Medical Image Analysis 109 (2026) 103942

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Predicting diabetic macular edema treatment responses using OCT: Dataset and methods of APTOS competition

Weiyi Zhang [a,1,2], Peranut Chotcomwongse [b,1,2], Yinwen Li [c,2], Pusheng Xu [a,2], Ruijie Yao [d], Lianhao Zhou [e], Yuxuan Zhou [f], Hui Feng [g], Qiping Zhou [h], Xinyue Wang [a,i,2], Shoujin Huang [j,2], Zihao Jin [k,2], Florence H T Chung [l,2], Shujun Wang [m,2], Yalin Zheng [n,2], Mingguang He [a,o,p,2,*], Danli Shi [a,o,2,*], Paisan Ruamviboonsuk [b,2,*]

[a] School of Optometry, The Hong Kong Polytechnic University, Hong Kong
[b] Department of Ophthalmology, College of Medicine, Rangsit University, Rajavithi Hospital, Thailand
[c] Department of Ophthalmology and Visual Sciences, Chinese University of Hong Kong, Hong Kong
[d] Duke University, USA
[e] A&M University, USA
[f] Department of Electronic Engineering, Tsinghua University, China
[g] Xidian University, China
[h] Svision Imaging Ltd, China
[i] Chien-Shiung Wu College, Southeast University, China
[j] Shenzhen Technology University, China
[k] Faculty of Engineering, The University of Hong Kong, Hong Kong
[l] Asia Pacific Tele-Ophthalmology Society, c/o GH164, The Hong Kong Polytechnic University, Hong Kong
[m] Department of Biomedical Engineering & Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong
[n] Faculty of Health and Life Sciences, University of Liverpool, United Kingdom
[o] Research Centre for SHARP Vision (RCSV), The Hong Kong Polytechnic University, Hong Kong
[p] Centre for Eye and Vision Research (CEVR), 17W Hong Kong Science Park, Hong Kong

ARTICLE INFO

Keywords:
Optical coherence tomography
Diabetic macular edema
Anti-VEGF therapy
Big data competition
Treatment response prediction

ABSTRACT

Diabetic macular edema (DME) significantly contributes to visual impairment in diabetic patients. Treatment responses to intravitreal therapies vary, highlighting the need for patient stratification to predict therapeutic benefits and enable personalized strategies. To our knowledge, this study is the first to explore pre-treatment stratification for predicting DME treatment responses. To advance this research, we organized the 2nd Asia-Pacific Tele-Ophthalmology Society (APTOS) Big Data Competition in 2021. The competition focused on improving predictive accuracy for anti-VEGF therapy responses using ophthalmic OCT images. We provided a dataset containing tens of thousands of OCT images from 2000 patients with labels across four sub-tasks. This paper details the competition's structure, dataset, leading methods, and evaluation metrics. The competition attracted strong scientific community participation, with 170 teams initially registering and 41 reaching the final round. The top-performing team achieved an AUC of 80.06 %, highlighting the potential of AI in personalized DME treatment and clinical decision-making.

* Corresponding authors.
E-mail addresses: wei-yi.zhang@connect.polyu.hk (W. Zhang), peranut.chot@gmail.com (P. Chotcomwongse), liyinwen-melody@link.cuhk.edu.hk (Y. Li), xupusheng977@163.com (P. Xu), ruijie.yao@duke.edu (R. Yao), leozhou@tamu.edu (L. Zhou), zhou-yx21@mails.tsinghua.edu.cn (Y. Zhou), fenghuiwk@gmail.com (H. Feng), qiping.zhou@intalight.com (Q. Zhou), 213222801@seu.edu.cn (X. Wang), Solor.pikachu@gmail.com (S. Huang), jinzh25@gmail.com (Z. Jin), florence.chung@asiateleophth.org (F.HT Chung), shu-jun.wang@polyu.edu.hk (S. Wang), Yalin.Zheng@liverpool.ac.uk (Y. Zheng), mingguang.he@polyu.edu.hk (M. He), danli.shi@polyu.edu.hk (D. Shi), paisan.trs@gmail.com (P. Ruamviboonsuk).

[1] These authors contributed equally to this work.
[2] These authors co-organized the competition. All others contributed methods of their algorithm(s) presented in the paper.

https://doi.org/10.1016/j.media.2026.103942
Received 7 May 2025; Received in revised form 5 December 2025; Accepted 9 January 2026
Available online 13 January 2026

1. Introduction

Diabetic retinopathy (DR) is a prevalent complication among individuals with diabetes, affecting approximately one-third of diabetic patients (Hashemi et al., 2022). Among patients with DR, diabetes macular edema (DME) stands out as a leading cause of visual impairment, with a growing global incidence affecting over 19 million individuals (Daruich et al., 2018). It is also the primary reason for vision loss among working-age adults worldwide. Effective and timely treatment is essential to preserving visual function and preventing disease progression. At present, intravitreal anti-vascular endothelial growth factor (anti-VEGF) injection has emerged as the standard therapy for treating DME (Solomon et al., 2017). This is attributed to its numerous clinical advantages, which encompass enhancements in both function and structure. Moreover, it has the potential to arrest or even reverse the progression of DR (Zhang et al., 2022). Nonetheless, patient responses vary considerably, with a significant proportion of patients showing suboptimal outcomes. According to the Diabetic Retinopathy Clinical Research (DRCR) Retina Network Protocol T study, it has been observed that 31.6 % to 65.6 % of patients with DME continue to experience persistent edema even after receiving regular intravitreal injections for a minimum of four times over a span of 24 weeks (Bressler et al., 2018).

Recently, optical coherence tomography (OCT) has become the gold standard for diagnosing and monitoring treatment responses in diabetic macular edema (DME) due to its high-resolution imaging and precise quantitative assessments of retinal thickness and microscopic pathological changes (Olson et al., 2013; Waheed and Duker, 2013). Compared to traditional methods such as slit-lamp biomicroscopy and fundus photography, OCT provides a more detailed evaluation, improving diagnostic accuracy and facilitating personalized therapeutic strategies (Koozekanani et al., 2000; Salz and Witkin, 2015). However, interpreting OCT scans requires specialized expertise, and the shortage of experienced ophthalmologists often leads to diagnostic inefficiencies. This underscores the need for automated OCT-based prediction systems to enhance patient stratification and optimize treatment decisions.

The integration of artificial intelligence (AI) into ophthalmology has driven advancements in automated diagnosis (Shi et al., 2023; Chen et al., 2024b,a). While OCT is the preferred imaging modality for DME diagnosis, assessing therapeutic responses remains challenging and often prone to variability. AI-powered automated analysis can alleviate the burden on ophthalmologists by improving diagnostic accuracy, standardizing assessments, and enhancing clinical efficiency (Sun et al., 2022). Based on this concept, numerous studies have focused on the use of AI for intelligent diagnosis of DME via OCT. Tang et al. proposed a deep learning system for the automated classification of DME based on both three-dimensional (3D) volume scans and 2D B-scan OCT images, yielding volume scan-level results for each eye (Tang et al., 2021). Midena et al. detected and quantified different OCT biomarkers in the eyes of DME patients, proposing a reliable and reproducible AI algorithm tool (Midena et al., 2023). Tripathi et al. employed an automatic method based on Generative Adversarial Networks (GANs) to generate OCT B-scan images of DME, to enhance the reliability of the DME detection system (Tripathi et al., 2023).

However, most existing studies focus on the classification of DME rather than predicting treatment outcomes for patients undergoing anti-VEGF therapy. Moreover, the majority of related datasets are private (Tang et al., 2017, 2021; Midena et al., 2023), and publicly available datasets are extremely limited in scale. For instance, the Duke Eye dataset (Srinivasan et al., 2014) includes only 15 subjects with DME, making it insufficient for robust deep learning research. A significant challenge in this field is the lack of large, publicly accessible datasets, which not only hinders the development and validation of deep learning models for predicting treatment responses, but also makes it difficult to fairly compare the performance of different approaches. This data scarcity restricts researchers' ability to design and evaluate algorithms effectively, ultimately limiting the clinical translation and advancement of AI-driven approaches for DME treatment prediction.

To address this gap, we collected and released a large-scale, publicly accessible OCT dataset specifically aimed at facilitating AI research for predicting treatment outcomes in DME patients undergoing anti-VEGF therapy. Additionally, we organized the 2nd Asia-Pacific Tele-Ophthalmology Society (APTOS) Big Data Competition in 2021. This competition aimed to foster advancements in AI-driven diagnostic algorithms for ophthalmic OCT, particularly enhancing predictive accuracy for patient responses to anti-VEGF treatments. The overarching objective of this initiative is to support ophthalmologists in developing individualized treatment plans and promoting timely and effective patient care. In this paper, we introduce the dataset, describe the competition setup, summarize the participating solutions, and present the results and discussion.

The remainder of this paper is organized as follows: Section 2 reviews related work, including previous challenges and AI methods for OCT image analysis. Section 3 introduces the OCT4DME dataset, while Section 4 describes the competition setup. Section 5 summarizes the participating solutions, and Section 6 reports the results. Section 7 discusses key findings, limitations, and clinical implications, followed by conclusions and future directions in Section 8.

2. Related work

In this section, we review previous challenges and associated datasets for retinal image analysis and summarize existing AI methodologies applied to OCT image analysis.

2.1. Review of previous challenges

Several public challenges have been organized for retinal image analysis (Table 1), with the majority focusing on color fundus photography (CFP). To the best of our knowledge, the Retinopathy Online Challenge (ROC) held in 2009 was the first competition targeting retinal images for automated microaneurysm detection (Niemeijer et al., 2009). Subsequent initiatives included the Diabetic Retinopathy Detection competition on Kaggle in 2015 (Dugas et al., 2015) and the APTOS 2019 Blindness Detection challenge (Kobat et al., 2022), both aimed at diabetic retinopathy (DR) severity grading. The Indian Diabetic Retinopathy Image Dataset (IDRiD) challenge (Porwal et al., 2020) extended the scope to include both DR grading and lesion segmentation, while DeepDRiD further incorporated image quality assessment and transfer learning from CFP to ultra-widefield (UWF) images (Liu et al., 2022). The DRAC 2022 (Diabetic Retinopathy Analysis Challenge) featured tasks including image quality grading, DR severity classification, and lesion segmentation on ultra-widefield OCTA images (Qian et al., 2024).

Beyond DR, other challenges addressed various ophthalmic conditions. The REFUGE (Retinal Fundus Glaucoma Challenge) and REFUGE2 targeted glaucoma detection and optic disc/cup segmentation using fundus images (Orlando et al., 2020; Fang et al., 2022b). The AIROGS (Artificial Intelligence for RObust Glaucoma Screening) challenge (De Vente et al., 2023) and JustRAIGS (Justified Referral in AI Glaucoma Screening) challenge (Lemij et al., 2023) focused on glaucoma detection, with the latter also involving multi-label classification. Multi-disease classification was explored in the Ocular Disease Intelligent Recognition (ODIR-2019) and Retinal Image Analysis for Multi-Disease Detection (RIADD) challenges (ODI, 2019; Pachade et al., 2025). In the context of pathological myopia, the PAthoLogic Myopia (PALM) challenge provided datasets for both disease detection and lesion segmentation (Fang et al., 2024). The ADAM (Automatic Detection challenge on Age-related Macular degeneration) challenge addressed age-related macular degeneration (AMD), including fovea localization and segmentation of optic disc and macular lesions (Fang et al., 2022a). The AGE (Angle-closure Glaucoma Evaluation) challenge focused on angle-closure classification and scleral spur localization in anterior segment OCT (AS-OCT) images

Table 1
Previous public challenges held for retinal image analysis. OCT = Optical Coherence Tomography (refers to posterior segment OCT unless otherwise specified); CFP = Color Fundus Photography; AS-OCT = Anterior Segment Optical Coherence Tomography; UWF = ultra-widefield; OCTA = Optical Coherence Tomography Angiography; FFA = Fundus Fluorescein Angiography.

