-scale input, multi-scale side output, re-designed skip connections from U-Net++ [179], and dual attention technique. Specifically, U-Net++ allows for encoding and decoding layers connected by a series of nested dense skip connections. This allows the semantic gap between encoding and decoding layers to be decreased. Similarly, the study [92] applied the recurrent neural network trained as a patch-based retinal boundary classifier with a graph search (RNN-GS) to segment seven retinal layers in OCT images from healthy children and three retinal layer edges in OCT images from patients with age-related macular degeneration (AMD). However, another study [127], suggested a novel method for automated segmentation of OCT which utilized deep learning, and combined fully convolutional networks with Gaussian processes (GP). The result confers the combined efforts of DenseNet [68] and regression-based post-processing where DenseNet allows each layer of the network to directly process outputs from all previous layers.

An automated retinal layer segmentation network called ReLayNet [135] successfully segmented a retinal OCT B-scan into 7 retinal layers and areas of fluid accumulation. This proposed approach utilizes convolutional encoding blocks to learn a hierarchy of contextual features that follow an expansive path of decoders for semantic segmentation. To enable automated investigation of abnormal maculae, Sun et al. [152] implemented an FCN to recognize retinal areas in OCT images. Gopinath et al. [58] used a combination of a CNN which extract image layers and edges of interest, and long short term memory (LSTM) [63] to trace the layer boundaries enabling segmentation of the retinal layers from OCT scans. In the study [151], the authors adopted a two-stage FCN method which trained sequentially to achieve better segmentation performance.

In the first stage, the OCT image was segmented using a trained FCN and the second stage was refined by another trained model with a decision mask to enhance the segmentation result.

However, the aforementioned machine learning-based image informatics methods have limitations relating to data annotation, model robustness when using OCT machines from different manufacturers, and the parameters of deep models required for clinical practice.

5.3 Data Preparation

Table 5.1 shows the detailed description of the ELM line dataset. In this section, we use only a subset of the entire dataset described in Section 2, captured in Sunderland Eye Infirmary, United Kingdom (UK), using the Heidelberg Spectralis (Heidelberg, Germany) as part of routine care, using the same imaging protocol. The individual OCT line scans were 768×496 pixels with the scaling varying slightly between datasets but typically equating to 5.47 microns per pixel in the X (horizontal) axis and 3.87 microns per pixel in the Y (vertical) axis. With 29-30 microns spacing between scans (Z-axis), there were 49 scans per dataset. Image dataset and clinical data on 107 eyes from 107 patients were analysed. The mean age was 70 years old (SD 6.6, range 48–84), 88 (82%) were female and 54 (50%) were right eyes. The mean minimum linear diameter of the holes was 384.7 microns and the median duration of symptoms was 6 months. All scans used a 16 automatic real-time setting enabling multi-sampling and noise reduction over 16 images.

Table 5.1 Illustration of ELM line dataset.

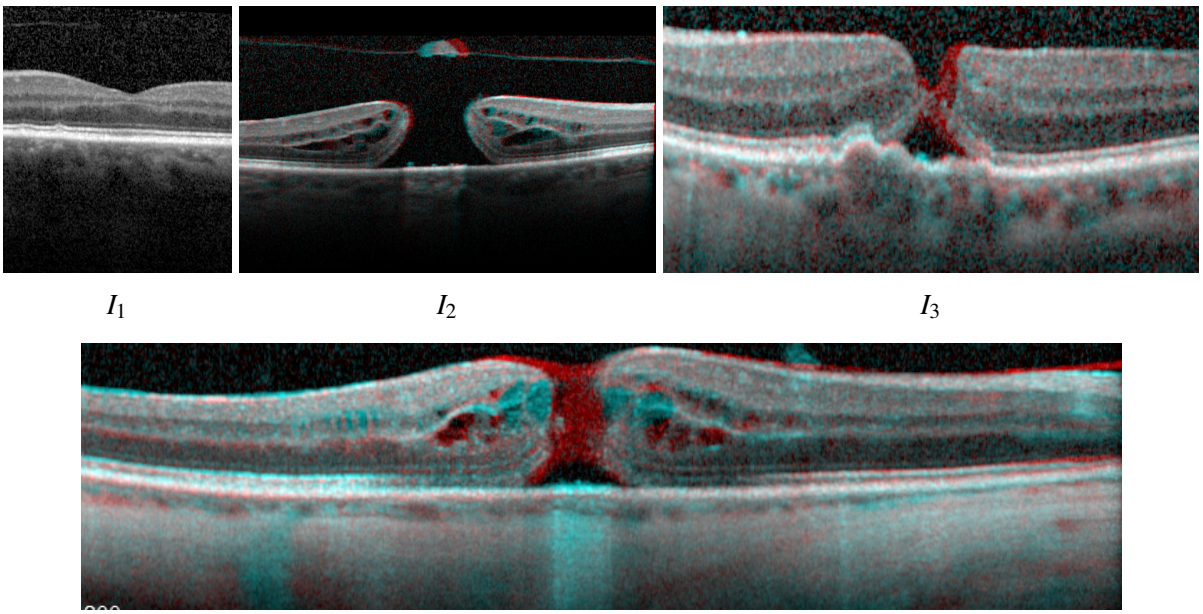
No. of 3D Volumes	No. of 2D Images
107	5243

5.3.1 Data Preprocessing and Anomaly Detection

In order to ensure the reliability and accuracy of our analysis, we employed image preprocessing, image quality assessment, and anomaly detection, as described in Sections 3.1.3, 3.1.3, and 3.1.3. By applying these methods, we ensured that the dataset used for analysis was of high quality and free from significant artifacts or outliers. The anomaly candidates shown in Fig. 5.3 and Fig. 5.2, based on the a subset of the entire dataset.

5.3.2 OCT Imaging Annotation Data

Annotation of the ELM line in MHs was performed by a data science researcher and two experienced ophthalmologists. Specifically, one academic ophthalmology trainee doctor with 2 years of experience, and one consultant ophthalmologist with over 20 years of clinical experience in their field. First, each slice of the OCT image was manually annotated by a researcher. To achieve robust annotation accuracy, all annotations were checked by the junior ophthalmologist. In challenging cases, the senior ophthalmologist was consulted and annotations were adjusted



I_1 I_2 I_3

I_4

Fig. 5.2 Remarkable samples showing high variation in motion between two neighbouring 2D slices in a SD OCT from the best (grey colour) to worst (red/blue colour) - I_1, I_2, I_3, and I_4.

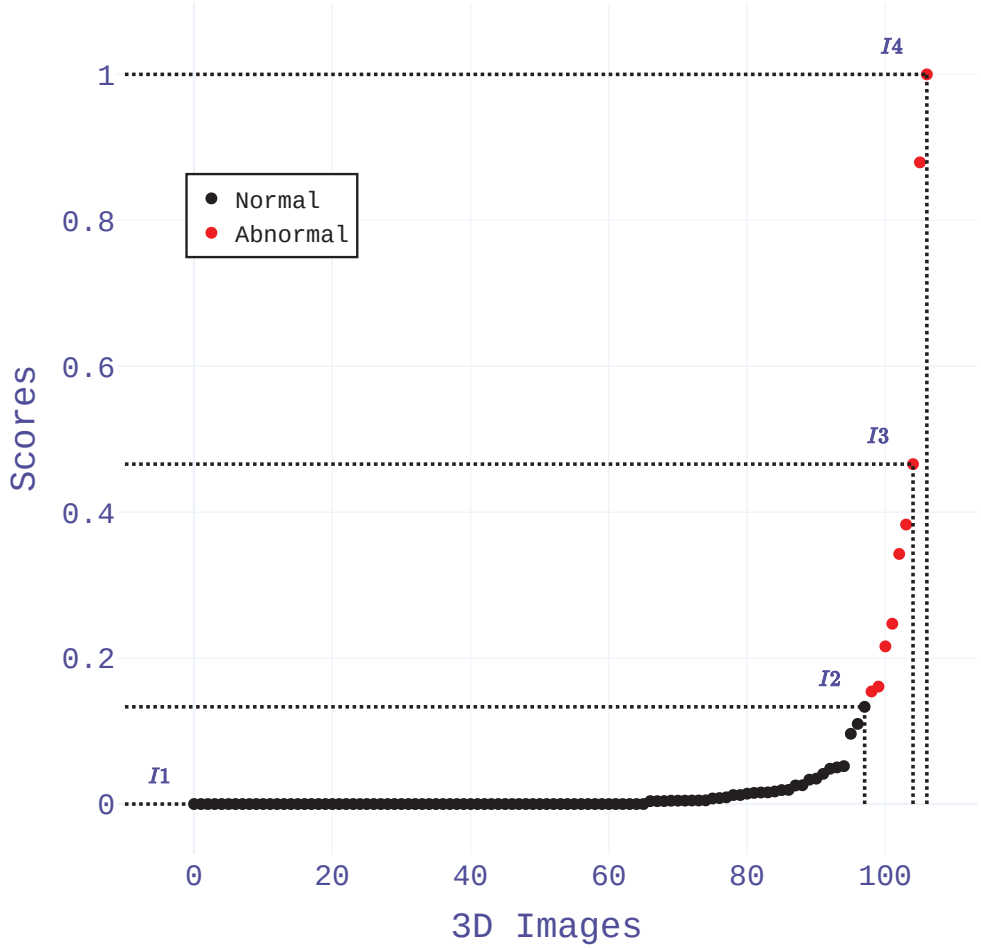


Fig. 5.3 A graph depicting the spectral domain optical coherence tomography 3D image anomaly detection results: black and red points represent normal and anomalous images respectively. I_1, I_2, I_3, and I_4 indicate randomly selected 3D images to be presented in Fig. 5.2.

accordingly. The slice-wise volumetric resultant annotations were exported as Tagged Image File Format (TIFF) files that contained a binary mask corresponding to [0, 255] for black and white pixels respectively.

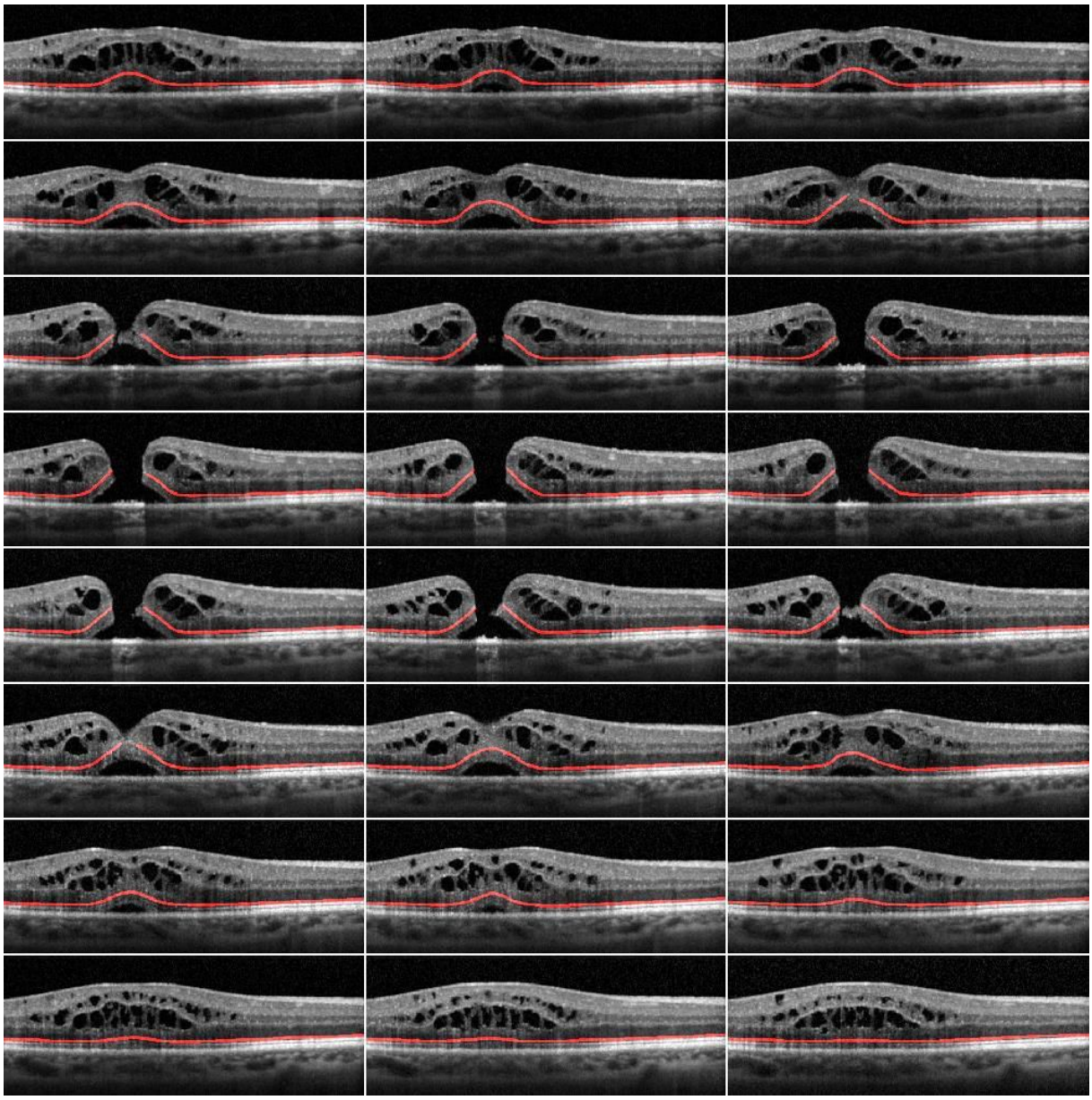


Fig. 5.4 Spectral domain optical coherence tomography (OCT) slices from a single idiopathic full-thickness macular hole (MH) from our dataset, with the external limiting membrane (ELM) shown in red colour.

