Multi-Modality Fusion Convolutional Network (MM-FCN) for Prediction of Retinal Vein Occlusion using Swept Source Optical Coherence Tomography Angiography

Xinyu Huo

THE UNIVERSITY OF
SYDNEY

Supervisor: Prof. Jinman Kim
Auxiliary Supervisor: Prof. Lei Bi

A thesis submitted in fulfilment of
the requirements for the degree of
Master of Philosophy

School of Computer Science
Faculty of Engineering
The University of Sydney
Australia

31 January 2025

ABSTRACT

Retinal vein occlusion (RVO) is one of the most complex vascular disorders affecting the retina, characterized by the occlusion of retinal veins and potentially leading to significant visual impairment. Inadequate evaluation and medical intervention of RVOs may lead to the extension of non-perfusion areas (NPAs, regions lacking adequate blood flow) with serious complications which underline the importance of distinguishing the progression and prognosis outcome of patients. RVOs can be diagnosed by the multi-modality Swept Source Optical Coherence Tomography Angiography (SS-OCTA) scanner that combines a B-scan and Angio-flow scan. B-Scan offers a cross-sectional view of the retina, allowing a detailed visualization of retinal layers and structural abnormalities, while Angio-flow scans complement with en-face projections of blood flow within different retinal and choroidal layers, enabling the visualization of microvascular and reveling of perfusion abnormalities. With enhanced imaging speed, depth penetration, and resolution, when compared to earlier RVO imaging methods, these advantages make SS-OCTA a valuable tool in detecting microvascular changes associated with retinal vascular diseases, especially for lesions caused by retinal vessel occlusion. Despite the advancements in SS-OCTA, unfortunately, its application to RVO has received little attention compared to other more common retinal diseases. This is likely due to the lower prevalence of the disease in the population, and the complexity of diagnosis that is limited by interpretation accuracy of imaging results.

In this thesis, the primary aim is to implement and analyze traditional machine learning and advanced deep learning methods for their computer-aided decision-making capabilities in RVO disease prognosis classification tasks using SS-OCTA. For the traditional machine learning, a SS-OCTA-based Radiomics pipeline is proposed which includes traditional machine learning for radiomics analysis and advanced ConvNeXt-based deep learning methods on both 2D and 3D image input, regarding the RVO

prognosis classification. In contrast, for the advanced deep learning method, a new 3D-based multi-modality alternating dynamic fusion ConvNeXt (mmDFC) method is proposed to alternatively train each modality for eliminating intermodal negative effects and optimally combine modality-specific information in both the B-scan and Angio-flow scan, as the 2D-based mmDFC is set as comparison. Here, convolutional network has been designed to extract subtle and long-range retinal features, coupled with a dynamic fusion method by leveraging the ConvNeXt architecture and multi-modality data for effective extraction and fusion of Angio-flow-related and B-scan-related structural information, resulting in superior performance compared to the traditional SS-OCTA-based Radiomics.

In addition, a 2D multi-modality Dual-Branch Correlation-driven Fusion ConvNeXt (mmDCFC) network is introduced as an enhanced cross-modal fusion strategy for 2D-based SS-OCTA datasets and also served as an extension of 2D mmDFC for enhancing cross-modal feature fusion. The separate extraction of modality-specific features and cross-modality base features that have been neglected in 2D mmDFC was conducted by a Convolutional-based encoder and Transformer-based encoder, with an overall performance better than the mmDFC. The outcome of this network found that the 2D mmDCFC network has effectively extracted the cross-modal common features at the low-frequency structural level and captured modality-specified features at the high-frequency detailed level. The fusion result of mmDCFC has been evaluated and visualized, in which the improvements in performance proved that this network can enhance the representation of different modalities and levels of features by emphasizing important patterns in different appearances, as the mmDCFC network achieved the best performance among all 2D networks proposed in this study.

Extensive experiments were conducted on SS-OCTA-based RVO datasets including single modality B-scan and Angio-flow datasets, demonstrating the validity and effectiveness of proposed methods for RVO prognosis prediction tasks. To be specific, as for the 2D SS-OCTA input, 2D traditional machine learning methods cooperate with

Radiomics method were firstly performed and compared to 2D ConvNeXt-based deep learning method (without modality fusion), further, these methods were set as comparisons to multi-modality-based 2D mmDFC network in which the mmDFC showed advanced performances. To enhance and evaluate the cross-modal feature fusion ability, the 2D mmDCFC was implemented and compared to 2D ConvNeXt-based deep learning (without modality fusion) as well as 2D mmDFC, in which the results of 2D mmDCFC gained the best performance among all 2D methods and demonstrated the feasibility of cross-modal feature fusion between SS-OCTA datasets. Regarding the 3D SS-OCTA input, 3D traditional machine learning methods integrated with Radiomics were compared to the 3D-based ConvNeXt method (without modality fusion), and also further compared to the 3D mmDFC, in which the 3D mmDFC achieved the best overall performance and highlighted the importance of 3D alternating training and dynamic multi-modality feature fusion methods for RVO prognosis classification task.

Acknowledgment

First and foremost, I would like to express my deepest gratitude to my supervisors, Professor Jinman Kim and A/Professor Lei Bi, for their invaluable guidance, unwavering support, and insightful feedback throughout this research journey. Their expertise and encouragement have been instrumental in shaping this thesis. No matter whether in life or in the process of research encountered difficulties, they always thought of me, provided various resources, tried to help and then encouraged me to move forward with confidence. As a beginner researcher with relatively little experience, they taught me a lot of more efficient and scientific research methods during the progress of my project, pointed out possible improvement directions for me, and made a lot of effort in my paper writing. Their perspectives have enriched my understanding and contributed the most to the development of this work.

I also want to thank my mentor, Dr. Yige Peng, for his dedicated guidance on all the details of my project. His rigor and extensive experience benefited me and greatly improved my development efficiency. It is because of him that my project can smoothly go from the beginning to the output. In addition, I would like to express my sincere thanks to Dr. Yupeng Xu, who is from Shanghai General Hospital, for providing me with useful advice and novel ideas. His expertise in research has motivated me and it's a great honor to work with him.

I would like to further appreciate my colleagues in our lab, for their help and suggestions from the start of my research period. In life, getting along with them makes me feel at ease with happiness. My sincere appreciation extends to the University of Sydney, for providing an intellectually stimulating environment and access to essential resources that facilitated my research.

I would send special thanks to my family and my friends for their unwavering support,

patience, and encouragement throughout this journey. They are my spiritual support, the source of happiness, and the biggest reliance in my life. Without their selfless devotion and love, I would have struggled mightily in times of difficulty. Although the length of life is so limited, it is enough to make me feel extremely satisfied.

Finally, I would like to acknowledge all the participants, collaborators, and individuals who contributed directly or indirectly to this research. Your generosity in sharing your time and insights has been invaluable. This thesis would not have been possible without the collective support of these individuals, to whom I am profoundly grateful.

Xinyu Huo

31 January 2025

List of Publications

The following publications were produced during my MPhil study. Publications marked with * are either directly or have major contribution to this thesis.

Journal Article (submitted)

- Y. Xu*, X. Huo*, et al., "**Advanced SS-OCTA Based Radiomics Machine Learning Pipeline for the Prognosis Classification of Retinal Vessel Occlusion**", submitted to *Journal of Translational Medicine.*

Journal Article (in preparation)

- X Huo et al., "**Foundation model guided dynamic multi-modality fusion for RVO prognosis predictions**", in preparation to submit to *Computer Methods and Programs in Biomedicine.*

Conference Article (Submitted)

- X Huo et al., **"mmDFC: Multi-modality Dynamic Fusion and Modal-decoupled Alternating Optimization for Retinal Vessel Occlusion Prognosis",** submitted to *International Conference on Digital Image Computing: Techniques and Applications (DICTA 2025).*

Contents

List of Figures

List of Tables

List of Abbreviation

Abbreviation: RVO	**Definition:** Retinal vein obstruction
Abbreviation: NPAs	**Definition:** non-perfusion areas
Abbreviation: OCT	**Definition:** Optical Coherence Tomography
Abbreviation: OCTA	**Definition:** Optical Coherence Tomography Angiography
Abbreviation: SS-OCTA	**Definition:** Swept-source OCTA
Abbreviation: mmDFC	**Definition:** multi-modality Alternating Dynamic Fusion ConvNeXt
Abbreviation: mRMR	**Definition:** Minimum Redundancy – Maximum Relevance
Abbreviation: PCA	**Definition:** Principal Component Analysis
Abbreviation: FAZ	**Definition:** multi-modality Dual-Branch Correlation-driven Fusion ConvNeXt
Abbreviation: Anti-VEGF	**Definition:** anti-vascular endothelial growth factor

Chapter 1. Introduction

1.1 Retinal Vessel Occlusion (RVO) Prognosis Outcome Classification

Retinal vein obstruction (RVO) is one of the most common vascular diseases affecting the retina and this can lead to severe visual impairment. RVOs can be classified into branch retinal vein occlusion (BRVO) and central retinal vein occlusion (CRVO) and often results in the development and extension of non-perfusion areas (NPAs), which refers to regions lacking adequate blood flow [1][2]. These NPAs arise due to vein blockage, with increased intraluminal pressure, capillary leakage, and subsequent closure, resulting in hypoxia and triggering pro-angiogenic factors like vascular endothelial growth factor (VEGF) [3].

As a type of severe retinal disease that can lead to a poor prognosis, assessing NPAs is crucial for the management of RVO. Traditional imaging techniques like fluorescein angiography (FA), Indocyanine Green Angiography (IGCA), and Color Fundus Photography (CFP) have been used to depict the NPAs, but all of these methods are invasive with limited retinal depth information provided and therefore are not efficient for the assessment of subtle microvascular structural features in RVO [4]. An evolved vivo imaging method of Optical Coherence Tomography (OCT) enables non-invasive assessment of the retinal structures and was shown to be able to identify some pathological changes, such as cornea thickness, anterior chamber depth, intraretinal fluid, and exudation, which are highly linked with retinal diseases [1][3]. The advent of OCT Angiography (OCTA) in recent years has revolutionized retinal vascular disease imaging by offering a non-invasive method with higher resolution and more efficient scanning speed to detect and quantify NPAs. OCTA allows detailed visualization of the retinal microvasculature and facilitate accurate assessment of the non-perfusion extent,

making OCTA one of the most advanced imaging methods in assessing retinal diseases, especially for microvascular-related diseases including RVO [5]. In addition, it provides improved sensitivity and larger imaging Field of View (FoV) compared to previous OCT and is widely adopted by clinical practices [6]. Swept-source OCTA (SS-OCTA) is an enhanced version of OCTA, which not only promotes and visualizes choroidal microvasculature, but also visualizes vessel-related pathologic feature changes, and thus enabling more acute observation of structural characteristics with better FoVs. The B-scan of SS-OCTA provides structural information of the retina and choroid regarding the cross-sectional direction, as presented in Fig. 1.1(a), and the Angio-flow of SS-OCTA gives high-resolution visualization of blood flow within the retinal and choroidal structures, as depicted in Fig. 1.1(b). New capabilities with SS-OCTA contain the separation of different retinal layers (in volume) to visualize retinal capillary plexuses and the choriocapillaris, including neo-vascularization related to RVO [7][8].

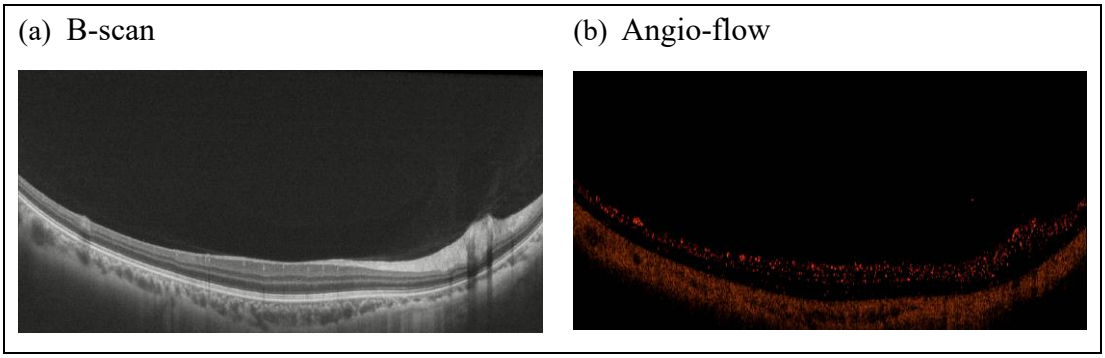

Fig 1.1 Examples of B-scan and Angio-flow slices obtained from SS-OCTA that belonged to one single patient

However, existing research on computer-aided retinal disease diagnosis and prognosis assessment is primarily concerned with other retinal diseases rather than RVO, and there is a lack of RVO-related research, particularly for the automatic analysis of OCTA and SS-OCTA-based data. This may be due to the diagnostic complexity and low prevalence in the population of the RVO [3]. Radiomics method, as a widely adopted computer-aided feature extraction tool in medical image analysis, is not yet investigated in RVO-related tasks. Radiomics is a concept that involves extracting a large number

of quantitative features from medical images that include texture, morphological, first-order, and filter-based features, for use in using machine learning algorithms for disease assessment [9]. Among them, the first-order features quantify basic statistical properties of the grayscale values such as brightness and heterogeneity, and they can reflect general intensity patterns of tissue of lesions. Texture features quantify the spatial arrangement and relationship of intensity values, offering insight into the structural heterogeneity of the tissue. Morphological features describe the geometry and spatial extent of the segmented lesion or anatomical structure. Filter-based features are derived by applying mathematical filters or transformations to the image before feature extraction, to emphasize specific spatial frequencies or patterns [10]. These radiomics features can uncover patterns and characteristics in images that are not easily visible to the naked eye but correlate with the prognosis results of patients. Radiomics features can be directly obtained from computer output for further feature selection and disease assessments, enabling the detection, quantitative description, and is expected to perform continuous monitoring of RVO progression.

Apart from paucity of traditional radiomics for RVO, there is also a lack of advanced machine learning methods for SS-OCTA data. Advanced methods such as Convolution Neural Networks have aided various retinal diseases' diagnosis and prognosis classification tasks but fewer are related to RVO. This is mainly attributed to the challenges in analyzing subtle RVO-related features that were limited by the imaging resolution and depth of ophthalmic scanning equipment, and the diagnostic complexity caused by its lower prevalence in the population [3]. To be specific, compared to some common retinal disease (e.g. Age-related Macular Degeneration (AMD), Diabetic macular edema (DME) and Diabetic Retinopathy (DR)), RVO is relatively rare, making it more difficult to collect large and high-quality datasets. RVO also has diverse subtypes and variable clinical presentations, which complicate accurate labeling, feature extraction, and model generalization. However, despite the applicability of SS-OCTA for RVO, the usage of this advanced scanning method is relatively recent, compared to former widely adopted methods, such as CFP, OCT, and OCTA, further

hinders the collection of large datasets. Existing methods mainly focused on using traditional OCT [11], as well as processing on 2D radiomics features [12][13] (features gained from processing 2D slice images) extracted from 3D volumetric images (stacks of 2D slices) from the scanner and therefore ignoring the subtle deep-layer and spatial information within the imaging slices. Deep learning has enabled end-to-end learning of semantic image features from large datasets, thereby reducing reliance on manual feature extraction. Xu et al. [14] developed a ResNet–based cross-attention model to classify retinal vein occlusion (RVO) subtypes from fundus images. Yim et al. [15] applied a 3D U-Net for tissue segmentation, followed by a CNN to predict exudative age-related macular degeneration (AMD) from OCT volumes. Holste et al. [16] introduced a longitudinal transformer to forecast open-angle glaucoma progression from sequential CFP images. Although effective, these models were designed for single-modality input based on data collected by scanners with relatively low resolution and limited depth. Thus, they cannot leverage the complementary information offered by multi-modal SS-OCTA, limiting their applicability in more complex diagnostic settings.

In this thesis, a new SS-OCTA Radiomics machine learning pipeline for RVO prognosis classification is proposed, by including traditional radiomics and advanced deep learning ConvNeXt method [17]. The depthwise convolutional neural networks (CNN) were leveraged in the ConvNeXt block for its ability to process spatial information efficiently to capture microvascular and structural abnormalities in SS-OCTA. Further, larger kernel sizes have been adopted in ConvNeXt structure to improve its receptive fields for the ability to obtain subtle contextual and spatial features. These optimizations make this network effective in analyzing detailed and structural information that is implicit in the retinal layers. The ConvNeXt structure also incorporates the concept of attention mechanism for the optimized extraction of deep features, by implementing normalization techniques and simplified architectural optimizations. In the proposed pipeline, by taking advantage of both ConvNeXt structure and Radiomics based on ‘late ‘fusion integration in the fully connected layer within deep model-extracted features

and radiomics features, the 3D-based Radiomics feature fusion model achieved the best overall performance among all the comparison methods.

The variety of available retinal imaging modalities underscores the need for efficient combination strategies to improve the extraction of disease-related features. As examples, in the research field of retinal disease prognosis classification, prior research has found that multiple modalities on Diabetic Retinopathy datasets that integrated imaging methods including CFP and OCT can improve CNN-based network performance of prognosis classification [18]. Multi-modality methods are also feasible for the fusion of OCT images and visual field-related features in retinal disease classification [19] for the diagnosis of Glaucoma based on Transformer network method. As a result, current research mainly focused on using traditional CNN structures and failed to utilize advanced SS-OCTA and perform on novel deep learning fusion structures. In addition, the multi-modality feature-fusion methods that are suitable for RVO datasets have not been specially customized, so that current methods have paid little attention to efficiently utilize the diverse progression-related information in two modalities within SS-OCTA, especially for RVO with complex prognosis features that are hard to analyze.

Traditional multi-modality fusion methods, such as early or late fusion, often fail to address differences in information entropy between modalities, in which the modality that contains less large-scale structural information such as Angio-flow may be hindered. The imbalance of information entropy can lead to one modality dominating [20][21] during training while others under-optimized, hence reducing the model performance. To overcome these challenges, a 3D-based multi-modality alternating dynamic fusion ConvNeXt (mmDFC) method is proposed. This method leverages the Alternating adaptive learning and Dynamic fusion method [21] designed to optimize single-modality individual training processes and perform dynamic fusion to enhance fusion weights between modalities during inference stages while balancing the contribution weights of each modality. Based on these methods, the mmDFC network

reduced the negative interaction between modalities that was caused by dominance of B-scan, and optimized the fusion weights based on classification uncertainty, which further enhanced the network's performance. Alternating adaptive training phase on single modality enabled independent optimization while preserving cross-modal interactions and gradient status via shared headers that were constructed by fully connected layers and gradient modification module. Then gradient modification mechanism ensured shared headers retain learning information by orthogonalizing the gradient directions and saving the gradient information in new gradient matrices, preventing the gradient forgetting across modalities. In the inference stage, the uncertainty-based dynamic fusion process adjusted fusion weights in each iteration to balance the contributions from all modalities for predictions. Further, the mmDFC network was modified as 3D version, in order to reduce the interference from non-lesion slices (that have no obvious signs related to RVO progression) in patient data with bad prognosis (NPA progression), since after data shuffling, these slices may cause high uncertainty to the training process, as they failed to reflect the lesion related to NPA progression correctly. Similarly, this 3D model can also mitigate the misjudgment caused by slices with lesions (in which some RVO progression signs occurred) that belong to patients with good prognosis (no NPA progression) in the training stage. Thus, the 3D-based mmDFC method achieves superior fusion and overall results compared to 2D-based mmDFC methods, and sum-fusion input (refers to setting B-scan and Angio-flow modalities as two channels in single image) ConvNeXt-based deep learning baseline. It also gained enhancement from the alternating training stage and dynamic fusion inference stage with better fusion performance compared to sum-fusion-input ConvNeXt baselines.