Year	Competition Name	Modality	Task	Dataset	Link
2015	Diabetic Retinopathy Detection (2015)	CFP	DR grading (5-level classification)	35k images	https://www.kaggle.com/c/diabetic-retinopathy-detection
2017	RETOUCH challenge	OCT	Detection and segmentation of retinal OCT fluid	112 OCT volumes (11,334 B-scans)	https://retouch.grand-challenge.org/
2017	Retinal OCT Classification Challenge (ROCC)	OCT	DR detection	165 images	https://rocc.grand-challenge.org/
2018	IDRiD - DR Segmentation & Grading (ISBI 2018)	CFP	Lesion segmentation, grading, disc/fovea localization.	54 images	https://idrid.grand-challenge.org
2018	REFUGE challenge	CFP	Glaucoma detection and disc/cup segmentation	1200 images	https://refuge.grand-challenge.org
2019	PALM: PAthoLogic Myopia Challenge	CFP	Pathological myopia detection (classification) and lesion segmentation (atrophy, etc.)	1200 images	https://palm.grand-challenge.org
2019	AGE challenge	AS-OCT	Angle closure classification and localization of scleral spur	4800 images	https://age.grand-challenge.org/
2019	APTOS 2019 Blindness Detection	CFP	DR grading (5-level classification)	3662 images	https://www.kaggle.com/competitions/aptos2019-blindness-detection
2019	Ocular Disease Intelligent Recognition (ODIR-2019)	CFP	Ocular disease recognition	10,000 images	https://odir2019.grand-challenge.org/
2020	REFUGE2 challenge	CFP	Glaucoma detection and disc/cup segmentation	2000 images	https://refuge.grand-challenge.org/
2020	ADAM challenge	CFP	Detection of AMD, localization of fovea, and segmentation of optic disc and lesions	1200 images	https://amd.grand-challenge.org/
2020	ISBI 2020: Diabetic Retinopathy-Grading and Image Quality Estimation Challenge	CFP + UWF	DR grading, image quality estimation and transfer learning (cross-domain DR grading)	2000 CFP and 256 UWF images	https://isbi.deepdr.org
2021	RIADD challenge	CFP	Ocular disease screening and classification	3200 images	https://riadd.grand-challenge.org/
2021	GAMMA Challenge	CFP + OCT	Glaucoma grading, fovea localization, and disc/cup segmentation	300 patients (CFP + OCT)	https://gamma.grand-challenge.org
2022	AIROGS challenge	CFP	Glaucoma detection	113,893 images	https://airogs.grand-challenge.org/
2022	DRAC 2022 Diabetic Retinopathy Analysis Challenge	OCTA	Image quality assessment, DR grading and lesion segmentation	1103 UWF-OCTA images	https://drac22.grand-challenge.org
2023	2023 APTOS Big Data Competition	FFA	Report generation from FFA images	50,000 images	https://2023.asiateleophth.org/big-data-competition/
2024	JustRAIGS challenge	CFP	Glaucoma detection and multi-label classification	110k images	https://justraigs.grand-challenge.org/
2024	2024 APTOS Big Data Competition	CFP + OCT	Generation of OCT images from CFP	1271 pairs of CFP and OCT images from 693 patients	https://2024.asiateleophth.org/big-data-competition/

(Fu et al., 2020). More recently, the 2023 APTOS Big Data Competition focused on report generation from fundus fluorescein angiography (FFA) images (Xu et al., 2025).

In addition to CFP-based challenges, a number of competitions have been devoted to OCT images. For instance, the Retinal OCT Fluid Challenge (RETOUCH) (Bogunović et al., 2019) addressed the detection and segmentation of intraretinal fluid (IRF), subretinal fluid (SRF), and pigment epithelial detachment (PED) in retinal OCT scans. A total of 112 macula-centered OCT volumes were included, with macular edema attributed to AMD in half of the cases and to retinal vein occlusion (RVO) in the other half. The Retinal OCT Classification Challenge (ROCC) (Rasti and Rabbani, 2017) focused on DR detection using OCT, while the GAMMA challenge (Wu et al., 2023) encompassed glaucoma grading, fovea localization, and optic disc/cup segmentation using both CFP and OCT modalities. Notably, the 2024 APTOS Big Data Competition emphasized OCT image generation from CFP (APT, 2024).

2.2. AI methods for OCT image analysis

The application of AI for OCT image analysis has progressed from traditional machine learning to deep neural networks (Akpinar et al., 2024; Sonobe et al., 2019), with a more recent focus on large fundamental models (Mihalache et al., 2024). Early OCT image analysis used traditional machine learning, manually extracting features(thickness of the retina, volume of the pathologies (Hussain et al., 2018)) for algorithms (logistic regression (Chen et al., 2006), PCA (Seeböck et al., 2018), SVMs, Random Forests (Hussain et al., 2018)) and use them for disease classification and basic segmentation. However, the performance of these methods was limited by the quality of feature engineering and generalization ability. Subsequently, deep learning with CNNs became dominant. Successful architectures like VGG (Awais et al., 2017; Feng et al., 2020) and ResNet (Wang et al., 2019; Kamble et al., 2018) can classify diseases, such as DR and AMD, in OCT images by extracting complex image features directly, often with better performance compared to older techniques. Furthermore, the development and application of specialized architectures such as U-Net and its variants have made great contributions to retinal layer detection and anatomical segmentation (Shen et al., 2023), providing remarkable performance (Ndipenoch et al., 2024) on commonly used datasets.

Recent AI advancements in retinal image analysis focus on large fundamental models, including pre-trained models (Shi et al., 2024), multimodal models (Chen et al., 2024a), and autoregressive models (Zhang et al., 2024), to improve generalizability and address data scarcity. Multimodal data integration(Shi et al., 2025) shows promise for better zero-shot capability in long-tail learning scenarios. GPT-4V and chatbot (Mihalache et al., 2024; Xu et al., 2024; Cappellani et al., 2024) interpretations also demonstrate significant potential for more comprehensive ophthalmic analysis, extending beyond image interpretation to case report generation and preliminary treatment recommendations, yet concerns regarding clinical applicability and interpretability persist.

3. Dataset: OCT4DME

3.1. Data overview

The OCT4DME dataset was utilized in the APTOS 2021 Big Data Competition, which was primarily centered on predicting the outcomes of anti-VEGF treatment in patients with diabetic macular edema (DME). This dataset consists of tens of thousands of OCT images from 2000 patients, both before and after undergoing initial anti-VEGF treatments. These treatments encompass a group of medications designed to reduce the growth of new blood vessels or swelling and are employed in the treatment of various eye conditions characterized by neovascularization or edema in the macular region of the retina. Authorized users of the APTOS cross-country datasets have the privilege of utilizing these datasets for constructing machine learning models and training their AI algorithms.

3.2. Data acquisition

The data was provided by the Asia Pacific Tele-Ophthalmology Society, Rajavithi Hospital (Thailand), and an anonymous Chinese cohort. Data was collected between January 1, 2015, and December 31, 2020. Inclusion criteria was as follows: (1) patients aged 18 years or older diagnosed with macular edema requiring intravitreal treatment; (2) patients who received at least one intravitreal anti-vascular endothelial growth factor (anti-VEGF) injection and completed a follow-up visit at least 6 months after the initial injection; and (3) patients who underwent OCT imaging within 3 months prior to the first anti-VEGF injection. Exclusion criteria included: (1) absence of recorded visual acuity; (2) a history of retinal treatment involving laser therapy or intraocular surgery; and (3) presence of coexisting ocular conditions that could affect visual acuity (e.g., advanced glaucoma, diabetic retinopathy, or significant media opacity).

OCT images with a resolution of 768×768 pixels were acquired using a 25-line raster scan protocol on the Heidelberg Spectralis OCT2 plus system (Heidelberg Engineering, Germany). In instances where raster scans were unavailable, 6-line radial scans were utilized as an alternative. Collection and annotation for OCT4DME were carried out in two stages. Specifically, during the first stage of the competition, images and labels from 2366 eyes were employed for training set, while an additional 361 eyes were designated for testing set. In the subsequent stage, training set involved scans from 221 eyes, with the remaining 342 eyes reserved for testing set. OCT4DME is distributed under CC BY-NC-SA 4.0.

3.3. Data statistics

Fig. 1 shows the two parts in each image from OCT4DME. The left is the fundus image, with green (or red) lines indicating the positions where the tomography images are taken. The right is the OCT image, which renders an in vivo cross-sectional view of the retina. The bright line with an arrow in the fundus image shows the position of the scan line for this OCT image. The full number of images (scans) for each eye (pre-treatment or post-treatment) is 6 with radial scan style or 19/25/31 with horizontal scan style. The detailed dataset characteristics are demonstrated in Table 2.

3.4. Data annotation

Automated central subfield thickness (CST) measurements were obtained using the Heidelberg Spectralis software. All automated segmentations were independently reviewed and confirmed by trained graders. In cases where only six-line raster scans were available rather than volume scans, central retinal thickness (CRT) was manually measured as the perpendicular distance from the internal limiting membrane (ILM) to the outer border of the retinal pigment epithelium (RPE). Image annotations were performed by two retina fellows-in-training under the supervision of a retinal specialist with over 25 years of experience (PR). Annotations were iteratively reviewed and refined until consensus was reached among the graders. For quality assurance, approximately 20 % of the annotated images were randomly selected and audited by PR. The labels for each eye are listed in Table 3. For eyes in the first stage, all the labels are collected in a per-eye style, where Intraretinal Fluid before treatment (preIRF), Subretinal Fluid before treatment (preSRF), Pigment Epithelial Detachment before treatment (prePED), Hyper Reflective Foci before treatment (preHRF), Intraretinal Fluid after treatment (IRF), Subretinal Fluid after treatment (SRF), Pigment Epithelial Detachment after treatment (PED) and Hyper Reflective Foci after treatment (HRF) are annotated once for the whole group of scans. However, in the second stage, IRF, SRF, PED, and HRF are annotated in a per-picture style. Each scan has its own annotations for these 4 items. The completeness of the annotations for the validation set in each of the two stages is guaranteed to facilitate evaluation. For eyes in the training set, some labels were not available due to a limited number of missing scans. To ensure training consistency and model robustness, we excluded these eyes with incomplete scans from the training dataset. Specifically, in the first stage, 1–3 images for an eye were missing, while only 2–10 horizontal scans were available for each eye in the subsequent stage.

4. Competition setup

4.1. Challenge organizers

The 2021 Asia Pacific Tele-Ophthalmology Society (APTOS) Big Data Competition was hosted by the APTOS and co-organized by Rajavithi Hospital in Thailand, Aravind Eye Hospital in India, and Zhongshan Ophthalmic Center at Sun Yat-sen University in China. It was sponsored and recognized by the Medical Services Department of the Ministry of Public Health of Thailand, and was exclusively supported by the Alibaba Cloud Tianchi platform.

4.2. Competition overview

The APTOS big data competition was officially launched from 28 Aug 2021 to 24 Jan 2022, as part of the 6th Asia-Pacific Tele-Ophthalmology Society (APTOS 2021) Symposium. The competition focused on predicting treatment outcomes for patients with DME undergoing anti-VEGF therapy based on OCT images. More details are available at https://2021.asiateleophth.org/big-data-competition/. Fig. 2 shows the overall competition organization workflow.

Participants with the best scores on the private leaderboard were eligible for cash prizes: US$8,000 for first place, US$5,000 for second place, and US$2,000 for third place. To qualify for the monetary awards, winning teams were required to present their solutions during the "Share & Tell" session of the APTOS Monthly Webinar on February 12, 2022. Members affiliated with the organizing institutions were allowed to participate in the competition; however, they were not eligible to receive any awards.

The competition comprised three stages: the preliminary round, the final round, and the final presentation. In the preliminary round, participants accessed the training dataset and preliminary test dataset from the APTOS website. They were required to debug their algorithms locally and submit their results. This round spanned three months, dur-

Table 2
Summary statistics of the OCT4DME dataset. The dataset consists of training and testing subsets across two competition stages.

Statistic	Train Set	Test Set
Dataset composition		
Stage 1: Eyes (train / test)	2366	361
Stage 2: Eyes (train / test)	221	342
Total eyes	2587	703
Scan protocol	Raster (25-line), Radial (6-line)	Raster (25-line), Radial (6-line)
Visit and scan statistics		
Visits per patient (mean ± std; median; min–max)	2.00 ± 0.00	2.00 ± 0.00
Scans per visit/case (mean ± std; median; min–max)	6.75 ± 3.09; median 6; range 1–25	6.57 ± 2.92; median 6; range 1–16
Demographics (Stage 2 counts)		
Age (mean ± std; range)	60.2 ± 10.6 (24–95)	61.6 ± 10.7 (26–92)
Gender distribution (% of Stage 2 eyes)	Female: 32.2 %; Male: 66.9 %; Unknown: 0.9 %	Female: 34.6 %; Male: 65.4 %
Diagnosis distribution (% of Stage 2 eyes)		
CNVM	30.8 %	33.3 %
DME	62.9 %	62.9 %
PCV	6.3 %	3.8 %
Treatment / anti-VEGF distribution (% of Stage 2 eyes)		
Accentrix	4.1 %	4.1 %
Avastin	77.4 %	78.4 %
Eylea	1.4 %	1.5 %
Ozurdex	1.9 %	2.9 %
Pagenax	0.5 %	–
Razumab	5.0 %	5.6 %
Tricort	10.0 %	7.6 %