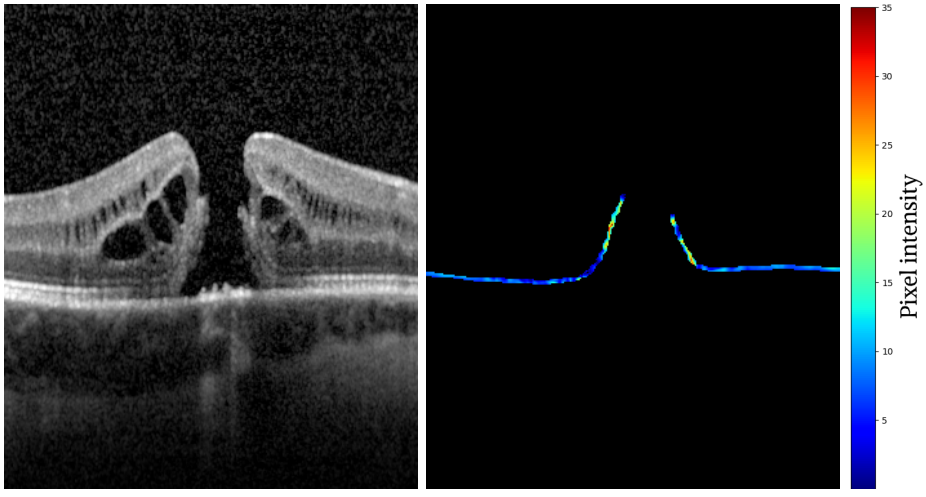
Annotation criteria

Annotating the ELM is challenging due to the presence of the neighbouring retinal layers and speckle noise. For each patient, a complete set of 49 2D tagged OCT image slices were annotated. However, the ELM line was not annotated if speckle-noise prevented the retinal layers from being accurately identified.

Annotations Uncertainty

Annotation data quality plays an important role in building accurate segmentation methods and facilitating the diagnosis and monitoring of disease. Therefore, we measured relevant annotation quality indices including gradient-based ELM line detection and MH detection in the annotated OCT image slices. Figure 5.5 shows an example of a gradient-based method that measures the intensity variation of the ELM line from a 3D OCT image. Yellow or red corresponds to the intensity variation of pixel change, whereas blue represents no change in pixel intensity. We used a Gaussian filter which calculates a multi-dimensional maximum filter on the edge of the ELM line. To measure the changes in the neighbouring pixels, connected components with five pixels compute the gradient local maxima and measure the variance. The main purpose of this method is to estimate the intensity variation that presents any uncertainty that could affect the annotation of the ELM line.

Idiopathic full-thickness macular holes were detected using the following three key steps: 1) we used a 3D annotated mask by estimating the maximum of an array or maximum along an axis towards the direction of depth; 2) we set the background label to zero to explicitly focus on the foreground ELM line; and 3) we separated the connected components of defined labels by taking the maximum from them. This helped to identify the continuous line (surface in 3D) representing the ELM line, and the one discontinuity corresponding to the presence of a macular hole. Each patient had only one macular hole present in their annotated images. Figure 5.4, shows multiple OCT slices of a single patient from our benchmark dataset, with its ELM line marked with red color.



(a) (b)

Fig. 5.5 Illustration of gradient-based intensity variation in the external limiting membrane (ELM) line. Here, *a* depicts the central slice of a full-thickness idiopathic macular hole using spectral domain optical coherence tomography (OCT), and *b* displays the corresponding corresponding gradient map. The colour bar displays pixel intensities, with red representing the maximum change in pixel intensity and blue referring to no variation in intensities of the 3D OCT image.

5.4 Methods

In this section, we present the architectural details of seven segmentation methods including fully convolutional networks (FCN) [108], U-Net [134], SegNet [14], Attention U-Net [124], recurrent residual convolutional neural networks (R2U-Net) [7], Efficient U-Net [17], and DeepLabv3+ [29] employed to segment the ELM line from OCT image.

5.4.1 Fully Convolutional Networks (FCN)

To perform semantic segmentation, [108] introduced the first deep learning-based fully convolutional network (FCN) method. We adopted FCN to provide an accurate ELM line segmentation. It involves four convolutional layers where the ResNet-18 (residual networks) [61] pre-trained network on ImageNet [40] is used to extract relevant features (i.e. edge, texture, and intensity, etc.) from OCT images. We upsampled the encoded features by applying bilinear interpolation to predict the segmentation mask. In addition, employing the skip connection to combine low and high layer features in the final output layer for fine-grained ELM line segmentation.

5.4.2 U-Net

We used U-Net [134] which includes five encoder and decoder layers. The encoder layer repeated two 3×3 convolutions and strides 1×1 followed by batch normalization and the non-linear ReLU activation function. These convolutional filters directly learn intrinsic retinal features (gray-level, texture, gradients, edges, etc.) from the OCT images. To down sample the feature maps, we applied a 2×2 max-pool operation with stride 2×2 after each encoder layer. Each decoder layer consists of an upsampling of the feature maps followed by a 2×2 up-convolution that divides the number of feature channels, a skip connection (concatenation) with the correspondingly cropped feature map from the encoder layers, and two 3×3 convolutions, each followed by a ReLU. This skip connection between encoding and decoding layers were added to preserve relevant information from the input features. At the final layer, a 1×1 convolution was adopted to map each 64 element feature vector into the binary segmentation task of the ELM line.

5.4.3 SegNet

We employed the SegNet [14] method to segment the ELM line. This network is followed by encoder and decoder blocks. Specifically, the encoder utilized a 13 convolution layer of the VGG16 [141] network. During the training process, convolution filters with size 3×3 extracted the set of feature maps from OCT images. Moreover, due to variance in the feature maps, batch normalization was applied to avoid overfitting. Non-linear ReLU activation functions were applied. To evade overfitting and reduce the input representation (feature dimensionality), max-pooling with size 2×2 and non-overlapping stride were 2×2 employed to the resultant feature maps. Likewise, the decoder has 13 deconvolutional layers that were used to up-sample the extracted feature maps of the ELM line. The high dimensional feature description at the

output of the final decoder was served to a sigmoid classifier to generate class probabilities for each pixel individually.

5.4.4 Attention Gates in U-Net (Attention U-Net)

Attention U-Net [124] was employed to provide ELM line segmentation from the OCT image dataset. It is the modified version of the very popular biomedical image segmentation model U-Net to advance model sensitivity to foreground pixels without needing complex design. This network architecture represents a similar structure of standard U-Net with five encoding and decoding layers. After each encoding layer, the attention module was applied. This module is employed by introducing a grid-based gating mechanism. Each encoder layer of attention-based features was passed to the corresponding decoder layer through the skip connection which evades disambiguating inappropriate and noisy features. This model helps to highlight more relevant ELM line features by ignoring other background regions and generate a precise output segmentation map.

5.4.5 Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net)

The R2U-Net [7] is a medical image segmentation model that provides an advancement from the classical U-Net method. It consists of five encoding and decoding layers. We used the recurrent residual convolution block in each encoding and decoding layer. Especially, this recurrent convolution was applied before the features downsampling and upsampling process in encoding and decoding layers respectively. Finally, this recurrent convolutional neural network (RCNN) block was also added before the final segmentation result. Similar to U-Net, R2U-Net utilized a skip connection between the encoding and decoding layers. The sigmoid activation function was used to calculate the final pixel probability of ELM line output.

5.4.6 Efficient U-Net

The EfficientNet [156] network was utilized as an encoder for feature extraction from OCT images and the decoder was used to provide the segmentation map of the ELM line. This EfficientNet has eight variations that include EfficientNetB0 to EfficientNetB7. Due to computational restriction, we applied EfficientNetB3 [156] to perform feature extraction inside the encoder layers of U-Net. It consists of mobile inverted bottleneck convolution (MBConv) [137] that combines the squeeze and excitation mechanism to highlight the more relevant features of the ELM line. Similarly, the decoder has kept a similar architecture to standard U-Net. Skip connection with element-wise feature concatenation was employed to preserve the high-level features and spatial information by filling the semantic gaps during the reconstruction process.

5.4.7 DeepLabv3+

To provide a precise semantic segmentation, we employed DeepLabv3+ [29], which is an extension of the previous method called DeepLabv3 [28]. During the training process, a pre-trained ResNet-18 model as a network backbone was used to extract rich feature information from the OCT dataset. Without adding any computational complexity, the extracted feature maps are inputted into the atrous convolution block that is used to control the resolution of obtained features from CNNs. Furthermore, it allows low and high-level features to be combined (ELM line, and its background). The atrous convolutional block also enables increments to capture multi-scale contextual information from OCT images. We used an output stride of 16 which enables a 1×1 convolution and three 3×3 dilated convolutions of varying dilation rates of 1, 6, 12, and 18 respectively. The last convolutional layer has a kernel size of 1×1 served by a global average pooling operation. To obtain high-level feature maps, an upsampling operation was applied after each atrous convolution layer, and then all extracted feature maps were concatenated. In the decoder, to recover the high-level and low-level feature maps bi-linear interpolation was used to upsample the features twice. To obtain rich information, we combined both high and low-level features. Ultimately, the bi-linear interpolation technique was employed for the final segmentation of ELM line output by upsampling the feature map by 2.

5.4.8 Cost Function

In this work, we used the sum of binary cross-entropy (BCE) and the Dice loss as a utilized cost function. Here, Dice loss $\mathscr{L}^{\text{Dice}}$ is the Dice coefficient *Dice* that can be expressed as follows:

$$\mathscr{L}^{\text{Dice}}(G,P) = 1 - Dice(G,P) = 1 - \frac{2|G| \cdot |P|}{|G|^2 + |P|^2} \tag{5.1}$$

where G is the ground truth image mask comprising an ELM line and P is the predicted mask from the image segmentation output.

The overall loss function is formulated as follows:

$$\begin{aligned} \mathscr{L}^{\text{OL}}(G,P) = \alpha(-(G \cdot log(P) + (1-G) \cdot log(1-P))) \\ + (1-\alpha)\mathscr{L}^{\text{Dice}}(G,P). \end{aligned} \tag{5.2}$$

Where, α is an empirical weighting factor.

5.5 Experimental design and results

5.5.1 Parameter Selection and Training

We implemented our network using Python 3.6, CUDA 10.2, cuDNN 7.0, PyTorch 1.8.1 running on a 64-bit Ubuntu operating system using a 3.4 GHz Intel Core-i9 with 32 GB of RAM and NVIDIA RTX 2080 Super GPU with a memory of 8 GB. We have trained our model by using the input resolution size of 256×256. During training, we used ADAM optimiser with $\beta_1 = 0.5$, $\beta_2 = 0.999$ and learning rate = 0.0002 with a batch size of two. We set the loss weighting factor

α to 0.5. Finally, we trained the model with 100 epochs and evaluated it on the best checkpoint saved by the highest Dice coefficient score. In order to generate the final segmentation result, we set the threshold value to 0.5.