However, 2D/3D mmDFC network failed to explicitly decompose and emphasize the cross-modal information, and this may cause a loss of modality-shared features, as the commonly presented morphology changes between B-scan and Angio-flow was not effectively represented by the network. The merits provided by the common features within both modalities can benefit the feature extraction performance of a fusion model

and can be used to solve the challenges in modeling desirable cross-modality information to maintain modality-specific features. The inner correlations between different modalities generated by both modalities that reflected their correlations in structure and vascular characteristics can be categorized into ‘modality-shared base’ features and their special morphology patterns can be included in ‘modality-specified detail’ features, in which base features tend to represent common long-range global characteristics, and detail features contain more specific and detailed local information. In this part, a 2D mmDCFC method is also proposed for RVO prognosis classification. This method is optimized by a correlation-driven dual-branch two-stage self-supervised fusion method [23] for its ability of decoupling and extracting long-range structural features and detailed local features, originally for image reconstruction. In 2D mmDCFC network, the base features inferred the structural interactions between the B-scan and the Angio-flow scans, while the detail features, especially in Angio-flow images, were used to represent subtle vessel changes highly linked with RVO progressions. According to the result conducted on single modality 2D SS-OCTA datasets including B-scan and Angio-flow, the 2D mmDCFC network was able to perform better than the comparison baselines including single modality performance, sum-fusion dataset input performance, and multi-modality performance given by 2D mmDFC. This cross-modal fusion method demonstrates the usability and importance of cross-modal feature extraction in the RVO prognosis classification task and provides inspiration for further enhancement on 3D multi-modality datasets, because directly converting the mmDFC network to 3D will incur excessively high computational costs.

1.2 Contributions

This study aims to explore and optimize the application of machine learning including advanced deep learning architectures in predicting NPA extension for RVO patients, on classifying the outcomes into good (no NPA progression) and poor (will have NPA progression) categories. SS-OCTA-based B-scan and Angio-flow modalities are

adopted for multi-modality learning, with advanced fusion strategies that can better leverage the cross-modal and modality-specified prognosis-related information performed. This thesis makes several key contributions:

- By leveraging the power of Radiomics and machine learning, a new prognosis classification model was developed that can aid in the early identification of high-risk patients of RVO. The results demonstrated that the combination of ConvNeXt-based deep learning model and traditional Radiomics features was effective for RVO prognosis classification, while 3D-based method achieved better performance when compared to 2D-based counterparts.

- A new fusion method multi-modality alternating dynamic fusion ConvNeXt (mmDFC) based on 3D input was proposed by alternating the training between the single-modalities to minimize interference and preserves cross-modal interactions. An uncertainty-based dynamic fusion mechanism was used to adaptively adjust weights of modalities for better fusion performance. In addition, the comparison of both 2D mmDFC and 3D mmDFC was performed, with 3D mmDFC achieving better fusion results.

- The 2D multi-modality Dual-Branch Correlation-driven Fusion ConvNeXt (mmDCFC) method for RVO prognosis classification was further proposed, as an extension of 2D mmDFC, to improve the extraction of modality-shared structural features (via Transformer encoders) and modality-specific detailed features (via CNN encoders). This outcome is an indication that the optimized interaction between B-scan and Angio-flow scan, regarding the common base features and specific detailed features, is effective for the RVO prognosis classification and reflects the importance of cross-modal feature fusion in this task, compared to the ConvNeXt-based 2D baselines on single modalities and sum-fusion dataset, as well as the 2D mmDFC. Visualization of the fusion results for mmDCFC also highlighted its ability to emphasize important

structural and detailed features.

- Comprehensive experiments were conducted on six RVO SS-OCTA datasets, including 2D B-scan and Angio-flow single modality datasets, 2D sum-fusion dataset, 3D B-scan and Angio-flow single modality datasets, as well as 3D sum-fusion dataset. The evaluation included mmDFC and mmDCFC methods, where were benchmarked to ConvNeXt-based deep learning network.

1.3 Thesis organizations

The remaining chapter of this thesis is organized as follows: in Chapter 2, related works about the assessment of RVO and multi-modality learning regarding retinal disease are presented. In Chapter 3, an SS-OCTA-based Radiomics machine learning pipeline is described, in which several machine learning-based Radiomics and advanced ConvNeXt-based networks on 2D and 3D SS-OCTA datasets are compared. Chapter 4 introduces the mmDFC network based on ConvNeXt-tiny structure with alternating adaptive learning and dynamic fusion mechanisms on 2D and 3D SS-OCTA single modality B-scan and Angio-flow datasets. Then, Chapter 4 presents the mmDCFC that integrates the correlation-driven dual-branch fusion method and ConvNeXt-tiny-based structure for better cross-modality feature extraction, on 2D and 3D SS-OCTA single modality datasets. Finally, Chapter 6 draws overall conclusions for this thesis and presents the future work regarding possible optimizations.

Chapter 2. Related Works

For RVO assessment tasks, key biomarkers that clinically unveiled structural features including foveal avascular zone (FAZ) area, vessel density (VD), and fractal dimension (FD) in both the superficial vessel plexus (SVP) and deep vessel plexus (DVP) have shown potential correlations with the NPAs [24], as they may be highly linked with the severity of RVO and future NPA progressions. To be specific, VD refers to the proportion of a given area occupied by blood vessels, typically expressed as a percentage or ratio. In OCTA and SS-OCTA, VD is calculated by measuring the total length or area of perfused vasculature within a defined region of interest (ROI), divided by the total area of that region [25]. SVP comprises a vasculature located in the inner retina, mainly the ganglion cell layer and nerve fiber layer, while DVP lies deeper, primarily within the inner nuclear layer. VD in SVP and DVP is used to quantify the degree of perfusion and vascular integrity [26]. Changes in vessel density in these layers are associated with retinal vascular diseases including RVO. FD is a quantitative metric used to describe the complexity and self-similarity of a vascular network. It reflects how completely a vascular pattern fills a 2D or 3D space. FD is typically computed using box-counting or skeletonization techniques applied to binarized vessel maps, it provides insight into geometrical organization and branching patterns of microvasculature. It serves as a biomarker of vascular health, with lower FD suggests reduced branching, vascular pruning, or ischemia [8]. However, multiple challenges persist in the clinical management of NPA in RVO, since there is a paucity of research that is directly about RVO progression assessment and longitudinal research evidence to validate their predictive role in the expansion of NPAs, hindering their precise and reproducible utilization. research based on advanced automatic methods, especially when using SS-OCTA data. Literature review of the field in RVO progression assessment resulted in no published work on measuring Radiomics machine learning pipeline.

Current OCTA-based research in the use of advanced machine learning to automate the assessment of RVO mainly focus on structural segmentation and vessel quantification. In one of early deep learning works, U-Net network has been adopted as the backbone for FAZ segmentation task based on a public OCTA dataset [27]. As shown in Fig. 2.1, the black hole in the center of the en-face image of the OCTA slice depicted by Fig. 2.1(a) represented the location of the FAZ, while the blue area shown in Fig. 2.1(b) represented the segmented FAZ area processed by the U-Net. Another U-Net-based model using OCT dataset that incorporates spatial constraints to optimize the segmentation of FAZ, ensuring that the detected fovea is located in biologically plausible regions, adopted AMD, DME, and RVO slices as one input dataset together with their spatial constraints to derive FAZ for each disease [28], and has improved the segmentation performance of FAZ. Similarly, VD and perfusion density that represent the degree of ischemia in RVO, were analyzed by the distance-threshold vector method that was calculated by the skeletonized image of retinal vessels based on distance-threshold vector method using OCTA [29]. These biomarker-based research demonstrated the importance of clinical features for auxiliary RVO progression assessment, and thus, the biomarkers including FAZ, VD, FD within the SVP and DVP were also adopted in our work to assess machine learning pipeline as comparison. However, current research failed to reflect extra retinal features that may imply disease-related characteristics apart from FAZ, VD, FD within the SVP and DVP, and there is limited number of machine learning methods that were conducted to investigate the performance of various biomarkers. This research highly focused on segmentation and quantification of important retinal structures without assessing the biomarkers' potential for disease outcome prediction, which showed limited concern on identifying their prognosis ability regarding RVO assessment, and neglected other implicit features that are difficult to notice by the naked eye.

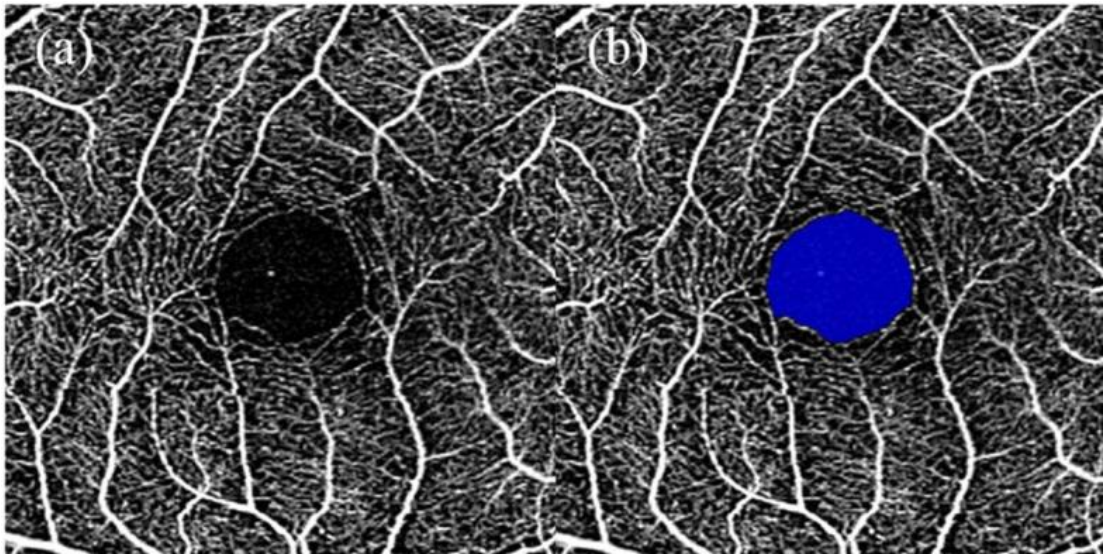

Fig 2.1 Automatic OCTA-based FAZ segmentation procedure [27]

2.1 Automatic Retinal Disease Assessment

At present, machine learning methods are recognized as having the potential to solve challenges in automatic diagnosis of ischemic RVO. Among all, however, most of the related research adopts backward scanning methods with narrow FoV, simple deep network structures, and/or trained and validated on small datasets.

Nagasato et al. [30] employed the VGG-16, a fundamental CNN model, for the detection of NPA and classification of different ischemic RVO types (Central RVO, Branch RVO) against normal eyes based on OCTA. However, this research used simple CNN architecture; it is expected that advanced deep learning architecture will improve performance, in particular with complicated prognosis outcomes. Ren et al. [31] introduced a modified CNN model for the classification of different types of RVO (Central RVO, Branch RVO) against normal eyes, utilizing more traditional CFP images. Visualization strategy of Grad-CAMs was used for interpretation of the network's region of interests. However, the CFP data may potentially be influenced by its ability in identifying RVO disease outcomes when compared to OCTA. A CNN-based Transfer learning network was developed for classifying Choroidal

neovascularization (CNV), Diabetic macular edema (DME), Multiple drusen, Age-related macular degeneration (AMD) and normal eyes, using OCT images (which only contain structural information) [32]; Similarly, another research built upon a modified CNN backbone utilizing Atrous Convolution and Residual Convolution block for the classification of AMD, DME, and CNV [33], evaluated on 2 OCT datasets. It can be seen from the literature that there is a scarcity of research on both the diagnostic and prognostic outcome of RVO, and existing studies rely on simple CNN architecture that are relatively common in retinal disease research, especially for RVO, as it can better extract detailed local structural features. However, there is also a lack of microvascular flow imaging modality in OCTA when only adopting CFP and OCT datasets, which may highly affect the accuracies in classifying NPA extension progressions.

2.2 Radiomics on Retinal Diseases

Radiomics refers to using advanced mathematical analysis methods for the extraction of quantitative imaging features from medical images, which can be utilized for machine learning models to improve diagnostic, prognostic, and classification performances of various diseases. Existing radiomics methods for the prediction of NPA extension mainly focus on using traditional retinal imaging methods and 2D radiomics features. Conventional machine learning techniques have been evaluated for the diagnostic accuracy of radiomic features extracted from OCT and OCTA images for diagnosing Diabetes Mellitus (DM) and Diabetic Retinopathy (DR) [12], suggested that radiomics combined with OCT/OCTA imaging can effectively aid in the diagnosis of diabetic retinal conditions. Meng et al. [13] introduced Radiomics method based on machine learning algorithms for the prognosis assessment of DME, performed on 2D OCT imaging. Current research indicated that Radiomics can accurately give assessment results regarding retinal disease for personalized treatment classification and pathology evaluation.

The integrated use of 2D and 3D radiomics features with advanced machine learning methods enables multi-source feature fusion from both radiomics side and deep feature side, is expected to enhance the predictability of NPA extension and improve our understanding for the prognosis of NPA in RVO. Despite its potential, there is no advanced fusion method that combined both radiomics and machine learning networks. Kar et al. [11] presents a 3D-based Radiomics methods on OCT for the ME (as secondary to DR and RVO) treatment outcome classification, in which ResNet-50 was adopted as comparison and proved the advancement of pure Radiomics method. However, this research only utilized classic 3D CNN-based deep learning method and assessed outcomes on OCT images, which cannot reflect subtle microvascular information related to NPA extensions that can be better depicted by OCTA.

2.3 Multi-modality Classification for RVO

Traditional machine learning algorithms have been introduced for the Branch RVO visual prognosis classification after anti-VEGF injection based on OCT [5]. By manually selecting and extracting key retinal structures on central fovea, Logistic Regression analysis was adopted for the prognosis prediction regarding to vision acuity. However, this research only utilized several medically defined features for decision making without leveraging the extra information implied in OCT images, especially features related to blood flow, disease-correlated structural characteristics, and other implicit features.

Based on the literature review conducted for this thesis, no relevant research about the prognosis based on Deep Learning related to the SS-OCTA RVO dataset has been found. There was however related research about other retinal diseases performed on OCT, CFP, OCTA, in which some research utilized multi-modality input performed on conventional imaging methods and deep learning networks. A deep multimodal fusion approach combining OCTA en-face flow images regarding different layers for enhanced

DR severity classification has been introduced [34], in which several traditional CNN structures were adopted as backbones and dense layers were implemented as fusion modules. The integration of structural and flow slice information as two modalities improved classification performance compared to single modalities, showing the value of multimodal imaging in retinal disease assessment. Vaghefi et al. [35] demonstrate that superior diagnostic accuracy can be achieved when Inception-ResNet is combined with multimodal images (OCT, OCTA, and CFP) analysis for AMD, while features were concatenated at the pooling layer. Hao et al. [36] proposed a novel deep learning framework, Eye-AD, to detect early-onset of Alzheimer's disease and mild cognitive impairment using retinal OCTA images which contained different layers of Angio-flow images, highlighting the potential of retinal imaging in neurological disease diagnosis. This study implemented CNN-based extractors and derived different scales of feature maps. The importance-based reweight mechanism was used to guide the importance-aware GNN in optimizing the feature maps and giving disease classification results. However, this study hasn’t performed conventional machine learning methods with clinical biomarkers as comparisons, and the SS-OCTA structural-based B-scans were not utilized for fusion, which may lead to a missing of meaningful retinal features.

Chapter 3. SS-OCTA Based Radiomics Machine Learning Pipeline

3.1 Introduction

In this chapter, the SS-OCTA based Radiomics pipeline is introduced, in which the Radiomics features for RVO prediction are processed by several feature selection and machine learning methods. The Radiomics features derived from traditional machine learning method were integrated with advanced ConvNeXt [17] deep learning-based feature extractor. The ConvNeXt architecture, inspired by the Transformer network [37] has higher efficiency and scalability with its ability to recognize global patterns in contrast to conventional CNNs. The proposed machine learning pipeline was evaluated on both 2D and 3D RVO SS-OCTA datasets, and experimental results indicated that the 3D ConvNeXt-based Radiomic fusion network and pure 3D ConvNeXt (without radiomics feature fusion) were more effective compared to 2D Radiomics-based machine learning methods and clinical-biomarkers-based machine learning method (where the clinically defined biomarkers' measurement values were utilized as input). The proposed SS-OCTA-based Radiomics machine learning pipeline has the following contributions:

- The traditional machine learning and advanced machine learning methods (deep learning) proposed in this pipeline adopted SS-OCTA-based scanning image that can exhibit superior resolution and imaging depth in which the retinal structural and microvascular information can be better captured when compared to conventional imaging methods, as SS-OCTA is considered as the most effective method to assess NPA progression in RVO. The result of this pipeline underscores the potential of advanced machine learning methods for complex retinal disease like RVO, and the significant improvement in classification performance on 3D-based models suggests that the volumetric data analysis in future diagnosis should

be prioritized.

- The ConvNeXt architecture was adopted to construct 2D and 3D versions of the Radiomics network and performed on OCTA datasets. Two types of fusion were applied: (i) ‘early-sum fusion’ used B-scan and Angio-flow modalities as the two input to the network and, (ii) ‘late fusion’ adopted the features from the input early-sum fusion dataset that were derived by the ConvNeXt-based model and then fused them with another different multi-modality dataset at the fully-connected layer: the radiomics features that contained various first-order, texture, shape and filter features. The late fusion of these multi-modal features based on ConvNeXt structure and Radiomics as evidenced by its highest AUC among all methods, highlights the importance of comprehensive data fusion techniques in RVO image analysis.

- Clinically defined biomarker features that have potential correlations with the NPAs were classified by conventional machine learning algorithms. These biomarkers include foveal avascular zone (FAZ) area, vessel density (VD), and fractal dimension (FD) in both the superficial vessel plexus (SVP) and deep vessel plexus (DVP). Experiments demonstrated that the best 3D machine learning method performed comparably to the clinically defined biomarker features but required less time cost on collecting data, while the best 2D and 3D deep learning method based on ConvNeXt achieved better overall performance compared to the clinically defined biomarkers, inferring the usefulness and effectiveness of advanced machine learning methods in RVO progression assessment.