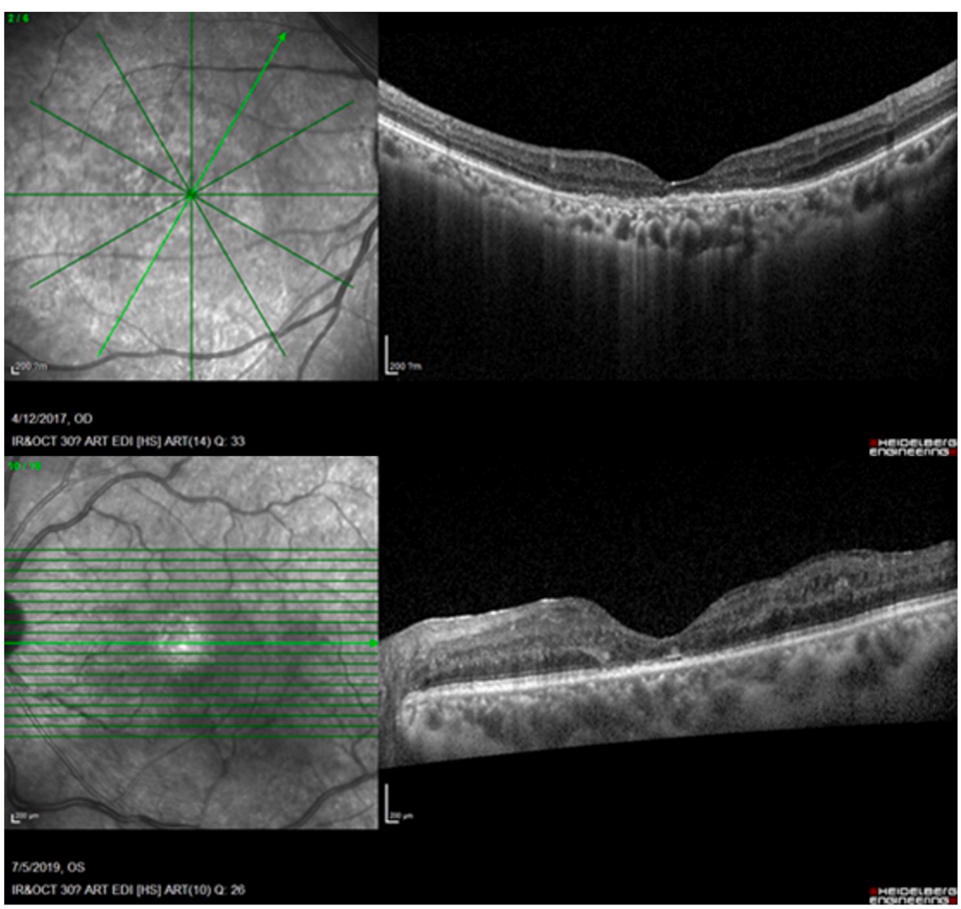

Fig. 1. An example from the dataset OCT4DME.

Table 3
Labels for each eye in the OCT4DME dataset.

Label	Annotation	Type
General Information		
patient ID	PatientID + EyeSide (L: left/R: right)	categorical
gender	1 = male, 2 = female	categorical
age	age of the patient	continuous
diagnosis	1 = wet AMD (CNVM), 2 = PCV, 3 = DME, 4 = RVO, 5 = CME, 6 = fellow eye, 9 = other diagnosis	categorical
anti-VEGF	1 = bevacizumab, 2 = ranibizumab, 3 = aflibercept, 4 = conbercept, 5 = triamcinolone, 6 = combination, 9 = other, 0 = none, (Ozurdex, Pagenax available in stage 2)	categorical
Pre-treatment biomarkers		
preVA	Visual Acuity (logMAR) before treatment	continuous
preCST	Central Subfield Thickness before treatment	continuous
preIRF	Intraretinal Fluid before treatment	binary
preSRF	Subretinal Fluid before treatment	binary
prePED	Pigment Epithelial Detachment before treatment	binary
preHRF	Hyper Reflective Foci before treatment	binary
Post-treatment biomarkers		
VA	Visual Acuity (logMAR) after treatment	continuous
CST	Central Subfield Thickness after treatment	continuous
IRF	Intraretinal Fluid after treatment	binary
SRF	Subretinal Fluid after treatment	binary
PED	Pigment Epithelial Detachment after treatment	binary
HRF	Hyper Reflective Foci after treatment	binary
Continue Injection (CI)	Whether anti-VEGF injection is continued	binary

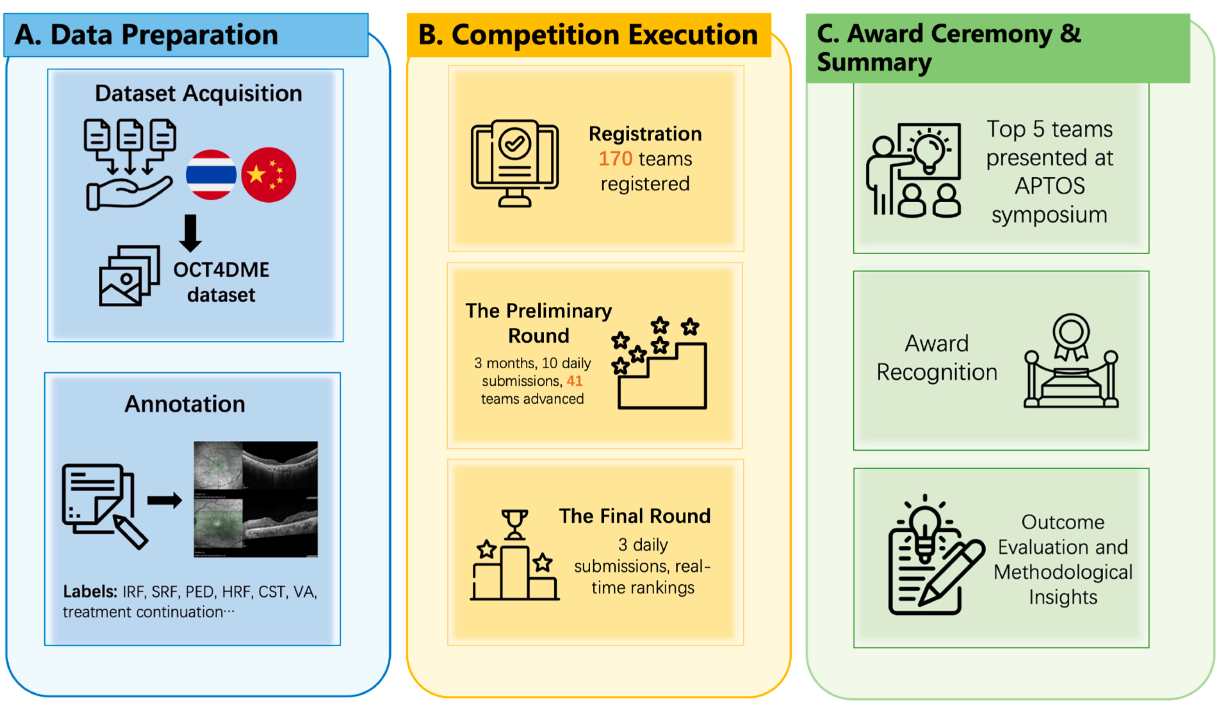

Fig. 2. The overall organization workflow of the 2021 Asia Pacific Tele-Ophthalmology Society (APTOS) Big Data Competition.

ing which each team had the opportunity to make up to 10 submissions per day. A total of 170 teams successfully submitted valid results to the challenge platform. The detailed timeline is as shown in Table 4.

All the 41 teams from the preliminary round advanced to the final round. Teams that successfully reached the final round received links and decryption passwords to download the dataset. In the final round, each team was allowed up to 3 submissions per day, with real-time rankings. The top 5 teams participated in the final presentation in January 2022. During this concluding event, teams presented insights into their AI model development process, training techniques, and innovative ideas, ultimately determining the final rankings and awards for the competition.

4.3. Competition tasks

The competition consisted of two rounds, each featuring distinct tasks focused on DME analysis using OCT. For detailed submission instructions, please refer to the official submission page.

Table 4
Timeline of the 2021 APTOS Big Data Competition.

Date	Event
August 28, 2021	Launch of Competition, registration opens
September 1, 2021	Submission portal opens; training and validation data released
November 19, 2021	Entry deadline, team merger deadline, and preliminary round submission deadline
December 31, 2021	Final submission deadline (Top 100 teams from the preliminary round eligible)
January 21, 2022	Code defense: top 5 finalists undergo code review and panel challenge
February 12, 2022	Prize presentation and winner showcase at the APTOS Monthly Webinar

4.3.1. The preliminary round:

Participants were provided with OCT images both before and after treatment, along with the visual acuity (VA) before treatment. For each patient ID in the dataset, they were tasked with:

1. Predicting the presence of Intraretinal Fluid (IRF), Subretinal Fluid (SRF), Pigment Epithelial Detachment (PED), and Hyperreflective Foci (HRF) in eyes (all referring to post-treatment conditions).
2. Predicting (measuring) the Central Subfield Thickness (CST) values both before and after treatment.
3. Predicting the VA after treatment.
4. Predicting the continuation or discontinuation of anti-VEGF injection.

Although predictions for IRF, SRF and HRF before treatment are not directly evaluated, participants are still required to predict them because predictions for continuing injection may heavily rely on improvements in these symptoms after treatment.

4.3.2. The final round:

The tasks in the final round mirrored those in the first round. However, for the presence of IRF, SRF, PED, and HRF after treatment, participants were required to predict the probability (ranging from 0 to 1) rather than providing a simple yes/no response.

4.4. Evaluation metrics

4.4.1. Rationale for evaluation metrics

To ensure a robust and clinically relevant assessment, the evaluation metrics were chosen based on a combination of machine learning standards and clinical practicality. Importantly, because the input/output formats differed between the Preliminary and Final Rounds, the interpretation of each metric also varies accordingly.

1. AUC for Binary Predictions: While the preliminary round required binary predictions for biomarkers (IRF, SRF, PED, and HRF), the Area Under the Curve (AUC) metric was used for evaluation. However, because both predictions and ground-truth labels were binary in the Preliminary Round, the ROC AUC metric degenerates into a single operating point rather than a full curve. In this constrained setting, the resulting AUC behaves similarly to balanced accuracy and does not reflect a ranking-based discriminative ability. This limitation originated from the official APTOS 2021 competition design, which required binary submissions during the qualification stage. Accordingly, AUC in this round should be interpreted only as a coarse screening metric rather than as a full evaluation of model performance. This constraint does not apply to the Final Round, where continuous probability scores were required and standard ROC and PR curve analyses were used.
2. Tolerance-Based Scoring for Continuous Variables: For continuous outcomes like CST and VA, a tolerance-based scoring system was adopted. This method converts a continuous prediction into a binary outcome (correct/incorrect) based on its proximity to the true value, which more closely reflects clinical decision-making. In a clinical setting, a prediction is considered valuable if it falls within a clinically acceptable margin of error, not just if it is perfectly precise.
3. Threshold Justification: The specific thresholds, such as the $\pm 7.5\,\%$ for CST and VA (Sugar et al., 2011), and the ± 0.05 for VA (Bressler et al., 2019), were chosen to represent a clinically relevant margin of error. These thresholds reflect the level of accuracy required for these predictions to be useful in guiding treatment decisions for DME patients.

4.4.2. The preliminary round:

In this phase, participants utilize the first stage dataset and submit predictions for PreCST, CST, VA, Continue Injection (CI), IRF, SRF, and HRF. The scoring criteria are as follows:

1. The Area Under the Curve (AUC) was used for scoring predictions for Continue Injection, IRF, SRF, and HRF.
2. For PreCST and CST, a score of 1 was awarded if the prediction was within a $\pm 7.5\,\%$ tolerance range of the true value. The score was 0 otherwise. Formally, for a prediction $\hat{y}$ and true value y, the score is 1 if $|\hat{y} - y| \leq 0.075 \times y$, and 0 otherwise.
3. In the case of VA predictions:
   - If the actual VA value is less than 1, a score of 1 is awarded for predictions within the range $[VA - 0.05, VA + 0.05]$. Formally, the score is 1 if $|\hat{y} - y| \leq 0.05$, and 0 otherwise.
   - If the actual VA value is greater than or equal to 1, a score of 1 is given for predictions within the range $[VA \times 0.925, VA \times 1.075]$. Formally, the score is 1 if $|\hat{y} - y| \leq 0.075 \times y$, and 0 otherwise.

4.4.3. The final round:

The Final Round score incorporates participants' results on the second stage dataset. In other words, the score represents the average of the 14 indices submitted (7 indices from the first stage dataset and 7 indices from the second stage dataset). The leaderboard is sorted based on the Final Round score and displays sub-scores for each index to assist participants in optimizing their models.

5. Summary of participating solutions

In this section, we compile the approaches employed by the top three winning teams in the competition. For ease of expression, we divided the 7 indices into three subtasks, as illustrated in Fig. 3.