The dataset was divided into training, validation, and testing sets by using five-fold cross-validation with a ratio of 70%, 10% and, 20% respectively. We evaluate the model performance on the independent test set. Note, we confirm that there are no patients shared within the splits and the folds of the cross-validation study. To provide variability and artificially increase the size of the ELM line dataset, various data augmentation techniques were applied. We scaled the images by changing the scaling variable from 0.5 to 1.5 with a step size of 0.25. Next, we applied gamma correction on the images by varying the gamma scaling constant from 0.5 to 1.5 with a step size of 0.5. Finally, we horizontally and vertically flipped the images, and rotated them with different angles of 15 degrees.

5.5.2 Quantitative Evaluation

In this study, we employ six evaluation metrics to evaluate the performance of each segmentation method, including Dice coefficient score (DSC), intersection over union (IoU), root mean square error (RMSE), Hausdorff distance (HD), sensitivity (SEN), and false-positive error (FPR). The segmentation results analysis relied on a confusion matrix that encompasses four different variables: true positive (TP), false positive (FP), true negative (TN), and false-negative (FN). Here, we present the mathematical formulations of the six metrics: DSC, IoU, SEN, FPR, RMSE, and HD.

Dice Coefficient: The Dice coefficient was used to measure the similarity between ground truth and the predicted mask.

$$DSC = \frac{2|P \cap G|}{|P| + |G|} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{5.3}$$

where G is the ground-truth mask and P is the predicted mask.

Intersection Over Union: The intersection over union calculates the percentage overlap between the ELM line ground truth and predicted mask.

$$IoU = \frac{TP}{TP + FP + FN} \tag{5.4}$$

Sensitivity: Sensitivity calculates the portion of positive pixels in the ground truth that is also identified as positive by the algorithm being evaluated. Sensitivity can be obtained by the Equation (5.5):

$$SEN = \frac{TP}{TP + FN} \tag{5.5}$$

False Positive Rate: This is the ratio of false positives to the sum of false positives and true negatives.

$$FPR = \frac{FP}{FP + TN} \tag{5.6}$$

Root Mean Square Error: RMSE measures the difference between the ground truth and the predicted mask of the ELM line (units in pixels). Here, a smaller value represents better segmentation output.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i \in T}(P(i) - G(i))}, \tag{5.7}$$

where $G(i)$ is the ground-truth of pixel i, $P(i)$ corresponding predicted mask, T is the set of valid pixels (i.e., both the ground-truth and predicted mask pixels that do not have mask values equal to zero or non-black regions), n is the number of elements in T.

Hausdorff Distance: This is a symmetric measure of distance between two masks (smaller is better) and is defined as [30]. This distance can be calculated in pixels. Here, The point sets S_G and S_P belong to the pixels of the ground truth mask and the prediction, respectively.

$$HD = max\left\{ \max_{S_G \in S(G)} d(S_G, S(P)), \max_{S_P \in S(P)} d(S_P, S(G)) \right\}. \tag{5.8}$$

Where, d refers to the distance between the two points.

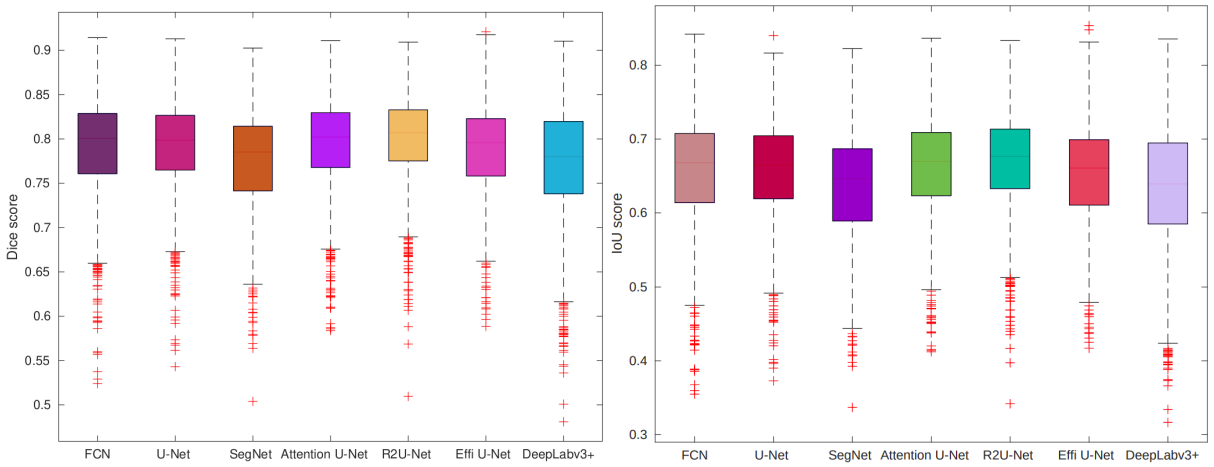
Table 5.2 Quantitative comparison of the seven segmentation methods with and without the effect of data augmentation on the ELM line test set by using six evaluation metrices i.e., DSC, IoU,RMSE, HD, SEN, and FPR. Statistically notable results are highlighted in bold font.

Models	With data augmentation						Without data augmentation					
	DSC ↑	**IoU ↑**	**RMSE ↓**	**HD ↓**	**SEN ↑**	**FPR ↓**	**DSC ↑**	**IoU ↑**	**RMSE ↓**	**HD ↓**	**SEN ↑**	**FPR ↓**
FCN [108]	77.307	63.562	0.070	7.990	71.008	0.149	75.851	61.782	0.072	8.719	68.506	0.013
U-Net [134]	78.026	64.358	0.069	6.099	70.989	0.142	77.430	**63.636**	0.071	8.190	70.170	0.141
SegNet [14]	76.501	62.435	0.071	8.266	69.140	0.141	75.289	60.917	0.073	8.917	67.550	0.144
Attention U-Net [124]	78.270	**64.711**	0.069	6.238	**71.547**	0.142	77.264	63.532	0.071	7.565	70.204	0.148
R2U-Net [7]	**78.299**	64.179	**0.069**	**5.188**	70.970	**0.131**	**78.036**	63.036	**0.069**	**5.991**	71.062	**0.141**
Efficient U-Net [17]	77.583	63.832	0.070	9.804	71.152	0.151	77.081	63.130	0.072	12.74	**71.145**	0.170
DeepLabv3+ [29]	76.328	62.277	0.072	6.230	70.843	0.177	75.742	61.256	0.075	7.982	69.531	0.194

5.5.3 State-Of-The-Art Method Comparisons

In Table 5.2, we present the quantitative results of seven segmentation methods on the ELM line dataset. These methods include the (FCN) [108], U-Net [134], SegNet [14], Attention U-Net [124], recurrent residual convolutional neural networks (R2U-Net) [7], Efficient U-Net [17], and DeepLabv3+ [29]. Initially, the experiments were performed without using any data augmentation techniques but later performed data augmentation on the segmentation results (see subsection 5.5.5). Experimental results demonstrated that R2U-Net shows effectiveness and superiority in DSC, RMSE, HD, and FPR scores of 78.03%, 0.069, 5.99, and 0.141%, respectively compared with other methodolgies. The recurrent convolutional neural network

(RCNN) leads to precise segmentation due to its temporal dependencies from OCT images. It extracts rich low-level features (i.e., edge, texture, and intensities, etc.) and high-level features. In turn, U-Net and Attention U-Net performed similarly and yielded 63.63% and 63.53% IoU scores, respectively. The attention mechanism allows capturing the most relevant foreground ELM line features and discards the irrelevant ones. SegNet scored comparatively lower segmentation results compared to the rest methods. Overall, we observed that the U-Net family models (i.e., U-Net, Attention U-Net, R2U-Net, and Efficient U-Net) generate similar ELM line segmentation results.



(a) Dice coefficient score.

(b) IoU score.

Fig. 5.6 Boxplots of Dice and IoU scores for all test samples of ELM line OCT dataset. Different colour boxes indicate the score range of several methods; the red line inside each box represents the median value, box limits include interquartile ranges Q2 and Q3 (from 25% to 75% of samples), upper and lower whiskers are computed as 1.5 times the distance of upper and lower limits of the box, and all values outside the whiskers are considered outliers, which are marked with the (+) symbol.

In Figure, 5.6a, 5.6b we have provided descriptive statistics of box-plot analysis for the most relevant DSC, and IoU scores. These measurements constitute each of the employed segmentation methods. We used different color boxes to indicate the score range of several methods; the red line inside each box represents the median value, box limits include interquartile ranges Q2 and Q3 (from 25% to 75% of samples), upper and lower whiskers are computed as 1.5 $\times$ the distance of upper and lower limits of the box, and all values outside the whiskers are considered outliers, which marked with the (+) symbol. As clearly shown, R2U-Net outperforms the other six methods with a lower standard deviation. Descriptive statistics also demonstrated that all the ELM line segmentation methods produced few outliers. These outliers may affect segmentation results for certain cases.

5.5.4 Qualitative Evaluation

Figure 5.7 presents the three qualitative comparisons of seven segmentation methods (i.e., FCN, SegNet, U-Net, Attention U-Net, R2U-Net, Efficient U-Net and DeepLabv3+) used to segment the ELM line. The white box provides a *zoom-in* visualization of the specific region, where

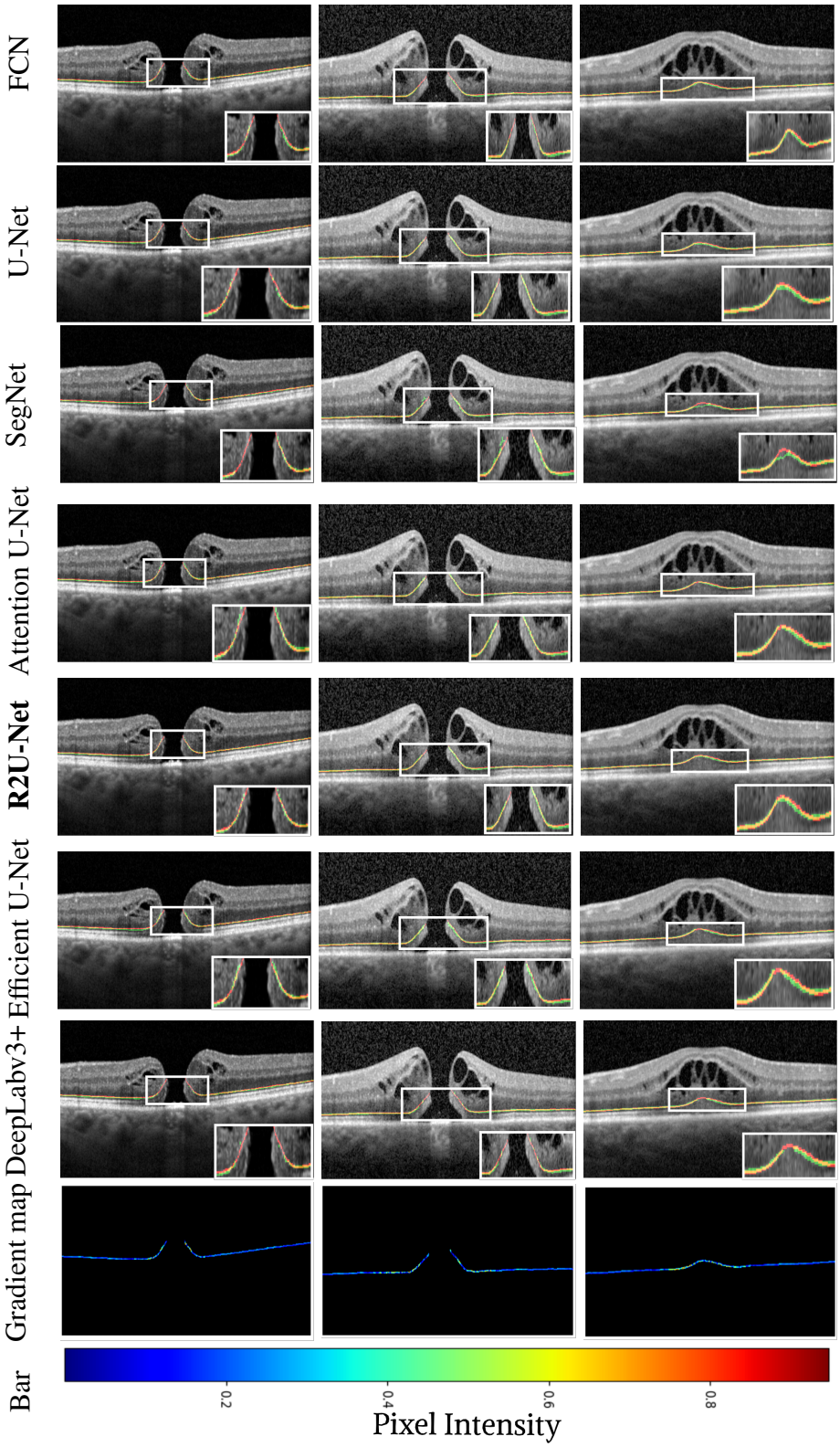


Fig. 5.7 Qualitative comparison of seven state-of-the-art segmentation methods evaluated on the ELM line test set. In a row, we present three examples used to compare all the segmentation methods. Here, we map the results into three colors: orange color refers to true positive, red color a false negative, and green color presents false positives. Further, gradient-based color maps depict the dark blue color that highlights no variation in the pixel intensities. In addition, a yellow or red color indicates increasing changes in intensity. The white box provides a *zoom-in* visualization of the specific region, where the compared segmentation methods failed to provide precise ELM line detection.

the compared segmentation methods failed to provide precise ELM line detection. We used colour-coding to visually inspect the segmentation results, i.e., orange colour represents true positives, red colour denotes the false negatives and green corresponds to false positives. It is evident that most of the compared ELM line segmentation methods performed similarly by generating only very few false positives due to the fuzzy boundaries and high variation in pixel intensity across the ELM line, as highlighted in the *zoom-in* box. To demonstrate this effect, we plotted the pixel intensity gradient map, where the blue colour corresponds to no variations in intensity, whilst certain areas, shown as a yellow or red colour, depict increasing changes in intensity.