3.2 Methods and Materials

3.2.1 Overview of the Framework

The framework shown in Fig. 3.1 shows the overall implementation procedure for the SS-OCTA-based Radiomics machine learning pipeline. The clinical concerns were first formulated: the lack of OCTA progression-related research evidence to validate the clinical biomarker's predictive role and the usability of radiomics and machine learning methods in the expansion of ischemic areas, setting obstacles to their further utilization and availability in prognostic assessment. It also indicates the necessity to investigate these methods as well as improve the efficiency of clinical-based biomarkers, Radiomics, and machine learning methods-extracted features' utilizations in RVO prognostic classification. As depicted in Fig 3.1 (a), the screening of clinical data, collection of RVO patients' SS-OCTA images and clinically defined biomarker data were conducted, with total 87 patients' data included and randomly divided into the proportion of 0.68: 0.32 as training and test dataset. Then, the flood fill algorithm was further implemented on 2D and 3D datasets to process the automatic Regions of Interest (ROI) segmentation as described in Fig. 3.1 (b), in order to restrict the valid size of the original SS-OCTA image to the area within the presence of retinal structures and minimize the impact of noise in the background area. By utilizing these images with ROIs, radiomics method can be conducted to gather shape, texture, filter-based and first-order features, then the feature selection strategies can help to further reduce the number of features by identifying the most relevant features while eliminating irrelevant or redundant features. In Fig 3.1 (c), SS-OCTA images were set as input to the deep learning model as an example, and the model-extracted feature maps were concatenated with corresponding patient-based top 50 radiomic features for each slice or volume. Results were compared to conventional machine learning methods regarding the NPA progression outcomes, which proved that the advanced 3D-based deep learning method gained the best efficiency among all methods in this study. For the model evaluation part in Fig 3.1 (d), our models were evaluated based on several

commonly used metrics, including Area Under the Receiver Operation Characteristic Curve (AUC), Accuracy, Sensitivity, Specificity, and F1 score. Additionally, Precision-Recall (P-R) Curves, Receiver Operation Characteristic Curves (ROC), and Decision Curves are presented as additional visualization results. Grad Cam [38] method was also implemented to visualize the areas of concern of the model in the decision procedure by generating heat maps for better machine learning method interpretability.

In this study, patients' data was separated to training and test dataset by the ratio of 0.68: 0.32. Class 0 refers to the stable non-perfusion group, while Class 1 means NPA group with lesion extension, each dataset has kept the class proportion of around 0.70: 0.30. (b) presented the Regions of interest (ROI) segmentation for denoising was conducted on SS-OCTA datasets, and the feature selection part was processed by Radiomics-based feature selection methods, with the bar chart showed top 50 radiomics features that have been utilized in this pipeline. (c) indicated the ConvNeXt-based and machine learning-based model implementation procedures combined with radiomics features. (d) showed an example of result evaluation for the best model and the 2D visualization implemented by Grad Cam heatmap algorithm [38].

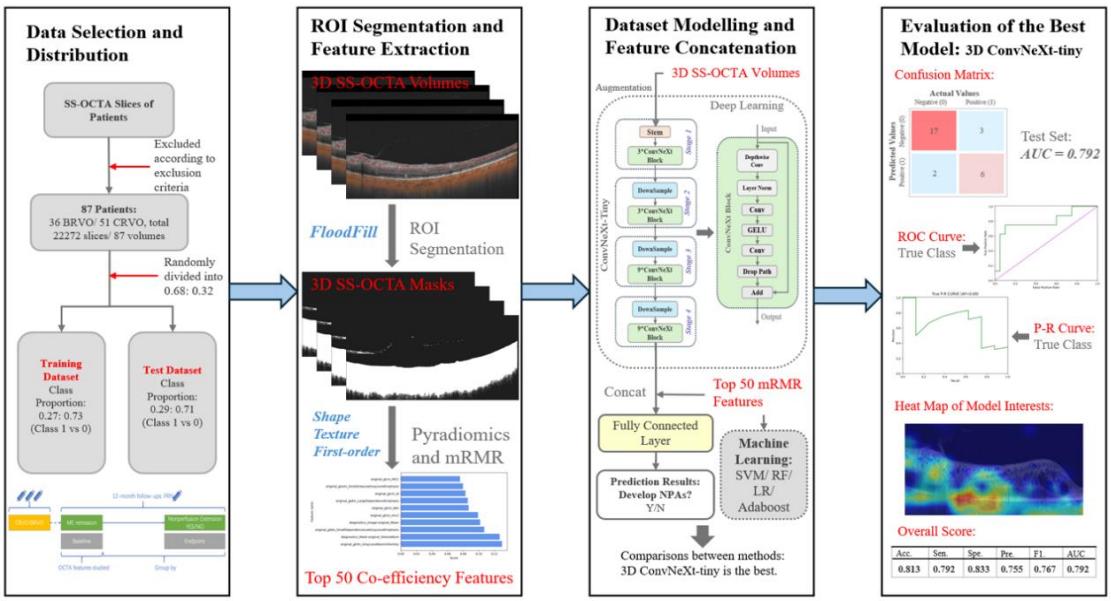

Fig 3.1 The overall framework of SS-OCTA-based Radiomics machine learning pipeline

3.2.2 Materials

This study utilized multiple SS-OCTA RVO datasets for the Radiomics machine learning pipeline to evaluate proposed methods. Both 2D and 3D image datasets were acquired from wide-field SS-OCTA machine (VG200D, SVision). A total of 87 patients were enrolled in this study; 36 patients had CRVO and 51 had BRVO. There were 63 (72.41%) and 24 (27.59%) patients with or without extension of non-perfusion during the follow-ups, respectively. No statistically significant differences were observed between the extended non-perfusion group (Class 1) or the stable non-perfusion group (Class 0) in gender, age, and eye laterality (OD/OS). The extension of non-perfusion was more common in BRVO than CRVO ($p = 0.031$) and the injection number of CVRO was bigger than BRVO in the 6-month follow-up. This study selected only 1 eye per patient (1 eye/1 patient) and the whole dataset included 87 patients ($n = 87$ eyes). The demographics of enrolled patients are shown in Table 3.1.

Table 3. 1 Demographics of enrolled patients

Characteristics	Level	Overall	Stable	Extended	p
n		87	63	24	
Age (mean (SD))		65.17 (14.19)	64.83 (15.53)	66.08 (10.09)	0.714
RVO Duration (mean (SD))		9.91 (4.62)	10.06 (4.73)	9.50 (4.39)	0.614
IOP (mean (SD))		14.84 (2.93)	14.81 (2.74)	14.92 (3.43)	0.875
DM (%)	0	87 (100.00)	63 (100.00)	24 (100.00)	NA
Hypertension (%)	0	28 (32.18)	21 (33.33)	7 (29.17)	0.908
Hyperlipidemia (%)	1	59 (67.82)	42 (66.67)	17 (70.83)	
	0	74 (85.06)	54 (85.71)	20 (83.33)	1
BCVA_new1 (mean (SD))	1	13 (14.94)	9 (14.29)	4 (16.67)	
		0.60 (0.38)	0.57 (0.36)	0.68 (0.44)	0.224
Disease (%)	BRVO	51 (58.62)	32 (50.79)	19 (79.17)	0.031
	CRVO	36 (41.38)	31 (49.21)	5 (20.83)	
Gender (%)	Female	47 (54.02)	33 (52.38)	14 (58.33)	0.797
	Male	40 (45.98)	30 (47.62)	10 (41.67)	

The partition strategies of the datasets are as follows: 28 patients (~29%) out of the whole dataset were randomly assigned to be the external independent test dataset, while 59 patients (~71%) were randomly assigned as the training dataset. For machine learning, a stratified four-fold cross-validation is performed based on the training dataset, where each fold contained randomly selected samples with equal distribution among the classes. For the deep learning part, a random four-fold cross-validation is used as the training dataset. For 2D datasets, the total number of images was 22,272, in which 256 continuous slices were acquired from each patient's eye scanning data. Regarding the 3D datasets, the overall number of volumes was 87, with each volume consisting of 256 slices (the same slices used in the 2D dataset). The patient allocation

in the training and test datasets was kept the same as the 2D dataset among all methods. Single B-scan and Angio-flow modality datasets with each of 87 volumes were also made and utilized in this study.

Fig 3.2 showed a set of samples of the sum-fusion dataset which contains two scan types (B-scan and Angio-flow) as two channels. Fig. 3.2(a) presents a slice from a Class 0 patient, while Fig. 3.2 (b) shows a Class 1 patient with NPA extension progression. The patterns depicted as red color that occurred in both (a) and (b) refer to Angio-flow information which were gained by the scanning of microvascular structures performed by SS-OCTA, and the color was set originally by the scanning equipment. The gray-scale structures in each sample were obtained by the B-scan of SS-OCTA and can reflect the morphology changes within the retina. The B-scan modality is similar to the conventional OCT but with higher resolution and imaging depth.

(a) Class 0 Patient Slice Sample (b) Class 1 Patient Slice Sample

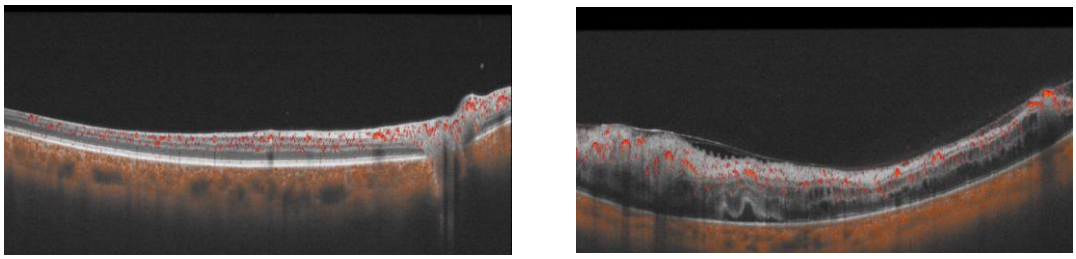

Fig 3. 2 Samples of early sum-fusion dataset

3.2.3 Radiomics and Machine Learning-based Methods

The standard procedure for Radiomics features extraction and deduction for machine learning-based classification tasks are as follows:

A. Region of Interest Segmentation

In this study, both 2D slices the of SS-OCTA dataset and 3D volumes of the SS-OCTA

dataset were adopted for both Region of Interest (ROI) segmentation and machine learning-based Radiomics feature extraction and selection. The Radiomics method was applied to both datasets in order to acquire high-throughput quantitative features, allowing machine learning methods to better represent the pathology characterization of the RVO. Regarding the preprocessing and ROI selection of the 2D dataset, first, the Flood Fill algorithm was adopted to extract ROIs, to acquire initial image masks for subsequent feature extraction. Second, the same strategies of crop and resize were implemented to make all the images including masks and original images the same size of 256×256 pixels. Then, the masks and image dataset were preprocessed. As for the 3D dataset part, the 2D slices first went through the Flood Fill algorithm to gain their masks. Secondly, both 2D masks and original images were cropped and resized to 256 ×256×256 pixels while maintaining dataset size consistency.

B. Radiomics Feature Extraction

Using the above preprocessed images and masks, a radiomics method for automatic feature extraction based on Pyradiomics [39] was performed. The extracted features for both 2D and 3D datasets consisted of shape information, texture information and first-order features, in which the Gray-level run length matrix (GLRLM), Gray-level co-occurrence matrix (GLCM), Gray-level size zone matrix (GLSZM), Gray-level dependence matrix (GLDM), Neighborhood gray-tone difference matrix (NGTDM) were included in the texture information category. In addition, voxel-based spatial information was captured by radiomics especially for the 3D dataset, which caused some small differences in feature types. As a result, these radiomics features served as a combination of quantitative information which provided a comprehensive representative for both 2D and 3D datasets, and they are suitable for machine learning methods to analyze and classify the different outcomes for each patient.

C. Radiomics Feature Selection

Before feeding the features into machine learning models, the radiomics features have gone through an excess procedure to reduce the dimensions. By using feature selection algorithms including mRMR (Minimum Redundancy – Maximum Relevance) [40], Lasso and PCA (Principal Component Analysis) for dimension reduction, the amount of redundant data was substantially reduced, to minimize the effect of noise data and the instability caused by correlation between features, while preventing a high possibility of overfitting. Among all, the mRMR algorithm enables the best selection of the smallest relevant feature subset for the appointed prediction task, with a balance of efficiency and performance, as well as result explainability. In addition, MID (Mutual Information Difference) and MIQ (Mutual Information Quotient) are two mostly adopted mRMR schemes for the combination of relevance and redundancy. After evaluating the above feature selection methods and combination schemes, mRMR (MID) gained the best overall performance on several optimized classifiers on both SS-OCTA-based 2D and 3D RVO datasets, thus is the most efficient method for feature selection in this work. After filtering by feature selection methods, these retinal features were standardized by z-score normalization to make the data conform to the standard normal distribution (mean value equals 0 and the standard deviation is 1). Then, the normalized SS-OCTA-based radiomics features were set as input for both the machine learning models and deep learning models.

D. Subset Evaluation

Several clinically defined biomarkers that consist of FAZ, fractal dimension features and superficial, deep vessel density features which are important for the medical practice in the evaluation of RVO progression, were tested based on the prediction performance. Multiple value combinations of these biomarkers are filtered by the best feature selection method, then set as input for the best machine learning classifier that is derived from Radiomics experiments, to identify whether these biomarkers

performed better than other set of features, according to metrics including Accuracy, Sensitivity, Specificity, Precision, F1-score and AUC. The results showed that when all clinical features for feature selection and model prediction, the performance was improved when compared with the classifications of single types of biomarkers. Using these biomarkers for identifying whether the non-perfusion area extended after anti-VEGF medications is more effective with better performance compared with 2D radiomics machine learning methods.

For these clinically defined biomarkers, a total of 117 features were selected including FAZ, fractal dimension features and superficial, deep vessel density features. All the combinations of features or single features have undergone grid search for SVM parameter selection, while the mRMR selection method has been performed for all features to compare with none-selection classification results, as well as single features. To be more specific, according to Table 2 below, the combination of FAZ + Superficial + Deep + Fractal Dimension features has been tested by selecting 50 features via the mRMR method with an accuracy of 0.65, which is prominently higher than any other feature types. The results showed that when utilizing all clinical features and doing feature selection for SVM classifications, the classification results were improved when compared with the classifications of single types of biomarkers.

3.2.4 Radiomics and Deep Learning-based Methods

For the prediction task performed by deep learning models, this study utilized ConvNeXt-tiny as baselines. ConvNeXt is an advanced convolutional neural network (CNN) model with simple concepts of design but high efficiency and scalability. The depthwise convolutions within ConvNeXt block and larger kernel size of 7 has improved receptive fields that enables obtaining subtle contextual and spatially features in images and volumes, makes it earned well performance on analyzing detailed and structural information. In order to select the best ConvNeXt type with a suitable volume

of hyper-parameters for a better performance on RVO datasets mentioned in Chapter 3.2.2, ConvNeXt-base model, small model, and ConvNeXt-tiny model were evaluated to search for the most efficient model. The ConvNeXt-tiny outperformed with the best accuracy of 0.728 and AUC of 0.645 and was selected as the 2D and 3D backbone of Radiomics deep learning method. The performance of larger ConvNeXt models were restricted by the size of dataset due to their high capacity with large volume of parameters, however, ConvNeXt-Tiny is less prone to overfitting problems on our dataset with fewer parameters and performed better without extensive regularization techniques. Detailed evaluation results are presented in Table 3.3.

The overall structure of ConvNeXt-tiny is shown in Fig 3.3. To fit the input of the 3D dataset, the ConvNeXt-tiny model was modified into a 3D version model to adapt the data format. While keeping the original kernel size but upgrading the convolution dimension into 3D, the 3D version of ConvNeXt was also tested on the external dataset and achieved the highest performance among all models in this study. In addition, with 3D extracted radiomics features gained from pyradiomics and mRMR, the deep features gained by the 3D ConvNeXt-tiny can be fused with these radiomics features before going through the fully connected layers and finally obtain the prediction results that enable a more comprehensive analysis of different feature types.

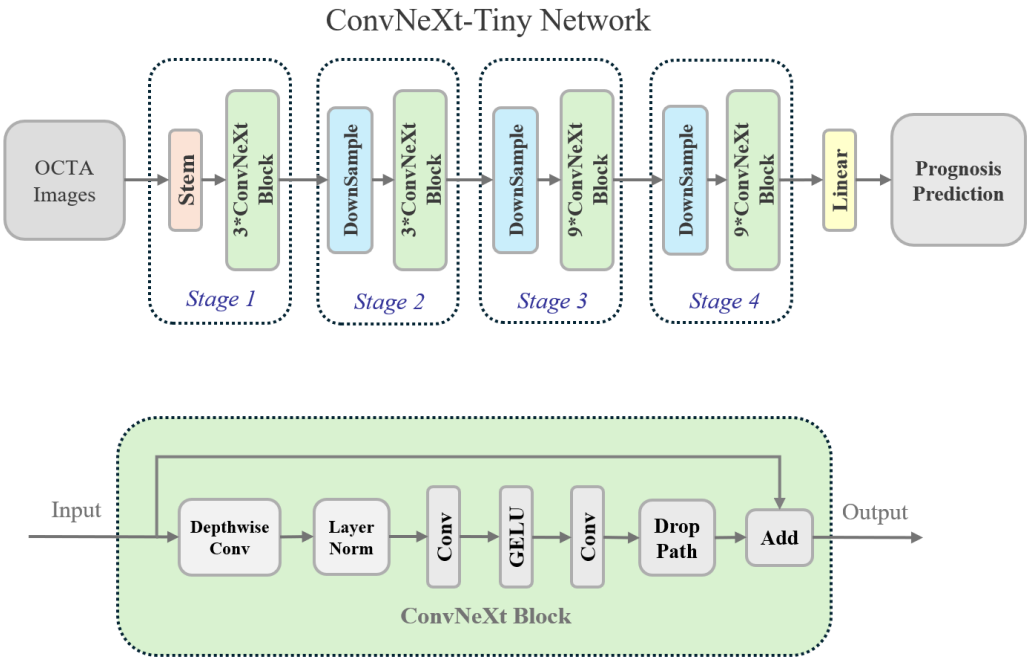

Fig 3. 3 Structure of ConvNeXt-tiny

3.2.5 Training of Machine Learning methods

This study includes both machine learning methods and deep learning methods for SS-OCTA-radiomics features. During the 2D data preprocess stage, images were cropped and resized to 256×256 pixels. Flood fill algorithm was then applied to extract the region of interest (ROI) from all the slices, and then the masks for these datasets were acquired for feature extraction. Afterwards, regarding the machine learning part, the Pyradiomics feature extraction procedure gained 107 2D Radiomics features for each patient's slices. These 2D features come from first-order features, Grey Level Cooccurrence Matrix (GLCM), Grey-Level Run Length Matrix (GLRLM), Grey-Level Size Zone Matrix (GLSZM), and Neighboring Grey Tone Difference Matrix (NGTDM). Additionally, the 3D dataset was preprocessed by stacking 2D slices that have been processed by ROI selection and resizing volumes to 256×256×256 pixels. Then, the volumes with ROI masks were fed into Pyradiomics to extract radiomics features. Except for the first-order features, Grey Level Cooccurrence Matrix (GLCM), Grey-Level Run Length Matrix (GLRLM), Grey-Level Size Zone Matrix (GLSZM), and

Neighboring Grey Tone Difference Matrix (NGTDM) that are commonly adopted in 2D radiomics, extra features in particular for 3D-based shape features were obtained with a total feature number of 118.