5.1. Binary classification tasks for IRF, SRF, PED, and HRF

5.1.1. Team BlueSky:

The champion team BlueSky encountered the challenge of transitioning from case-level to image-level labels within the dataset due to feature differences between the two-stage datasets. To address this hurdle, they employed a combination of Swin-Transformer (Liu et al., 2021) and Multi-Instance Learning (MIL) (Carbonneau et al., 2018). The team treated the classification task as an MIL problem, defining a "bag" as either encompassing all images within a case (for the first stage dataset) or each individual image (for the second stage dataset). For input, they extracted macular OCT image parts from the raw images. The Swin-Transformer was chosen as the backbone network, with a focus on macular OCT images and the utilization of max pooling to yield the final results. Additionally, they observed an imbalance in category distribution, particularly for IRF and HRF, between the first and the second

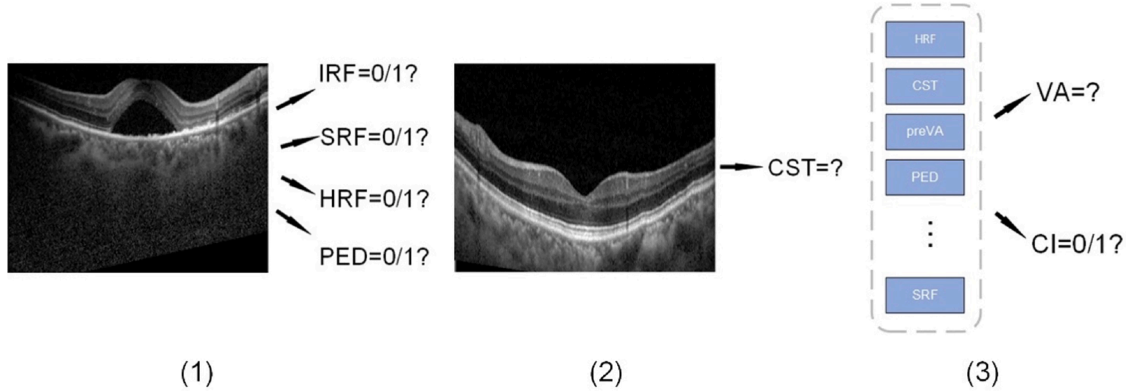

Fig. 3. Three sub-tasks of the competition, which are diagnosing the symptoms of the eye, predicting the Central Subfield Thickness, and judging whether to continue treatment.

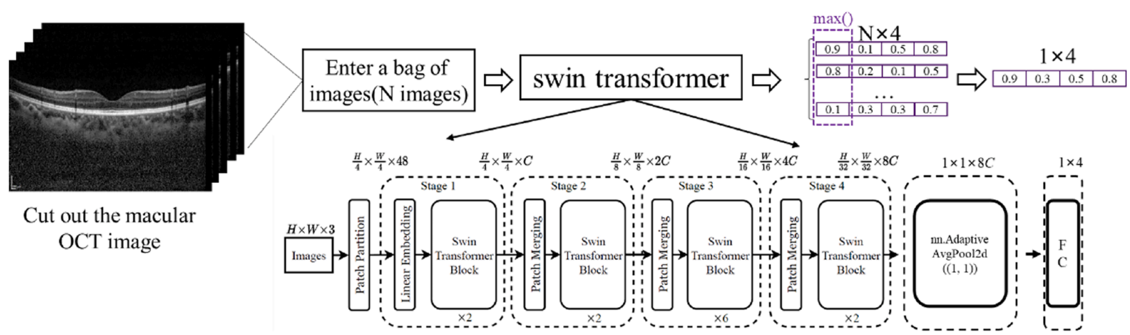

Fig. 4. Solution for Task 1 by the champion team BlueSky. This illustrates the architecture of the Multi-Instance Learning network proposed by the team. Given a bag containing n images, the network outputs an N x 4-dimensional vector, where 4 represents the categories: IRF, SRF, PED, and HRF.

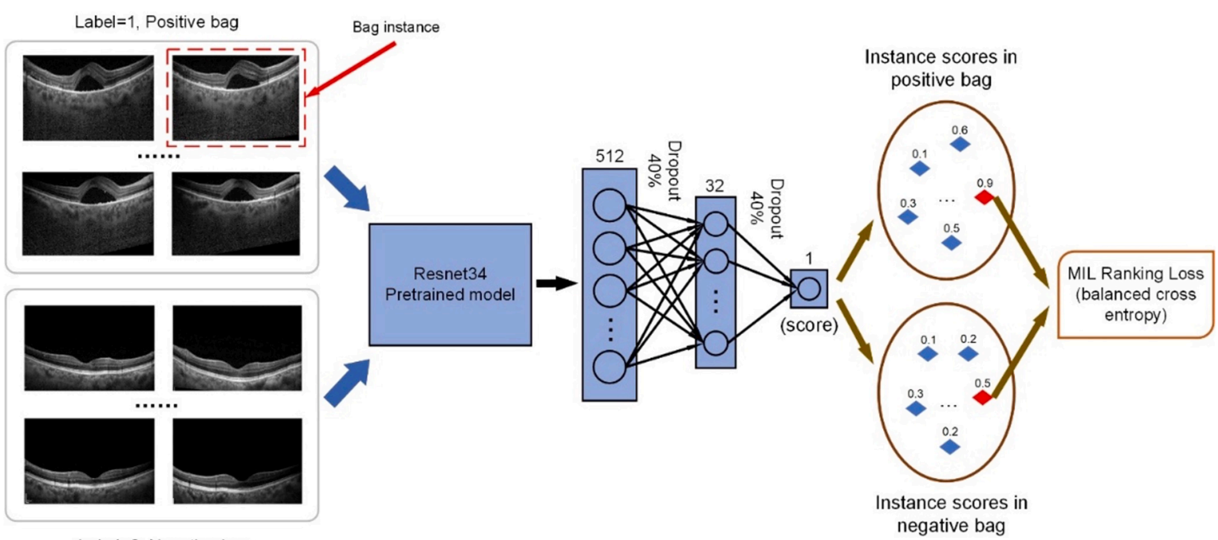

Fig. 5. Solution for Task 1 by the runner-up team LightRain. They used a weakly supervised learning framework for classification.

stage datasets. Taking this into account, the team strategically balanced the category distribution in the merged dataset, contributing to an overall enhanced performance. The network architecture for this section is illustrated in Fig. 4.

5.1.2. Team LightRain:

The runner-up team LightRain addressed the imbalance between positive and negative samples using a weak supervision framework for MIL, as shown in Fig. 5, similar to the approach employed by team BlueSky. They utilized ResNet34 (He et al., 2016) as the base classifier and incorporated dropout to reduce overfitting. They introduced data augmentation to increase the diversity of training data, including random flipping, random cropping, and JPEG compression (Wallace, 1992). The MIL Ranking Loss function (Sultani et al., 2018) and balanced cross-entropy loss were employed for optimization, so that only the samples with the highest score in the bag were used for optimization:

$$L(B) = -\alpha y \log(\max_{i \in B} f(V_i)) - (1-\alpha)(1-y)\log(1 - \max_{i \in B} f(V_i)),\ \alpha \in (0,1)$$

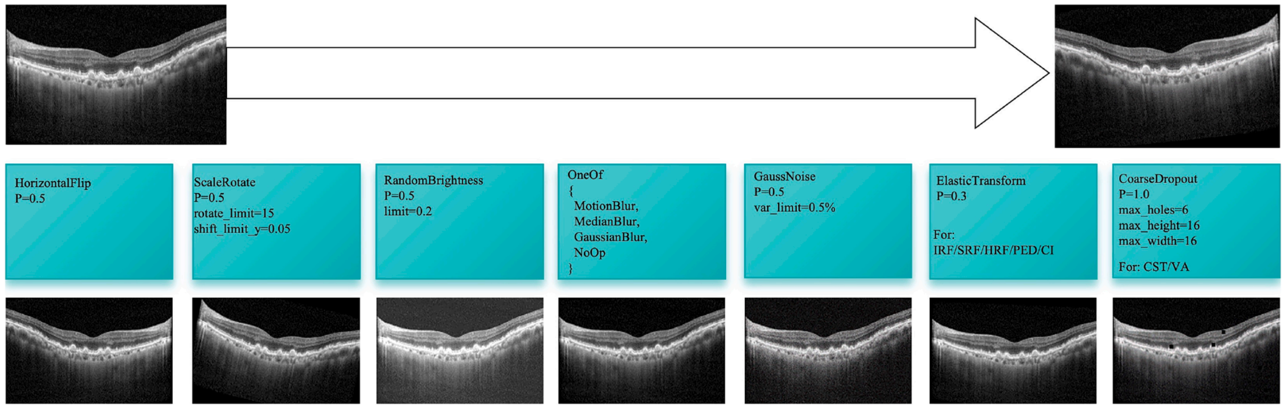

Fig. 6. The flow of data augmentation, proposed by the third-place team DarkStyle.

(1)

The team also implemented Test Time Augmentation (TTA) (Shanmugam et al., 2021) for improved detection.

5.1.3. Team DarkStyle:

Before conducting the experiments, the third-place team DarkStyle used an image augmentation flow to improve the generalization of the models, as shown in Fig. 6. For the prediction of IRF/SRF/HRF/PED, they utilized EfficientNet-B0 (Tan and Le, 2019) as the backbone. A global average pool layer and multilayer perceptron layers were added, and binary cross-entropy loss with label smoothing set to 0.1 was employed to combat overfitting. For each task, they trained a separate model, resulting in a total of four classification models. Particularly, in the prediction of PED, they conducted a gamma transform for image preprocessing. The optimization used the Adam optimizer with a learning rate of 0.001. The team experimented with various backbones, including ResNet50 (He et al., 2016), MobileNetV2 (Sandler et al., 2018), and EfficientNet-B0 (Tan and Le, 2019) in the preliminary round, with EfficientNet-B0 demonstrating the best balance between accuracy and efficiency. Their approach included rigorous validation techniques including the 5-fold cross-validation to ensure robust model performance.

5.2. Regression tasks for central subfield thickness (CST)

5.2.1. Team BlueSky:

In tackling the regression task of CST values before and after treatment, the team BlueSky faced the complexities arising from diverse scanning methods (radial and horizontal) and data irregularities. They devised a robust approach for identifying the central image by computing the sum of rows of the G channel in fundus images. For CST regression, they opted for ResNet101 (He et al., 2016) as the backbone, and introduced a tailored loss function that integrated mean absolute percentage error with smooth L1 loss:

$$M = \left| \frac{\hat{y}_i - y_i}{y_i} \right| \tag{2a}$$

$$\text{loss} = \begin{cases} 0.5 \times \frac{M^2}{\beta}, & M < \beta, \quad \beta = 0.075 \\ M - 0.5 \times \beta, & M \geq \beta, \quad \beta = 0.075 \end{cases} \tag{2b}$$

During data input, they focused on the central third of the macular OCT image and performed random cropping within a range of 50 pixels horizontally. To improve model generalization, the training process encompassed distinct stages for radial and horizontal scans. The ultimate model was refined through iterative training steps, as illuminated in Fig. 7, culminating in a robust and computationally efficient solution. In the initial phase, both radial and horizontal scans were initially trained together to ensure an ample sample size. Team BlueSky resized the image to 224, employed Imagenet (Deng et al., 2009) pre-training parameters, and trained for 230 epochs to achieve quick convergence, resulting in model 0. Subsequently, they resized the image to 512, used the parameters from model 0, and trained for 140 epochs to enhance image resolution and reduce regression errors, yielding model 1. In the second phase, they conducted separate training for radial and horizontal scans. Using an image size of 512 and the parameters from model 1, they trained the two scans independently. The predictions from the two scans were then combined to generate the final result.

5.2.2. Team LightRain:

For the regression task on CST, the runner-up team LightRain utilized Densenet169 as the backbone (Huang et al., 2017). They encountered challenges related to the quality of labeled data. To address this, they implemented a refinement strategy: during a preliminary test, samples with substantial discrepancies (exceeding 15 %) between the predicted and actual values were identified and removed from the training set. The CST values were normalized, and the Mean Absolute Error (MAE) loss served as the loss function. Random inversion was applied as a data augmentation technique to mitigate CST's sensitivity to target size.

5.2.3. Team DarkStyle:

For CST regression, the third-place team DarkStyle applied a key strategy to improve the performance by excluding horizontal scan images far from the center point. This involved analyzing the left part of the original OCT images and applying distance and selection criteria to filter out relevant images. The optimal parameters for training and prediction were determined separately for the images from the first and the second stage datasets. EfficientNet-B0 was utilized as the backbone.

5.3. Prediction of patients' visual acuity and treatment continuation

5.3.1. Team BlueSky:

In predicting patients' VA and treatment continuation, BlueSky integrated the outcomes from the previous tasks. Recognizing the potential impact of errors in both classification and regression tasks, they implemented a robust approach using a voting regressor for enhanced stability, as shown in Fig. 8. The decision-making process considered 15 features, including classification results, regression results, and sponsor-provided features. The ensemble strategy involved multiple regressors, such as Gradient Boosting Regressor (Friedman, 2001), Random Forest Regressor (Segal, 2004), and Linear Regression, contributing to the overall stability and reliability of the model.