5.5.5 Ablation Study

In recent years, ablation studies are involved in various relevant scientific publications to assess deep learning performance [113]. We provide a broad range of ablation studies that involve the effect of data augmentation, the effect of the loss function, and the computational complexity of each segmentation method.

Effect of Data Augmentation: Table 5.2 shows the effect of data augmentation with each of the employed methods. As can be seen, adding the additional diversity in feature representation leads to a significant improvement in segmentation outcomes. These varying imaging features (e.g., transformation, rotation, contrast changes, and flipping operations) helps to fill the semantic gap and make network training more generalized. This strategy served to improve the model performance by around $1.5\% - 2.5\%$. Moreover, extensive testing confirms rotation of the OCT images by more than 15 degrees did not result in any significant deterioration in its performance.

Effect of Loss Function: Figure 5.8 presents the effects of various loss functions (i.e., BCE, Dice loss, IoU loss, BCE+IoU, and BCE+Dice loss) on ELM line test set. We evaluated all the seven segmentation methods underlying different loss functions. Experimental results confirm that the R2U-Net has outperformed other methods with a combination of BCE and Dice loss functions. This combination of loss functions allows faster loss convergence during training and achieves more precise ELM line segmentation. Further, it helps in reducing the number of false positives from the segmented mask remarkably. This ablation study verifies the utilized loss function (BCE + Dice loss) yields a better increment at pixel-level in the ELM line segmentation results.

Computational Complexity: In Table 5.3, we present the computational complexity of each method in terms of its parameters (M) and MACs (G). FCN follows a large number of convolution layers that contained 135.53M parameters with 50.13 MACs. However, DeepLabv3+ showed the second-highest complexity with 59.34M parameters compared to the rest methods. Further, U-Net and Attention U-Net have very comparable trainable parameters and achieved similar segmentation results. But, R2U-Net has 39.09M parameters with the highest 152.94 MACs that increased by the recurrent neural network operation. Besides, Efficient U-Net only contained 21.44M parameters with the lowest 9.69 MACs. This allows the network to train faster and achieved comparable results with other state-of-the-art segmentation methods.

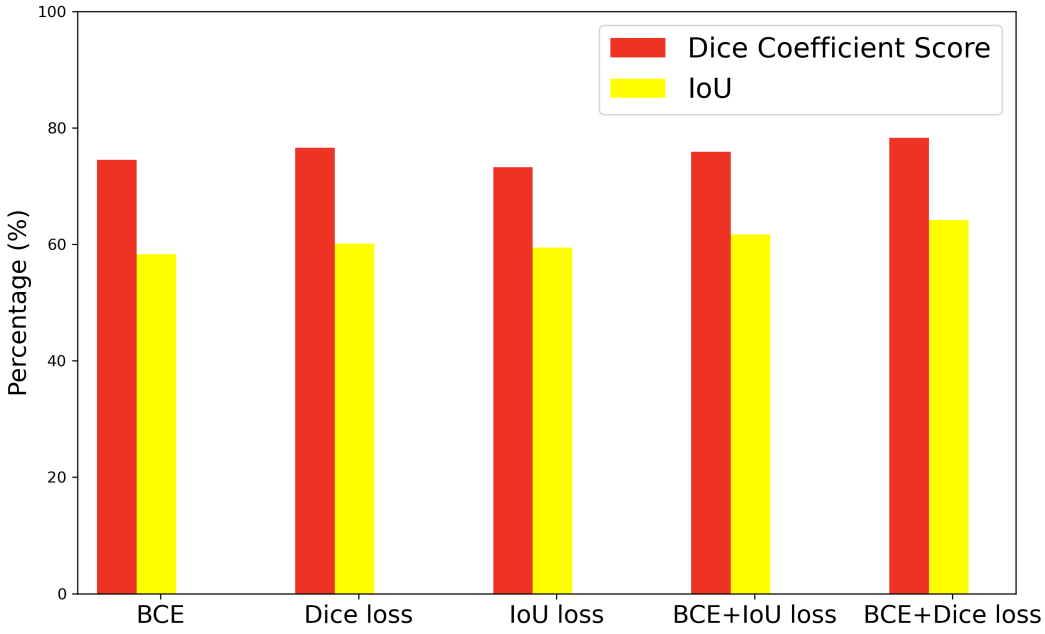


Fig. 5.8 Illustration of various loss functions performance (i.e., BCE, Dice, IoU, BCE+IoU, and BCE+Dice) comparison by the best results achieved by R2U-Net method. The dice coefficient and IoU scores were adapted to measure the quantitative changes.

Table 5.3 Comparison of various model trainable parameters and multiply–accumulate operation (MACs).

Models	Params (M)	MACs (G)
FCN	135.53	50.13
SegNet	29.44	40.10
U-Net	34.53	65.53
Attention U-Net	34.88	66.64
R2U-Net	39.09	152.94
Efficient U-Net	21.44	9.69
DeepLabv3+	59.34	22.21

5.5.6 Limitations

Accuractly identifying the ELM line using OCT images is a challenging assignment due to artifacts and variation in pixel intensities leads to more false-positive results. We accept that our study has limitations. First, we performed an analysis on a dataset compiled from one hospital with a single manufacturers' OCT device on a predominantly white population. Therefore, the experimental results may not be generalizable to other populations or OCT appliances. Second, the benchmark dataset only includes images from patients with idiopathic full-thickness macular holes, a particularly challenging scenario for ELM line detection. Lastly, in this study, we have not calculated any clinically applicable measurements from the images and compared them to human-derived ones, which is our next aim.

5.6 Conclusion

In this section, we have presented an ELM line segmentation using OCT image dataset. Specifically, the dataset contains images from one eye of 107 patients with idiopathic full-thickness macular holes, with precise ELM line annotations. The OCT images have variation in image contrast, motion, brightness, and speckle noise, typical of routinely collected clinical data. We performed a detailed analysis of image and annotation quality, using a range of standard scores, and gradient mapping of the ELM line. We also performed an extensive ablation study that included the effect of the loss function, the effect of data augmentation, and computational complexity and compared the results of seven state-of-the-art image segmentation methods. All the segmentation methods achieved greater than 75% Dice coefficient score, however, there remains potential to improve the methodology with further study. The ability to detect and measure the ELM is an important part of retinal assessment in a number of blinding diseases. In particular in eyes with macular holes, where the ELM can be challenging to evaluate, it can be used to select the optimal surgical techniques and assess prognosis. In future work, we will assess the use of automated ELM detection to derive clinically relevant measures to predict outcomes in patients undergoing surgery. Furthermore, while this study focused on 2D segmentation, future work will explore 3D architectures to leverage volumetric context and incorporate uncertainty prediction to improve clinical interpretability. These advancements will require expanded datasets with 3D annotations and agreement from multiple experts to create reliable training labels. The key takeaway from Chapter 5 is that accurate and automated ELM segmentation is achievable on real-world clinical OCT data despite image quality variability. This directly supports the overall aim of this thesis by providing a reliable structural biomarker that can contribute to predictive models of surgical outcomes in macular hole patients.

5.7 Summary

This chapter focused on the segmentation of the retinal ELM in patients with idiopathic full-thickness macular holes using a subset of an OCT image dataset. The chapter emphasized the importance of accurate ELM detection for clinical assessments and the challenges posed by

OCT image quality variations. It highlighted the potential of computer-aided diagnosis (CAD) systems to assist ophthalmologists by providing an automated and reliable means of segmenting the ELM. The limitations of classical image analysis techniques are discussed, leading to the exploration of machine learning-based methods for better performance.

The study introduced an automated ELM detection framework using the OCT dataset. Extensive experiments were conducted, comparing the performance of seven state-of-the-art image segmentation methods using six evaluation metrics. The results showed that all methods achieved a Dice coefficient score greater than 75%, indicating promising improvement. The chapter concluded by acknowledging the significance of accurate ELM detection for retinal assessment, especially in macular holes. Although this chapter focused on segmentation, it contributes to the overall aim of this thesis by enabling precise extraction of structural features relevant to macular hole prognosis. Clinically, successful restoration of the ELM is strongly associated with macular hole closure and improved visual acuity. As such, the methods developed here can inform future predictive models of surgical outcomes and visual function.

Chapter 6

Synthesis: Deep Learning Framework for Macular Hole Assessment

Contents

6.1 Introduction

his chapter aims to consolidate the key contributions of this thesis by synthesising the findings from Chapters 3, 4, and 5. It will clarify how the developed framework for VA prediction, the uncertainty quantification method, and the approach to ELM layer segmentation collectively and significantly enhance the comprehensive assessment of MHs. Furthermore, it will underscore the significance of the novel 3D SD-OCT imaging benchmark dataset in facilitating these advancements, subsequently contributing to improved clinical decision-making and patient care.

6.2 Integrated Framework for Macular Hole Assessment

6.2.1 Visual Acuity Prediction as a Foundation

Chapter 3 presented a deep learning framework for predicting postoperative VA in MH patients. The accurate prediction of VA is crucial for surgical planning and patient expectations. By employing 2D and 3D CNN models, this framework provides valuable insights into potential visual recovery following surgery. While the core algorithms are established, the framework's contribution lies in its robust implementation, comparative analysis, and optimisation for the specific task of MH VA prediction. Integrating accurate VA prediction with detailed anatomical information, such as ELM layer segmentation, can provide a more comprehensive understanding of the patient's condition and inform more tailored surgical strategies.

6.2.2 Uncertainty Quantification for Enhanced Reliability

Chapter 4 addressed the inherent uncertainty in OCT images and DL model predictions. The introduction of the U-ARM model allows for more reliable predictions by quantifying the uncertainty associated with VA predictions. This is particularly important in clinical settings, where a prediction's confidence level is critical for decision-making. Although uncertainty quantification methods exist, their application and adaptation within the context of MH and OCT image analysis is a significant contribution, as presented in this thesis. The evaluation of U-ARM demonstrates the importance of incorporating reliability measures into DL-based predictive models. Furthermore, quantifying uncertainty in VA prediction adds a crucial layer of reliability, enabling clinicians to make more informed decisions, especially in cases where the model expresses higher uncertainty.

6.2.3 ELM Layer Segmentation for Anatomical Insights

Chapter 5 focused on the segmentation of the retinal ELM layer. Accurate segmentation of the ELM layer provides crucial anatomical information that can aid in surgical planning and prognosis. Specifically, the detailed knowledge of the ELM layer morphology, facilitated by accurate segmentation, can inform the selection of appropriate surgical techniques for macular hole closure. This is because the ELM's integrity and its relationship to the surrounding retinal structures are key factors influencing the surgical approach and, ultimately, postoperative visual

acuity scores. While the segmentation models used are established, the chapter contributes a detailed analysis of their performance in the context of MH, addressing the challenges posed by OCT image quality. The automated ELM detection framework enhances the objectivity and efficiency of this crucial clinical task, providing surgeons with valuable data to optimize surgical strategies and improve patient outcomes.