Then, utilize Principal Component Analysis (PCA), Lasso, and mRMR for selection of 50 main features regarding the machine learning methods. By performing Grid Search pipeline for selecting and training the best set of parameters, SVM, Logistic Regression (LR), and Adaboost classifiers were evaluated, as shown in Table 3. 2, with a stratified four-fold cross-validation performed based on the training dataset. The evaluation results were derived from the best parameters on the common external test set. 2D features were extracted for each 256 slices of the patients, but for the 3D dataset, volume-based features were acquired for per patient. 3D feature selection on Lasso, mRMR and PCA were implemented and further classified by Support Vector Machine (SVM), Adaboost and LR (Logistic Regression) as well. By training the input radiomics features data and evaluating external test dataset features on each 2D and 3D model, as well as selecting the best combination of parameters, the most efficient model was chosen as representative for each type of classifier or model.

To be more specific, for the SVM model, the kernel was poly-type with the degree set to 2, while the penalty factor was 50. The LR model has an L2 penalty with the penalty factor C equal to 1.0. Then, the estimator for the Adaboost model was set to a decision tree classifier, with the max depth being 8, the min samples leaf was 5, the number of estimators was 50, and the learning rate was set to 0.8. Each of these machine learning models has gone through the stratified four-fold cross-validation on a training dataset, and then evaluated on an external test dataset, and the final fitted parameters for each classifier are shown in Table 3. 2. To equally analyze the model performance, metrics that consist of Accuracy, Sensitivity, Specificity, Precision, F1-score and AUC were calculated on the train and test dataset. In addition, ROC Curves and Confusion matrices were plotted for the visualizations of machine-learning evaluation results.

Table 3.2 Best combination of parameters for machine learning models

Feature Selection + Classifier	Parameters
mRMR + SVM	C=50, degree=2, kernel='poly'
mRMR + LR	Penalty='L2', C=1.0
mRMR + Adaboost	Estimator=DecisionTreeClassifier, max_depth=8, min_samples_leaf=5, n_estimators=50 learning_rate=0.8
Lasso + SVM	C = 50, kernel='linear'
Lasso+ LR	Penalty='L2', C=0.1
Lasso + Adaboost	Estimator=DecisionTreeClassifier, max_depth=8, min_samples_leaf=3, n_estimators=50 learning_rate=0.8
PCA + SVM	C = 0.1, kernel='rbf', gamma = 0.1
PCA + LR	Penalty='L2', C = 1.0
PCA + Adaboost	Estimator=DecisionTreeClassifier, max_depth=8, min_samples_leaf=5, n_estimators=50 learning_rate=0.6

As for the ConvNeXt-tiny that was selected as backbone of Radiomics and deep learning methods, it consists of 3, 3, 6, 3 times of stacking depth-wise ConvNeXt blocks in each convolutional layers (while ConvNeXt-base and ConvNeXt-small have larger numbers of input channels with 3, 3, 27, 3 times of stacking ConvNeXt blocks in each layer, totally 4 layers), connected with Average Poling, Layer Normalization, and Linear Layer, in which the model size is the most suitable for the SS-OCTA RVO dataset. Regarding the ConvNeXt block, the kernel size was set to 7×7 in 2D model, GELU was selected for activation function, and additional Layer Scale and Drop Path were utilized to improve training stability. In addition, the ConvNeXt-tiny model was also modified into a 3D version to fit the 3D RVO data input, while keeping the original kernel size but upgrading the convolution dimension into 3D (7×7×7). As for the fusion

of Radiomics, with 3D extracted radiomics features gained from pyradiomics and mRMR, the late fusion of deep features gained by the 3D ConvNeXt-tiny-based network and radiomics features can be performed by concatenation and fed into the fully connected layers, then the final prediction results can be obtained. The training stage was performed on the ConvNeXt-tiny backbone, and both 2D and 3D dataset involved in the experiments. The P-R Curves of the best machine learning model and deep learning model were also visualized, as well as the ROC curves for these models. In addition, according to clinical practice, Decision Curves were performed to show the net benefits caused by the 2D and 3D deep learning models. For further evaluations on analyzing clinically defined biomarkers, and BRVO/CRVO subgroup model performance, the above metrics are also applicable, and the best feature selection method and machine learning classifier were tested.

Table 3.3 Evaluations on ConvNeXt-base, ConvNeXt-small and ConvNeXt-tiny models

Model	**Acc.**	**Sen.**	**Spe.**	**Pre.**	**F1.**	**AUC**
ConvNeXt-base	0.710	0.517	0.539	0.623	0.461	0.612
ConvNeXt-small	0.712	0.522	0.550	0.638	**0.675**	0.611
ConvNeXt-tiny	**0.728**	**0.573**	**0.621**	**0.675**	0.565	**0.645**

3.3 Experiments and Results

3.3.1 Evaluation Metrics

The evaluation metrics in this research are Accuracy (Acc), Sensitivity (Sen), Specificity (Spe), Precision (Pre), F1-score (F1) and Area Under the ROC Curve (AUC), defined as following:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3.\ 1)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (3.\ 2)$$

$$Specificity = \frac{TN}{TN+FP} \quad (3.\ 3)$$

$$Precision = \frac{TP}{TP+FP} \quad (3.4)$$

$$F1\ score = \frac{2TP}{2TP+FP+FN} \quad (3.5)$$

We also calculated the AUCs for each method and visualized the best AUC plots on 2D and 3D datasets. A higher AUC score indicates a better ability for classification in our research. The Precision-recall (PR) curve is also used for comparisons between the best 2D and 3D methods as it is commonly adopted for classification task [13].

3.3.2 Results

In this research, both 2D and 3D datasets with sum fusion of single modalities including B-scan and Angio-flow are evaluated. As a result, the mRMR selection method with Mutual Information Quotient (MID) evaluation criteria, returned a minimum set of the most useful features with high stableness on 2D sum-fusion dataset, compared to Lasso and PCA, and achieved the best performance when classified by SVM. The mRMR integrated with SVM gained the best machine learning performance for the 3D RVO SS-OCTA dataset, with the highest AUC of 61% and Acc of 61%. Fig 3.4 (a) shows the Top 10 features with their co-efficient scores selected by mRMR based on the 3D dataset, with the 3D dataset achieving the best AUC. These results indicated the usefulness of mRMR feature selection method, and the effectiveness of SVM with poly kernel that can better modeling non-linear relations between radiomics features.

Fig 3.4 (a) showed visualization of the co-efficient features distribution of top 10 radiomics features processed by mRMR. Fig 3.4 (b) depicted the violin Plots on external test dataset for the 2D and 3D ConvNeXt-tiny model with prediction confidence score distributions by performing the comparison of BRVO and CRVO types. It shows that the 3D ConvNeXt has received better overall scores on BRVO and CRVO with significantly larger proportion of high-confidence model testing samples. (c, d) shows the Decision Curves for the best 2D and 3D prediction models on an

external test dataset.

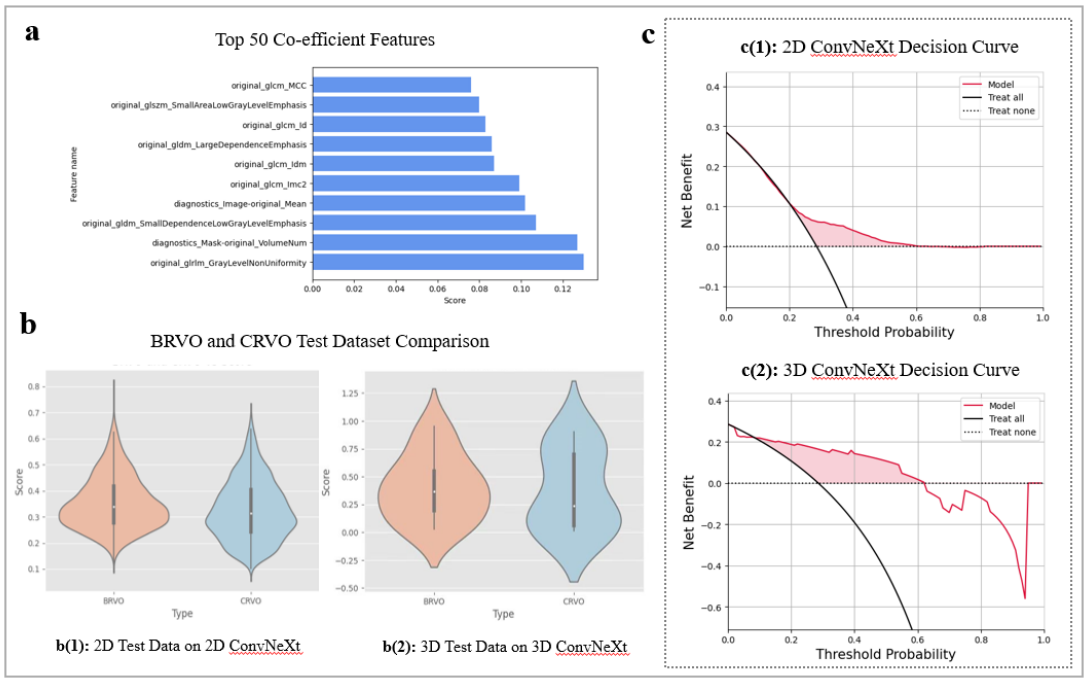

Fig 3.4 2D and 3D ConvNeXt-based network decision curves, dataset comparisons and sorted co-efficient radiomics features

Additionally, evaluations on clinically defined biomarker features performed on the above deduction methods and classifiers were completed. Single types of biomarkers and combinations of biomarkers were tested to find the best set of features and also undergone the comparisons between Radiomics and machine learning methods. In total, 117 clinical features were selected including FAZ, fractal dimension features, and superficial, deep vessel density features. All the combinations of features or single features have conducted Grid Search for SVM parameter selection, while the mRMR selection method has been performed for all features to compare with non-selection classification results, as well as single features. To be more specific, according to Table 3.4 below, the combination of FAZ + Superficial + Deep + Fractal Dimension features has been tested by selecting 50 features via the mRMR method with an accuracy of 0.65, which is notably higher than any other feature types. As a result, when utilizing all clinical features for SVM classifications, the results were improved when compared with the classifications of single types of biomarkers. Using these biomarkers for

identifying whether the non-perfusion area extended after anti-VEGF medications is more effective with a better performance compared with 2D radiomics machine learning methods as they also contain high medical diagnostic significance. On the contrary, adopting a 3D volume dataset directly for machine learning might be a more efficient way, with less feature collection time cost and better performance for an improvement of 3% on AUC.

Table 3.4 Classification results for clinical biomarkers. The best results were highlighted.

Evaluation Method	Acc.	Sen.	Spe.	Pre.	F1.	AUC.
FAZ	0.590	0.460	0.650	0.330	0.370	0.550
Superficial	0.590	0.450	0.640	0.340	0.380	0.550
Deep	0.640	0.310	**0.780**	0.370	0.320	0.540
Fractal Dimention	0.460	**0.830**	0.310	0.320	**0.460**	0.570
FAZ + Super + Deep + Fractal Dimension	0.520	0.500	0.540	0.330	0.400	0.520
FAZ+Super+Deep+Fractal+mRMR	**0.650**	0.400	0.760	**0.390**	0.390	**0.580**

As for the ConvNeXt-tiny model for RVO prognosis classification, preprocessed 2D SS-OCTA data served as input for training with a random four-fold cross-validation strategy and then evaluating on external test dataset which is in parallel to the Radiomics pipeline part. The 2D machine learning methods are compared in Table 3.5. 3D SS-OCTA dataset on 3D ConvNeXt-tiny is evaluated as comparisons. In addition, the 2D ConvNeXt-tiny network was fused with 2D slice-based or 3D patient-based Radiomics mRMR feature selection data before fully connecting layers. The 3D ConvNeXt-tiny further tested the fusion performance with 3D patient-based Radiomics data as well. According to the results shown in Table 3.6, 3D-based ConvNeXt achieved the best accuracy and AUC on both tests. Still, when fused with 3D patient-based Radiomics features with 3D ConvNeXt-tiny features, it achieved the most advanced results among all the models mentioned in this research, with the accuracy of 81.3% and AUC of 79.2%.

Decision Curve, as a visualization tool used to evaluate the clinical usefulness of deep learning models, provides insight into how well a model performs in terms of decision-making by balancing the benefits (true positives) against the potential harms (false positives). It is particularly useful in medical evaluations where decision-making involves trade-offs between risks and benefits. As for the net benefits of 2D and 3D ConvNeXt-based models revealed by the Decision Curve shown in Fig. 3.4 c (1~2), in both 2D and 3D ConvNeXt-tiny-based models, the performances are superior to the "Treat all" strategy mostly across the range of the threshold. Compared with the "Treat none" strategy, each ConvNeXt-tiny-based model is better across a threshold probability less than 0.6 for the 2D decision curve and 0.62 for the 3D decision curve, which means the model is beneficial between the thresholds within the red zone. Overall, the red zone beneath the "Treat none" and restricted by the "Treat all" and 3D ConvNeXt-tiny-based Radiomics model benefits curve is significantly larger than the red zone of 2D ConvNeXt-tiny-based curve, which indicates 3D ConvNeXt-tiny-based Radiomics model can produce a higher net benefit over a given range of thresholds (e.g. 0.2 ~ 0.6) is generally preferable, with a better ability of balancing the benefits and risks of assessment more effectively.

Table 3.5 Comparisons among machine learning methods and deep learning methods on 2D OCTA dataset. The best model was highlighted with the best AUC score.

Model	Acc.	Sen.	Spe.	Pre.	F1.	AUC
2D mRMR + SVM	0.560	0.670	0.610	0.450	0.530	0.580
2D mRMR + LR	0.540	0.430	0.590	0.340	0.330	0.520
2D mRMR + RF	0.610	0.220	0.760	0.310	0.230	0.510
2D mRMR + Adaboost	0.570	0.520	0.680	0.400	0.390	0.580
2D ConvNeXt-tiny	0.728	**0.573**	0.621	0.675	**0.565**	**0.645**
2D ConvNeXt-tiny + 3D mRMR Features	**0.730**	0.552	**0.967**	**0.681**	0.530	0.552

Table 3.6 Comparisons among machine learning methods and deep learning methods on 3D OCTA dataset. The best results and model were highlighted.

Model	Acc.	Sen.	Spe.	Pre.	F1.	AUC.
3D mRMR + SVM	0.610	0.620	0.600	0.400	0.480	0.610
3D mRMR + LR	0.540	0.430	0.590	0.340	0.330	0.520
3D mRMR + RF	0.660	0.170	0.880	0.330	0.220	0.520
3D mRMR + Adaboost	0.630	0.540	0.670	0.390	0.450	0.600
3D ConvNeXt-tiny	0.750	0.778	0.722	0.714	0.719	0.779
3D ConvNeXt-tiny + 3D mRMR Features	**0.821**	**0.792**	**0.833**	**0.755**	**0.768**	**0.794**

Compared with conventional machine learning methods, ConvNeXt-based network gained better scores in most of the metrics apart from sensitivity and specificity on 2D dataset, showing that the deep learning approach may have better feature extraction and prediction ability on 2D data for 11.8% improvement on Accuracy compared to the best machine learning method, and improved 6.5% on AUC. Regarding the 3D ConvNeXt-tiny, it can be inferred that 3D data for RVO may be more effective for prognosis assessment, resulting in a 2.2% better classification performance on accuracy, and 14.9% on AUC. Further, late fusion method for different features between deep learning model-derived features and Radiomics features helped the model with an enhanced performance of 7.1% on Acc and 1.5% on AUC, especially on the 3D dataset.

3.3.3 Visualization of Results

Regarding the ROC, P-R Curves of both 2 classes, and confusion matrices for machine learning models, visualization plots are shown in Fig 3. 5. The ROCs of both 2D and 3D machine learning models based on SVM and mRMR (the best of 2D methods and 3D methods) are included, while the SVM plus mRMR on 3D datasets gained the best performance with a higher AUC of 0.610. The P-R curves of both 2 classes also indicate this evaluation results as 3D data achieved higher APs. Fig 3.5 a (1) ~ (4) showed the Confusion Matrix, ROC and class-based P-R curves regarding the best 2D machine

learning model: Radiomics-based mRMR features classified by SVM, in which the AUC of this model is 0.58 and the Average Precision (AP) values (refers to the area under the P-R curve). Fig 3.5 b (1) ~ (4) showed the visualization performance of the best 3D model: 3D-based mRMR features classified by SVM. The higher AUC and AP value indicates the better NPA extension predicting ability of this method.

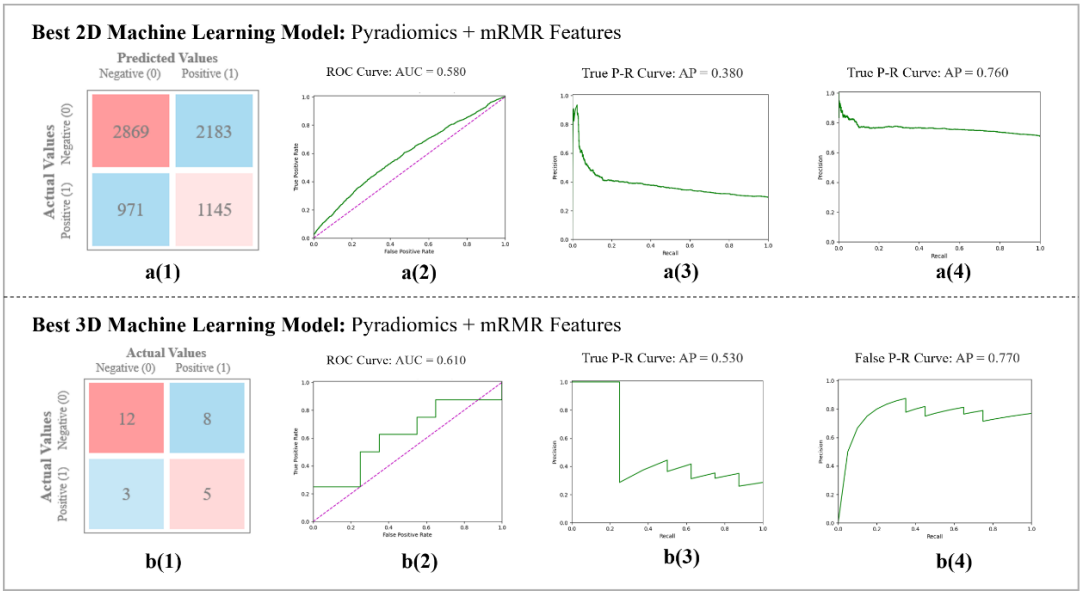

Fig 3. 5 ROC and P-R curves for best 2D and 3D machine learning models

For Comparisons between 2D ConvNeXt-tiny-based and 3D ConvNeXt-tiny-based network, ROC and P-R Curves are shown below in Fig 3.6, as well as the Confusion matrices of these deep learning models. Fig 3.6 a (1) ~ (4) depicted the visualization results for the best overall 2D model: ConvNeXt-tiny. Fig 3.6 b (1) ~ (4) showed the visualization results for the best model in this chapter: 3D ConvNeXt-tiny concatenated with Radiomics-based 3D mRMR features, with significantly higher AUC and APs, it can successfully identify the data with higher proportion of True Negative and True Positive according to the confusion matrices, inferring better accuracy on predicting NPA extension.