5.3.2. Team LightRain:

The team LightRain also combined results from the first two tasks with patient information to form the input feature vectors. For

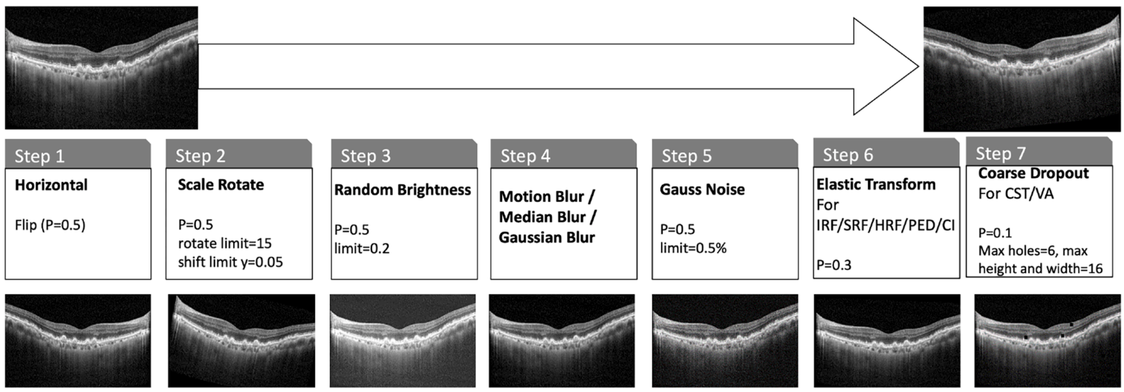

Fig. 7. Solution for Task 2 by the champion team BlueSky. They used a two-step training progress for Central Subfield Thickness (CST) regression.

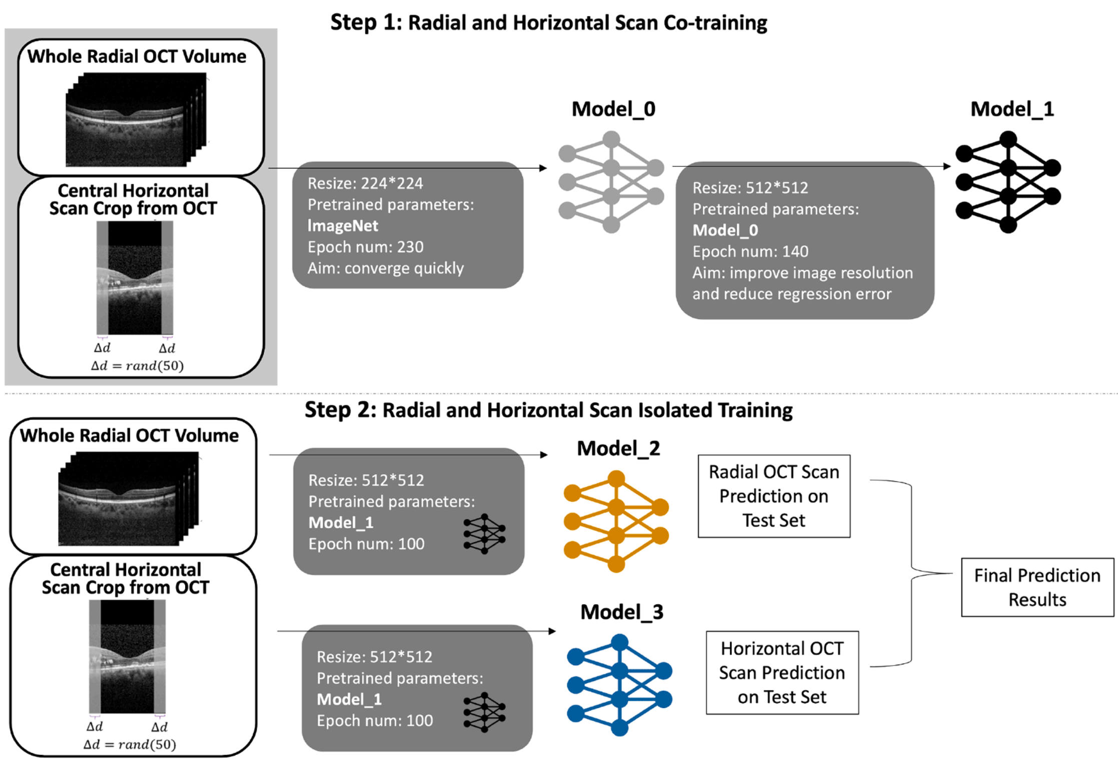

Fig. 8. The ensemble strategy for predicting visual acuity and treatment continuation proposed by the champion team BlueSky.

preprocessing, they conducted one hot encoding for the disease types and injection drug types. A three-layer neural network was utilized to predict VA changes, and an integrated learning framework Gradient Boosting was applied to predict whether patients need continuous injection therapy.

5.3.3. Team DarkStyle:

The third-place team approached VA prediction as a classification task for the first-stage data and a regression task for the second-stage data. Both stages jointly used image features and patient metadata. To predict the need for continued injections, this team performed an in-depth analysis of metadata, dropping low correlation columns such as gender. They leveraged the diagnosis value equal to 6 as an indicator of a fellow eye, transforming CST values accordingly. The overall architecture is illustrated in Fig. 9. The team explored simple patterns based on observations in training metadata and introduced a deep learning model for image prediction and an XGBoost Regression model for metadata prediction in the first stage. For the second stage, eight different models were used for metadata, and the results were weighted and combined to enhance accuracy. The comparison of the key methods adopted by the top three teams across the three tasks can be found in Table 5.

6. Results

6.1. Data distribution

The disparity in category distribution between the preliminary and final datasets, particularly in the cases of IRF and HRF, raises interesting considerations. Fig. 10 reveals a substantial shift in the proportions, with IRF increasing from 32.2 % in the first-stage dataset to 84.1 % in the second stage, and finally stabilizing at 52.4 % after the combination. Likewise, HRF starts at 30.3 % in the first stage, surges to 95.7 % in the second stage, and then settles at 55.8 % post-combination. This observed shift highlights the impact of dataset merging on achieving a more

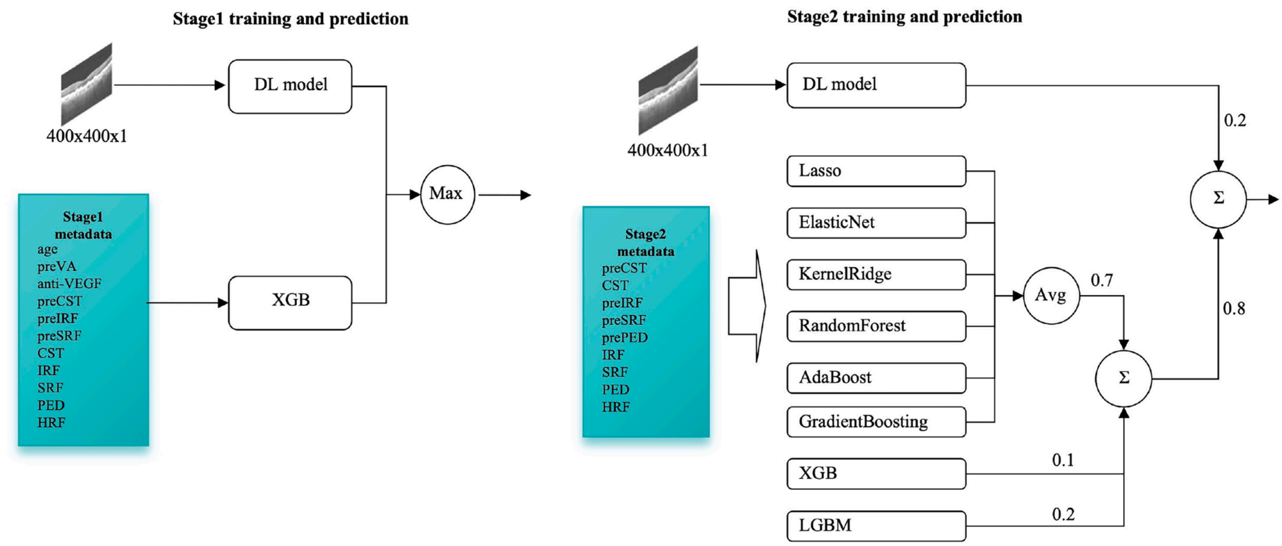

Fig. 9. Solution for predicting continued injection by the third-place team DarkStyle. They used a two-step training progress for Central Subfield Thickness (CST) regression.

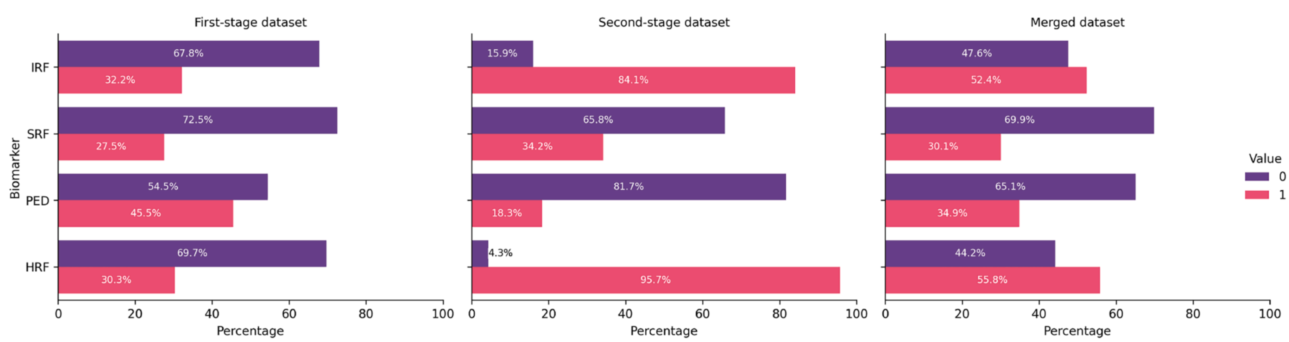

Fig. 10. The distribution of data for IRF (Intraretinal Fluid), SRF (Subretinal Fluid), PED (Pigment Epithelium Detachment), and HRF (Hyperreflective Foci) in datasets at different stages.

Table 5
Summary of the key methods adopted by the top three teams across the three tasks.

Task	BlueSky (1st Place)	LightRain (2nd Place)	DarkStyle (3rd Place)
Task 1: Binary Classification (IRF, SRF, PED, HRF)	Swin-Transformer + MIL; macular cropping; class rebalancing	ResNet34 + Weakly-supervised MIL; ranking loss; augmentation; TTA	EfficientNet-B0; separate binary classifiers; label smoothing; gamma transform for PED
Task 2: Regression (CST)	ResNet101; two-stage training (radial/horizontal); hybrid loss (MAPE + smooth L1); cropping central region	DenseNet169; outlier removal (>15% error); MAE loss; random inversion augmentation	EfficientNet-B0; exclude off-center horizontal scans; task-specific parameter tuning
Task 3: Prediction (VA & Treatment Continuation)	Ensemble (Gradient Boosting, RF, Linear Regression); 15 features from previous tasks; voting regressor	Three-layer NN + Gradient Boosting; one-hot metadata encoding; integration of Task 1 & 2 results	Multi-stage pipeline; metadata refinement (drop low-corr cols); DL model + XGBoost ensemble; stage-specific training

balanced sample distribution. It prompts exploration into the implications of category distribution dynamics on the performance of models across different stages of the competition.

6.2. Overall results

Table 6 displays the top ten teams based on their overall scores. We present their performances in both the preliminary round (evaluation on the first stage test set) and the final round (evaluation on the merged test set). A noteworthy observation is the substantial increase in scores for all teams from the preliminary round to the final round. Teams including BlueSky, GuardianHairline, c_c, HowToImproveScore, and UnityTeam maintain or improve their rankings on the test dataset, indicating their robustness compared to other methods. Conversely, teams with lower rankings in the final round generally exhibit poorer generalization capabilities. Specifically,

Table 6
Grading results for the 2021 APTOS Big Data Competition. Teams are ranked based on the overall score in the final round. The numbers in parentheses after the scores indicate the team's ranking in the preliminary round, with bolded numbers indicating a rise or no change in the ranking between the preliminary and final rounds, and non-bolded numbers indicating a fall in the rankings.

Rank	Team	Organization	Preliminary	Final
1	BlueSky		0.53 (12)	0.7440
2	LightRain	Xi'an Electronic Science and Technology University	0.56 (1)	0.7399
3	DarkStyle	Visual Microimaging	0.56 (1)	0.7354
4	ChubbyFirst		0.56 (1)	0.7332
5	GuardianHairline	Hangzhou Electronic Science and Technology University	0.46 (31)	0.7323
6	c_c	Chengdu University of Information Technology	0.54 (10)	0.7296
7	MedEngTogether	Tsinghua University	0.56 (1)	0.7261
8	HowToImproveScore	Beijing University of Technology	0.55 (8)	0.7228
9	MedImgProc	South China University of Technology	0.55 (8)	0.7166
10	UnityTeam	Visual Microimaging	0.44 (50)	0.7084

Table 7
Detailed results of the top 3 teams for each subtask, grouped by stage.