6.3 The Role of the Novel 3D SD-OCT Imaging Benchmark Dataset

The novel 3D SD-OCT imaging benchmark dataset introduced in this thesis is a key enabler of the advancements described in Chapters 3, 4, and 5. Existing datasets often suffer from limitations such as small sample size, lack of standardization, and insufficient image quality, hindering the development of robust DL models. The new dataset addresses these limitations by providing:

- A large and well-curated collection of 3D SD-OCT images specifically focused on MH.
- Standardised imaging protocols and quality control measures.
- Detailed annotations, including ELM layer segmentations and corresponding visual acuity measurements.

This dataset has facilitated the development, training, and validation of the proposed framework, enabling more accurate and reliable MH assessment. It also provides a valuable resource for future research in this field, promoting the development of even more sophisticated and clinically relevant tools.

6.4 Conclusion

This chapter has demonstrated how the individual components of this thesis (*the VA prediction framework, UQ, and ELM layer segmentation*) synergistically contribute to a more comprehensive and reliable assessment of macular holes. By integrating functional predictions with anatomical insights and reliability measures, the framework offers a powerful tool to support clinical decision-making. While Chapters 3 and 5 utilise existing algorithms, their value is enhanced by their application within this integrated framework and their evaluation using the novel dataset. The proposed UQ method in Chapter 4 enhances the clinical utility of the framework. The novel 3D SD-OCT imaging benchmark dataset is a crucial contribution that underpins the validity and impact of the entire work, paving the way for improved diagnosis, surgical planning, and, ultimately, better outcomes for patients with macular holes.

Chapter 7

Conclusions and Future Work

Contents

7.1 Overview

This chapter concludes the thesis by comprehensively summarising the research and its key findings. It revisits the research questions that guided this study and outlines how each question was addressed. Furthermore, this chapter discusses the broader implications of the research outcomes for predicting VA following MH surgery and the automated segmentation of the retinal ELM layer. Finally, it identifies potential directions for future work to build upon the advancements made in this thesis and expand the applicability of the developed methodologies.

7.2 Revisit the RQs and Research Conclusions

In conclusion, this thesis presents a robust and comprehensive framework for enhancing the prediction of VA outcomes and the segmentation of the retinal ELM layer in patients undergoing MH surgery. This section revisits the core RQs and research conclusions that guided this study, summarising key findings, methodological insights, and implications.

Addressing the Foundation: Existing Data and the Novel Dataset

1. *RQ1: How have idiopathic full-thickness macular holes (MHs) been quantitatively assessed in the published literature, and what are the limitations of existing datasets?*
   RQ2: How does the newly introduced 3D SD-OCT imaging benchmark dataset contribute to advancing DL models for MH analysis?

   Chapter 1 and 2 of this thesis directly address these foundational questions. A comprehensive review of the existing literature (RQ 1) identified the various methods used to to assess MHs quantitatively and critically evaluated the limitations of currently available datasets for training and validating DL models. These limitations include limited sample size, lack of standardised imaging protocols, variability in image quality, and the absence of detailed 3D volumetric data. To overcome these challenges and advance the retinal imaging analysis, this thesis introduces a novel 3D SD-OCT imaging benchmark dataset curated for MH analysis (RQ 2). This dataset aims to provide a more comprehensive and standardised resource, facilitating the development and evaluation of more robust and accurate DL models for tasks such as VA prediction and retinal structure segmentation.

DL Models for VA Prediction

2. *RQ3:How do different state-of-the-art DL models perform in predicting postoperative VA from preoperative OCT images?*
   RQ4: What relationships exist between preoperative OCT images, preoperative VA, and postoperative VA, and how can DL-based predictive models leverage these relationships?
   RQ5: How can an automated OCT image analysis framework for ophthalmologists be developed using DL algorithms trained on preoperative OCT images and postoperative VA outcomes?

Chapter 3 explores 2D and 3D CNN models addressing *RQ 3, 4, 5*. The 2D CNN models demonstrated impressive predictive accuracy with an MAE of 6.47 ETDRS letters, effectively handling various image quality issues. These models offer a computationally efficient approach to VA prediction. The 3D CNN models generally provided superior R-squared and Pearson correlation coefficient performance, indicating they capture more nuanced spatial information. However, this comes at the cost of increased computational complexity. This comparative analysis presents the importance of balancing computational efficiency and spatial resolution based on the specific requirements of the clinical application (*RQ 3*). This study also contributes to understanding the relationships between preoperative OCT images, preoperative VA, and postoperative VA (*RQ 4*) by establishing a DL-based approach to predict postoperative VA. The development of these models contributes to *Rq 5*, which showcases the potential for automated VA prediction tools for ophthalmologists.

Uncertainty Quantification with U-ARM

3. *RQ6: What preprocessing, image quality assessment, and anomaly detection techniques enhance the robustness of DL-based OCT analysis?*
   RQ7: What are the common sources of uncertainty in OCT images, and how can they be effectively represented and quantified in DL-based predictive models?
   RQ8: How does the proposed UQ method compare with commonly used UQ approaches in improving the reliability of 2D and 3D DL-based predictive models?

   Building on this foundation, Chapter 4 introduced the uncertainty-aware regression model (U-ARM) to address *RQ 6, 7, 8*. U-ARM enhances the transparency and reliability of predictions by explicitly quantifying uncertainty, effectively preventing over-confidence and the over-inflation of uncertainty. Through extensive evaluations of UQ methods, U-ARM demonstrated superior performance in predicting postoperative VA and estimating associated uncertainties compared to other UQ methods (*RQ 7 and 8*). The ability of U-ARM to generalise well to out-of-sample data, including low-quality images and unseen instances, showcases its robustness and adaptability, improving its potential for clinical application. This advancement emphasises the potential of U-ARM as a more trustworthy and reliable tool for clinicians. The incorporation of uncertainty measures is critical for enhancing the reliability and clinical utility of predictive models. Furthermore, the preprocessing, image quality assessment and anomaly detection techniques used in this chapter contribute to addressing *Rq 6*, which improves the robustness of DL-based OCT analysis.

ELM Layer Segmentation

4. *RQ 9: How do different state-of-the-art DL-based segmentation models perform in detecting the ELM layer in OCT images?*
   RQ 10: How can automated ELM layer segmentation be integrated into clinical workflows, and what are the latest advancements in DL-based ELM layer detection?

Chapter 5 focused on the segmentation of the retinal ELM layer in patients with idiopathic full-thickness MHs. The comparison of seven state-of-the-art image segmentation methods highlighted the challenges posed by variations in OCT image quality and the limitations of classical image analysis techniques *RQ 9*. The introduction of an automated ELM detection framework, which achieved a Dice coefficient score greater than 75% across all methods, demonstrates the potential of the CAD systems to provide accurate and reliable segmentation. This capability is crucial for integrating automated ELM layer segmentation into clinical workflows, as explored in *RQ 10*. The automated framework can assist clinicians in accurately identifying and assessing the ELM layer, potentially leading to improved surgical planning and outcome prediction.

Overall, this thesis offers valuable insights into the application of advanced DL techniques for OCT image analysis in the context of MH surgery. The findings highlight the importance of selecting appropriate models based on clinical needs, the benefits of incorporating uncertainty quantification, and the potential of automated segmentation frameworks.

7.3 Future Work

While this thesis established a robust framework for improving visual acuity prediction and retinal structure segmentation in MH surgery using advanced DL techniques, several promising works for future research remain.

Further Optimization and Exploration of Models:

Future work should further optimise the developed 2D and 3D CNN models. This could explore more advanced CNN architectures, such as Transformers or hybrid CNN-Transformer models, to capture long-range dependencies and improve predictive accuracy and segmentation performance. Research on different loss functions, optimisation algorithms, and regularisation techniques could also enhance DL model generalisation and robustness of DL models. Furthermore, exploring the benefits of incorporating attention mechanisms within the models could improve their ability to focus on clinically relevant features within the OCT images.

Hybrid Approaches and Multi-Modal Integration:

Hybrid approaches that combine the strengths of 2D and 3D models could be beneficial in predicting VA following MH surgery. For instance, a multi-stream architecture that processes 2D slices and 3D volumes and fuses their features could potentially leverage complementary information. Additionally, future research could investigate the integration of other relevant clinical data, such as patient demographics, preoperative visual acuity, and MH characteristics (e.g., minimum linear diameter, basal diameter, height), with the OCT image data to develop more comprehensive predictive models.

Expanding Applicability to Other Retinal Conditions:

The methodologies developed in this thesis have shown promise for MH analysis. Future work should explore the applicability and adaptability of these models and frameworks to other retinal conditions characterised by structural changes observable in OCT images, such as epiretinal membranes, vitreomacular traction, and age-related macular degeneration. This would involve fine-tuning the existing models or developing new ones tailored to the specific characteristics of these conditions.

Longitudinal Studies and Temporal Analysis:

This thesis primarily focused on predicting outcomes based on preoperative data. Future research could investigate using longitudinal OCT data to predict the trajectory of visual recovery or the risk of complications over time after MH surgery. Developing models that can analyze the temporal sequences of OCT images could provide valuable insights into the dynamic changes in the retina post-surgery.

Clinical Translation and Explainability:

A critical direction for future work is the translation of these advanced DL tools into clinically usable systems. This would involve rigorous external validation on multi-centre datasets, development of user-friendly interfaces, and assessment of their impact on clinical decision-making workflows. Furthermore, the explainability and interpretability of the DL models is crucial for building trust among clinicians and patients. Attention map techniques could provide insights into the model predictions and segmentation results.

Addressing Data Imbalance and Rare Cases:

Macular hole surgery outcomes and the prevalence of specific morphological features can be imbalanced in clinical datasets. Future work should explore techniques to address data imbalance, such as advanced sampling strategies or synthetic data generation, to improve the performance of the models, particularly for less frequent but clinically significant cases.

Application of the Segment Anything Model (SAM) for ELM Segmentation:

An up-and-coming area of future research is the application of the recently developed Segment Anything Model (SAM) for ELM layer segmentation. SAM's ability to perform zero-shot segmentation with minimal prompting could overcome the limitations of current supervised learning approaches that require large amounts of labelled data. Investigating how SAM can be effectively adapted and fine-tuned for accurate and robust ELM segmentation in OCT images, including those with varying image quality and pathologies, is a crucial direction for future work. This could lead to more efficient and generalizable ELM segmentation tools, reducing the need for extensive manual annotation.

By pursuing these future research directions, the field can further advance the application of DL in ophthalmology, ultimately leading to improved patient outcomes and more effective clinical management of macular hole and other retinal conditions.

References

[1] Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., et al. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297.

[2] Abramoff, M. D., Garvin, M. K., and Sonka, M. (2010). Retinal imaging and image analysis. *IEEE reviews in biomedical engineering*, 3:169–208.

[3] Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., and Folk, J. C. (2018). Pivotal trial of an autonomous ai-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ digital medicine*, 1(1):39.

[4] Abràmoff, M. D., Garvin, M. K., and Sonka, M. (2010). Retinal imaging and image analysis. *IEEE Reviews in Biomedical Engineering*, 3:169–208.

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

[6] Ali, F. S., Stein, J. D., Blachley, T. S., Ackley, S., and Stewart, J. M. (2017). Incidence of and risk factors for developing idiopathic macular hole among a diverse group of patients throughout the United States. *JAMA Ophthalmology*, 135(4):299–305.

[7] Alom, M. Z., Hasan, M., Yakopcic, C., and Taha, Tarek M.and Asari, V. K. (2018). Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation. *arXiv preprint arXiv:1802.06955*.

[8] Alqudah, A. M. (2020). Aoct-net: a convolutional network automated classification of multiclass retinal diseases using spectral-domain optical coherence tomography images. *Medical & Biological Engineering & Computing*, 58(1):41–53.

[9] Amini, A., Schwarting, W., Soleimany, A., and Rus, D. (2020). Deep evidential regression. *Advances in Neural Information Processing Systems*, 33:14927–14937.

[10] Anantrasirichai, N., Achim, A., Morgan, J. E., Erchova, I., and Nicholson, L. (2013). Svm-based texture classification in optical coherence tomography. In *IEEE International Symposium on Biomedical Imaging*, pages 1332–1335.