The comparison shows that the 3D ConvNeXt-tiny has superior performance on both ROC and P-R Curve, while the ConvNeXt-tiny 2D gained an overall better accuracy than all the 2D machine learning methods, and surpassed the 3D machine learning methods on the most important metrics: accuracy and AUC. Thus, it can be inferred that 3D ConvNeXt-tiny with mRMR features calculated by Radiomics is the most effective method to predict the progression of RVO patients among all the tested methods. The confusion metrics shown in Fig 3.6 a(1) and b(1) are derived from the test set evaluation results from both deep learning networks.

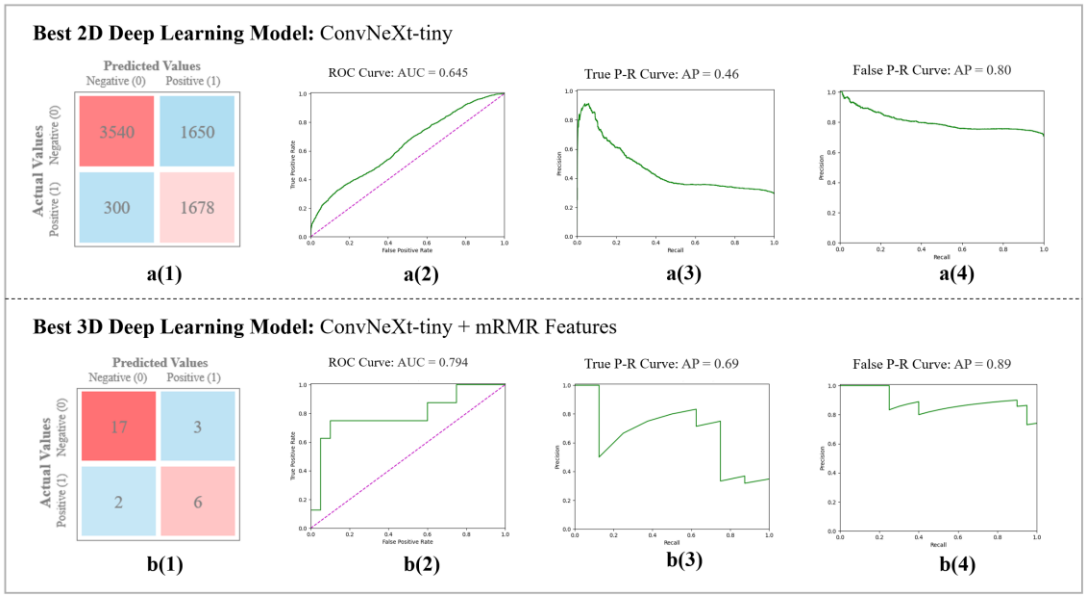

Fig 3. 6 P-R and ROC curves for best 2D and 3D deep learning model

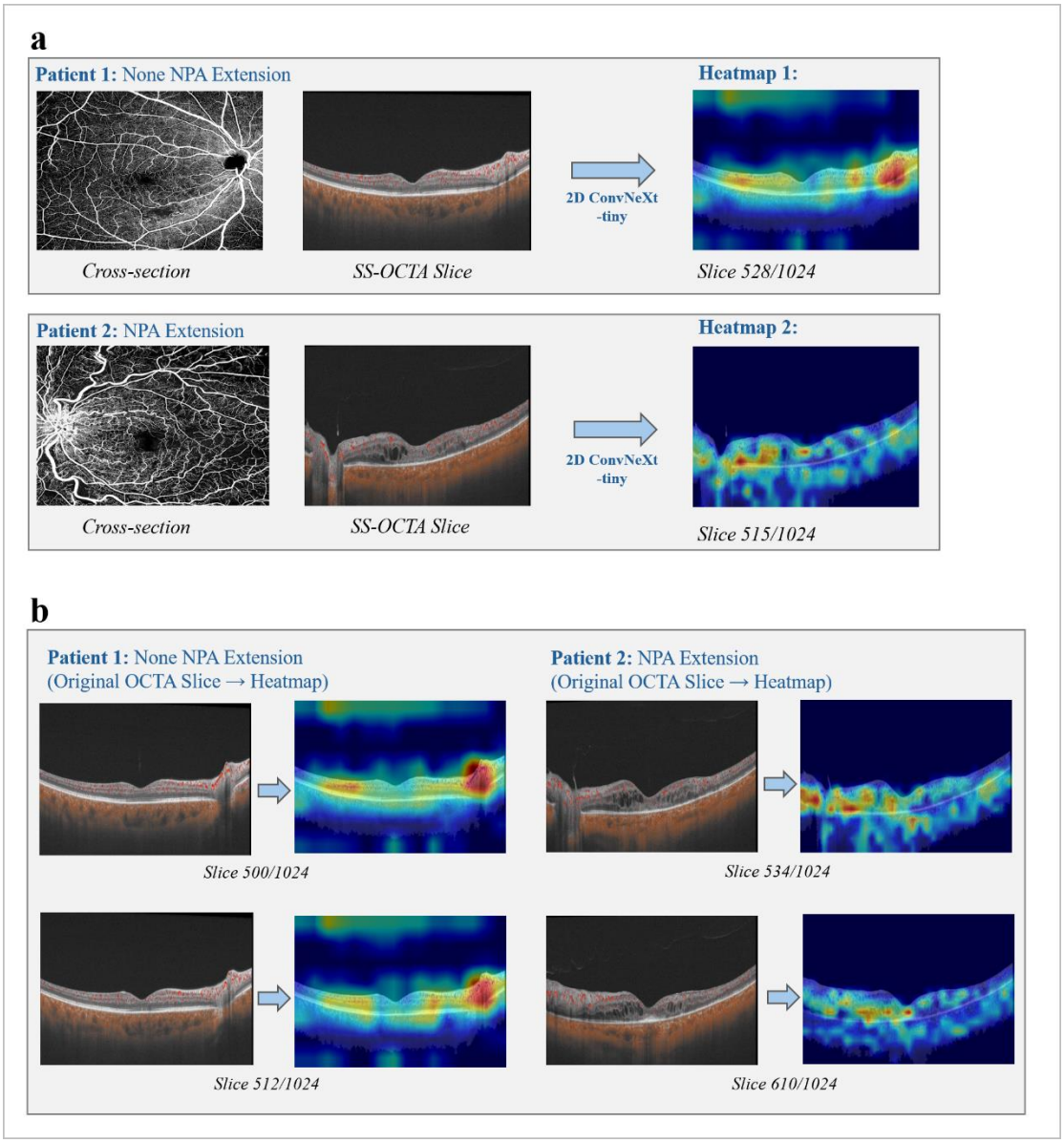

Fig 3.7 Heatmaps processed by the best 2D model**:** ConvNeXt-tiny-based model. (a) shows the comparisons between cross-section Angio-flow-based SS-OCTA images and original SS-OCTA slices in this study, and the heatmaps were generated by SS-OCTA slices in both types of patient samples. (b) presented SS-OCTA sum-fusion data slices and its heatmaps for each type of patient with relatively high degree of confidences.

3.4 Discussions

The proposed SS-OCTA-based machine learning pipeline showed promising capability of ConvNeXt-tiny-based network on 2D and 3D RVO sum-fusion dataset and the advantages of utilizing machine learning for Radiomics analysis regarding the RVO prognosis classification. In general, 3D ConvNeXt-tiny with 3D dataset gained superior performances on all the metrics compared to 3D Radiomics-based mRMR and SVM, in which when fused with Radiomics-derived features in the last linear layer, the performances reached the best and surpassed the pure 3D ConvNeXt in AUCs. The ability of 3D ConvNeXt-based network emphasized the benefits of extra spatial information in prognosis classification, as this information is lost in 2D slices. Moreover, the Radiomics-derived features offered complementary strengths to the 3D ConvNeXt. Radiomics involves handcrafted, quantitative descriptors, such as shape, intensity, and texture, derived from medical images using domain knowledge. These features are interpretable and clinically valuable, especially where explainability is critical. In contrast, ConvNeXt automatically extracted hierarchical deep learning-based features, capturing complex patterns that are often missed by traditional methods. Thus, as the improvement on 3D ConvNeXt fused with Radiomics shows, combining Radiomic and deep features leverages both interpretability and the high representational power of deep models. As for the 2D-based ConvNeXt-tiny, it exceeded the performance of best 2D machine learning methods, indicating the effectiveness of deep learning models, particularly for the complex disease conditions related to RVO. In addition, the 2D slice-based Radiomics features were not adopted in evaluations in Table 3.5, due to the ineffective performance that has lower Acc and AUC (Acc: 0.729, AUC: 0.550), when fused with the 2D ConvNeXt-tiny. In 2D-based dataset, the risk of label-noise mismatch may affect the outcomes, for instance, slices with no lesion may appear in patients that were labeled as bad case, this will cause confusion to the model and bring noise to the training part. In contrast, using patient-level 3D features ensures the radiomics input is aligned with the label, reducing noise in the learning process. Moreover, for machine learning, a stratified four-fold cross-validation is performed

based on the training dataset and ensures fair evaluation between methods in scenarios as machine learning models are sensitive to class imbalances. For the deep learning part, a random four-fold cross-validation is used as the training dataset. The robustness and scalability of deep learning indicate that it is resistant to slightly imbalanced datasets (e.g. in this study, the proportion of class 0 and 1 is about 7:3 and the dataset size is over 22k), and label smooth cross entropy method for loss calculation was utilized in deep learning models to reduce the side effects caused by label imbalance.

Various OCTA-based parameters that mainly focus on measuring Angio-flow-related clinically defined biomarkers, such as the foveal avascular zone (FAZ), vessel density (VD), and fractal dimension (FD) of the superficial vascular plexus (SVP) and deep vascular plexus (DVP), have been developed to quantitatively monitor retinal ischemia and non-perfusion areas. However, in this study, the predictive performance of these parameters was relatively poor. This could be attributed to the fact that the NPA extensions were relatively small (one disc area) and that the BRVO and CRVO patients in this study were all classified as non-ischemic. Anatomic analyses revealed that more venous components were detected in the DVP, while more arterial components were present in the SVP, underscoring the important role of the DVP in RVO pathologies. VD from the DVP demonstrated higher accuracy than the SVP, indicating that changes occur early in the venous components of RVO. Additionally, the branching complexity of the capillary network, as measured by FD, showed significantly higher sensitivity in detecting the extension of non-perfusion areas, suggesting that these capillary changes occur early before non-perfusion area extension. The performance of all the clinically defined biomarkers exceeded the 2D Radiomics machine learning method in accuracy, suggesting its advancement in depicting progression-related features. However, the superior performance gained by 3D Radiomics machine learning method indicated the greater importance of representing spatial features in RVO progression.

Heatmaps can help to identify whether the high response areas are concentrated in the most distinctive parts clinically. We choose the final convolution layer in the

ConvNeXt-tiny for one representative in each class (NPA extension or non-extension), and the zones with high-temperature colors inferred its importance for the model to make decisions. Compared with the cross-section Angio-flow images for each example of slice heatmap, the voids in the vascular vein (e.g. the middle position in the cross-section slice in Fig 3.7 (a)) and the location of the FAZ (e.g. the right position in the cross-section slice in Fig 3.7 (a)) correspond to the middle and right structural regions of the SS-OCTA image, and both of these areas have highlighted regions. These areas which contain blood flow and morphological information, are highly associated with the progression of NPA, and this suggests that the model successfully recognized clinically distinctive areas linked with RVO progression.

3.5 Summary

In this chapter, both machine learning and advanced deep learning methods were adopted in this Radiomics machine learning pipeline to predict the extension of NPA in patients with retinal vein occlusion (RVO) after anti-VEGF treatment. The results demonstrated that deep learning methods can effectively improve the prediction performance compared to traditional radiomics or clinical biomarkers when combined with machine learning methods. In addition, the results of both deep learning and machine learning methods support the conclusion that utilization of 3D datasets can achieve better prediction performance on 3D-based models in contrast to 2D datasets on 2D models. Particularly, adopting Radiomics method with feature selection algorithm for the fusion with deep learning method can further magnify the performance gap, and offers a synergistic approach that leverages the strengths of both methodologies to enhance the analysis and interpretation of RVO images. As an advanced non-invasive imaging method, SS-OCTA integrated with Radiomics and deep learning, can be of considered as an auxiliary for assessing retinal pathology and predicting individualized RVO progression.

Chapter 4. Multi-modality Alternating Dynamic Fusion ConvNeXt Network for RVO Prognosis Classification

4.1 Introduction

In this chapter, a multi-modality alternating dynamic fusion ConvNeXt (mmDFC) network is introduced in which the ConvNeXt structure is leveraged to separately extract B-scan only and Angio-flow only disease related features. The multi-modality alternating dynamic fusion method [21] aims to reduce the interfere between different modalities, allowing each modality to be optimized effectively without modality-dominance issues caused by the modality (e.g. B-scan) that contains richer information than another modality (e.g. Angio-flow). Compared to 2D SS-OCTA single modality performance and 2D SS-OCTA modality-fusion data performance for the RVO prognosis classification task, the mmDFC network reached the best overall results, indicating the usefulness of this method. Notably, when utilized the 3D version of mmDFC network on 3D SS-OCTA dataset, the performance of this network further enhanced, which proved the effectiveness of 3D-based mmDFC method. The main contributions of this study are as follows:

- By alternating the training stages with gradient preserving method and enhancing modality reweight strategy based on uncertainty, the 2D mmDFC network outperformed the single-modality B-scan and Angio-flow ConvNeXt-based baselines, as well as the sum-fusion dataset input ConvNeXt-based baseline, thus demonstrating the fusion effectiveness of alternating dynamic fusion method.

- The proposed method demonstrated the feasibility of using 3D data in 3D mmDFC network for improving classification performance. The revised 3D ConvNeXt-

based encoder enabled the alternating dynamic fusion method to make use of the implicit spatial information in 3D RVO datasets that are not available in 2D datasets. Compared to the 3D ConvNeXt-based baseline on sum-fusion dataset, this method leveraged advanced fusion mechanism by optimizing single modalities separately, rather than simply carrying out early sum fusion and neglecting the suboptimization of Angio-flow modality.

4.2 Methods and Materials

4.2.1 Overview of the Framework

A. The structure of mmDFC network

The multimodality learning strategy based on alternating adaptation learning and dynamic fusion forms the main section of the mmDFC. The training stage of mmDFC regarding iterations that implemented alternating unimodal adaptation method is presented in Fig 4.1. In every iteration, a batch of Angio-flow images was set as model input, and first sent to the image encoder 1 (based on ConvNeXt-tiny) to get modality-based deep features. Meanwhile, the prediction loss was calculated by the shared head as a fully connected layer. The loss was also transferred to the gradient modification block for weight correction matrix transformation, and the gradients for this modality can be preserved. Next, the B-scan images were fed into the network and went through encoder 2 in the next iteration and were processed individually to get the loss and preserved in the gradient modification block. Fig 4.1 also showed the structure of ConvNeXt-tiny that was set as the encoders for deep feature extraction. ConvNeXt-tiny was selected and implemented for its effectiveness in classifying the prognosis outcomes on SS-OCTA dataset as demonstrated in Chapter 3.2.2. In addition, the B-scan modality was input into a ConvNeXt-tiny-based encoder with larger kernel size of 7, while the Angio-flow modality was fed into a ConvNeXt-tiny-based encoder with a

kernel size of 3. Due to the modality difference, the B-scan images tend to represent layer structural information and RVO complications such as edema, these characteristics often correlate with microvascular information that appeared on Angio-flow but with more obvious occurrences, thus are suitable for bigger kernel size with larger receptive fields in larger scale for a better representation ability of global features. On the contrary, Angio-flow images fit smaller kernel size because the vessel information related to NPAs represents as red points and are in a much smaller scale, thus, more effective for ConvNeXt-tiny with smaller kernel size and receptive field to capture detailed features in Angio-flow images.

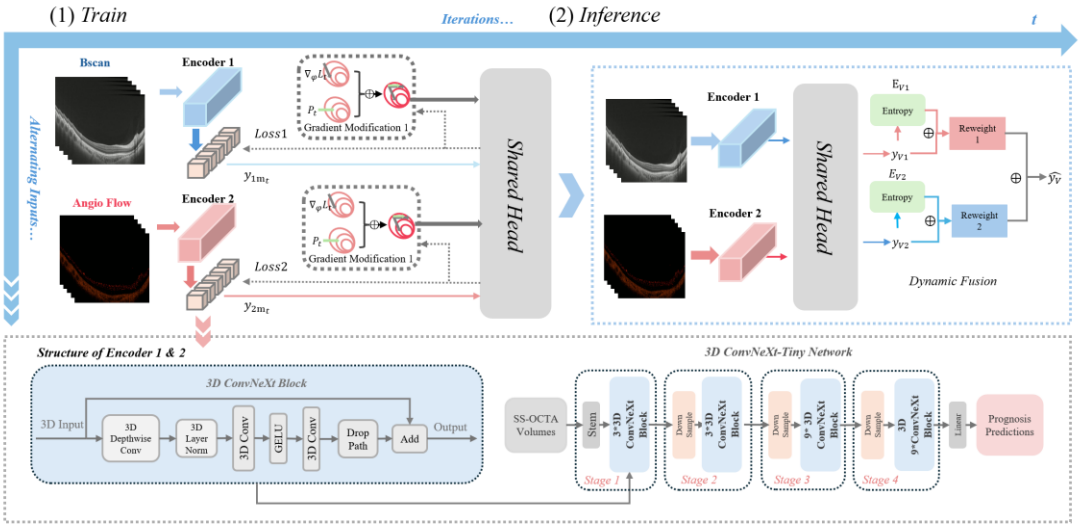

Fig 4. 1 Overall framework of the mmDFC training stage

Next, the inference stage of the mmDFC network is also depicted in Fig 4.1. By using the encoders in the training stage, the Angio-flow and B-scan images were processed separately to get the share head output y_V. Then the uncertainty value E_V was gained based on the softmax output regarding the network's classification result. Reweight 1 and 2 refers to the importance of weights for each modality and is multiplied with the output y_V. Reweighted features for the B-scan and Angio-flow modalities were added together to obtain the final dynamic fusion output $\widehat{yv}$. By calculating the fusion loss and prediction output of $\widehat{yv}$, the performance of the Alternating adaptive learning and dynamic fusion mechanism with ConvNeXt backbone can be obtained.