Rank	Team	Score	Stage 1								Stage 2						
			preCST	VA	CI	CST	IRF	SRF	HRF	–	preCST	VA	CI	CST	IRF	SRF	PED
1	BlueSky	0.7440	0.5643	0.3216	0.7026	0.5906	0.9293	0.9218	0.8593	–	0.6871	0.5234	0.7828	0.6930	0.9338	0.9264	0.9796
2	LightRain	0.7399	0.5468	0.3216	0.6852	0.5994	0.9221	0.9311	0.8772	–	0.6462	0.5234	0.7800	0.6842	0.9313	0.9300	0.9796
3	DarkStyle	0.7354	0.5965	0.3421	0.6983	0.5994	0.9058	0.8854	0.8485	–	0.6520	0.5234	0.8006	0.6430	0.9074	0.9174	0.9749

LightRain, DarkStyle, ChubbyFirst, MedEngTogether, and MedImgProc experienced a decline in rankings on the final test set.

6.3. Individual results

The detailed results of the top 3 teams for each subtask are demonstrated in Table 7. The champion team, BlueSky, showcased a remarkable overall performance with a score of 0.7440. Particularly noteworthy was their strength in CST during the first round, where they achieved a high score of 0.7026, demonstrating the effectiveness of their regression approach.

The runner-up team, LightRain, achieved a commendable score of 0.7399. They exhibited strength in predicting whether to Continue Injection or not in the second stage with a score of 0.7800, highlighting the high clinical applicability of their method. Furthermore, they maintained high scores in the first stage SRF (0.9311) and the first stage IRF (0.9221), indicating their excellent performance in classification tasks.

The third-place team, DarkStyle, secured a score of 0.7354 and demonstrated noteworthy strengths in the second stage prediction. They predicted the second stage CST with a score of 0.8006. They also displayed strong results in the second stage SRF (0.9174) and in the second stage IRF (0.9074), contributing to their solid overall performance and ranking.

Overall, each team displayed unique strengths in various aspects of the competition. BlueSky excelled in CST and SRF, LightRain showcased competitiveness in CI, and DarkStyle demonstrated strengths in CST and SRF. These three winning teams maintained a commendable overall performance throughout the competition, highlighting their consistency and expertise.

The results showcase exceptional performance in various tasks: The decision consistency for Continued Injection is notable, with AI aligning with ophthalmologists in 80.06 % of the cases. In PED detection, an impressive AUC of 97.96 % was achieved. SRF detection and IRF detection exhibited high AUC rates of 93.68 % and 93.38 %, respectively. For CST prediction before treatment and after treatment, the model achieved AUCs of 68.71 % and 69.30 %, respectively. VA prediction after treatment maintained a commendable AUC of 55.56 %.

In summary, the models developed by the participants exhibited outstanding performance across a range of crucial parameters, underscoring their versatility and effectiveness in tackling intricate OCT tasks. This underscores the potential of utilizing machine learning to address the challenges associated with DME, particularly in predicting patient responses before treatment, enabling personalized treatment plans, and improving patient management.

7. Discussions

In this section, we summarize the characteristics, experiences, and insights gained from the methods employed during the competition. Additionally, we analyze the value of artificial intelligence and machine learning approaches in DME intervention, along with feasible future works.

7.1. Data preprocessing and analysis

A solid foundation lies in data analysis. The top two teams conducted meticulous analyses of the dataset distribution before training and took steps to equalize sample distribution. Balanced datasets ensure that the model does not develop biases toward more frequent classes, improving both the generalizability and reliability of predictions. This is particularly important in medical applications, where model errors could directly impact patient care and treatment plans. The runner-up team LightRain implemented a two-stage training approach to pre-filter noise samples, ensuring enhanced training quality.

To maximize data utilization, all winning teams designed their data augmentation methods, tailoring them for different tasks. The third-place team DarkStyle introduced seven distinct augmentation techniques, creating an augmentation flow for diverse and enriched training data, as in Fig. 6.

Notably, the avoidance of image compression in regression tasks and the preference for larger image sizes, especially in medical datasets, emerge as valuable insights gleaned from the competition.

7.2. Model innovations and generalization

The competition results demonstrate that models need not be overly complex; rather, well-established deep learning architectures can effectively address the competition tasks. Importantly, several methodological insights proved particularly well suited to OCT-based treatment response prediction. Multi-instance learning (MIL) consistently reconciled image- and case-level supervision. Metadata fusion strategies, such as incorporating fellow-eye indicators or baseline CST values, provided

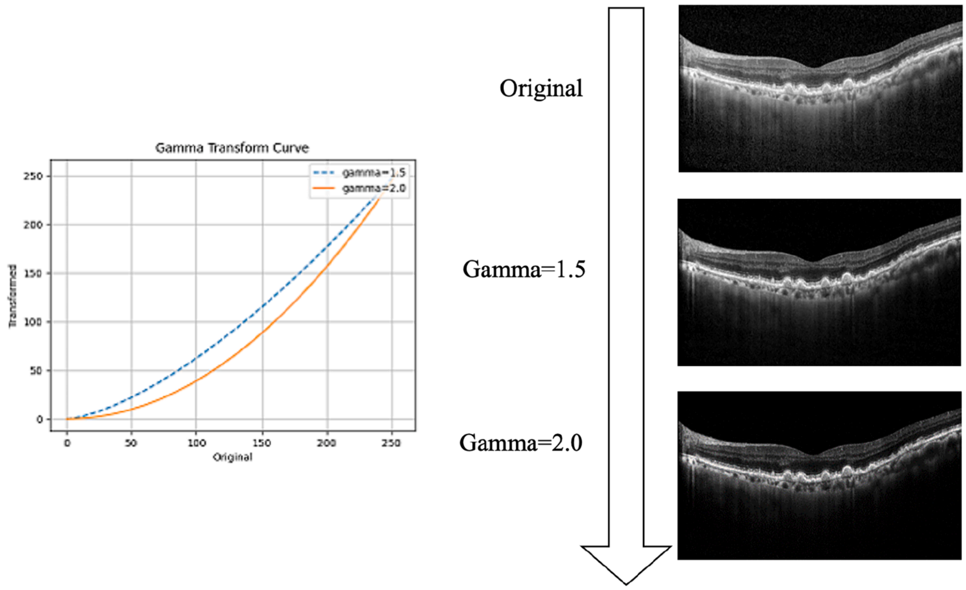

Fig. 11. Gama transform of the OCT images.

complementary signals that improved prediction accuracy. Moreover, task-aware augmentation and ranking-based loss functions effectively addressed label sparsity and enhanced robustness.

For example, the champion team BlueSky combined Swin-Transformer with an MIL framework to reconcile image-level and case-level annotations, while also addressing class imbalance through dataset rebalancing. The runner-up team LightRain similarly adopted MIL with a ResNet34 backbone, complemented by weak supervision, extensive augmentation, and a ranking-based loss. Their use of test-time augmentation further improved robustness. In contrast, the third-place team DarkStyle emphasized systematic preprocessing, tailored augmentation pipelines, and lightweight backbones such as EfficientNet-B0, highlighting the importance of balancing accuracy with computational efficiency.

Across tasks, ensemble methods proved robust for integrating heterogeneous predictions, particularly for visual acuity and treatment continuation. Both BlueSky and LightRain leveraged multi-stage pipelines or ensemble strategies to stabilize predictions, demonstrating that thoughtful integration of intermediate task outputs can substantially enhance downstream performance. Taken together, these results suggest that top teams succeeded not through complexity, but through methodological precision, careful data management, and task-aware model design.

7.3. Clinical findings and implications

In the context of predicting PED, the team DarkStyle observed a notable enhancement in the AUC by approximately 0.3 % with the application of a gamma transform. The gamma transform, characterized by a transform coefficient greater than 1.0, accentuates dark pixels while dimming light pixels, as illustrated in Fig. 11. This improvement can be attributed to that PED symptoms predominantly manifest in the bottom-lightest part of the retina. By applying the gamma transform, these subtle features become more distinguishable, facilitating more accurate identification of PED.

To unravel inter-variable relationships, a correlation matrix was plotted for all metadata columns, and low correlation columns, such as gender, were subsequently dropped, as depicted in Fig. 12. Notably, a diagnosis value of 6 corresponds to the fellow eye, signifying a healthy eye or one not severe enough to necessitate treatment. Analysis revealed that the mean value of the fellow eye's CST is approximately 250, suggesting deviations from this value may indicate anomalies. Thus, the original CST values were transformed by subtracting 250.

Further exploration involved seeking simple patterns based on training metadata observations, such as:

1. If IRF, SRF, HRF, and PED all disappeared after treatment, does this imply that continuing injections is entirely unnecessary? Contrary to this assumption, 38 cases in the first-stage training data required continued injections, indicating that these features alone are not sufficient for decision-making.
2. Similarly, if IRF, SRF, HRF, and PED all appeared after treatment, does it signify a 100 % unnecessary continuation of injection, suggesting the treatment had no effect? Again, the answer is negative, with 44 continued injection cases found in the first-stage training data.

Exploration extended to more intricate assumptions, including changes in VA, yet no straightforward rules were discovered. This suggests that crucial information influencing the decision to continue injection is embedded in the OCT images, beyond what can be inferred from simple metadata patterns. Future work may benefit from combining metadata analysis with deeper image feature extraction to improve treatment decision-making.

7.4. Advancing towards clinical application

A pivotal question emerges: Can these models be effectively employed in real clinical settings to autonomously predict whether patients with DME should undergo anti-VEGF treatment? While this question remains open, the development of an automated, objective diagnostic system holds promise in mitigating individual biases and conserving significant time for human experts. OCT, being non-invasive and cost-efficient, is extensively utilized by clinical experts for DME monitoring (Panozzo et al., 2003). Leveraging AI methods with OCT for automated anti-VEGF therapy decisions in patients with DME appears to be a viable solution for assisting diagnosis under limited conditions (Wang et al., 2023; Padmasini et al., 2023).

In our competition, these models achieved promising performance, with the best system reaching an AUC of 80.06 %. However, three limitations must be acknowledged. First, the dataset is derived from Asian populations (Thailand, India, and China), which, while clinically valuable, may limit generalizability to more diverse populations. Second, although an AUC above 0.8 indicates potential for decision support (Çorbacıoğlu and Aksel, 2023), it falls short of the higher thresholds (e.g.,

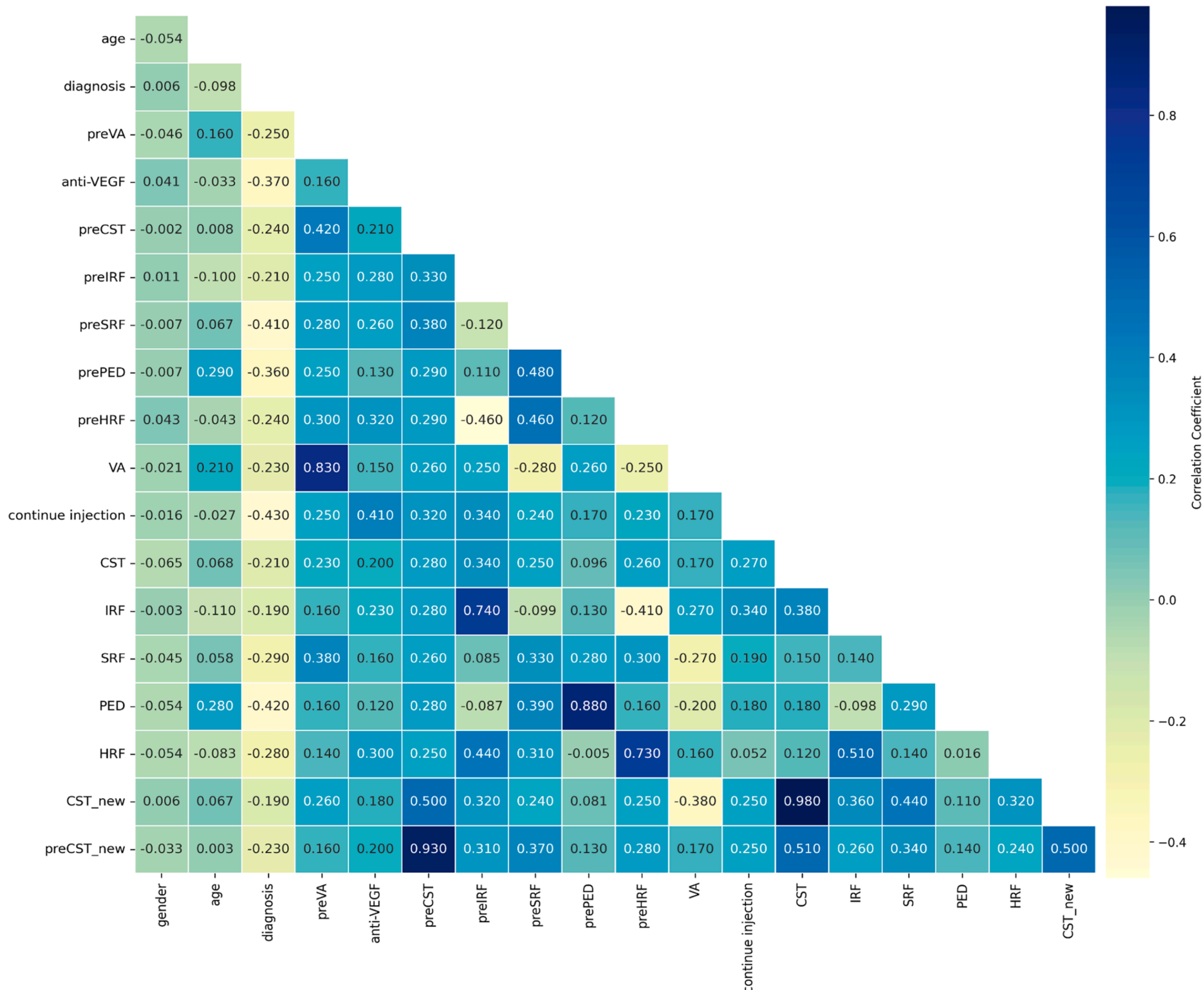

Fig. 12. The correlation matrix for all metadata columns in the dataset.