[11] Araújo, T., Aresta, G., Mendonça, L., Penas, S., Maia, C., Carneiro, Â., Mendonça, A. M., and Campilho, A. (2020). Dr| graduate: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. *Medical Image Analysis*, 63:101715.

[12] Ayhan, M. S., Kühlewein, L., Aliyeva, G., Inhoffen, W., Ziemssen, F., and Berens, P. (2020). Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection. *Medical Image Analysis*, 64:101724.

[13] Badar, M., Haris, M., and Fatima, A. (2020). Application of deep learning for retinal image analysis: A review. *Computer Science Review*, 35:100203.

[14] Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495.

[15] Baghaie, A., Yu, Z., and D'Souza, R. M. (2015). State-of-the-art in retinal optical coherence tomography image analysis. *Quantitative Imaging in Medicine and Surgery*, 5(4):603.

[16] Bagheri, A. B., Rouzi, M. D., Koohbanani, N. A., Mahoor, M. H., Finco, M., Lee, M., Najafi, B., and Chung, J. (2023). Potential applications of artificial intelligence (ai) and machine learning (ml) on diagnosis, treatment, outcome prediction to address health care disparities of chronic limb-threatening ischemia (clti). In *Seminars in Vascular Surgery*. Elsevier.

[17] Baheti, B., Innani, S., Gajre, S., and Talbar, S. (2020). Eff-unet: A novel architecture for semantic segmentation in unstructured environment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 358–359.

[18] Bai, F., Gibson, S. J., Marques, M. J., and Podoleanu, A. (2018). Superpixel guided active contour segmentation of retinal layers in OCT volumes. In *Canterbury Conference on OCT with Emphasis on Broadband Optical Sources*, volume 10591, pages 38–47.

[19] Bai, Y., Mei, J., Yuille, A., and Xie, C. (2021). Are transformers more robust than CNNs? In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

[20] Bansal, R., Raj, G., and Choudhury, T. (2016). Blur image detection using Laplacian operator and Open-CV. In *International Conference System Modeling & Advancement in Research Trends*, pages 63–67.

[21] Baroni, M., Fortunato, P., and Laezza, I. (2007). Automatic detection of the retinal layer structures on oct images. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4444–4447. IEEE.

[22] Bezryadin, S., Bourov, P., and Ilinih, D. (2007). Brightness calculation in digital image processing. In *International Symposium on Technologies for Digital Photo Fulfillment*, volume 2007, pages 10–15.

[23] Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). LOF: identifying density-based local outliers. *ACM SIGMOD International Conference on Management of Data*, 29(2):93–104.

[24] Budai, A., Michelson, G., and Hornegger, J. (2010). Multiscale blood vessel segmentation in retinal fundus images. In *Bildverarbeitung für die Medizin*, pages 261–265.

[25] Cabrera Fernández, D., Salinas, H. M., and Puliafito, C. A. (2005). Automated detection of retinal layer structures on optical coherence tomography images. *Optics express*, 13(25):10200–10216.

[26] Carass, A., Lang, A., Hauser, M., Calabresi, P. A., Ying, H. S., and Prince, J. L. (2014). Multiple-object geometric deformable model for segmentation of macular OCT. *Biomedical Optics Express*, 5(4):1062–1074.

[27] Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2018a). DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 40(04):834–848. https://doi.org/10.48550/arXiv.1606.00915.

[28] Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

[29] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018b). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*, pages 801–818.

[30] Chen, X., Williams, B. M., Vallabhaneni, S. R., Czanner, G., Williams, R., and Zheng, Y. (2019). Learning active contour models for medical image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11632–11640.

[31] Chen, Y., Mancini, M., Zhu, X., and Akata, Z. (2022). Semi-supervised and unsupervised deep visual learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[32] Chen, Y., Nasrulloh, A., Wilson, I., Caspar, G., Maged, H., Obara, B., and Steel, D. (2020a). Macular hole morphology and measurement using an automated three dimensional image segmentation algorithm. *British Medical Journal Open Ophthalmology*.

[33] Chen, Y., Nasrulloh, A. V., Wilson, I., Geenen, C., Habib, M., Obara, B., and Steel, D. H. (2020b). Macular hole morphology and measurement using an automated three-dimensional image segmentation algorithm. *BMJ Open Ophthalmology*, 5(1):e000404.

[34] Chiu, S. J., Allingham, M. J., Mettu, P. S., Cousins, S. W., Izatt, J. A., and Farsiu, S. (2015). Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema. *Biomedical Optics Express*, 6(4):1172–1194.

[35] Chiu, S. J., Izatt, J. A., O'Connell, R. V., Winter, K. P., Toth, C. A., and Farsiu, S. (2012). Validated automatic segmentation of AMD pathology including drusen and geographic atrophy in SD-OCT images. *Investigative Ophthalmology & Visual Science*, 53(1):53–61.

[36] Chiu, S. J., Li, X. T., Nicholas, P., Toth, C. A., Izatt, J. A., and Farsiu, S. (2010). Automatic segmentation of seven retinal layers in sdoct images congruent with expert manual segmentation. *Optics express*, 18(18):19413–19428.

[37] Commons, W. (2024). File:human eye cross section detached retina-es.svg — wikimedia commons, the free media repository. [Online; accessed 5-August-2024].

[38] D'Amico, D. J. (1994). Diseases of the retina. *New England Journal of Medicine*, 331(2):95–106.

[39] Dehghan Rouzi, M., Moshiri, B., Khoshnevisan, M., Akhaee, M. A., Jaryani, F., Salehi Nasab, S., and Lee, M. (2023). Breast cancer detection with an ensemble of deep learning networks using a consensus-adaptive weighting method. *Journal of Imaging*, 9(11):247.

[40] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

[41] Ding, M., Kong, K., Chen, J., Kirchenbauer, J., Goldblum, M., Wipf, D., Huang, F., and Goldstein, T. (2021). A closer look at distribution shifts and out-of-distribution generalization on graphs. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*.

[42] Dolezal, J. M., Srisuwananukorn, A., Karpeyev, D., Ramesh, S., Kochanny, S., Cody, B., Mansfield, A. S., Rakshit, S., Bansal, R., Bois, M. C., et al. (2022). Uncertainty-informed deep learning models enable high-confidence predictions for digital histopathology. *Nature Communications*, 13(1):6572.

[43] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

[44] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

[45] Draelos, R. L., Carin, L., et al. (2018). Machine learning-based screening of age-related macular degeneration from oct using retinal layer thickness profiles. In *Ophthalmic Technologies XXVIII*, volume 10579, page 105790K. SPIE.

[46] Duan, J., Xie, W., Liu, R. W., Tench, C., Gottlob, I., Proudlock, F., and Bai, L. (2018). Oct segmentation: Integrating open parametric contour model of the retinal layers and shape constraint to the mumford-shah functional. In *International Workshop on Shape in Medical Imaging*, pages 178–188.

[47] Emam, K. E., Rodgers, S., and Malin, B. (2015). Anonymising and sharing individual patient data. *British Medical Journal*, 350.

[48] Falcão, F., Zoppi, T., Silva, C. B. V., Santos, A., Fonseca, B., Ceccarelli, A., and Bondavalli, A. (2019). Quantitative comparison of unsupervised anomaly detection algorithms for intrusion detection. In *ACM/SIGAPP Symposium on Applied Computing*, pages 318–327.

[49] Felfeli, T. and Mandelcorn, E. D. (2019). Macular hole hydrodissection: surgical technique for the treatment of persistent, chronic, and large macular holes. *Retina*, 39(4):743–752.

[50] Flaxman, S. R., Bourne, R. R., Resnikoff, S., Ackland, P., Braithwaite, T., Cicinelli, M. V., Das, A., Jonas, J. B., Keeffe, J., Kempen, J. H., et al. (2017). Global causes of blindness and distance vision impairment 1990–2020: a systematic review and meta-analysis. *The Lancet Global Health*, 5(12):e1221–e1234.

[51] Francis, L. and Sreenath, N. (2019). Pre-processing techniques for detection of blurred images. In *International Conference on Computational Intelligence and Data Engineering*, pages 59–66.

[52] Fu, Y., Lei, Y., Wang, T., Curran, W. J., Liu, T., and Yang, X. (2021). A review of deep learning based methods for medical image multi-organ segmentation. *Physica Medica*, 85:107–122. https://doi.org/10.1016/j.ejmp.2021.05.003.

[53] Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Balcan, M. F. and Weinberger, K. Q., editors, *The 33rd International Conference on Machine Learning*, volume 48 of *Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

[54] Garvin, M. K., Abramoff, M. D., Wu, X., Russell, S. R., Burns, T. L., and Sonka, M. (2009). Automated 3-d intraretinal layer segmentation of macular spectral-domain optical coherence tomography images. *IEEE transactions on medical imaging*, 28(9):1436–1447.

[55] Geenen, C., Murphy, D. C., Sandinha, M. T., Rees, J., and Steel, D. H. (2019). Significance of preoperative external limiting membrane height on visual prognosis in patients undergoing macular hole surgery. *Retina*, 39(7):1392–1398.

[56] Goldstein, M. and Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS One*, 11(4):e0152173.

[57] Gong, X. and Bansmer, S. (2015). Horn–schunck optical flow applied to deformation measurement of a birdlike airfoil. *Chinese Journal of Aeronautics*, 28(5):1305–1315.

[58] Gopinath, K., Rangrej, S. B., and Sivaswamy, J. (2017). A deep learning framework for segmentation of retinal layers from OCT images. In *IAPR Asian Conference on Pattern Recognition*, pages 888–893.

[59] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *jama*, 316(22):2402–2410.

[60] Gupta, V., Mahle, R., and Shriwas, R. S. (2015). Image denoising using wavelet transform method. In *International Conference on Wireless and Optical Communications Networks*, pages 1–4.

[61] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

[62] Hee, M. R., Puliafito, C. A., Wong, C., Duker, J. S., Reichel, E., Schuman, J. S., Swanson, E. A., and Fujimoto, J. G. (1995). Optical coherence tomography of macular holes. *Ophthalmology*, 102(5):748–756.

[63] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

[64] Horn, B. K. and Schunck, B. G. (1981). Determining optical flow. In *Techniques and Applications of Image Understanding*, volume 281, pages 319–331.

[65] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*. https://doi.org/10.48550/arXiv.1704.04861.

[66] Hu, Y., Xiao, Y., Quan, W., Zhang, B., Wu, Y., Wu, Q., Liu, B., Zeng, X., Fang, Y., Hu, Y., et al. (2021). A multi-center study of prediction of macular hole status after vitrectomy and internal limiting membrane peeling by a deep learning model. *Annals of Translational Medicine*, 9(1).

[67] Huang, D., Swanson, E. A., Lin, C. P., Schuman, J. S., Stinson, W. G., Chang, W., Hee, M. R., Flotte, T., Gregory, K., Puliafito, C. A., et al. (1991). Optical coherence tomography. *Science*, 254(5035):1178–1181.

[68] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2261–2269.

[69] Huang, P., He, P., Tian, S., Ma, M., Feng, P., Xiao, H., Mercaldo, F., Santone, A., and Qin, J. (2022). A vit-amc network with adaptive model fusion and multiobjective optimization for interpretable laryngeal tumor grading from histopathological images. *IEEE Transactions on Medical Imaging*, 42(1):15–28.

[70] Huang, P., Tan, X., Zhou, X., Liu, S., Mercaldo, F., and Santone, A. (2021). Fabnet: fusion attention block and transfer learning for laryngeal cancer tumor grading in p63 ihc histopathology images. *IEEE Journal of Biomedical and Health Informatics*, 26(4):1696–1707.

[71] Huang, P., Zhou, X., He, P., Feng, P., Tian, S., Sun, Y., Mercaldo, F., Santone, A., Qin, J., and Xiao, H. (2023). Interpretable laryngeal tumor grading of histopathological images via depth domain adaptive network with integration gradient cam and priori experience-guided attention. *Computers in Biology and Medicine*, 154:106447.

[72] Ishikawa, H., Stein, D. M., Wollstein, G., Beaton, S., Fujimoto, J. G., and Schuman, J. S. (2005). Macular segmentation with optical coherence tomography. *Investigative Ophthalmology & Visual Science*, 46(6):2012–2017.

[73] Jackson, T. L., Donachie, P. H., Sparrow, J. M., and Johnston, R. L. (2013). United Kingdom national ophthalmology database study of vitreoretinal surgery: report 2, macular hole. *Ophthalmology*, 120(3):629–634.