Overall, the training phase enables the encoders to alternately learn each mode while maintaining cross-modality information through shared heads. A gradient correction mechanism is introduced to prevent shared heads from forgetting previously learned modal information. At the inference phase, the dynamic fusion of multimodal information by assessing the uncertainty of each modal prediction is utilized to optimize the fusion weights of each modality. Thus, the three key stages in this mechanism are: alternating adaptive single-modality-based learning, gradient modification that prevents modal forgetting, and dynamic fusion based on uncertainty at the inference stage.

B. Alternating Adaptive Learning

First, in multimodal learning scenarios, modality inertia occurs because some modalities are under-optimized when learning with other modalities, which further leads to suboptimal fusion performance. To solve this problem, the Alternating adaptive learning paradigm was proposed, and each modality was optimized independently and while eliminating the interference caused by information entropy and converging speed between them, so that the less optimized (e.g. Angio-flow) modality with less information entropy contained can reach its full potential without being overshadowed by more informative modality (e.g. B-scan).

During training, a 3D-based modality input *M* is denoted by m_t, with V_m as its data volume. Each modality uses a dedicated encoder *Hm*, while a shared linear head g handles cross-modal prediction. At each step, the model trains the data from a single modality, defined by m_t [21]:

$$m_t = t \bmod M, where\ t < M. \qquad (4.\ 1)$$

Then, in every training stage *t,* by minimizing the prediction loss of one single modality,

the model can iterate and optimize its parameters.

$$Loss_t = E_{(X,Y)\sim D_{m_t}}\left[l\left(g\left(H_{m_t}(X;\,\theta_{m_t})\right);\varphi\right),Y\right] \qquad (4.2)$$

In which θ_{m_t}, φ are the learnable parameters in encoder H_{m_t} and share head g, and D_{m_t} refers to the dataset of M modality. This procedure enables the model to extract multiple features by avoiding modality laziness caused by low entropy modality with less texture information. In the RVO dataset, the information entropy of B-scan images tends to be higher than Angio-flow images, due to its representation of complex structural texture information. But according to the clinically based research [3] [4] [20] and single modality baseline results in Chapter 4.3.3, the Angio-flow modality is more important for reflecting RVO-related microvascular features with better baseline performance on ConvNeXt-based methods and should be weighted more in the prediction tasks. By means of multi-modality Alternating adaptive learning mechanism, this model was able to surpass the performance of ConvNeXt-tiny-based baseline and avoid modality dominance which may affect the optimization.

C. Shared head

In order to prevent modality inertia by optimizing each mode independently, capturing information about interactions across modes is another key aspect. Although the modality inertia problem can be mitigated by independently optimizing the learning process for each modality, in addition to preserving the multimodal representation, an ideal multimodal model should also be able to capture information about the interactions between the modalities. In Formula 4.2, the mmDFC framework uses a shared head g for all modalities to capture cross-modality interaction information through the entire process. However, this sequential optimization process presents a new challenge: when learning a new modality, g tends to forget previously trained modality information, which is called modality forgetting. This problem can significantly weaken the effectiveness of multi-modality learning.

D. Gradient Modification

Inspired by the weight correction method, a gradient correction matrix G_t, is introduced in each iteration to correct the gradient of the parameter φ of the shared head g before starting to learn a new modality, for preserving previous modality information loss and avoid modality forgetting problem. This gradient correction ensures that the direction of the parameter update is orthogonal to the cross-plane of the previous modality features [21]:

- During iteration *t*, the parameters φ of share head *g* can be defined as:

$$\varphi\_t = \varphi\_(t-1) - \gamma \begin{cases} \nabla_{\varphi} Loss_t, & if\ t = 0, \\ P_t \nabla_{\varphi} Loss_t, & if\ t > 0. \end{cases} \tag{4. 3}$$

  In which $Loss_t$ is the loss in Function (4. 2).

- To gain the gradient modification matrix, using the least recursive squares algorithm. Specifically, the average output of the individual encoder is defined as:

$$\overline{H_{mt}(X)} = \frac{1}{N_{mt}} \sum_{k=1}^{V_{mt}} H_{mt}(X_{mt}, k) \tag{4. 4}$$

- Define *s* as encoder output dimension, then in each iteration *t*, the gradient modification matrix can be gained by the following function:

$$G_t = G_{t-1} - I_t [\overline{H_{mt}(X)}]^T G_{t-1} \tag{4. 5}$$

  and

$$I_t = \frac{G_{t-1} \overline{H_{mt}(X)}}{\alpha + [\overline{H_{mt}(X)}]^T G_{t-1} \overline{H_{mt}(X)}} \tag{4. 6}$$

  In which α is a hyperparameter used to prevent the denominator from being zero. The modified matrix is initialized to the identity matrix before training. By introducing a gradient orthogonalization process to correct the weight update direction of the shared head *g*, interference between continuous modalities is mitigated and more efficient modality information saving is facilitated.

E. Dynamic Fusion

After the mmDFC network alternatively learned features from modality-specific

encoders and shared heads during training, the effectiveness of integrating method for multimodal information should be strengthened. To achieve weight-optimized fusion method, a weighted new combination of each modality regarding the prediction uncertainty was used. Specifically, given a volumetric validation sample x_V (e.g. Angio-flow), the prediction procedure can be defined as follows:

$$\hat{y}_V = \sum_{m=1}^{M} \beta_{m,V} * g \circ H_{m_t}(x_{m,V};\ \theta_m^*, \varphi^*) \quad (4.\ 7)$$

In which $\beta_{m,V}$ refers to the weight of importance of modality m when predicting the sample *V*'s label. θ_m^* and φ^* is the related optimization parameters in encoder hm and share head *g*.

$$Entropy_{m,V} = -P_{m,V}^T \log p_{m,V} \quad (4.\ 8)$$

$$p_{m,V} = Softmax(g \circ H_{m_t}(x_{m,V};\ \theta_m^*,\ \varphi^*)) \quad (4.\ 9)$$

In order to determine $\beta_{m,V}$, the dynamic fusion mechanism assumes that when the prediction result of a modality exhibits higher uncertainty, it is more likely to produce false result and affect the overall assessment (in this study, this modality refers to B-scan). Therefore, prediction uncertainty was used as a proxy to measure the importance of each modality. It is worth noting that each modality, whether dominant or minor, can reflect strong uncertainties. Uses the information entropy of each output of modalities to evaluate the uncertainties (4. 8). Softmax function was used to transfer logits to probabilities $p_{m,V}$ according to function (4. 9). A higher entropy indicates a lower prediction confidence and can be given smaller importance weights. The calculation of the importance of weight for modality *m* can be represented as:

$$\beta_{m,V} = \frac{exp(max_{m1=1,\dots,M}\ e_{m1,V} - e_{m1,V})}{\sum_{m2=1}^{M} exp(max_{m=1,\dots,M}\ e_{m1,V} - e_{m2,V})} \quad (4.\ 10)$$

By introducing dynamic fusion mechanism at inference section that explicitly considers the predictive uncertainties associated with each modality, the mmDFC network was able to better handle the imbalance of mode-specific information contained in B-scan and Angio-flow images, as the Angio-flow images were usually calculated with lower entropy and tend to be under optimized. By giving higher weights of importance in inference fusion stage, the extracted features gained by ConvNeXt-based encoder on

Angio-flow data can be set to higher weights for late sum fusion, thereby enhancing the effectiveness of multimodality fusion by balancing the difference between modalities.

4.2.2 Materials

The datasets contain both 2D and 3D format of SS-OCTA RVO imaging data. For 2D datasets, the total number of image pairs was 22,272 from 87 patients as mentioned in Chapter 3.2.2, in which per patient contains 256 slices including two scan types: B-scan and Angio-flow, and the training and test procedure were conducted on the same training and validation datasets as mentioned in Chapter 3.2.2. The sum fusion dataset that combined B-scan and Angio-flow modalities was set as comparison to evaluate the usefulness of mmDFC network.

For 3D datasets, the overall number of volumes is 87, which corresponds to the total number of 87 patients. The 3D sum-fusion dataset was also generated. Image standardization was first applied to the SS-OCTA dataset. Image normalization process was adopted after the standardization. Further, data augmentation strategies including random rotation and random flips were used to improve the robustness of the model.

4.3 Experiments and Results

4.3.1 Experimental Setting

The following experiments have been conducted on the 2D and 3D RVO datasets for the mmDFC network:

- Training and performance analysis. Each training stage took around 4 hours to train on 12 epochs with a batch size of 16 on a A6000 GPU, in which the model converged before the 10th epoch. The learning rate of the training process was set to 5×10^{-5}, and the Adam optimizer with step linear learning rate scheduler was

adopted in this process.

- Comparison of the overall performance of the proposed mmDFC method with the original ConvNeXt-based baseline mentioned in Chapter 3.2.2 for 2D and 3D RVO datasets. Specifically, the mmDFC adopted 2D and 3D two single modality input as it performs multi-modality fusion, while the ConvNeXt-based baseline used the direct input of sum-fusion dataset, but not performed any kind of feature fusion.

- Comparison of the overall performance of the proposed method with the original multi-modality alternating adaptive learning and dynamic fusion method which was based on ResNet-34 model and SS-OCTA RVO dataset with single modality input. This step enables further evaluation of ConvNeXt-based structure when integrated with mmDFC on RVO prognosis classification task.

- Separate evaluation of main methods in mmDFC network including Alternating adaptive learning and Dynamic fusion was conducted to quantify the improvement that each method provided to enhance the fusion result of mmDFC. The Alternating adaptive learning integrated with ConvNeXt-based encoder on two single modality input and the Dynamic fusion method that optimized the inference stage of fusion with ConvNeXt-based encoder were evaluated as comparisons.

4.3.2 Evaluation Metrics

The evaluation metrics used are Accuracy (Acc), Sensitivity (Sen), Specificity (Spe), Precision (Pre), F1-score (F1) and Area Under the ROC Curve (AUC). These metrics adopted the same calculation functions as introduced in Chapter 3.3.1 and were based on macro average calculation. In addition, Precision-recall (PR) curves and ROC curves are also used for comparisons between the best 2D and 3D methods, consistent to published classification study [11][12].

4.3.3 Results

Table 4.1 demonstrates that mmDFC improved existing ConvNeXt-based baseline, multi-modality Alternating adaptive learning method, and multi-modality Dynamic fusion method on 2D-based RVO prognosis prediction task. When compared to ConvNeXt-based baseline model constructed by ConvNeXt-tiny and performed on single modality datasets, this mmDFC model achieved a superior result with greater than 2% increment. As for the results on the sum-fusion dataset of two modalities, the mmDFC resulted in an increase of 0.8% on classification accuracy. In addition, when compared with multi-modality Alternating adaptive learning and Dynamic fusion network with ResNet-34 as backbone, this network with ConvNeXt-based backbone improved by ~2% performance on accuracy regarding the 2D single modality dataset.

In Table 4.1, there are three 2D ConvNeXt-tiny baseline evaluations on two single modality including B-scan dataset and Angio-flow dataset. The sum fusion dataset was processed based on early sum fusion of B-scan and Angio-flow modalities by setting these modalities as different input channels, as introduced in Chapter 3.2.2, then this fusion dataset was fed into the 2D ConvNeXt-tiny baseline model and got evaluation results. On the contrary, the original multi-modality alternating adaptive learning and dynamic fusion method with its ResNet-34 backbone and mmDFC method with ConvNeXt-based structure as new backbone were evaluated on separate input of two individual RVO modalities.

Table 4.1 2D Evaluation results and model comparisons on 2D SS-OCTA datasets (input: 256×256)

Model	Dataset	Acc.	Sen.	Spe.	Pre.	F1.	AUC
2D ConvNeXt-tiny	Bscan	0.689	0.550	0.520	0.623	0.501	0.596
2D ConvNeXt-tiny	Angio-flow	0.709	0.562	0.590	0.667	0.550	0.610
2D ConvNeXt-tiny	Fusion	0.728	0.573	0.621	0.675	0.565	0.645
2D mmDFC (ResNet)	Multi	0.689	0.560	0.507	0.604	0.580	0.599
2D Dynamic (ConvNeXt)	Multi	0.732	0.570	0.622	0.700	0.625	0.647
2D Alternating (ConvNeXt)	Multi	0.733	0.569	0.623	0.698	0.629	0.650
2D mmDFC (ConvNeXt)	**Multi**	**0.735**	**0.573**	**0.625**	**0.701**	**0.630**	**0.651**

The 3D-based datasets were also tested on 3D models as indicated in Table 4.2. Totally 5 models were evaluated on sum-fusion dataset input and the individual single modality input. The 3D mmDFC network shows a significant increase of ~3% accuracy and AUC score on 3D individual modality input datasets. In addition, when compared with original multi-modality alternating adaptive learning and dynamic fusion method with ResNet-34 as backbone model, this mmDFC network based on ConvNeXt backbone improves ~11% AUC performance on 3D dataset.

Table 4.2 3D Evaluation results and model comparisons on 3D SS-OCTA datasets. The best results are bolded, and second-best results are underlined (input: 256×256×256).

Model	Dataset	Acc.	Sen.	Spe.	Pre.	F1.	AUC
3D ConvNeXt-tiny	Bscan	0.678	0.697	0.600	0.682	0.662	0.670
3D ConvNeXt-tiny	Angio-flow	0.714	0.721	0.701	0.699	0.684	0.701
3D ConvNeXt-tiny	Fusion	0.750	0.778	0.722	0.714	0.719	0.718
3D mmDFC (ResNet)	Multi	0.710	0.701	0.682	0.677	0.674	0.690
3D Dynamic (ConvNeXt)	Multi	0.768	0.786	<u>0.726</u>	0.721	0.740	0.785
3D Alternating (ConvNeXt)	Multi	<u>0.775</u>	<u>0.788</u>	0.724	<u>0.722</u>	<u>0.744</u>	<u>0.791</u>
3D mmDFC (ConvNeXt)	**Multi**	**0.781**	**0.792**	**0.730**	**0.728**	**0.749**	**0.799**

Results in Table 4.1 and 4.2 was based on disease classification that included both Branch retinal vein occlusion (BRVO) and central retinal vein occlusion (CRVO) disease types. However, there are subtle differences between these types where CRVO patients in this study required more injections than BRVO patients, resulting in less non-perfusion extension, and therefore more difficult to correctly classify.

4.4 Discussions

The evaluation results of the mmDFC network show that it achieved the highest accuracy among all other baseline methods (as in Chapter 3) on both 2D and 3D datasets. This is attributed to the benefits of using mmDFC method, and can be summarized as follows: (1) The ConvNeXt as backbone successfully extracted subtle RVO progression-related features from the B-scan and Angio-flow modalities; (2) The Alternating adaptive training and Dynamic fusion reweighting inference framework of mmDFC enables the optimized fusion method of balancing two RVO modalities results in a further enhanced progression prediction ability; (3) The 3D-based input and 3D mmDFC network leveraged the implicit spatial information which is important for RVO structural and detail feature extraction, however, this information is lost in 2D dataset and networks. This optimization made the performance improve by over 2% in prediction AUC and 3.1% in accuracy.

As introduced in Chapter 3, ConvNeXt-tiny structure is the most effective backbone compared to all the ConvNeXt types, due to its suitable size of network parameters and it is less prone to overfitting problem with better overall performance. However, the original alternating dynamic fusion method with ResNet-34 as the backbone resulted in lower performance compared to the ConvNeXt-based mmDFC. The SS-OCTA RVO datasets with superior resolutions and various types of multi-modality features contain subtle retinal abnormalities that require strong feature extraction capabilities of the machine learning model to capture the structural features in OCT images which tend to have larger size in appearance whereas the Angio-flow information is more subtle and detailed are harder to extract. The limited depth and scalability of ResNet-34 without large kernels, layer scaling, and advanced normalization techniques that appeared in ConvNeXt-tiny, as well as its smaller size of parameters, restricted the effectiveness on SS-OCTA datasets. ConvNeXt-tiny's advanced architecture is more adept at identifying these nuanced patterns and resulted in better classification results which improved ResNet-34 structure for 10.9% on AUC and 7.1% on accuracy.

The Alternating adaptive learning module in the mmDFC network reduced the interaction between B-scan and Angio-flow, with Angio-flow usage being sub-optimized with it having lower information entropy with less large-scale structural information observed in the training stage. This is despite Angio-flow modality achieved better classification results compared to B-scan on both 2D and 3D single modality datasets (Tables 4.1 and 4.2), inferring its importance on identifying RVO progression outcomes [3][4][20]. Thus, the sum-fusion of two modalities did not mitigate the negative effects caused by sub-optimization when utilized on single-modality networks such as single ResNet, or ConvNeXt. Regarding the mmDFC network that has integrated alternating training procedure and enabled separate training on ConvNeXt-tiny, these two drawbacks are attenuated and reached better overall performance compared to baseline evaluation results, in which the mmDFC outperformed the best baseline ConvNeXt-tiny (on sum-fusion dataset) by 2.1% on AUC and further indicated its effectiveness for RVO prognosis classification.

The single-modality evaluations in Tables 4.1 and 4.2 demonstrates that the B-scan modality reached lower scores on all the metrics compared to Angio-flow, which indicates higher classification uncertainty regarding the ground truth that may lead to worse classification performance. In addition to the inference procedure from mmDFC network, the uncertainty values of each data from each modality were calculated based on the classification probabilities processed by Softmax and were further linked with new fusion strategy where lower probabilities derived by the model results in higher uncertainty and was assigned to a lower fusion weight. This reweighting strategy (namely Dynamic fusion method) improved the performances of proposed network as the 2D and 3D mmDFC showed compared to the 2D/3D Alternating methods and Dynamic fusion methods, inferring that the uncertainty of the B-scan due to its weak ability of progression interpretation may affect the overall performance and can be optimized by reducing its fusion weight at the last Softmax layer. Notably, consistent to the comparisons of 2D/3D ConvNeXt results in Chapter 3.3.2, it's worth to focus on processing and analysing 3D-based RVO dataset including single-modality and multi-

modalities for disease assessment, as it has been proved in this chapter that 3D mmDFC also surpassed the 2D mmDFC for 14.8% enhancement on AUC.