AUC > 0.9) that are typically required for autonomous clinical adoption (Wu et al., 2022). These considerations highlight the need for external validation across broader demographics and prospective, multi-center studies to ensure safety, reliability, and clinician trust. Besides, a potential limitation of leaderboard-driven competitions is the risk of participants overfitting to the test set through multiple submissions. Although this risk was mitigated by limiting submissions to a maximum of 10 per day in the preliminary round and 3 per day in the final round, and by requiring finalists to present and justify their methods, we acknowledge this as a general limitation of such challenge formats.

To move from retrospective validation towards real-world clinical use, a stepwise pathway is envisioned. This includes validating models on external multi-center datasets, conducting prospective studies to assess real-time performance, and integrating decision-support tools into clinical workflows. Interpretability and uncertainty estimation will be key for clinician trust, while regulatory approval and post-deployment monitoring are essential to ensure safety and effectiveness.

Analysis of submitted solutions suggests several key insights for future development. First, subtle image preprocessing strategies, such as the gamma transform applied by team DarkStyle, can enhance detection of features like PED by emphasizing low-intensity regions in OCT scans (Fig. 11). Second, careful metadata utilization improves model performance; for example, the diagnosis value indicating the fellow eye and baseline CST values were informative when integrated into the prediction pipeline. Third, simple rule-based patterns from metadata alone are insufficient for treatment decision-making, highlighting the necessity of combining rich OCT image features with clinical data.

Future directions include: (1) combining advanced image feature extraction with structured metadata to enhance treatment outcome predictions, (2) exploring model interpretability techniques, such as Integrated Gradients, to visualize regions of interest and increase clinician trust (Qi et al., 2019), (3) validating models across diverse populations to improve generalizability, and (4) addressing practical challenges such as data quality, workflow integration, and bias mitigation. Rigorous post-deployment monitoring will be essential to ensure the safety, reliability, and clinical utility of these AI-based systems.

8. Conclusion

We organized a competition focused on DME analysis using OCT data, with the key aim of predicting patient responses to treatment, specifically the decision of whether to continue anti-VEGF injection therapy. The competition addresses the unique challenges of DME analysis through OCT. This paper presents the dataset we released tailored for this task, outlines the competition's process and evaluation framework, and presents top-performing algorithms. Offering the first in-depth exploration of OCT multi-task DME analysis, our work lays the foundation for further research in this clinically relevant domain. Researchers are encouraged to leverage the openly accessible dataset to advance the field further.

Data and code availability

The dataset is available to the public on the official website at APTOS2021 Dataset. Interested applicants are required to fill out the application form on the APTOS website, and sign the Data Use Agreement. The application will undergo review, and APTOS will respond within 7 working days.

To enhance reproducibility, we provide the publicly available code repositories from participating teams at https://github.com/Michi-3000/APTOS2021_codes.

Our team is actively expanding upon the OCT4DME dataset. A follow-up competition, the Asia Pacific Tele-Ophthalmology Society (APTOS-2024) Challenge, has been organized to advance the field of ophthalmic AI by focusing on the synthesis of 3D OCT images from 2D fundus photography. Information about this challenge and the associated new datasets is made available in the competition report (Liu et al., 2025).

Fundings

The study was supported by the Global STEM Professorship Scheme (P0046113) and Henry G. Leong Endowed Professorship in Elderly Vision Health.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Weiyi Zhang: Investigation, Writing – original draft, Writing – review & editing; **Peranut Chotcomwongse:** Data curation; **Yinwen Li:** Writing – original draft; **Pusheng Xu:** Writing – original draft; **Ruijie Yao:** Methodology; **Lianhao Zhou:** Methodology; **Yuxuan Zhou:** Investigation; **Hui Feng:** Methodology; **Qiping Zhou:** Methodology; **Xinyue Wang:** Writing – original draft; **Shoujin Huang:** Writing – review & editing; **Zihao Jin:** Writing – review & editing; **Florence H T Chung:** Project administration; **Shujun Wang:** Writing – review & editing; **Yalin Zheng:** Writing – review & editing; **Mingguang He:** Supervision, Funding acquisition; **Danli Shi:** Supervision, Writing – review & editing; **Paisan Ruamviboonsuk:** Supervision, Funding acquisition.

Acknowledgments

We thank the InnoHK HKSAR Government for providing valuable supports.

This research received support from the JC STEM Lab of Innovative Light Therapy for Eye Diseases funded by The Hong Kong Jockey Club Charities Trust. The sponsor or funding organization did not participate in the design or implementation of this study.

Appendix A. APTOS 2021 team introduction

BlueSky: **1.** Ruijie Yao (ruijie.yao@duke.edu) is a first-year Ph.D. student in the Mechanical Engineering and Materials Science (MEMS) program at Duke University. He received his master's degree in 2024 and his bachelor's degree in 2021, both from the Department of Electronic Engineering at Tsinghua University. His research interests lie in the areas of computer vision, machine learning, and medical intelligence. **2.** Lianhao Zhou (leozhou@tamu.edu) is a first-year Ph.D. student in Computer Science at Texas A&M University. He received both his Bachelor's and Master's degrees from the Department of Electronic Engineering at Tsinghua University. His research interests include AI for Science and LLM Agent. **3.** Yuxuan Zhou (zhou-yx21@mails.tsinghua.edu.cn) is a fourth-year Ph.D. student at the department of Electronic Engineering, Tsinghua University. He also received his Bachelor's degree from the department of Electronic Engineering at Tsinghua University. His research interests include LLM medical evaluation and domain-specific training.

LightRain: Hui Feng(fenghuiwk@gmail.com) is from Xidian University. His research interests are computer vision, video anomaly detection, multimodal generation, and applications of large language models.

DarkStyle: DarkStyle team consists of Qiping Zhou (qiping.zhou@intalight.com), an Algorithm Engineer at Svision Imaging Ltd., specializing in deep learning solutions for medical image analysis.

References

2024 APTOS big data competition. 2024. https://2024.asiateleophth.org/big-data-competition/.

Akpinar, M.H., Sengur, A., Faust, O., Tong, L., Molinari, F., Acharya, U.R., 2024. Artificial intelligence in retinal screening using OCT images: a review of the last decade (2013–2023). Comput. Methods Programs Biomed. 254, 108253. https://doi.org/10.1016/j.cmpb.2024.108253

Awais, M., Müller, H., Tang, T.B., Meriaudeau, F., 2017. Classification of sd-oct images using a deep learning approach. In: 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA). IEEE, pp. 489–492.

Bogunović, H., Venhuizen, F., Klimscha, S., Apostolopoulos, S., Bab-Hadiashar, A., Bagci, U., Beg, M.F., Bekalo, L., Chen, Q., Ciller, C., et al., 2019. Retouch: the retinal oct fluid detection and segmentation benchmark and challenge. IEEE Trans. Med. Imaging 38 (8), 1858–1874.

Bressler, N.M., Beaulieu, W.T., Glassman, A.R., Blinder, K.J., Bressler, S.B., Jampol, L.M., Melia, M., 3rd Wells, , J.A., 2018. Persistent macular thickening following intravitreous aflibercept, bevacizumab, or ranibizumab for central-involved diabetic macular edema with vision impairment: a secondary analysis of a randomized clinical trial. JAMA Ophthalmol. 136 (3), 257–269. https://doi.org/10.1001/jamaophthalmol.2017.6565

Bressler, N.M., Odia, I., Maguire, M., Glassman, A.R., Jampol, L.M., MacCumber, M.W., Shah, C., Rosberger, D., Sun, J.K., et al., 2019. Association between change in visual acuity and change in central subfield thickness during treatment of diabetic macular edema in participants randomized to aflibercept, bevacizumab, or ranibizumab: a post hoc analysis of the protocol t randomized clinical trial. JAMA Ophthalmol. 137 (9), 977–985.

Cappellani, F., Card, K.R., Shields, C.L., Pulido, J.S., Haller, J.A., 2024. Reliability and accuracy of artificial intelligence chatGPT in providing information on ophthalmic diseases and management to patients. Eye 38 (7), 1368–1373.

Carbonneau, M.-A., Cheplygina, V., Granger, E., Gagnon, G., 2018. Multiple instance learning: a survey of problem characteristics and applications. Pattern Recognit. 77, 329–353.

Chen, H.-Y., Huang, M.-L., Hung, P.-T., 2006. Logistic regression analysis for glaucoma diagnosis using stratus optical coherence tomography. Optom. Vis. Sci. 83 (7), 00023. https://doi.org/10.1097/00006324-200607000-00023

Chen, R., Zhang, W., Song, F., Yu, H., Cao, D., Zheng, Y., He, M., Shi, D., 2024a. Translating color fundus photography to indocyanine green angiography using deep-learning for age-related macular degeneration screening. NPJ Dig. Med. 7 (1), 34.

Chen, X., Zhang, W., Xu, P., Zhao, Z., Zheng, Y., Shi, D., He, M., 2024b. Ffa-gpt: an automated pipeline for fundus fluorescein angiography interpretation and question-answer. NPJ Dig. Med. 7 (1), 111.

Çorbacıoğlu, Ş.K., Aksel, G., 2023. Receiver operating characteristic curve analysis in diagnostic accuracy studies: a guide to interpreting the area under the curve value. Turk. J. Emerg. Med. 23 (4), 195–198.

Daruich, A., Matet, A., Moulin, A., Kowalczuk, L., Nicolas, M., Sellam, A., Rothschild, P.R., Omri, S., Gélizé, E., Jonet, L., Delaunay, K., De Kozak, Y., Berdugo, M., Zhao, M., Crisanti, P., Behar-Cohen, F., 2018. Mechanisms of macular edema: beyond the surface. Prog. Retin Eye Res. 63, 20–68. https://doi.org/10.1016/j.preteyeres.2017.10.006

De Vente, C., Vermeer, K.A., Jaccard, N., Wang, H., Sun, H., Khader, F., Truhn, D., Aimyshev, T., Zhanibekuly, Y., Le, T.-D., et al., 2023. Airogs: artificial intelligence for robust glaucoma screening challenge. IEEE Trans. Med. Imaging 43 (1), 542–557.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, pp. 248–255.

Dugas, E., Jared, Jorge, Cukierski, W., 2015. Diabetic retinopathy detection. Kaggle https://kaggle.com/competitions/diabetic-retinopathy-detection.

Fang, H., Li, F., Fu, H., Sun, X., Cao, X., Lin, F., Son, J., Kim, S., Quellec, G., Matta, S., et al., 2022a. Adam challenge: detecting age-related macular degeneration from fundus images. IEEE Trans. Med. Imaging 41 (10), 2828–2847.

Fang, H., Li, F., Wu, J., Fu, H., Sun, X., Orlando, J.I., Bogunović, H., Zhang, X., Xu, Y., 2024. Open fundus photograph dataset with pathologic myopia recognition and anatomical structure annotation. Sci. Data 11 (1), 99.

Fang, H., Li, F., Wu, J., Fu, H., Sun, X., Son, J., Yu, S., Zhang, M., Yuan, C., Bian, C., et al., 2022b. Refuge2 challenge: a treasure trove for multi-dimension analysis and evaluation in glaucoma screening. arXiv preprint arXiv:2202.08994 .

Feng, D., Chen, X., Zhou, Z., Liu, H., Wang, Y., Bai, L., Zhang, S., Mou, X., 2020. A preliminary study of predicting effectiveness of anti-VEGF injection using OCT images based on deep learning. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, pp. 5428–5431.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat., 1189–1232.

Fu, H., Li, F., Sun, X., Cao, X., Liao, J., Orlando, J.I., Tao, X., Li, Y., Zhang, S., Tan, M., et al., 2020. Age challenge: angle closure glaucoma evaluation in anterior segment optical coherence tomography. Med. Image Anal. 66, 101798.

Hashemi, H., Rezvan, F., Pakzad, R., Ansaripour, A., Heydarian, S., Yekta, A., Ostadimoghaddam, H., Pakbin, M., Khabazkhoob, M., 2022. Global and regional prevalence of diabetic retinopathy; a comprehensive systematic review and meta-analysis. Semin. Ophthalmol. 37 (3), 291–306. https://doi.org/10.1080/08820538.2021.1962920

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708.