[74] Jin, W., Li, X., Fatehi, M., and Hamarneh, G. (2023). Guidelines and evaluation of clinical explainable ai in medical image analysis. *Medical Image Analysis*, 84:102684.

[75] Jordan, M. (2009). The exponential family: Conjugate priors.

[76] Kajić, V., Považay, B., Hermann, B., Hofer, B., Marshall, D., Rosin, P. L., and Drexler, W. (2010). Robust segmentation of intraretinal layers in the normal human fovea using a novel statistical model based on texture and shape analysis. *Optics Express*, 18(14):14730–14744.

[77] Kauer-Bonin, J., Yadav, S. K., Beckers, I., Gawlik, K., Motamedi, S., Zimmermann, H. G., Kadas, E. M., Haußer, F., Paul, F., and Brandt, A. U. (2022). Modular deep neural networks for automatic quality control of retinal optical coherence tomography scans. *Computers in Biology and Medicine*, 141:104822.

[78] Kavitha, S. and Ramakrishnan, S. (2012). Automatic detection of drusen in retinal images. *Computerized Medical Imaging and Graphics*, 36(4):329–341.

[79] Kawczynski, M. G., Bengtsson, T., Dai, J., Hopkins, J. J., Gao, S. S., and Willis, J. R. (2020). Development of deep learning models to predict best-corrected visual acuity from optical coherence tomography. *Translational Vision Science & Technology*, 9(2):51–51.

[80] Ker, J., Wang, L., Rao, J., and Lim, T. (2018). Deep learning applications in medical image analysis. *IEEE Access*, 6:9375–9389.

[81] Kiani, A., Uyumazturk, B., Rajpurkar, P., Wang, A., Gao, R., Jones, E., Yu, Y., Langlotz, C., Ball, R., Montine, T., et al. (2020). Impact of a deep learning assistant on the histopathologic classification of liver cancer. npj digital medicine 3. *Nature Research*, 3.

[82] Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations, San Diego, CA, USA, May 7-9*. https://doi.org/10.48550/arXiv.1412.6980.

[83] Kniestedt, C. and Stamper, R. L. (2003). Visual acuity and its measurement. *Ophthalmology Clinics of North America*, 16(2):155–70.

[84] Kobayashi, H. and Kobayashi, K. (1999). Correlation of quantitative three-dimensional measurements of macular hole size with visual acuity after vitrectomy. *Graefe's archive for clinical and experimental ophthalmology*, 237(4):283–288.

[85] Kong, H., Akakin, H. C., and Sarma, S. E. (2013). A generalized laplacian of gaussian filter for blob detection and its applications. *IEEE Transactions on Cybernetics*, 43(6):1719–1733.

[86] Koozekanani, D., Boyer, K., and Roberts, C. (2001). Retinal thickness measurements from optical coherence tomography using a markov boundary model. *IEEE Transactions on Medical Imaging*, 20(9):900–916.

[87] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105.

[88] Kucukgoz, B., Yapici, M. M., Murphy, D. C., Spowart, E., Steel, D. H., and Obara, B. (2024a). Deep learning using preoperative optical coherence tomography images to predict visual acuity following surgery for idiopathic full-thickness macular holes. *IEEE Access*.

[89] Kucukgoz, B., Yapici, M. M., Steel, D. H., and Obara, B. (2023). Evaluation of 2D and 3D deep learning approaches for predicting visual acuity following surgery for idiopathic full-thickness macular holes in spectral domain optical coherence tomography images. In *2023 International Symposium on Image and Signal Processing and Analysis*, pages 1–6.

[90] Kucukgoz, B., Zou, K., Murphy, D. C., Steel, D. H., Obara, B., and Fu, H. (2024b). A clinician's guide to predict postoperative visual acuity in patients with macular holes. *Computerized Medical Imaging and Graphics*.

[91] Kugelman, J., Alonso-Caneiro, D., Read, S. A., Vincent, S. J., Chen, F. K., and Collins, M. J. (2020). Effect of altered oct image quality on deep learning boundary segmentation. *IEEE Access*, 8:43537–43553.

[92] Kugelman, J., Alonso-Caneiro, D., Read, S. A., Vincent, S. J., and Collins, M. J. (2018). Automatic segmentation of OCT retinal boundaries using recurrent neural networks and graph search. *Biomedical Optics Express*, 9(11):5759–5777.

[93] Kusuhara, S. and Negi, A. (2014). Predicting visual outcome following surgery for idiopathic macular holes. *Ophthalmologica*, 231(3):125–132.

[94] Lachance, A., Godbout, M., Antaki, F., Hébert, M., Bourgault, S., Caissie, M., Tourville, É., Durand, A., and Dirani, A. (2022). Predicting visual improvement after macular hole surgery: a combined model using deep learning and clinical features. *Translational Vision Science & Technology*, 11(4):6–6.

[95] Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017a). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

[96] Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017b). Simple and scalable predictive uncertainty estimation using deep ensembles. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, page 6405–6416. Curran Associates, Inc.

[97] Landa, G., Gentile, R., Garcia, P., Muldoon, T., and Rosen, R. (2012). External limiting membrane and visual outcome in macular hole repair: spectral domain oct analysis. *Eye*, 26(1):61–69.

[98] Leandro, J. J. G., Cesar, R. M., and Jelinek, H. F. (2006). Blood vessel segmentation in retinal images using wavelets and mathematical morphology. *IEEE Transactions on Medical Imaging*, 25(9):1214–1222.

[99] Lemaitre, G., Rastgoo, M., Massich, J., Sankar, S., Mériaudeau, F., and Sidibé, D. (2015). Classification of sd-oct volumes with lbp: application to dme detection. *Ophthalmic Medical Image Analysis International Workshop 2*.

[100] Li, B., Han, Z., Li, H., Fu, H., and Zhang, C. (2022). Trustworthy long-tailed classification. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6970–6979.

[101] Li, H., Lim, J. H., Liu, J., Wong, T. Y., and Wong, H. K. (2011). Automated detection of macular edema with optical coherence tomography. *IEEE Transactions on Biomedical Engineering*, 58(2):263–266.

[102] Li, X., Shen, L., Shen, M., Tan, F., and Qiu, C. S. (2019). Deep learning based early stage diabetic retinopathy detection using optical coherence tomography. *Neurocomputing*, 369:134–144.

[103] Liu, P., Sun, Y., Dong, C., Song, D., Jiang, Y., Liang, J., Yin, H., Li, X., and Zhao, M. (2016). A new method to predict anatomical outcome after idiopathic macular hole surgery. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 254(4):683–688.

[104] Liu, W., Sun, Y., and Ji, Q. (2020). MDAN-UNet: Multi-scale and dual attention enhanced nested U-Net architecture for segmentation of optical coherence tomography images. *Algorithms*, 13(3):60.

[105] Liu, X., Song, L., Liu, S., and Zhang, Y. (2021). A review of deep-learning-based medical image segmentation methods. *Sustainability*, 13(3):1224. https://doi.org/10.3390/su13031224.

[106] Liu, Y.-Y., Chen, M., Ishikawa, H., Wollstein, G., Schuman, J. S., and Rehg, J. M. (2011). 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*, 15(5):748–759.

[107] Lois, N., Burr, J., Norrie, J., Vale, L., Cook, J., McDonald, A., Boachie, C., Ternent, L., and McPherson, G. (2011). Internal limiting membrane peeling versus no peeling for idiopathic full-thickness macular hole: a pragmatic randomized controlled trial. *Investigative Ophthalmology & Visual Science*, 52(3):1586–1592.

[108] Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440.

[109] Martinez-Perez, M. E., Hughes, A. D., Thom, S. A., Bharath, A. A., and Parker, K. H. (2007). Segmentation of blood vessels from red-free and fluorescein retinal images. *Medical Image Analysis*, 11(1):47–61.

[110] Matsui, T. and Lee, S. (1993). Automatic recognition of blood vessels in fundus images using morphological processing. *Pattern Recognition*, 26(3):405–411.

[111] McCannel, C. A., Ensminger, J. L., Diehl, N. N., and Hodge, D. N. (2009). Population-based incidence of macular holes. *Ophthalmology*, 116(7):1366–1369.

[112] Meinhardt-Llopis, E., Pérez, J. S., and Kondermann, D. (2013). Horn-schunck optical flow with a multi-scale strategy. *Image Processing on Line*, 3:151–172.

[113] Meyes, R., Lu, M., de Puiseau, C. W., and Meisen, T. (2019). Ablation studies in artificial neural networks. *arXiv preprint arXiv:1901.08644*.

[114] Mishra, Z., Ganegoda, A., Selicha, J., Wang, Z., Sadda, S. R., and Hu, Z. (2020). Automated retinal layer segmentation using graph-based algorithm incorporating deep-learning-derived information. *Scientific Reports*, 10(1):9541.

[115] Motozawa, N., An, G., Takagi, S., Kitahata, S., Mandai, M., Hirami, Y., Yokota, H., Akiba, M., Tsujikawa, A., Takahashi, M., et al. (2019). 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*, 8(4):527–539.

[116] Mujat, M., Chan, R. C., Cense, B., Park, B. K., Joo, C., Akkin, T., Chen, T.-W., and de Boer, J. F. (2010). Retinal nerve fiber layer thickness map determined from optical coherence tomography images. *Optics Express*, 18(10):10259–10273.

[117] Murphy, D. C., Nasrulloh, A., Lendrem, C., Lendrem, C., la Cour, M. A. M., Obara, B., and Steel, D. (2020). Predicting post-operative vision for macular hole with automated image analysis. *Ophthalmology Retina*.

[118] Nasrulloh, A., Willcocks, C., Jackson, P. T., Geenen, C., Habib, M. S., Steel, D. H., and Obara, B. (2018). Multi-scale segmentation and surface fitting for measuring 3d macular holes. *IEEE Transactions on Medical Imaging*, 37(2):580–589.

[119] Nasrulloh, A. V., Willcocks, C. G., Jackson, P. T., Geenen, C., Habib, M. S., Steel, D. H., and Obara, B. (2017). Multi-scale segmentation and surface fitting for measuring 3-d macular holes. *IEEE Transactions on Medical Imaging*, 37(2):580–589.

[120] Niemeijer, M., van Ginneken, B., Russell, S. R., Suttorp-Schulten, M. S., and Abràmoff, M. D. (2007). Automated detection and differentiation of drusen, exudates, and cotton-wool spots in digital color fundus photographs for diabetic retinopathy diagnosis. *Investigative ophthalmology & visual science*, 48(5):2260–2267.

[121] Novosel, J., Thepass, G., Lemij, H. G., de Boer, J. F., Vermeer, K. A., and van Vliet, L. J. (2015). Loosely coupled level sets for simultaneous 3D retinal layer segmentation in optical coherence tomography. *Medical Image Analysis*, 26(1):146–158.

[122] Obata, S., Ichiyama, Y., Kakinoki, M., Sawada, O., Saishin, Y., Ito, T., Tomioka, M., and Ohji, M. (2021). Prediction of postoperative visual acuity after vitrectomy for macular hole using deep learning–based artificial intelligence. *Graefe's Archive for Clinical and Experimental Ophthalmology*, pages 1–11.

[123] Oishi, A., Hata, M., Shimozono, M., Mandai, M., Nishida, A., and Kurimoto, Y. (2010). The significance of external limiting membrane status for visual acuity in age-related macular degeneration. *American journal of ophthalmology*, 150(1):27–32.

[124] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., et al. (2018). Attention U-Net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*.

[125] Orlando, J. I., Seeböck, P., Bogunović, H., Klimscha, S., Grechenig, C., Waldstein, S., Gerendas, B. S., and Schmidt-Erfurth, U. (2019). U2-net: A bayesian u-net model with epistemic uncertainty feedback for photoreceptor layer segmentation in pathological oct scans. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1441–1445. IEEE.

[126] Park, C., Rouzi, M. D., Atique, M. M. U., Finco, M., Mishra, R. K., Barba-Villalobos, G., Crossman, E., Amushie, C., Nguyen, J., Calarge, C., et al. (2023). Machine learning-based aggression detection in children with adhd using sensor-based physical activity monitoring. *Sensors*, 23(10):4949.

[127] Pekala, M., Joshi, N., Liu, T. A., Bressler, N. M., DeBuc, D. C., and Burlina, P. (2019). Deep learning based retinal OCT segmentation. *Computers in Biology and Medicine*, 114:103445.