Regarding the evaluations of main components in the mmDFC method, the Alternating adaptive training method and Dynamic fusion method have been analyzed individually to quantify their contributions to the final classification results. The separate evaluations for Alternating adaptive-only training stage and dynamic fusion-only in inference stage in Table 4.2 showed that the Alternating adaptive learning method improves more significantly on multi-modality dataset compared to dynamic fusion method, but the enhancement on 2D dataset is marginal compared to 3D dataset. This result infers that the improvements caused by mmDFC method does not offset the shortcomings of the 2D dataset itself such as slices without obvious lesions. However, the performance on 3D dataset of mmDFC indicates that Alternating training and Dynamic fusion can progressively improve accuracy.

4.5 Summary

In this study, the mmDFC network successfully optimized feature fusion mechanism and the final prediction performance for RVO multi-modality dataset. This network has leveraged the advantages of the advanced structure: ConvNeXt-tiny, enhanced the fusion weight calculation strategy, and separated modality training procedure to reduce the interference between B-scan and Angio-flow modality while preventing the model from losing its gradient. On both 2D and 3D multi-modality dataset, this fusion method improves the ConvNeXt-based baseline models as mentioned in Chapter 3 on sum-fusion input, as well as single modality performance. Notably, the 3D version of this dynamic fusion network achieved the highest accuracy of 78.1% and exceeded the 3D baseline for 2.1%. In addition, the journal paper "Foundation model guided dynamic multi-modality fusion for RVO prognosis predictions" that is in preparation to submit to *Computer Methods and Programs in Biomedicine*, is based on the methodology extension of 2D model in this chapter.

Chapter 5. Correlation-driven Dual-branch Fusion Network for RVO Prognosis Classification

5.1 Introduction

In this chapter, a 2D multi-modality Dual-Branch Correlation-driven Fusion ConvNeXt (mmDCFC) method for RVO prognosis classification is proposed. The mmDFC network introduced in Chapter 4 successfully isolated the interaction between two modalities to better optimize each modality without being affected by single-modality dominance; however, it was not able to support intermediate fusion between modalities that is potentially important for RVO prognosis classification, as it has been investigated the occurrence of microvascular flow signals is associated with the location of lesions which may also lead to retinal structural changes [3][41]. In this chapter, Correlation-Driven Dual-Branch Feature Fusion (CDD-Fuse) [42] is adopted to the mmDFC to successfully separate and fuse the modality-shared structural features (via Transformer encoders) and modality-specific detailed features (via CNN encoders), inferring the potentials in optimizing the interaction between B-scan and Angio-flow regarding the common base features and specific detailed features. As an extension of the 2D mmDFC, the mmDCFC method introduced capability to capture correlated features across modalities, which resulted in enhanced multimodal fusion while ensuring the maintenance of the modality-specific information. The main contributions are as follows:

- Compared to previous research on RVO-related tasks, the proposed 2D mmDCFC feature fusion method leveraged the advanced correlation-based detail and basic feature extraction and fusion methods, as well as ConvNeXt-tiny structure for RVO-related classification, in which the main architecture utilized Transformer-

based encoder that enables more effective representations of different level of RVO features (mainly on B-scan as it contains more structural features), resulting in a better performance on 2D multi-modality RVO progression classification task and potentially makes a better use of B-scan.

- mmDCFC demonstrated the importance of intermediate feature extraction and fusion which can improve the representation on 2D multi-modal datasets, and achieved superior performance when compared to single modalities and sum-fusion method. By utilizing the extracted features of detailed local information and long-range structural information, the ConvNeXt-tiny-based method was integrated in the features output, surpassing the 2D mmDFC method.

5.2 Methods and Materials

5.2.1 Overview of the Framework

The purpose of the CDD-Fuse decomposition mechanism that was leveraged in mmDCFC network is to better combine and separate the common and specialized characteristics of each modality, such as highlight different level of areas and texture details to aid the classifier for better performance. The CDD-Fuse mechanism is a combination of Transformer and CNN structures, which can further enhance the Angio-flow and B-scan fusion representation results as they can focus on long-range dependence and local details in each modality. Base features between modalities that contain structural and macroscopic information in RVO imaging which are more commonly shared, and detailed features in individual modalities that include modality-specified features, such as microvascular changes, can be extracted by Transformer and CNN in two branches and more customized for RVO prognosis classification task.

The pipeline can be divided into 2 stages. In the first stage shown in Fig 5.1, CDD-Fuse

mechanism in mmDCFC network first uses Restormer block to extract cross-modal shallow features, and then introduces a two-branch Transformer-CNN feature extractor, where Lite Transformer (LT) block utilizes long-range attention to process low-frequency global features. Invertible Neural Networks (INN) blocks are used to extract high-frequency local features. Based on the embedded semantic information, low-frequency features should be correlated, and high-frequency features should be uncorrelated. In this procedure, the losses of these two branches are isolated and backward separately, which is similar to the Alternating adaptive training stage in mmDFC. Therefore, a correlation-driven loss function is proposed to enable the network to decompose features more efficiently. In the second training stage shown in Fig 5.2, the aforementioned LT and INN modules output fused images regarding the base and detail features, and then further processed by the ConvNeXt-based classifier to get the final classification decision.

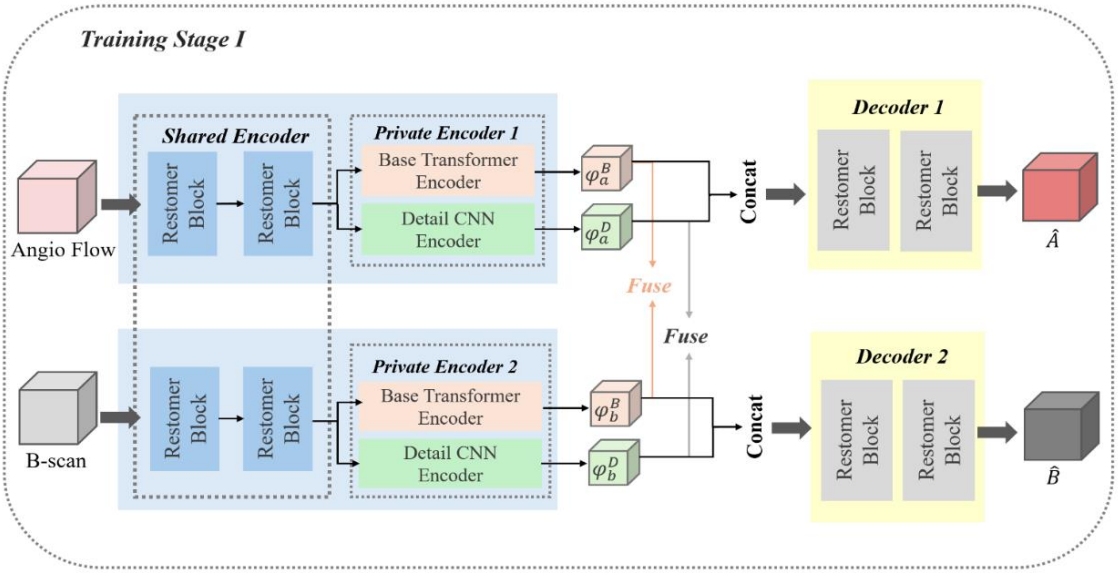

Fig 5.1 Overall framework for training stage I

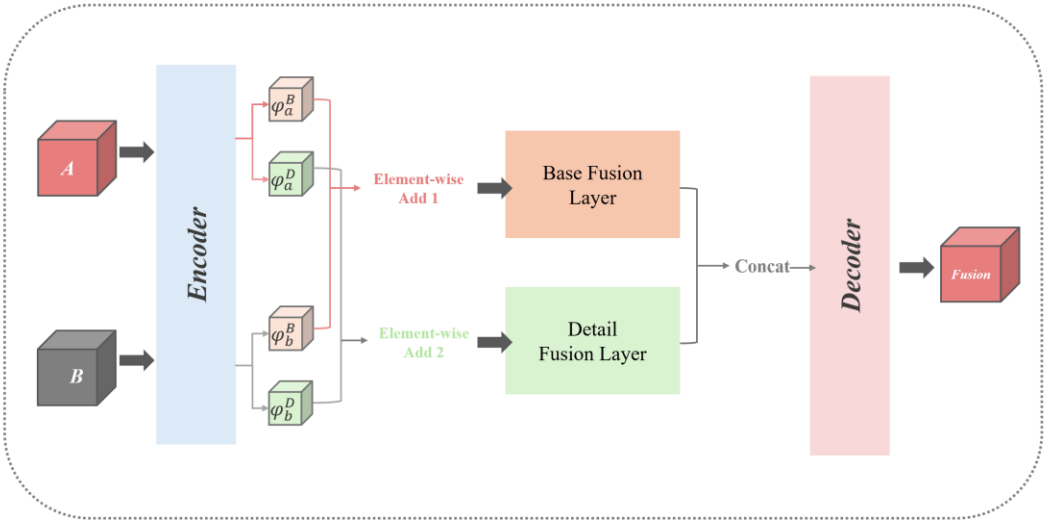

Fig 5.2 Overall framework for training stage II

A. Restormer Encoder

The encoders of this network contain three parts: Restormer block-based share feature encoder (SFE), Lite Transformer (LT) block-based base transformer encoder (BTE), Restormer Block-based Share Feature Encoder (SFE), Lite Transformer (LT) Invertible Neural networks (INN) block-based detail CNN encoder (DCE), where BTE and DCE together form a short-distance encoder. For the input 3 channel Angio-flow input images and single-channel B-scan image, the SFE, BTE and DCE modules are represented by S, B and D respectively. First, for the SFE module used to extract common features, its goal is to extract shallow features, as follows.

$$\varphi_I^S = S(I), \qquad \varphi_V^S = S(V),$$

The reason for using Restormer block in SFE is to extract global features using a self-attention mechanism between dimensions, so cross-modal shallow features can be extracted without additional computing power.

BTE is used to extract low-frequency base features from common features in B-scan and Angio-flow, which tend to include structural information, as follows.

$$\varphi_I^B = B(\varphi_I^S), \qquad \varphi_V^B = B(\varphi_V^S),$$

In order to extract distance dependency, transformers with spatial self-attention is used. In order to balance effect and operation efficiency, LT block is adopted as the base unit of BTE, which can ensure effect while reducing the number of parameters. DCE, as opposed to BTE, is used to extract high frequency details, as follows.

$$\varphi_I^D = D(\varphi_I^S), \qquad \varphi_V^D = D(\varphi_V^S),$$

$$\varphi^B = F_B(\varphi_I^B, \varphi_I^B), \qquad \varphi^D = F_D(\varphi_I^D, \varphi_I^D)$$

Considering that edge texture information is also important in the fusion task, it is expected that DCE will save as much detailed information as possible. The INN module ensures that input information is preserved as much as possible by generating input and output from each other and can therefore be used for lossless feature extraction in DCE by coupling INN with affine coupling layers. The transformation of each reversible layer is as follows.

$$\varphi_{I,k+1}^S[c+1:C] = \varphi_{I,k}^S[c+1:C] + I_1\big(\varphi_{I,k}^S[1:C]\big),$$

$$\varphi_{I,k+1}^S[1:c] = \varphi_{I,k}^S[1:c] \odot exp\,(I_2\big(\varphi_{I,k+1}^S[c+1:C]\big) + I_3(\varphi_{I,k+1}^S[c+1:C]),$$

$$\varphi_{I,k+1}^S = C * A\tau\big\{\varphi_{I,k+1}^S[1:c], \varphi_{I,k+1}^S[c+1:C]\big\}$$

B. Base and Detail Fusion Layer

This layer is used to fuse the base and detail features extracted by the two encoders. Considering that the inductive bias of base and detail feature fusion should be the same as that of the encoder's base and detail feature extraction, the base and detail fusion layer is implemented using LT and INN blocks, as follows.

$$\varphi^B = F_B(\varphi_I^B, \varphi_V^B), \qquad \varphi^D = F_D(\varphi_I^D, \varphi_V^D)$$

C. Decoder

The decoder first concatenates the decomposed features in the channel dimension as input, then takes the source graph as output in the training stage I, and the fusion graph

as output in the training settlement II, as follows.

Stage I: $\hat{I} = DC(\varphi_I^B, \varphi_I^D)$, $\hat{V} = DC(\varphi_V^B, \varphi_V^D)$;

Stage II: $F = DC(\varphi^B, \varphi^D)$

Since the input is characterized by cross-modal and multi-band characteristics, keeping the decoder structure consistent with SFE means Restormer block as the base unit.

D. Two-stage training

In the first stage, B-scan and Angio-flow images are used as the input of SFE to extract shallow features, and BTE and DCE extract high and low frequency features. Then the Angio-flow base and detail features are splicing, after that, the B-scan base and detail features are also concatenated, later these features are all sent to the decoder to rebuild the fusion features. Finally, these features are fed into ConvNeXt-tiny-based classifier to give final classification result.

The first stage loss function is as follows.

$$L_{total}^I = L_{Angio} + \alpha_1 L_{Bscan} + \alpha_2 L_{decomp}$$

The first two are reconstruction losses in Angio-flow and B-scan and the third is feature decomposition losses. One stage loss overall is so that information is not lost during encoding and decoding. The first Angio-flow reconstruction loss form is as follows.

$$L_{Angio} = L_{int}^I(I, \hat{I}) + \mu L_{SSIM}(I, \hat{I})$$

$$L_{int}^I = \left\| I - \hat{I} \right\|_2^2$$

$$L_{SSIM}(I, \hat{I}) = 1 - SSIM(I, \hat{I})$$

$$L_{decomp} = \frac{(L_{CC}^D)^2}{L_{CC}^B} = \frac{(CC(\varphi_I^D, \varphi_V^D))^2}{CC(\varphi_I^B, \varphi_V^B) + \epsilon}$$

CC in the formula is the correlation coefficient commonly used in fusion. This loss is to make the distance between common features as close as possible and the distance between unique features as far as possible, and the correlation coefficient can measure the distance between features, so the low-frequency base feature is taken as the denominator, and the high-frequency detail feature is taken as the numerator.

$$L_{total}^{II} = L_{int}^{II} + \alpha_3 L_{grad} + \alpha_4 L_{decomp}$$

$L_{int}^{II} = \frac{1}{HW} \left\| I_f - max(I_{Angio,} I_{Bscan}) \right\|_1$ and $L_{grad} = \frac{1}{HW} \left\| |\nabla I_f| - max(|\nabla I_{Angio}|, |\nabla I_{Bscan}|) \right\|_1$.

In which ∇ indicates the Sobel gradient operator, α_3 and α_4 are tuning parameters.

5.2.2 Materials

The dataset used in this study is 2D format of two modality SS-OCTA RVO imaging data including B-scan and Angio-flow. The total number of image pairs is 22,272 from 87 patients, while the external test dataset was set as the same as mentioned in Chapter 3.2.2. The ConvNeXt-based baselines on sum-fusion dataset and single modality datasets were set as comparisons to evaluate the performance of mmDCFC. The preprocessing of multi-modality datasets first utilized image normalization strategy to standardize the datasets. Data augmentation strategies including random rotation and random flips were used to improve the robustness of the model.

5.3 Experiments and Results

5.3.1 Experiment Setup

The following experiments have been performed on the 2D multi-modality RVO datasets for the mmDCFC network. Each experiment was performed on the same training and validation datasets as mentioned in Chapter 3.2.2:

- The mmDCFC model training and performance analysis. Each training stage took around 6 hours to train on 10 epochs with a batch size of 8 (4 of B-scan and 4 of Angio-flow) on an A6000 GPU, in which the model converged before the 10th epoch. The learning rate of the training process was set to $5*10^{-5}$, and the Adam optimizer with a step linear learning rate scheduler was adopted in this process.
- Comparison of the overall performance of mmDCFC method and mmDFC method for two modalities input.
- Visualization of mmDCFC method regarding the multi-modality fusion results.

5.3.2 Results

Table 5.1 showed that the mmDCFC fusion method improved the existing ConvNeXt or Multi-modality Dynamic fusion baseline methods on 2D-based RVO prognosis prediction task. When compared with ConvNeXt baseline model performed on both B-scan and Angio-flow single modality datasets, as well as the fusion dataset of these two modalities, this model surpassed all the single modality results and gained an increase of 0.9% accuracy on 2D early sum fusion dataset. Compared to the 2D mmDFC based on ConvNeXt introduced in Chapter 3, the 2D mmDCFC surpassed its accuracy and AUC for 0.2%. The best model results of 2D mmDCFC were highlighted in Table 5.1 and the second-best result mainly produced by model 2D mmDFC were underlined.

Table 5.1 Evaluation results of mmDCFC and model comparisons on 2D SS-OCTA datasets**.** The best results were highlighted, and the second-best results were underlined.

Model	Dataset	Acc.	Sen.	Spe.	Pre.	F1.	AUC.
2D ConvNeXt-tiny	Bscan	0.689	0.550	0.520	0.623	0.501	0.596
2D ConvNeXt-tiny	Angio-flow	0.709	0.562	0.590	0.667	0.520	0.610
2D ConvNeXt-tiny	Fusion	0.728	0.573	0.621	0.675	<u>0.565</u>	0.645
<u>2D mmDFC</u>	<u>Multi</u>	<u>0.735</u>	<u>0.577</u>	<u>0.625</u>	<u>0.701</u>	0.563	<u>0.651</u>
2D mmDCFC	**Multi**	**0.737**	**0.581**	**0.650**	**0.693**	**0.578**	**0.653**

Fig. 5.3 shows two visualization samples of the mmDCFC results. After the model automatic reweight procedure for different levels of information, the left side of Fig. 5.3(a) shows several Angio-flow points that were labeled as yellow by setting the weight of each image channel, which infers a microvascular level of detail features that are quite different from the base structural and long-range information that are colored blue. The visualizations show that blood flow information (yellow pattern) is prominent; this is consistent to clinical-based research [3][4] which suggests that blood flow information is more relevant to the prognosis of RVO. Notably, not all parts of Angio-flow information were noted as yellow color, because the model defined these areas as low-frequency features and tend to interact with the base structural information and was colored in blue. Meanwhile, in Fig. 5.3(b), detailed features mainly related to Angio-flow from another patient are depicted in yellow on the right side of the image, as they probably indicate the FAZ characteristics in NPA progression. The visualization results in Fig 5.3 also intuitively showed expected fusion effect, making it easier for the decoder to distinguish important low-contrast areas from the surrounding structures, as the modality-specified and modality-shared information were depicted as different colors. High and low frequency information are visualized by channel reweighting, which refers to setting the three channels of the output matrix with different weights to

make the features visualized with specified colors.