Hussain, M.A., Bhuiyan, A., Luu, C.D, Theodore Smith, R., Guymer, R.H., Ishikawa, R., Schuman, J.S., Ramamohanarao, K., 2018. Classification of healthy and diseased retina using SD-OCT imaging and random forest algorithm. PLoS ONE 13 (6), 1–17. https://doi.org/10.1371/journal.pone.0198281

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

Kobat, S.G., Baygin, N., Yusufoglu, E., Baygin, M., Barua, P.D., Dogan, S., Yaman, O., Celiker, U., Yildirim, H., Tan, R.-S., et al., 2022. Automated diabetic retinopathy detection using horizontal and vertical patch division-based pre-trained denseNET with digital fundus images. Diagnostics 12 (8), 1975.

Koozekanani, D., Roberts, C., Katz, S.E., Herderick, E.E., 2000. Intersession repeatability of macular thickness measurements with the humphrey 2000 OCT. Invest. Ophthalmol. Vis. Sci. 41 (6), 1486–1491.

Lemij, H.G., de Vente, C., Sánchez, C.I., Vermeer, K.A., 2023. Characteristics of a large, labeled data set for the training of artificial intelligence for glaucoma screening with fundus photographs. Ophthalmol. Sci. 3 (3), 100300.

Liu, B., Zhang, W., Chotcomwongse, P., Chen, X., Chen, R., Pakaymaskul, P., Arjkongharn, N., Vongsa, N., Cheng, X., Ge, Z., et al., 2025. Aptos-2024 challenge report: generation of synthetic 3d oct images from fundus photographs. arXiv preprint arXiv:2506.07542 .

Liu, R., Wang, X., Wu, Q., Dai, L., Fang, X., Yan, T., Son, J., Tang, S., Li, J., Gao, Z., et al., 2022. Deepdrid: diabetic retinopathy-grading and image quality estimation challenge. Patterns 3 (6).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022.

Midena, E., Toto, L., Frizziero, L., Covello, G., Torresin, T., Midena, G., Danieli, L., Pilotto, E., Figus, M., Mariotti, C., Lupidi, M., 2023. Validation of an automated artificial intelligence algorithm for the quantification of major OCT parameters in diabetic macular edema. J. Clin. Med. 12 (6). https://doi.org/10.3390/jcm12062134

Mihalache, A., Huang, R.S., Popovic, M.M., Patil, N.S., Pandya, B.U., Shor, R., Pereira, A., Kwok, J.M., Yan, P., Wong, D.T., Kertes, P.J., Muni, R.H., 2024. Accuracy of an artificial intelligence chatbot's interpretation of clinical ophthalmic images. JAMA Ophthalmol. 142 (4), 321–326. https://doi.org/10.1001/jamaophthalmol.2024.0017

Ndipenoch, N., Miron, A., Li, Y., 2024. Performance evaluation of retinal oct fluid segmentation, detection, and generalization over variations of data sources. IEEE Access 12, 31719–31735.

Niemeijer, M., Van Ginneken, B., Cree, M.J., Mizutani, A., Quellec, G., Sánchez, C.I., Zhang, B., Hornero, R., Lamard, M., Muramatsu, C., et al., 2009. Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs. IEEE Trans. Med.Imaging 29 (1), 185–195.

, 2019. Peking university international competition on ocular disease intelligent recognition (ODIR-2019). https://odir2019.grand-challenge.org/.

Olson, J., Sharp, P., Goatman, K., Prescott, G., Scotland, G., Fleming, A., Philip, S., Santiago, C., Borooah, S., Broadbent, D., Chong, V., Dodson, P., Harding, S., Leese, G., Styles, C., Swa, K., Wharton, H., 2013. Improving the economic value of photographic screening for optical coherence tomography-detectable macular oedema: a prospective, multicentre, UK study. Health Technol. Assess 17 (51), 1–142. https://doi.org/10.3310/hta17510

Orlando, J.I., Fu, H., Breda, J.B., Van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P.-A., Kim, J., Lee, J., et al., 2020. Refuge challenge: a unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Med. Image Anal. 59, 101570.

Pachade, S., Porwal, P., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Luo, Z., Han, F., Sun, Z., Qihan, L., Kamata, S.-i., et al., 2025. Rfmid: retinal image analysis for multi-disease detection challenge. Med. Image Anal. 99, 103365.

Padmasini, N., Umamaheswari, R., Sikkandar, M.Y., Sindal, M.D., 2023. Comparison of Deep CNNs in the Identification of DME Structural Changes in Retinal OCT Scans. Elsevier. pp. 35–51.

Panozzo, G., Gusson, E., Parolini, B., Mercanti, A., 2003. Role of OCT in the diagnosis and follow up of diabetic macular edema. In: Seminars in Ophthalmology. Taylor & Francis, pp. 74–81.

Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Son, J., Bae, W., Liu, L., Wang, J., Liu, X., Gao, L., et al., 2020. Idrid: diabetic retinopathy–segmentation and grading challenge. Med. Image Anal. 59, 101561.

Qi, Z., Khorram, S., Li, F., 2019. Visualizing deep networks by optimizing with integrated gradients. In: CVPR Workshops. Vol. 2, pp. 1–4.

Qian, B., Chen, H., Wang, X., Guan, Z., Li, T., Jin, Y., Wu, Y., Wen, Y., Che, H., Kwon, G., et al., 2024. Drac 2022: a public benchmark for diabetic retinopathy analysis on ultra-wide optical coherence tomography angiography images. Patterns 5 (3).

Rasti, R., Rabbani, H., 2017. Retinal OCT classification challenge (ROCC). https://rocc.grand-challenge.org/.

Salz, D.A., Witkin, A.J., 2015. Imaging in diabetic retinopathy. Middle East Afr. J. Ophthalmol. 22 (2), 145–150. https://doi.org/10.4103/0974-9233.151887

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., 2018. Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520.

Seeböck, P., Waldstein, S.M., Klimscha, S., Bogunovic, H., Schlegl, T., Gerendas, B.S., Donner, R., Schmidt-Erfurth, U., Langs, G., 2018. Unsupervised identification of disease marker candidates in retinal OCT imaging data. IEEE Trans. Med. Imaging 38 (4), 1037–1047.

Segal, M.R., 2004. Machine learning benchmarks and random forest regression. Technical Report, Center for Bioinformatics and Molecular Biostatistics, University of California, San Francisco. https://escholarship.org/uc/item/35x3v9t4.

Shanmugam, D., Blalock, D., Balakrishnan, G., Guttag, J., 2021. Better aggregation in test-time augmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1214–1223.

Shen, Y., Li, J., Zhu, W., Yu, K., Wang, M., Peng, Y., Zhou, Y., Guan, L., Chen, X., 2023. Graph attention u-net for retinal layer surface detection and choroid neovascularization segmentation in oct images. IEEE Trans. Med. Imaging 42 (11), 3140–3154.

Shi, D., Zhang, W., Chen, X., Liu, Y., Yang, J., Huang, S., Tham, Y.C., Zheng, Y., He, M., 2024. Eyefound: a multimodal generalist foundation model for ophthalmic imaging. arXiv preprint arXiv:2405.11338 .

Shi, D., Zhang, W., He, S., Chen, Y., Song, F., Liu, S., Wang, R., Zheng, Y., He, M., 2023. Translation of color fundus photography into fluorescein angiography using deep learning for enhanced diabetic retinopathy screening. Ophthalmol. Sci. 3 (4), 100401.

Shi, D., Zhang, W., Yang, J., Huang, S., Chen, X., Xu, P., Jin, K., Lin, S., Wei, J., Yusufu, M., et al., 2025. A multimodal visual–language foundation model for computational ophthalmology. npj Dig. Med. 8 (1), 381.

Solomon, S.D., Chew, E., Duh, E.J., Sobrin, L., Sun, J.K., VanderBeek, B.L., Wykoff, C.C., Gardner, T.W., 2017. Diabetic retinopathy: a position statement by the american diabetes association. Diab. Care 40 (3), 412–418. https://doi.org/10.2337/dc16-2641

Sonobe, T., Tabuchi, H., Ohsugi, H., Masumoto, H., Ishitobi, N., Morita, S., Enno, H., Nagasato, D., 2019. Comparison between support vector machine and deep learning, machine-learning technologies for detecting epiretinal membrane using 3d-OCT. Int. Ophthalmol. 39 (8), 1871–1877. https://doi.org/10.1007/s10792-018-1016-x

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

Sugar, E.A., Jabs, D.A., Altaweel, M.M., Lightman, S., Acharya, N., Vitale, A.T., Thorne, J.E., MUST Trial Research Group , et al., 2011. Identifying a clinically meaningful threshold for change in uveitic macular edema evaluated by optical coherence tomography. Am. J. Ophthalmol. 152 (6), 1044–1052.

Sultani, W., Chen, C., Shah, M., 2018. Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6479–6488.

Sun, J., Wei, D., Wang, L., Zheng, Y., 2022. Lesion guided explainable few weak-shot medical report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention.

Tan, M., Le, Q., 2019. Efficientnet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR, pp. 6105–6114.

Tang, F., Wang, X., Ran, A.R., Chan, C.K.M., Ho, M., Yip, W., Young, A.L., Lok, J., Szeto, S., Chan, J., Yip, F., Wong, R., Tang, Z., Yang, D., Ng, D.S., Chen, L.J., Brelén, M., Chu, V., Li, K., Lai, T.H.T., Tan, G.S., Ting, D.S.W., Huang, H., Chen, H., Ma, J.H., Tang, S., Leng, T., Kakavand, S., Mannil, S.S., Chang, R.T., Liew, G., Gopinath, B., Lai, T.Y.Y., Pang, C.P., Scanlon, P.H., Wong, T.Y., Tham, C.C., Chen, H., Heng, P.A., Cheung, C.Y., 2021. A multitask deep-learning system to classify diabetic macular edema for different optical coherence tomography devices: a multicenter analysis. Diab. Care 44 (9), 2078–2088. https://doi.org/10.2337/dc20-3064

Tang, F.Y., Ng, D.S., Lam, A., Luk, F., Wong, R., Chan, C., Mohamed, S., Fong, A., Lok, J., Tso, T., et al., 2017. Determinants of quantitative optical coherence tomography angiography metrics in patients with diabetes. Sci. Rep. 7 (1), 2575.

Tripathi, A., Kumar, P., Mayya, V., Tulsani, A., 2023. Generating OCT b-scan DME images using optimized generative adversarial networks (GANs). Heliyon 9 (8), e18773. https://doi.org/10.1016/j.heliyon.2023.e18773

Waheed, N.K., Duker, J.S., 2013. Oct in the management of diabetic macular edema. Curr. Ophthalmol. Rep. 1 (3), 128–133. https://doi.org/10.1007/s40135-013-0019-z

Wallace, G.K., 1992. The JPEG still picture compression standard. IEEE Trans. Consum. Electron. 38 (1), xviii–xxxiv.

Wang, J., Deng, G., Li, W., Chen, Y., Gao, F., Liu, H., He, Y., Shi, G., 2019. Deep learning for quality assessment of retinal OCT images. Biomed. Opt. Express 10 (12), 6057–6072.

Wang, X., Tang, F., Chen, H., Cheung, C.Y., Heng, P.-A., 2023. Deep semi-supervised multiple instance learning with self-correction for DME classification from OCT images. Med. Image Anal. 83, 102673.

Wu, J., Fang, H., Li, F., Fu, H., Lin, F., Li, J., Huang, Y., Yu, Q., Song, S., Xu, X., et al., 2023. Gamma challenge: glaucoma grading from multi-modality images. Med. Image Anal. 90, 102938.

Wu, J.-H., Nishida, T., Weinreb, R.N., Lin, J.-W., 2022. Performances of machine learning in detecting glaucoma using fundus and retinal optical coherence tomography images: a meta-analysis. Am. J. Ophthalmol. 237, 1–12.

Xu, P., Chen, X., Zhao, Z., Shi, D., 2024. Unveiling the clinical incapabilities: a benchmarking study of GPT-4v (ision) for ophthalmic multimodal image analysis. Br. J. Ophthalmol. 108 (10), 1384–1389.

Xu, P., Chotcomwongse, P., Zhang, W., Chen, X., Wu, X., Chung, F. H.T., Zhang, X., He, M., Shi, D., Ruamviboonsuk, P., 2025. Angioreport: dataset and baseline methods for fundus angiography report generation. Br. J. Ophthalmol. 109, 1283–1288.

Zhang, J., Zhang, J., Zhang, C., Zhang, J., Gu, L., Luo, D., Qiu, Q., 2022. Diabetic macular edema: current understanding, molecular mechanisms and therapeutic implications. Cells 11 (21). https://doi.org/10.3390/cells11213362

Zhang, W., Huang, S., Yang, J., Chen, R., Ge, Z., Zheng, Y., Shi, D., He, M., 2024. Fundus2video: cross-modal angiography video generation from static fundus photography with clinical knowledge guidance. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 689–699.