[128] Poplin, R., Varadarajan, A. V., Blumer, K., Liu, Y., McConnell, M. V., Corrado, G. S., Peng, L., and Webster, D. R. (2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature biomedical engineering*, 2(3):158–164.

[129] Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[130] Rasti, R., Rabbani, H., Mehridehnavi, A., and Hajizadeh, F. (2017). Macular oct classification using a multi-scale convolutional neural network ensemble. *IEEE Transactions on Medical Imaging*, 37(4):1024–1034.

[131] Rizzo, S., Savastano, A., Lenkowicz, J., Savastano, M. C., Boldrini, L., Bacherini, D., Falsini, B., and Valentini, V. (2021). Artificial intelligence and oct angiography in full thickness macular hole. new developments for personalized medicine. *Diagnostics*, 11(12):2319.

[132] Romo-Bucheli, D., Erfurth, U. S., and Bogunović, H. (2020). 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*, 24(12):3456–3465.

[133] Ronneberger, O., Fischer, P., and Brox, T. (2015a). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer.

[134] Ronneberger, O., Fischer, P., and Brox, T. (2015b). U-Net: convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 234–241.

[135] Roy, A. G., Conjeti, S., Karri, S. P. K., Sheet, D., Katouzian, A., Wachinger, C., and Navab, N. (2017). Relaynet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks. *Biomedical Optics Express*, 8(8):3627–3642.

[136] Ryu, J., Rehman, M. U., Nizami, I. F., and Chong, K. T. (2023). Segr-net: A deep learning framework with multi-scale feature fusion for robust retinal vessel segmentation. *Computers in Biology and Medicine*, 163:107132.

[137] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520.

[138] Secretariat, M. A. (2009). Optical coherence tomography for age-related macular degeneration and diabetic macular edema: An evidence-based analysis. *Ontario Health Technology Assessment Series*, 9(13):1.

[139] Sedai, S., Antony, B., Mahapatra, D., and Garnavi, R. (2018). Joint segmentation and uncertainty visualization of retinal layers in optical coherence tomography images using bayesian deep learning. In *Computational Pathology and Ophthalmic Medical Image Analysis: First International Workshop, COMPAY 2018, and 5th International Workshop, OMIA 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16-20, 2018, Proceedings 5*, pages 219–227. Springer.

[140] Simone, G., Pedersen, M., and Hardeberg, J. Y. (2012). Measuring perceptual contrast in digital images. *Journal of Visual Communication and Image Representation*, 23(3):491–506.

[141] Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[142] Singh, S. P., Wang, L., Gupta, S., Goli, H., Padmanabhan, P., and Gulyás, B. (2020). 3D deep learning on medical images: a review. *Sensors*, 20(18). https://doi.org/10.3390%2Fs20185097.

[143] Singh, V. K., Kucukgoz, B., Murphy, D. C., Xiong, X., Steel, D. H., and Obara, B. (2022). Benchmarking automated detection of the retinal external limiting membrane in a 3d spectral domain optical coherence tomography image dataset of full thickness macular holes. *Computers in Biology and Medicine*, 140:105070.

[144] Smiddy, W. E. and Flynn Jr, H. W. (2004). Pathogenesis of macular holes and therapeutic implications. *American journal of ophthalmology*, 137(3):525–537.

[145] Srinivasan, P. P., Kim, L. A., Mettu, P. S., Cousins, S. W., Comer, G. M., Izatt, J. A., and Farsiu, S. (2014a). Automated detection of diabetic macular edema and age-related macular degeneration in oct images. *Biomedical optics express*, 5(10):3568–3577.

[146] Srinivasan, P. P., Kim, L. A., Mettu, P. S., Cousins, S. W., Comer, G. M., Izatt, J. A., and Farsiu, S. (2014b). Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images. *Biomedical Optics Express*, 5(10):3568–3577.

[147] Steel, D., Donachie, P., Aylward, G., Laidlaw, D., and Williamson, T. (2021a). Macular hole outcome group. factors affecting anatomical and visual outcome after macular hole surgery: findings from a large prospective uk cohort. *Eye*, 35(1):316–325.

[148] Steel, D., Donachie, P., Aylward, G., Laidlaw, D., Williamson, T., and Yorston, D. (2021b). Factors affecting anatomical and visual outcome after macular hole surgery: findings from a large prospective uk cohort. *Eye*, 35(1):316–325.

[149] Steel, D. H. and Lotery, A. J. (2013). Idiopathic vitreomacular traction and macular hole: a comprehensive review of pathophysiology, diagnosis, and treatment. *Eye*, 27(1):S1–S21.

[150] Subramanian, M., Kumar, M. S., Sathishkumar, V., Prabhu, J., Karthick, A., Ganesh, S. S., Meem, M. A., et al. (2022). Diagnosis of retinal diseases based on bayesian optimization deep learning network using optical coherence tomography images. *Computational Intelligence and Neuroscience*, 2022.

[151] Sun, Y., Fu, Z., Stainton, S., Barney, S., Hogg, J., Innes, W., and Dlay, S. (2019). Automated retinal layer segmentation of OCT images using two-stage fcn and decision mask. In *IEEE International Symposium on Signal Processing and Information Technology*, pages 1–6.

[152] Sun, Z. and Sun, Y. (2019). Automatic detection of retinal regions using fully convolutional networks for diagnosis of abnormal maculae in optical coherence tomography images. *Journal of Biomedical Optics*, 24(5):1–9.

[153] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826.

[154] Takahashi, A., Yoshida, A., Nagaoka, T., Takamiya, A., Sato, E., Kagokawa, H., Kameyama, D., Sogawa, K., Ishiko, S., and Hirokawa, H. (2012). Idiopathic full-thickness macular holes and the vitreomacular interface: a high-resolution spectral-domain optical coherence tomography study. *American Journal of Ophthalmology*, 154(5):881–892.

[155] Tam, A. L., Yan, P., Gan, N. Y., and Lam, W.-C. (2018). The current surgical management of large, recurrent, or persistent macular holes. *Retina*, 38(7):1263–1275.

[156] Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR.

[157] Tan, M. and Le, Q. (2021). Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning*, pages 10096–10106. PMLR.

[158] Tian, J., Varga, B., Somfai, G. M., Lee, W.-H., Smiddy, W. E., and DeBuc, D. C. (2015). Real-time automatic segmentation of optical coherence tomography volume data of the macular region. *PloS One*, 10(8):e0133908.

[159] Tian, L., Hunt, B., Bell, M. A. L., Yi, J., Smith, J. T., Ochoa, M., Intes, X., and Durr, N. J. (2021). Deep learning in biomedical optics. *Lasers in Surgery and Medicine*, 53(6):748–775. https://doi.org/10.1002/lsm.23414.

[160] Tian, Y., Zeng, H., Hou, J., Chen, J., and Ma, K.-K. (2020). Light field image quality assessment via the light field coherence. *IEEE Transactions on Image Processing*, 29:7945–7956.

[161] Trinh, M., Khou, V., Zangerl, B., Kalloniatis, M., and Nivison-Smith, L. (2021). Modelling normal age-related changes in individual retinal layers using location-specific oct analysis. *Scientific Reports*, 11(1):558.

[162] Tsuji, T., Hirose, Y., Fujimori, K., Hirose, T., Oyama, A., Saikawa, Y., Mimura, T., Shiraishi, K., Kobayashi, T., Mizota, A., et al. (2020). Classification of optical coherence tomography images using a capsule network. *BMC Ophthalmology*, 20(1):1–9.

[163] Ullrich, S., Haritoglou, C., Gass, C. t., Schaumberger, M., Ulbig, M., and Kampik, A. (2002). Macular hole size as a prognostic factor in macular hole surgery. *British Journal of Ophthalmology*, 86(4):390–393.

[164] Upadhyay, U., Karthik, S., Chen, Y., Mancini, M., and Akata, Z. (2022). Bayescap: Bayesian identity cap for calibrated uncertainty in frozen neural networks. In *European Conference on Computer Vision*, pages 299–317. Springer.

[165] Venhuizen, F. G., van Ginneken, B., Bloemen, B., van Grinsven, M. J., Philipsen, R., Hoyng, C., Theelen, T., and Sánchez, C. I. (2015). Automated age-related macular degeneration classification in oct using unsupervised feature learning. In *Medical Imaging 2015: Computer-Aided Diagnosis*, volume 9414, page 94141I.

[166] Walter, T., Klein, J.-C., Massin, P., and Erginay, A. (2002). A contribution of image processing to the diagnosis of diabetic retinopathy-detection of exudates in color fundus images of the human retina. *IEEE transactions on medical imaging*, 21(10):1236–1243.

[167] Wang, M., Lin, T., Wang, L., Lin, A., Zou, K., Xu, X., Zhou, Y., Peng, Y., Meng, Q., Qian, Y., et al. (2023). Uncertainty-inspired open set learning for retinal anomaly identification. *Nature Communications*, 14(1):6757.

[168] Wang, Z., Zhong, Y., Yao, M., Ma, Y., Zhang, W., Li, C., Tao, Z., Jiang, Q., and Yan, B. (2021). Automated segmentation of macular edema for the diagnosis of ocular disease using deep learning method. *Scientific Reports*, 11(1):13392.

[169] Wen, H., Zhao, J., Xiang, S., Lin, L., Liu, C., Wang, T., An, L., Liang, L., and Huang, B. (2022). Towards more efficient ophthalmic disease classification and lesion location via convolution transformer. *Computer Methods and Programs in Biomedicine*, 220:106832.

[170] Xie, X., Niu, J., Liu, X., Chen, Z., Tang, S., and Yu, S. (2021). A survey on incorporating domain knowledge into deep learning for medical image analysis. *Medical Image Analysis*, 69:101985.

[171] Xie, Y., Nguyen, Q. D., Hamzah, H., Lim, G., Bellemo, V., Gunasekeran, D. V., Yip, M. Y., Lee, X. Q., Hsu, W., Lee, M. L., et al. (2020). Artificial intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: an economic analysis modelling study. *The Lancet Digital Health*, 2(5):e240–e249.

[172] Xiong, X., Steel, D. H., and Obara, B. (2022). Evaluation of 2D and 3D deep learning approaches for automatic segmentation of the retinal external limiting membrane in spectral domain optical coherence tomography images. In *International Conference on Intelligent Informatics and Biomedical Science*, volume 7, pages 341–345. IEEE. https://doi.org/10.1109/ICIIBMS55689.2022.9971669.

[173] Xu, D., Yuan, A., Kaiser, P. K., Srivastava, S. K., Singh, R. P., Sears, J. E., Martin, D. F., and Ehlers, J. P. (2013). A novel segmentation algorithm for volumetric analysis of macular hole boundaries identified with optical coherence tomography. *Investigative Ophthalmology & Visual Science*, 54(1):163–169.

[174] Yanagihara, R. T., Lee, C. S., Ting, D. S. W., and Lee, A. Y. (2020). Methodological challenges of deep learning in optical coherence tomography for retinal diseases: a review. *Translational Vision Science & Technology*, 9(2):11–11.

[175] Yang, M., Han, J., Park, J. I., Hwang, J. S., Han, J. M., Yoon, J., Choi, S., Hwang, G., and Hwang, D. D.-J. (2023). Prediction of visual acuity in pathologic myopia with myopic choroidal neovascularization treated with anti-vascular endothelial growth factor using a deep neural network based on optical coherence tomography images. *Biomedicines*, 11(8):2238.

[176] Yim, J., Chopra, R., Spitz, T., Winkens, J., Obika, A., Kelly, C., Askham, H., Lukic, M., Huemer, J., Fasler, K., et al. (2020). Predicting conversion to wet age-related macular degeneration using deep learning. *Nature Medicine*, 26(6):892–899.

[177] Zhang, Y., Zhang, B., Coenen, F., Xiao, J., and Lu, W. (2014). One-class kernel subspace ensemble for medical image classification. *EURASIP Journal on Advances in Signal Processing*, 2014(1):1–13.

[178] Zhou, X., Tang, C., Huang, P., Tian, S., Mercaldo, F., and Santone, A. (2023). Asi-dbnet: an adaptive sparse interactive resnet-vision transformer dual-branch network for the grading of brain cancer histopathological images. *Interdisciplinary Sciences: Computational Life Sciences*, 15(1):15–31.

[179] Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. (2018). UNet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11.

[180] Zou, K., Chen, Z., Yuan, X., Shen, X., Wang, M., and Fu, H. (2023). A review of uncertainty estimation and its application in medical imaging. *Meta-Radiology*, 1(1):100003.