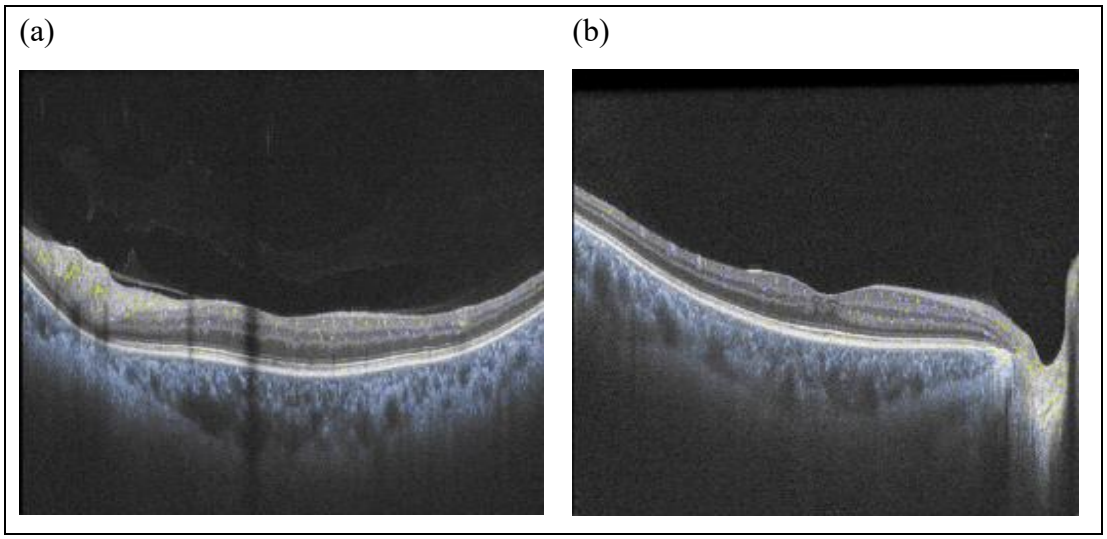

Fig 5.3 Visualization for B-scan and Angio-flow images processed by mmDCFC

5.4 Discussions

The mmDCFC is based on 2D RVO dataset and is designed as an extension of 2D mmDFC for an enhanced cross-modal feature fusion ability. The evaluations in Table 5.1 presented that the mmDCFC improved the performance of mmDFC, especially in accuracy and AUC, by 0.2%. In mmDFC, the network prevents modality dominance but doesn't explicitly decompose and emphasize the cross-modal information. In RVO morphologically, the occurrence of microvascular flow signals is associated with the location of lesions, such as edema and NPAs, and these symptoms may also lead to retinal structural changes [3][42]. Thus, the lack of extraction of common structural basic features that correlated across B-scan and Angio-flow, and the redundant base information that is implicit in each single modality may cause inefficiency in performance. On the contrary, the mmDCFC can separate modality-specific features as well, while enforcing correlation features between modalities by a Transformer-based encoder. Transformers are specialized in extracting long-range dependencies that are highly linked to modality-shared structural information and are useful in processing common features between B-scan and Angio-flow. In addition, the Transformer based

structure also better leveraged the potential in B-scan as this modality hasn't been fully utilized in mmDFC and was given a lower fusion weight. Thus, the shared structural features and the usability of B-scan were neglected in mmDFC as it focuses on only one modality in one iteration and decreased the contributions of B-scan, which will possibly lead to a lack of meaningful cross-modality structural representations and worse overall performance.

There are several limitations of this study. It has been demonstrated that the mmDCFC network is useful in capturing modality-specified information and combining modality-shared information, but it hasn't fully leveraged the feature extraction capability of ConvNeXt-based architecture, as the ConvNeXt-tiny was only set to work as the classifier for the extracted and fused features. The CNN-based module in mmDCFC can be further optimized by the ConvNeXt-based architecture by replacing the CNN as the new encoder, to leverage its advantages in extracting subtle features. Second, the enhancement of mmDCFC is expected to be amplified when applied to 3D RVO multi-modality dataset, akin to the advantage of 3D over 2D as in Chapter 4.3.3 where 3D datasets can eliminate the adverse effects caused by 2D no-lesion slices. However, convert the mmDCFC to 3D version will cause high computation costs, as several Transformer-based Restormer blocks were implemented as extractors in this network and the GPU memory cost of 2D mmDCFC with low batch size of 4 has almost reached the limit.

5.5 Summary

The 2D mmDCFC network introduced in this study leveraged both the CDD-Fuse mechanism and ConvNeXt-tiny for an enhanced representation of cross-modal and modal-specified information that are implicit in the RVO multi-modality dataset and served as an extension of 2D mmDFC. As compared to previous research and methods introduced in Chapters 3 and 4 on RVO-related tasks, this method utilizes an advanced feature fusion method and decoding structure: correlation-based dual-branch basic and detailed feature extraction and fusion mechanism for better modality-shared feature capture while preserving the modality-specific features. In addition, the mmDCFC network is a combination of Transformer and CNN architectures that was adapted for different feature types: large-scale structural information and subtle Angio-flow information, enabling customized optimization for specific multi-modality RVO task, resulting in a better representation of features and overall performance.

This mmDCFC method proves that the separate extraction of low-frequency and high-frequency information are beneficial to a better extraction and interpretation of key RVO features, and also showed superior performance compared to non-fusion baselines on sum-fusion dataset. In addition, it provides inspirations for the exceeding optimizations of subsequent networks.

Chapter 6. Conclusions and Future Works

6.1 Conclusions

In this study, a SS-OCTA-based Radiomics machine learning pipeline was first proposed, for an assessment of traditional machine learning models, advanced deep learning models, and machine learning on clinical-defined disease-related biomarkers, as well as the radiomics feature fusion with deep learning. We demonstrated that the deep learning method with 3D ConvNeXt-based structure can improve the prediction performance compared to various machine learning models in both 2D and 3D, especially when fused with radiomics features.

Secondly, in order to make full use of the B-scan and Angio-flow modalities and extract the specific characteristics implied in these imaging modalities, we proposed a 2D/3D mmDFC network to reduce the interference between modalities and enhance the weighting strategies when fusing these modalities. Both 2D and 3D networks showed improved results when compared to machine learning baseline models.

Thirdly, we introduce a 2D mmDCFC method for RVO prognosis classification task as a new intermediate feature extraction and fusion strategy to extend upon the mmDFC. This network aims to strengthen the common features and specialized features from cross-modal low-frequency and modality-differed high-frequency RVO images. The result for mmDCFC method showed an improvement of ~1% when compared to ConvNeXt baseline with a sum-fusion input and gained superior performance than 2D mmDFC. Visualization results also indicate that the fusion strategy emphasizes special feature textures and depicts them with different appearances. Overall, the 2D mmDCFC method showed effectiveness among all 2D models especially by leveraging the implicit cross-modality information, indicating the usefulness of the modal-shared features regarding the multi-modality learning for RVO progression classification.

6.2 Limitations and Future Works

In this work, the 2D mmDCFC method successfully enhanced the cross-modal representation while preserving modality-specified information between modalities. However, this approach may have dismissed some spatial RVO structural features and correlations that exist across the B-scan and Angio-flow volumes, which are partially hidden in each modality and can reflect the inner correlation between RVO-related retinal layer morphological changes and microvascular changes across the tissue. In RVO disease manifestation, the various complications caused by ischemic phenomena can lead to thickness, volume, shape and vessel density changes that are interrelated to NPAs [43], and some features that can only be extracted from 3D volumes, such as FAZ that occurs on the orthogonal faces that are hard to reflect on 2D slices adopted in this research. It can be inferred that there are some extra spatial characteristics in morphology that were lost in 2D mmDCFC, and these significant structural changes are also highly linked to the bad prognosis of RVO.

Moreover, the obstruction caused by some 2D slices (e.g. sample labeled as good prognosis may have disease-related lesion that infers a bad prognosis) cannot be resolved by mmDCFC. To be more specific, the 3D ConvNeXt-based models in Chapter 3 and the 3D-based mmDCF in Chapter 4 resulted in better performance compared to their 2D benchmark models, but the improvement of 2D mmDCFC is marginal, even if it has surpassed the 2D mmDFC. This evidence may infer that the interference and inefficiency caused by some 2D slices greatly limited the model's performance, and this drawback may also affect the mmDCFC network, making it unable to better exert its ability in cross-modal feature extraction. Obviously, converting to a 3D network can solve the 2D interference problem and provide more useful positional correlation information as well.

Thus, my future work is to modify the mmDCFC network and design a new 3D version of mmDCFC fusion network. This is also due to the good alignment of B-scan and

Angio-flow in our SS-OCTA dataset. Regarding this, the early fusion of these two modalities is possible and promising to enhance the input of prognosis decision deep-learning models. The 2D mmDCFC cannot be directly promoted to 3D because of the high computational costs caused by Transformer-based Restormer blocks, and it's better to take advantage of this structure for a lightweight effective cross-modal fusion approach. In addition, in order to benefit from the dynamic fusion strategy and ConvNeXt structure, the final fusion of the new 3D mmDCFC will be further modified to an uncertainty-based fusion structure, and the ConvNeXt with an advanced feature extraction ability will be integrated into this 3D mmDCFC network to replace the original CNN encoder, for an optimized fusion strategy for base and detailed features, enhanced local features extraction strategies, and is expected to produce better classification performance.

References

[1] I. U. Scott, P. A. Campochiaro, N. J. Newman, and V. Biousse, “Retinal vascular occlusions,” *The Lancet*, vol. 396, no. 10266, p. 1927, 2020, doi: 10.1016/S0140-6736(20)31559-2.

[2] P. A. Campochiaro et al., "Vascular Endothelial Growth Factor Promotes Progressive Retinal Nonperfusion in Patients with Retinal Vein Occlusion," Ophthalmology, vol. 120, no. 4, pp. 795–802, 2013.

[3] S. S. Hayreh, “Photocoagulation for retinal vein occlusion,” *Prog. Retin. Eye Res.*, vol. 85, p. 100964, Nov. 2021, doi: 10.1016/j.preteyeres.2021.100964.

[4] T. E. de Carlo, A. Romano, N. K. Waheed, and J. S. Duker, “A review of optical coherence tomography angiography (OCTA),” *Int. J. Retin. Vitr.*, vol. 1, p. 5, 2015, doi: 10.1186/s40942-015-0005-8.

[5] D. Seknazi *et al*., “Optical coherence tomography angiography in retinal vein occlusion: Correlations between macular vascular density, visual acuity, and peripheral nonperfusion area on fluorescein angiography,” *Retina*, vol. 38, no. 8, p. 1562, 2017, doi: 10.1097/IAE.0000000000001737.

[6] F. Zheng *et al*., “Advances in swept-source optical coherence tomography and optical coherence tomography angiography,” *Adv. Ophthalmol. Pract. Res.*, vol. 3, no. 2, p. 67, 2022, doi: 10.1016/j.aopr.2022.10.005.

[7] I. Laíns *et al*., “Retinal applications of swept source optical coherence tomography (OCT) and optical coherence tomography angiography (OCTA),” *Prog. Retin. Eye Res.*, vol. 84, p. 100951, Sep. 2021, doi: 10.1016/j.preteyeres.2021.100951.

[8] D. Le, T. Son, and X. Yao, “Machine learning in optical coherence tomography angiography,” *Exp. Biol. Med.*, vol. 246, no. 20, p. 2170, 2021, doi: 10.1177/15353702211026581.

[9] Z. Wang *et al*., “Radiomic and deep learning analysis of dermoscopic images for skin lesion pattern decoding,” *Sci. Rep.*, vol. 14, no. 1, pp. 1–13, 2024, doi: 10.1038/s41598-024-70231-x.

[10] R. Cattell, S. Chen, and C. Huang, “Robustness of radiomic features in magnetic resonance imaging: Review and a phantom study,” *Vis. Comput. Ind. Biomed. Art*, vol. 2, no. 1, pp. 1–16, 2019, doi: 10.1186/s42492-019-0025-6.

[11] S. S. Kar *et al*., “Multi-compartment spatially-derived radiomics from optical coherence tomography predict anti-VEGF treatment durability in macular edema secondary to retinal vascular disease: Preliminary findings,” *IEEE J. Transl. Eng. Health Med.*, vol. 9, p. 1000113, 2021, doi: 10.1109/JTEHM.2021.3096378.

[12] L. Carrera-Escalé *et al*., “Radiomics-based assessment of OCT angiography images for diabetic retinopathy diagnosis,” *Ophthalmol. Sci.*, vol. 3, no. 2, p. 100259, 2022, doi: 10.1016/j.xops.2022.100259.

[13] Z. Meng *et al*., “Machine learning and optical coherence tomography-derived radiomics analysis to predict persistent diabetic macular edema in patients undergoing anti-VEGF intravitreal therapy,” *J. Transl. Med.*, vol. 22, p. 358, 2024, doi: 10.1186/s12967-024-05141-7.

[14] W. Xu *et al*., “Development and application of an intelligent diagnosis system for retinal vein occlusion based on deep learning,” *Dis. Markers*, vol. 2022, p. 4988256, 2021, doi: 10.1155/2022/4988256.

[15] J. Yim *et al*., “Predicting conversion to wet age-related macular degeneration using deep learning,” *Nat. Med.*, vol. 26, no. 6, pp. 892–899, Jun. 2020, doi: 10.1038/s41591-020-0867-7.

[16] G. Holste *et al*., “Harnessing the power of longitudinal medical imaging for eye disease prognosis using transformer-based sequence modeling,” *NPJ Digit. Med.*, vol. 7, no. 1, pp. 1–13, 2024, doi: 10.1038/s41746-024-01207-4.

[17] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A ConvNeXt for the 2020s,” *arXiv preprint arXiv:2201.03545*, 2022.

[18] I. Laíns *et al*., “Retinal applications of swept source optical coherence tomography (OCT) and optical coherence tomography angiography (OCTA),” *Prog. Retin. Eye Res.*, vol. 84, p. 100951, Sep. 2021, doi: 10.1016/j.preteyeres.2021.100951.

[19] D. S. Kermany *et al*., “Identifying medical diagnoses and treatable diseases by image-based deep learning,” *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[20] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, “Balanced multimodal learning via on-the-fly gradient modulation,” *arXiv preprint arXiv:2203.15332*, 2022.

[21] X. Zhang, J. Yoon, M. Bansal, and H. Yao, “Multimodal representation learning by alternating unimodal adaptation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 27456–27466.

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

[23] Z. Zhao *et al*., “CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5906–5916.

[24] Y. Hirano *et al*., “Multimodal imaging of microvascular abnormalities in retinal vein occlusion,” *J. Clin. Med.*, vol. 10, no. 3, p. 405, Jan. 2021, doi: 10.3390/jcm10030405.

[25] F. Coscas *et al*., “Optical coherence tomography angiography in retinal vein occlusion: Evaluation of superficial and deep capillary plexa,” *Am. J. Ophthalmol.*, vol. 161, pp. 160–171, 2016, doi: 10.1016/j.ajo.2015.09.026.

[26] R. F. Spaide *et al*., “Quantitative analysis of OCT angiography features in diabetic retinopathy,” *Prog. Retin. Eye Res.*, vol. 64, pp. 34–52, 2018, doi: 10.1016/j.preteyeres.2017.11.002.

[27] M. Guo, M. Zhao, A. M. Y. Cheong, H. Dai, A. K. C. Lam, and Y. Zhou, “Automatic quantification of superficial foveal avascular zone in optical coherence tomography angiography implemented with deep learning,” *Vis. Comput. Ind. Biomed. Art*, vol. 2, no. 1, p. 21, Dec. 2019, doi: 10.1186/s42492-019-0031-8.

[28] S. Schurer-Waldheim, P. Seebock, H. Bogunovic, B. S. Gerendas, and U. Schmidt-Erfurth, “Robust fovea detection in retinal OCT imaging using deep learning,” *IEEE J. Biomed. Health Inform.*, vol. 26, no. 8, pp. 3927–3937, Aug. 2022, doi: 10.1109/JBHI.2022.3166068.

[29] P. Lauermann *et al*., “Distance-thresholded intercapillary area analysis versus

vessel-based approaches to quantify retinal ischemia in OCTA," *Transl. Vis. Sci. Technol.*, vol. 8, no. 4, p. 28, 2019, doi: 10.1167/tvst.8.4.28.

[30] D. Nagasato *et al*., "Automated detection of a nonperfusion area caused by retinal vein occlusion in optical coherence tomography angiography images using deep learning," *PLoS ONE*, vol. 14, no. 11, p. e0223965, 2019, doi: 10.1371/journal.pone.0223965.

[31] X. Ren *et al*., "Artificial intelligence to distinguish retinal vein occlusion patients using color fundus photographs," *Eye*, vol. 37, no. 10, pp. 2026–2032, Jul. 2023, doi: 10.1038/s41433-022-02239-4.

[32] D. S. Kermany *et al*., "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, Feb. 2018, doi: 10.1016/j.cell.2018.02.010.

[33] S. A. Kamran, S. Saha, A. S. Sabbir, and A. Tavakkoli, "Optic-Net: A novel convolutional neural network for diagnosis of retinal diseases from optical tomography images," in *Proc. 18th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Boca Raton, FL, USA, 2019, pp. 964–971, doi: 10.1109/ICMLA.2019.00165.

[34] Y. Li *et al*., "Hybrid fusion of high-resolution and ultra-widefield OCTA acquisitions for the automatic diagnosis of diabetic retinopathy," *Diagnostics*, vol. 13, no. 17, p. 2770, 2023, doi: 10.3390/diagnostics13172770.

[35] E. Vaghefi, S. Hill, H. M. Kersten, and D. Squirrell, "Multimodal retinal image analysis via deep learning for the diagnosis of intermediate dry age-related macular degeneration: A feasibility study," *J. Ophthalmol.*, vol. 2020, p. 7493419, 2020, doi: 10.1155/2020/7493419.

[36] J. Hao *et al*., "Early detection of dementia through retinal imaging and trustworthy AI," *NPJ Digit. Med.*, vol. 7, no. 1, pp. 1–15, 2024, doi: 10.1038/s41746-024-01292-5.

[37] A. Vaswani *et al*., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 6000–6010.

[38] R. R. Selvaraju, M. Cogswell, A. Das, *et al.*, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, pp. 336–359, 2020, doi: 10.1007/s11263-019-01228-7.

[39] J. J. M. van Griethuysen *et al*., “Computational radiomics system to decode the radiographic phenotype,” *Cancer Res.*, vol. 77, no. 21, pp. e104–e107, 2017, doi: 10.1158/0008-5472.CAN-17-0339.

[40] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, 2005.

[41] A. H. Kashani, S. Y. Lee, A. Moshfeghi, M. K. Durbin, and C. A. Puliafito, “Optical coherence tomography angiography of retinal venous occlusion,” *Retina*, vol. 35, no. 11, pp. 2323–2331, 2015.

[42] Z. Zhao *et al*., “CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5906–5916.

[43] C. L. Spina, U. D. Benedetto, M. B. Parodi, G. Coscas, and F. Bandello, “Practical management of retinal vein occlusions,” *Ophthalmol. Ther.*, vol. 1, no. 1, p. 3, 2012, doi: 10.1007/s40123-012-0003-y.